

Agentes en Claude sin quemar el credito.

CATEGORÍA · AGENTES Y AUTOMATIZACIÓN

Lo que cambia el 15 de junio de 2026 y como optimizar tu uso de OpenClaw, Hermes y cualquier agente externo antes de que te llegue la factura.

QUÉ VAS A ENCONTRAR

- 01 Los dos buckets de facturación y que entra en cada uno
- 02 Cuanto cubre tu crédito según tu plan (el math real)
- 03 5 estrategias para usar más tokens sin gastar más
- 04 La zona gris que Anthropic aún no ha documentado

FICHA DEL RECURSO

Qué es, para qué sirve y cómo se implementa

FUNCIÓN

Lo que cambia el 15 de junio de 2026 y como optimizar tu uso de OpenClaw, Hermes y cualquier agente externo antes de que te llegue la factura.

PARA QUÉ SIRVE

Antes de abril de 2026, tu suscripción de Claude cubría todo: el chat, Claude Code y los tokens que consumían tus agentes externos. OpenClaw, Hermes o cualquier app construida sobre el Agent SDK corrían sobre el mismo crédito. No pagabas aparte. Algunos usuarios pagaban \$20 al mes y consumían miles de dólares en tokens a través de esos agentes. Anthropic absorbía la diferencia. En abril, Anthropic bloqueó el acceso por completo: las suscripciones dejaron de cubrir agentes externos. Ahora los reactivaron, pero con una condición nueva. Desde el 15 de junio, la facturación se divide en dos buckets que no se mezclan. El primero cubre tu uso interactivo: chat, Claude Code en terminal, Cowork. El

DÓNDE APLICARLO

Cubre: Los dos buckets de facturación y que entra en cada uno; Cuanto cubre tu crédito según tu plan (el math real); 5 estrategias para usar más tokens sin gastar más; La zona gris que Anthropic aún no ha documentado.

CÓMO IMPLEMENTARLO

- 1 01 / El Cambio: Desde el 15 de junio, hay dos buckets.
- 2 02 / El Math Real: Cuanto cubre tu crédito exactamente.
- 3 03 / Optimización: 5 movimientos antes del 15 de junio.
- 4 04 / La Zona Gris: Lo que Anthropic aún no ha documentado.

01 / EL CAMBIO

Desde el 15 de junio, hay dos buckets.

Antes de abril de 2026, tu suscripción de Claude cubría todo: el chat, Claude Code y los tokens que consumían tus agentes externos. OpenClaw, Hermes o cualquier app construida sobre el Agent SDK corrían sobre el mismo crédito. No pagabas aparte. Algunos usuarios pagaban \$20 al mes y consumían miles de dólares en tokens a través de esos agentes. Anthropic absorbía la diferencia. En abril, Anthropic bloqueó el acceso por completo: las suscripciones dejaron de cubrir agentes externos. Ahora los reactivaron, pero con una condición nueva. Desde el 15 de junio, la facturación se divide en dos buckets que no se mezclan. El primero cubre tu uso interactivo: chat, Claude Code en terminal, Cowork. El segundo es exclusivo para uso programático: agentes y apps de terceros. Si uno se agota, no puede tomar del otro.

LOS DOS BUCKETS

BUCKET 1 - Suscripción normal (sin cambios):

Chat Claude.ai, Claude Code interactivo, Cowork

BUCKET 2 - Agent SDK credits (nuevo):

claude -p, OpenClaw, Hermes, Conductor, Zed,

Jean, Claude Code GitHub Actions, Agent SDK propio

No se acumulan. No se mezclan. No hay rollover.

02 / EL MATH REAL

Cuanto cubre tu credito exactamente.

El crédito Agent SDK equivale al precio mensual de tu plan y se cobra a precios de API reales. Pro recibe \$20, Max 5x recibe \$100, Max 20x recibe \$200. Si los agotas antes del fin de mes, tus trabajos en background se detienen. No hay continuidad automática. El modelo que eliges para tus agentes determina cuánto dura ese crédito. Opus 4.7 además tokeniza entre 32% y 47% más que versiones anteriores para el mismo texto, lo que acelera el consumo más de lo que parece.

TOKENS POR \$200 DE CREDITO (MIX 50/50)

Claude Opus 4.7: ~13.3M tokens

Claude Sonnet 4.6: ~22M tokens (+65%)

Claude Haiku 4.5: ~67M tokens (+5x)

Nota: Opus 4.7 tokeniza 32-47% mas que versiones previas para el mismo texto.

03 / OPTIMIZACION

5 movimientos antes del 15 de junio.

El objetivo no es usar menos. Es que cada token haga más trabajo. Estos cinco ajustes reducen el consumo sin tocar la calidad del output.

LOS 5 MOVIMIENTOS

1. Audita tu consumo actual. Calcula cuántos tokens quemas por mes a precios API. Si es menos de tu crédito, probablemente estás mejor que antes.
2. Baja el modelo cuando no necesitas Opus.
Routing, clasificación y resumen: Sonnet o Haiku.
Reserva Opus para razonamiento complejo real.
3. Comprime el contexto. Si tu agente carga el mismo prompt largo en cada turno, pagas esos tokens repetidamente. System prompts cortos, contexto relevante al turno solamente.
4. Activa prompt caching. Si tu agente repite el mismo system prompt, el caching reduce hasta 90% el costo de esos tokens. Disponible vía API.
5. Configura el toggle de overage con un cap.
Viene desactivado por defecto. Si lo activas sin límite, la factura puede sorprenderte.

04 / LA ZONA GRIS

Lo que Anthropic aun no ha documentado.

Hay tres casos que Anthropic no ha clarificado oficialmente al momento de publicar esta guía. Si tu flujo depende de alguno de estos, espera la documentación antes de asumir que entra en el bucket interactivo. Si gastas sistemáticamente más de tu crédito mensual, hay dos alternativas concretas. Cursor Ultra cuesta \$200 por mes e incluye \$400 en crédito SDK, el doble que Max 20x por el mismo precio. Las rutas híbridas combinan Claude para razonamiento con Haiku o GPT-4o-mini para tareas repetitivas de bajo costo.

SIN CONFIRMAR POR ANTHROPIC

- Hooks ejecutados desde sesiones interactivas
 - > no se sabe a qué bucket van
 - Subagentes lanzados desde sesiones interactivas
 - > pendiente de aclaración oficial
 - Herramientas MCP
 - > sin documentación aún
- Revisa el Help Center de Anthropic antes del 15/06.

SIGUIENTE PASO

Sígueme en

Publico guías como esta cada semana. Todo lo que comparto lo pruebo en mis propios proyectos y con mis clientes. Si esto te fue útil, compártelo con alguien que use agentes de Claude.

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