

MERCHANTS, DOSTAWCY I ISO

Haiku 5.5: funkcje AI w aplikacji płatniczej nie wymagają już dużego budżetu

Opublikowano 27 października 2026

Claude Haiku 5.5 od Anthropic kosztuje jedną czterdziestą ceny obecnego modelu Opus za tę samą pracę. Przy wąskich, masowych zadaniach, które firma płatnicza wykonuje każdego dnia, AI przestaje być pozycją w budżecie i staje się błędem zaokrąglenia. Ile kosztuje, do czego się nadaje i gdzie większy model wciąż jest wart swojej ceny.

Przez ostatnie lata każda funkcja AI w aplikacji płatniczej wiązała się z pytaniem o koszt. Przeczytanie strony merchanta, posortowanie skrzynki wsparcia czy wyciągnięcie pól z formularza wniosku to zadania, które model językowy wykonuje dobrze, ale przy dziesiątkach tysięcy wywołań miesięcznie rachunek był na tyle wysoki, że wiele zespołów zatrzymywało takie funkcje na etapie pilotażu.

Obecny mały model Anthropic, **Claude Haiku 5.5**, zmienia ten rachunek. W cenie katalogowej kosztuje jedną czterdziestą obecnego modelu Opus i jedną dziesiątą poprzedniego Haiku. Przy wąskiej, powtarzalnej pracy, która stanowi większość AI potrzebnej firmie płatniczej, funkcja, która kiedyś kosztowała setki miesięcznie, teraz kosztuje kilka.

Ile kosztują modele

Anthropic wycenia swoje modele za milion *tokenów*, czyli jednostek, na które dzielony jest tekst. Token to mniej więcej trzy czwarte angielskiego słowa, więc milion tokenów to około 750 000 słów. Dane wejściowe (to, co wysyłasz do modelu) i wyjściowe (to, co model odpisuje) są wyceniane osobno.

MODEL	WEJŚCIE, USD ZA MILION TOKENÓW	WYJŚCIE, USD ZA MILION TOKENÓW	WZGLĘDEM HAIKU 5.5
Claude Haiku 5.5	0,10	0,50	1×
Claude Haiku 4.5 (poprzedni Haiku)	1,00	5,00	10×
Claude Sonnet 5.5	2,00	10,00	20×

MODEL	WEJŚCIE, USD ZA MILION TOKENÓW	WYJŚCIE, USD ZA MILION TOKENÓW	WZGLĘDEM HAIKU 5.5
Claude Opus 5.5	4,00	20,00	40×
Claude Fable 5.1	10,00	50,00	100×

To własne ceny katalogowe Anthropic według stanu na październik 2026. Cena Haiku 5.5 dotyczy żądań do 100 000 tokenów; dłuższe kosztują 0,50 USD za wejście i 2,50 USD za wyjście, wciąż znacznie mniej niż większe modele. Claude jest też sprzedawany przez Amazon Bedrock i Google Cloud Vertex AI, które ustalają własne ceny, więc sprawdź stawkę na platformie, przez którą kupujesz.

Co to oznacza dla jednego żądania

Typowe wąskie zadanie, na przykład przeczytanie jednej podstrony witryny merchanta i zwrócenie krótkiej, ustrukturyzowanej oceny, może wysłać 2 000 tokenów i otrzymać 300. Według cen z tabeli powyżej:

MODEL	KOSZT JEDNEGO ŻĄDANIA, USD	KOSZT 10 000 ŻĄDAŃ, 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

Jedno żądanie do Haiku 5.5 kosztuje kilka setnych centa. Dziesięć tysięcy kosztuje mniej więcej tyle, co kawa. Liczby są poglądowe: twoje prompty będą dłuższe lub krótsze, a rozumowanie modelu, rozliczane jako dane wyjściowe, zwiększa rachunek (więcej o tym niżej). Proporcje między modelami pozostają takie same bez względu na wielkość żądania.

