Pricing AI 2026 Lo que tu CFO no está viendo

CATEGORÍA · NEGOCIO, MONETIZACIÓN E INDUSTRIAS

55x de diferencia entre los modelos que tu equipo usa hoy. La mayoría de empresas LATAM está pagando 3x lo necesario sin saberlo.

QUÉ VAS A ENCONTRAR

01 Tabla completa: DeepSeek V4, Opus 4.7, Sonnet 4.6, Haiku 4.5, GPT-5.5, Gemini 2.5

02 El descuento de cache del 90% que casi nadie configura

03 Matriz de decisión: qué modelo para cada caso de uso

04 5 errores de arquitectura que inflan tu factura cada mes

05 Caso real: cliente que bajó de 8K a 2.4K USD/mes sin perder calidad

FICHA DEL RECURSO

Qué es, para qué sirve y cómo se implementa

FUNCIÓN

55x de diferencia entre los modelos que tu equipo usa hoy. La mayoría de empresas LATAM está pagando 3x lo necesario sin saberlo.

PARA QUÉ SIRVE

La mayoría de empresas LATAM montó su stack de AI con un solo modelo top: Opus, GPT, o Gemini. Funciona, da resultados, nadie cuestiona. El problema aparece cuando llega la factura. 5K, 8K, 15K USD al mes en API. Mucho de eso es tráfico que un modelo 50x más barato resuelve idéntico. Este PDF te muestra exactamente dónde está esa grasa. Datos públicos, cálculo simple, decisión clara. 55x 68x 70% DIFERENCIA EN INPUT POR MILLÓN DIFERENCIA EN OUTPUT POR MILLÓN DEL TRÁFICO TÍPICO ES BAJO PERFIL

DÓNDE APLICARLO

Cubre: Tabla completa: DeepSeek V4, Opus 4.7, Sonnet 4.6, Haiku 4.5, GPT-5.5, Gemini 2.5; El descuento de cache del 90% que casi nadie configura; Matriz de decisión: qué modelo para cada caso de uso; 5 errores de arquitectura que inflan tu factura cada mes; Caso real: cliente que bajó de 8K a 2.4K USD/mes sin perder calidad.

CÓMO IMPLEMENTARLO

1. Por Qué Este Pdf Existe: Pagas por capacidad que tu uso no necesita.
2. La Decisión Clave: No es qué modelo elegir.
3. 01 / Tabla Maestra: Precios públicos abril 2026.
4. 02 / Análisis Input: 55x de diferencia.
5. 03 / Análisis Output: 68x de diferencia.
6. 04 / Cache: El descuento del 90% que casi nadie configura.
7. 05 / Matriz De Decisión: Qué modelo para cada tarea.

POR QUÉ ESTE PDF EXISTE

Pagas por capacidad que tu uso no necesita.

La mayoría de empresas LATAM montó su stack de AI con un solo modelo top: Opus, GPT, o Gemini. Funciona, da resultados, nadie cuestiona. El problema aparece cuando llega la factura. 5K, 8K, 15K USD al mes en API. Mucho de eso es tráfico que un modelo 50x más barato resuelve idéntico. Este PDF te muestra exactamente dónde está esa grasa. Datos públicos, cálculo simple, decisión clara. 55x 68x 70% DIFERENCIA EN INPUT POR MILLÓN DIFERENCIA EN OUTPUT POR MILLÓN DEL TRÁFICO TÍPICO ES BAJO PERFIL

LA DECISIÓN CLAVE

No es qué modelo elegir.

Es cómo mezclarlos. Pensar la arquitectura AI como una sola elección de modelo es el error que más dinero quema en LATAM. Tu sistema tiene tareas heterogéneas: clasificar, extraer, conversar, razonar. Cada una merece el modelo correcto. Un solo modelo para todo es como pagar Uber Black para hacer mandados. Llega, sirve, pero pagas 5x sin razón.

PRINCIPIO

Mezcla modelos por tarea, no por preferencia.
El modelo top va al razonamiento que decide.
El modelo barato va al volumen repetitivo.

01 / TABLA MAESTRA

Precios públicos abril 2026.

Precios por millón de tokens en USD. Cache-R = costo de leer un token cacheado. Verifica la cifra exacta con el provider antes de calcular tu factura.

PRECIOS POR MILLON DE TOKENS USD

MODELO	INPUT	OUTPUT	CACHE-R	CONTEXTO
DeepSeek V4	$0.27	$1.10	$0.07	128K
Haiku 4.5	$0.80	$4.00	$0.08	200K
GPT-5.5	$1.25	$10.00	$0.13	400K
Sonnet 4.6	$3.00	$15.00	$0.30	200K
Gemini 2.5	$3.50	$10.50	$0.35	1M
Opus 4.7	$15.00	$75.00	$1.50	200K

02 / ANÁLISIS INPUT

55x de diferencia.

El factor que casi nadie mira. Input es lo que cuesta leer datos: el PDF, el prompt, el contexto, el historial. En sistemas de producción, input suele ser 70 a 90 por ciento del costo total. DeepSeek V4 cobra $0.27. Opus 4.7 cobra $15.00. Misma tarea de leer un documento, 55x de diferencia. Si tu volumen es alto, esto define si tu margen es 40 por ciento o 4 por ciento.

EJEMPLO

10M tokens de input al mes (volumen medio LATAM).
DeepSeek V4: $2.70
Opus 4.7: $150.00
Diferencia: $147.30 mensuales solo por input.

