

Estira tu crédito al doble

CATEGORÍA · PRIMEROS PASOS CON CLAUDE

El checklist para que tus agentes de Claude rindan el doble con el mismo crédito.

QUÉ VAS A ENCONTRAR

- 01 Por qué tu crédito de agentes ahora tiene tope
- 02 El modelo correcto para cada tarea
- 03 Como bajar el costo del contexto hasta 10 veces
- 04 Los 4 ajustes que duplican tu crédito

FICHA DEL RECURSO

Qué es, para qué sirve y cómo se implementa

FUNCIÓN

El checklist para que tus agentes de Claude rindan el doble con el mismo crédito.

PARA QUÉ SIRVE

Desde el 15 de junio, automatizar con Claude por código corre contra un crédito aparte, incluido en tu plan pero con tope. Cuando se acaba, tus agentes se detienen, y si quieres seguir, pagas extra a precio de API. La buena noticia: con cuatro ajustes puedes hacer que ese mismo crédito te rinda el doble. Aquí están. 2x 10x MÁS TAREAS CON EL MISMO CRÉDITO MÁS BARATO EL CONTEXTO CACHEADO

DÓNDE APLICARLO

Cubre: Por qué tu crédito de agentes ahora tiene tope; El modelo correcto para cada tarea; Como bajar el costo del contexto hasta 10 veces; Los 4 ajustes que duplican tu crédito.

CÓMO IMPLEMENTARLO

- 1 Por Qué Esto Importa: Tu crédito de agentes ahora tiene tope.
- 2 01 / El Modelo Correcto: No uses Opus para todo.
- 3 02 / Contexto: Menos contexto, menos gasto.
- 4 03 / Prompt Caching: Reutiliza el contexto fijo.
- 5 04 / Paralelo Y Effort: No dispares de más.
- 6 Antes De Cada Run: Checklist para no quemar crédito

POR QUÉ ESTO IMPORTA

Tu crédito de agentes ahora tiene tope.

Desde el 15 de junio, automatizar con Claude por código corre contra un crédito aparte, incluido en tu plan pero con tope. Cuando se acaba, tus agentes se detienen, y si quieres seguir, pagas extra a precio de API. La buena noticia: con cuatro ajustes puedes hacer que ese mismo crédito te rinda el doble. Aquí están. 2x 10x MÁS TAREAS CON EL MISMO CRÉDITO MÁS BARATO EL CONTEXTO CACHEADO

01 / EL MODELO CORRECTO

No uses Opus para todo.

El error que más crédito quema es correr el modelo caro en tareas que un modelo barato hace igual. Opus para razonar, Sonnet o Haiku para ejecutar.

CÓMO HACERLO

1. Reserva Opus para lo difícil: planear, decidir, depurar lógica compleja.
2. Manda a Sonnet o Haiku lo repetitivo: escribir código, correr comandos, formatear.
3. En subagentes, fija el modelo barato con model: sonnet o model: haiku.

REGLA

Opus piensa. Sonnet y Haiku ejecutan.

02 / CONTEXTO

Menos contexto, menos gasto.

Cada token que entra cuesta. Si le metes el proyecto entero al agente cuando solo necesita tres archivos, estás pagando de más en cada llamada.

CÓMO HACERLO

1. Dale solo los archivos que la tarea necesita, no el repo completo.
2. Cierra o resume las conversaciones largas antes de seguir.
3. Evita pegar documentos enteros cuando basta un fragmento.

REGLA

Antes de cada run pregúntate:

¿esto necesita ver todo, o solo una parte?

03 / PROMPT CACHING

Reutiliza el contexto fijo.

Si tu agente manda el mismo bloque de instrucciones o contexto en cada llamada, el prompt caching te deja reutilizarlo y baja el costo de esos tokens de entrada hasta una décima parte.

CÓMO HACERLO

1. Identifica la parte de tu prompt que no cambia entre llamadas: instrucciones, contexto base.
2. Márcala para cache según la documentación de Anthropic.
3. Las llamadas siguientes leen ese bloque cacheado a una fracción del precio.

REGLA

Contexto que se repite = candidato a cache.

Úsalo y paga ese bloque hasta 10 veces menos.

04 / PARALELO Y EFFORT

No dispares de más.

Correr diez agentes en paralelo y dejar el effort en máximo se siente productivo, pero multiplica el gasto. Ajusta los dos a lo que la tarea de verdad pide.

CÓMO HACERLO

1. No corras todo en paralelo por costumbre. Encadena lo que puede ir en orden.
2. Baja el nivel de effort o thinking en tareas simples.
3. Sube a máximo solo cuando el problema lo amerita.

REGLA

Paralelo y effort alto:

solo cuando el problema lo pide.

ANTES DE CADA RUN

Checklist para no quemar crédito

MODELO

¿Es tarea de Opus, o la hace Sonnet o Haiku?

Subagentes con el modelo barato fijado

CONTEXTO

Solo los archivos que la tarea necesita

Conversaciones largas resumidas

Bloque fijo marcado para cache

EJECUCIÓN

Paralelo solo si de verdad hace falta

Effort acorde a la dificultad

SIGUIENTE PASO

Que rinda el doble.

El cambio de Anthropic no tiene que costarte más, si sabes dónde se va el crédito. Aplica estos cuatro comparaciones y sistemas que puedes usar el mismo día. Más recursos para tu negocio en El que mide su crédito rinde el doble que el que lo gasta a ciegas.

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