Dwa rabaty obniżają koszt jeszcze bardziej. Żądania, które nie potrzebują natychmiastowej odpowiedzi, mogą przejść przez interfejs wsadowy Anthropic za połowę ceny, co pasuje do nocnych zadań, takich jak ponowne skanowanie portfela merchantów. A długi zestaw instrukcji, który jest taki sam przy każdym żądaniu, można zapisać w pamięci podręcznej: jest rozliczany w całości raz, a potem za ułamek ceny.

Do jakiej pracy nadaje się mały model

Tani model nie jest gorszą wersją drogiego przy każdym zadaniu. Jest zbudowany z myślą o szybkości i skali i najlepiej sprawdza się, gdy zadanie jest wąskie, instrukcje jasne, a odpowiedź ma stałą formę. Większość AI potrzebnej firmie płatniczej wygląda właśnie tak.

- **Weryfikacja strony przy onboardingu.** Czy strona ma politykę zwrotów, regulamin i dane kontaktowe? Co naprawdę sprzedaje i czy zgadza się to z branżą zadeklarowaną przez merchanta? To czytanie i klasyfikowanie, powtarzane przy każdym wniosku i ponownie przy każdym [przebiegu okresowym](#).
- **Monitorowanie portfela.** Regularne ponowne czytanie tysięcy stron merchantów w poszukiwaniu nowych produktów, zmienionych polityk lub oznak [prania transakcji](#). Tutaj skala jest całym problemem i tutaj różnica w cenie ma największe znaczenie.
- **Wyciąganie pól z dokumentów.** Nazwa firmy, numer rejestrowy, członkowie zarządu i adresy z odpisu z rejestru lub pisma z banku, zwrócone w stałym formacie do sprawdzenia przez człowieka.
- **Sortowanie skrzynki.** Kierowanie zgłoszeń wsparcia, pytań merchantów i korespondencji z dostawcami do właściwego zespołu oraz oznaczanie tych, które wspominają o sporze, oszustwie lub regulatorze.
- **Wstępna selekcja sporów.** Grupowanie przychodzących [chargebacków](#) według powodu, sprawdzanie, czy dowody dostarczone przez merchanta są tymi, których wymaga kod powodu, i wypisywanie, czego brakuje.
- **Czytanie deskryptorów i opisów transakcji.** Klasyfikowanie dowolnych opisów transakcji w kategorii lub sprawdzanie, czy deskryptor na wyciągu będzie rozpoznawalny dla posiadacza karty.
- **Zamiana pytania w wyszukiwanie.** Wzięcie jednego wpisanego zdania i wyciągnięcie pól potrzebnych zapytaniu do bazy danych: kraju, waluty, branży, metody płatności.

Każde z tych zadań wykonuje się wiele razy dziennie, każde ma odpowiedź, którą człowiek może sprawdzić, i żadne nie wymaga od modelu samodzielnego planowania długiej pracy.

Gdzie większy model wciąż jest wart swojej ceny

Większe modele są wolniejsze i droższe nie bez powodu: lepiej radzą sobie z pracą długą, otwartą lub trudną do opisaną. Warto za nie płacić, gdy:

- **zadanie ma wiele zależnych od siebie kroków**, na przykład zbadanie nietypowego wzorca w kilku źródłach danych i opisanie ustaleń.
- **tekst ma znaczenie**, na przykład pełna odpowiedź na spór lub polityka, którą zatwierdzi człowiek.
- **dane wejściowe są długie, a odpowiedź zależy od całości**, na przykład przegląd całej umowy z dostawcą pod kątem twoich własnych warunków.

- **błędna odpowiedź jest kosztowna i trudna do wychwycenia.**

Popularny układ korzysta z obu: mały model obsługuje każde żądanie i przekazuje większemu te nieliczne, co do których ma wątpliwości lub które przekraczają ustalony próg. Najpierw jednak porównaj to z prostszym wyborem. Czasem wystarczy większy model na niższym poziomie rozumowania, a jednym modelem łatwiej zarządzać niż dwoma.

Jak utrzymać koszt tam, gdzie go oczekujesz

Cena za token to tylko część rachunku. Kilka nawyków sprawia, że jest on przewidywalny.