03 / ANÁLISIS OUTPUT

68x de diferencia.

Cuándo Opus se justifica. Output es lo que el modelo genera. Cuesta más que input siempre. Opus 4.7 cobra $75 por millón. DeepSeek V4 cobra $1.10. La pregunta no es si Opus es mejor: es si tu tarea necesita esa calidad. Opus se justifica en: razonamiento legal complejo, análisis financiero con múltiples variables, código que requiere arquitectura, escritura larga premium. Para clasificar tickets o extraer datos de PDFs, es overkill caro.

REGLA DE DECISIÓN

Si la tarea se puede validar con un humano en menos de 30 segundos:
no necesita Opus. Modelo barato es suficiente.
Si la tarea requiere experticia profunda para validar:
ahí es donde Opus o Sonnet pagan su precio.

04 / CACHE

El descuento del 90% que casi nadie configura.

Cuando tu sistema repite el mismo contexto varias veces (system prompt, instrucciones, ejemplos), puedes cachearlo. La primera vez paga precio normal. Cada lectura siguiente paga 10 por ciento del precio. En Opus 4.7: input cacheado cuesta $1.50 en vez de $15.00. En DeepSeek: $0.07 en vez de $0.27. Ahorro inmediato del 90 por ciento sobre la parte repetida. La mayoría de equipos LATAM no lo tiene activado. Quemamos 90 por ciento del presupuesto de input sin razón técnica.

DÓNDE APLICAR CACHE

Sistema con system prompt grande y consultas frecuentes.
Agentes con instrucciones detalladas que se repiten.
RAG con few-shot examples constantes.
Cualquier flujo donde el 30 por ciento o más del input se repite.

05 / MATRIZ DE DECISIÓN

Qué modelo para cada tarea.

Esta es la matriz que aplicamos en cada auditoría Cnvierte. No es teoría: son las decisiones que ya hemos validado con clientes en producción.

TAREA · MODELO RECOMENDADO

Clasificacion de tickets / leads: DeepSeek V4 o Haiku 4.5
Extraccion de datos de PDFs: DeepSeek V4 o Haiku 4.5
Chat al cliente final: GPT-5.5 o Sonnet 4.6
Resumen de reuniones: Sonnet 4.6
Analisis financiero / legal: Opus 4.7
Codigo arquitectura compleja: Opus 4.7
Code completion en IDE: Sonnet 4.6 o GPT-5.5
Generacion alto volumen / batch: DeepSeek V4
Investigacion con contexto largo: Gemini 2.5

06 / ERRORES COMUNES

5 errores de arquitectura que inflan tu factura.

Uno: usar el mismo modelo para todo el sistema. Es la causa número uno de facturas infladas. Dos: no cachear. El descuento del 90 por ciento está esperando, nadie lo activa. Tres: no setear max_tokens en output. Modelos generan respuestas más largas de lo necesario y tú pagas cada token. Cuatro: prompts sin optimizar. System prompts de 5000 tokens cuando 800 hacen el mismo trabajo. Cinco: no medir por tarea. Sin telemetría por endpoint no puedes saber dónde está el desperdicio real.

07 / CASO REAL

Cliente Cnvierte:

8K a 2.4K USD/mes. Empresa LATAM con sistema de soporte automatizado. Stack original: 100 por ciento Opus 4.7. Factura mensual: 8K USD aproximadamente. La auditoría reveló el desglose real del tráfico. 70 por ciento era clasificación automática de tickets entrantes (urgente, común, descarte). 20 por ciento era extracción de datos de PDFs adjuntos. 10 por ciento era razonamiento sobre casos complejos escalados. Migración: el 70 por ciento de clasificación pasó a DeepSeek V4. El 20 por ciento de extracción pasó a Haiku 4.5. El 10 por ciento crítico se quedó en Opus 4.7. Activamos cache para los system prompts. Resultado: la factura bajó a 2.4K mensuales. Mismo output al usuario final. 5.6K USD recuperados cada mes. ROI de la auditoría en menos de una semana.

AUDITORÍA RÁPIDA

Revisa tu stack en 10 minutos

VISIBILIDAD

Sabes cuánto te cuesta cada modelo de tu stack al mes
Tienes telemetría por tarea, no solo costo total
Identificas el top 3 de endpoints que más consumen
ARQUITECTURA
Tu sistema usa al menos 2 modelos diferentes por tier de tarea
Las tareas de alto volumen corren en el modelo más barato viable
El modelo top está reservado para razonamiento real
CACHE Y LÍMITES
Tienes cache activo en system prompts de más de 1000 tokens
Cada llamada tiene max_tokens explícito en el output
Los prompts están optimizados, no inflados con contexto irrelevante

AUDITORÍA GRATIS CNVIERTE

Si pagas $5K+/mes en AI te debemos una auditoría.

En Cnvierte hacemos auditorías gratis a empresas LATAM que pagan 5K USD o más al mes en AI. Revisamos tu stack, identificamos dónde estás quemando dinero, y te entregamos el plan de migración concreto. Sin venta dura: si tu setup ya está óptimo, te lo decimos.

SIGUIENTE PASO

Aplícalo esta semana.

Elige el primer paso de la sección de implementación y hazlo hoy.

MÁS GUÍAS GRATIS EN 377CREATIVE.COM/RECURSOS