

Elige el modelo correcto de AI

CATEGORÍA · PRIMEROS PASOS CON CLAUDE

La matriz que uso con mis clientes para no gastar de mas en modelos que no necesitan.

QUÉ VAS A ENCONTRAR

- 01 Por que pagar 10X no significa recibir 10X
- 02 Los 3 niveles de modelo y cuando usar cada uno
- 03 La regla de decision que aplica en cualquier caso

FICHA DEL RECURSO

Qué es, para qué sirve y cómo se implementa

FUNCIÓN

La matriz que uso con mis clientes para no gastar de mas en modelos que no necesitan.

PARA QUÉ SIRVE

Recibes lo mismo. Gemini 2.5 Flash cuesta \$0.30 por millon de tokens. Claude Sonnet 4.6 cuesta \$3.00. GPT-4o cuesta \$2.50. En el benchmark mas exigente de la industria, Flash saca 69.8 puntos. Sonnet saca 68.3. Un punto y medio de diferencia. Diez veces el precio. La mayoría de equipos usa el flagship para todo: revision de facturas, respuestas de soporte, clasificacion de correos. Tareas donde Flash produce el mismo resultado al 10% del costo.

DÓNDE APLICARLO

Cubre: Por que pagar 10X no significa recibir 10X; Los 3 niveles de modelo y cuando usar cada uno; La regla de decision que aplica en cualquier caso.

CÓMO IMPLEMENTARLO

- 1 01 / El Problema: Pagas 10X mas.
- 2 02 / La Matriz: 3 niveles de modelo.
- 3 03 / La Regla: Una pregunta.

01 / EL PROBLEMA

Pagas 10X mas.

Recibes lo mismo. Gemini 2.5 Flash cuesta \$0.30 por millon de tokens. Claude Sonnet 4.6 cuesta \$3.00. GPT-4o cuesta \$2.50. En el benchmark mas exigente de la industria, Flash saca 69.8 puntos. Sonnet saca 68.3. Un punto y medio de diferencia. Diez veces el precio. La mayoría de equipos usa el flagship para todo: revision de facturas, respuestas de soporte, clasificacion de correos. Tareas donde Flash produce el mismo resultado al 10% del costo.

02 / LA MATRIZ

3 niveles de modelo.

3 casos distintos. Nivel 1, Volumen: Flash a \$0.30/M. Clasificación, extracción, resumen, soporte básico. El 80% de las tareas operativas de un negocio entra aquí. Nivel 2, Análisis: Sonnet o GPT-4o a \$2.50-3.00/M. Contratos, estrategia, código para producción. Cuando el output necesita razonamiento en contexto complejo. Nivel 3, Crítico: Opus a \$15/M. Decisiones de alto impacto, razonamiento multi-paso. Cuando el error tiene consecuencias reales para el negocio.

REFERENCIA DE PRECIOS (JUNIO 2026)

Gemini 2.5 Flash: \$0.30 / millón de tokens

GPT-4o: \$2.50 / millón de tokens

Claude Sonnet 4.6: \$3.00 / millón de tokens

Claude Opus 4.7: \$15.00 / millón de tokens

03 / LA REGLA

Una pregunta.

Una decisión. Antes de elegir el modelo, una sola pregunta: si este modelo comete un error, cuánto cuesta arreglarlo? Si el costo del error es mayor que el costo del modelo, sube un nivel. Si el costo del error es menor, baja un nivel. Para el 80% de las tareas operativas de un negocio, Flash alcanza. El flagship queda para cuando el error tiene consecuencias reales.

LA REGLA EN UNA LINEA

Costo del error > costo del modelo: sube un nivel.

Costo del error < costo del modelo: baja un nivel.

Duda: empieza con Flash y sube solo si el resultado no alcanza.

SIGUIENTE PASO

Aplica esto desde hoy.

MÁS GUÍAS GRATIS EN [377CREATIVE.COM/RECURSOS](https://377creative.com/recursos)

377 CREATIVE

Agencia digital en español · Recursos de descarga libre, comparte citando la fuente.