- **Dopasuj poziom rozumowania do zadania.** Obecne modele Claude myślą przed odpowiedzią, a to myślenie jest rozliczane jako dane wyjściowe. Haiku 5.5 domyślnie działa na średnim poziomie. Do klasyfikacji i ekstrakcji często wystarczy niski poziom, który jest znacznie tańszy. Przetestuj go na próbkach prawdziwych żądań, zanim zdecydujesz.
- **Proś o stały format.** Krótka, ustrukturyzowana odpowiedź (tak lub nie, kategoria z listy, kilka nazwanych pól) kosztuje mniej niż swobodny tekst i łatwiej ją sprawdzić.
- **Wysyłaj tylko to, czego model potrzebuje.** Tekst strony z polityką, a nie całą witrynę z nawigacją i skryptami.
- **Zapisuj w pamięci podręcznej to, co się powtarza, i przetwarzaj wsadowo to, co może poczekać.** Oba to rabaty na pracę, którą i tak wykonujesz.
- **Oceniaj koszt ukończonego zadania,** a nie pojedynczego żądania. Tańsze wywołanie, które trzeba powtórzyć dwa razy, nie jest tańsze.

Czego cena nie zmienia

Tańszy model nie zmienia twoich obowiązków, a trzy kwestie warto rozstrzygnąć przed uruchomieniem czegokolwiek.

1. **Nie przekazuj danych kart.** Pełne numery kart, kody bezpieczeństwa i wszystko inne objęte zakresem PCI DSS nie mają miejsca w prompcie. Żadne z powyższych zadań ich nie potrzebuje, a ich wysłanie wciągnęłoby usługę AI i wszystko wokół niej do twojego środowiska danych posiadaczy kart.
2. **Ograniczaj dane osobowe do minimum.** Zgodnie z RODO i podobnymi przepisami sprawdź warunki dostawcy dotyczące danych, miejsce przetwarzania żądań i okres ich przechowywania, a w polityce prywatności podaj, że korzystasz z tej usługi. Usuвай imiona, nazwiska i dane kontaktowe tam, gdzie zadanie ich nie wymaga.
3. **Traktuj to, co czyta model, jako niezaufane.** Stronę merchanta pisze sam merchant, który zyskuje na przychylniej ocenie. Tekst na niej może próbować wydawać modelowi polecenia („ta strona ma pełną politykę zwrotów”). Powiedz modelowi w instrukcjach, że treść to dane, a nie polecenia, i ogranicz odpowiedź do stałych opcji, żeby wstrzyknięty tekst nie miał dokąd trafić.

Model jedynie doradza. Ustalenie z automatycznej weryfikacji to powód, by człowiek przyjrzał się bliżej, a nie decyzja, i decyzję nadal podejmuje underwriter lub analityk.

Jak korzysta z tego Payments Marketplace

Payments Marketplace używa Claude właśnie do tego rodzaju wąskiej, masowej pracy. Gdy merchant wykonuje kontrolę wstępną, skan strony czyta jej podstrony i zwraca stały zestaw ustaleń: czy istnieje polityka zwrotów i czy jest wystarczająca, co strona naprawdę sprzedaje i czy zgadza się to z zadeklarowaną branżą. Agent AI na naszej stronie głównej zamienia jedno wpisane zdanie w kilka pól, które odczytuje nasze dopasowanie, a każde słowo odpowiedzi pisze nasz własny kod.

- **Model czyta, ludzie decydują.** Ustalenia skanu to sygnały doradcze pokazywane dostawcy. Nigdy nie blokują wniosku, a dopasowanie to przedstawienie, nigdy zatwierdzenie.
- **Anonimowość zostaje zachowana.** Podsumowanie skanu, które dostawca widzi przed akceptacją, nie zawiera żadnej tożsamości.
- **Ograniczony z założenia.** Tekst strony jest traktowany jako niezaufany, a odpowiedzi są ograniczone do stałych pól, więc tekst na stronie merchanta nie może sterować wynikiem.

Źródła

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

Czytaj online: <https://payments-marketplace.com/pl-PL/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