— GUÍA DE MODELOS

4 modelos de Claude *explicados.*

→ Haiku, Sonnet, Opus y Fable — cuándo usar cada uno.

→ Y cómo ponerlos a trabajar juntos en un proyecto.

La guía para dejar de elegir al azar y empezar a elegir con criterio. Menos tokens quemados, mejores resultados.

AUTOR

Jorge Riquelme

Fundador · Otherway Labs

SOBRE

Modelos Claude

Haiku · Sonnet · Opus · Fable

MODELOS

4

— INTRODUCCIÓN

Elegir mal *cuesta caro.*

Abres Claude y ves un selector con cuatro nombres: Haiku, Sonnet, Opus y Fable. La mayoría elige uno al azar y se queda ahí para siempre. Ese es el primer error.

Cada modelo tiene un punto dulce. Usar el más caro para todo te quema el plan. Usar el más barato para todo te entrega resultados mediocres.

Acá va mi criterio real: qué modelo uso para cada cosa, cuándo salto de uno a otro, y cómo los combino en un mismo proyecto para gastar menos y llegar más lejos.

— EN ESTA GUÍA

01 Haiku — el que casi nunca uso 03
02 Sonnet — el caballo de batalla 03
03 Opus — el que razona y programa 04
04 Fable — el que sienta las bases 04
05 Cómo ponerlos a trabajar juntos 05
06 El caso real (y cuánto cuesta) 06
07 Checklist de decisión rápida 07

Modelos	Lectura	Nivel
4	7 min	Base

01

MODELO 01 · VELOCIDAD

Haiku, el *más barato.*

Rápido y liviano. Yo casi no lo toco.

Haiku es el modelo más rápido, más simple y más barato de la familia. Responde casi al instante y consume una fracción de lo que gastan los demás.

El problema no es el precio, es el techo. Le falta razonamiento para casi todo lo del día a día. Yo personalmente no lo ocupo nunca — ni siquiera para las tareas chicas.

→ TIP Reserva Haiku para **volumen**: clasificar cientos de mensajes, etiquetar, extraer datos simples. Tareas donde la velocidad importa más que el criterio.

02

MODELO 02 · DÍA A DÍA

Sonnet, tu *default.*

El caballo de batalla del 90% del día.

Sonnet es mi modelo de las tareas diarias. Lo uso para lo pequeño y lo repetitivo: redactar, resumir, ordenar, responder. Todo lo que no pide mucho razonamiento.

Es el equilibrio entre rapidez, costo y calidad. Si una tarea no te hace dudar de si el modelo "va a entender", Sonnet basta y sobra. No pagues Opus por esto.

→ TIP Haz de Sonnet tu **default**. Empieza toda tarea acá y sube de modelo solo cuando el resultado te quede corto — no antes.

03

MODELO 03 · RAZONAMIENTO

Opus, el que *piensa.*

Para lo complejo, lo que va paso a paso.

Cuando necesito mayor razonamiento, salto de inmediato a Opus. Programar, resolver algo enredado, planear una tarea larga en varios pasos: ese es su terreno.

Rinde mucho más que Sonnet en problemas difíciles, y cuesta más. Por eso no es para todo. Es para el momento en que Sonnet empieza a equivocarse o a quedarse corto.

→ TIP Sube a Opus cuando la tarea tenga **varios pasos encadenados** o cuando Sonnet ya falló dos veces en lo mismo. Ahí el costo extra se paga solo.

04

MODELO 04 · POTENCIA

Fable, el *más poderoso.*

El último de Anthropic. Úsalo con cuidado.

Fable es el modelo más nuevo y más potente que sacó Anthropic — también el más caro. Entra cuando Opus no llega al resultado o cuando estás iterando demasiado sin avanzar.

Yo lo uso para arrancar proyectos de programación: solo la partida, para que siente las bases de forma analítica. Después me salto a Opus o Sonnet para ejecutar y gastar menos.

→ TIP No dejes Fable corriendo todo el proyecto. Úsalo para **diseñar la base**, luego baja de modelo. Es un bisturí, no un martillo.

— EL SISTEMA

Ponlos a trabajar *en conjunto.*

El truco no es elegir un modelo. Es usar varios en el mismo proyecto, cada uno en su fase. Aprovechas la potencia donde importa y ahorras donde no.

— EL FLUJO DE UN PROYECTO

Fase 01 → Fable

Sienta las bases

Arranca el proyecto. Define arquitectura y plan de forma analítica. Solo la partida.

Fase 02 → Opus

Ejecuta lo difícil

Construye las partes complejas paso a paso, con las bases ya puestas.

Fase 03 → Sonnet

Mantiene el ritmo

Cierra lo repetitivo y los ajustes. Gastas una fracción del total.

— SEÑALES PARA CAMBIAR DE MODELO

Sube de modelo cuando...

- El modelo se equivoca dos veces en lo mismo.
- La tarea tiene varios pasos encadenados.
- Arrancas algo de cero y necesitas criterio estructural.

Baja de modelo cuando...

- Ya está la arquitectura y solo toca ejecutar.
- La tarea es repetitiva o mecánica.
- Necesitas velocidad, no criterio.

→ REGLA DE TOKENS

Nunca uses el modelo más caro para lo que el más barato resuelve igual. La potencia se **reserva**, no se derrocha.

— LA EVIDENCIA

Esto no es teoría. Ya lo *probaron.*

Lo de repartir el trabajo entre modelos no me lo inventé yo. Anthropic lo usa en su propio producto — y midió el resultado.

+90%

más efectivo que Opus en solitario

Opus dirige, *Sonnet ejecuta.*

Anthropic armó su sistema de investigación con esta misma lógica: un modelo potente —Claude Opus— dirige y reparte el trabajo, y varios modelos más baratos —Claude Sonnet— lo ejecutan en paralelo. Esa combinación superó a Opus trabajando solo en un **90,2%** en su evaluación interna. Es una prueba de Anthropic sobre tareas de investigación, no un benchmark público. Pero deja algo claro: repartir rinde más que cargarle todo al modelo caro.

15×

más tokens que una conversación

Pero no sale *gratis.*

El mismo equipo midió que un esquema así consume cerca de **15 veces** más tokens que una conversación normal de chat. Por eso no lo apliques a todo. Conviene cuando la tarea vale el gasto: proyectos grandes, con partes que se reparten en paralelo, o información que no cabe en una sola conversación. Si la pega es chica, un solo modelo bien elegido te sale más barato y más simple.

Fuente — Anthropic Engineering · «How we built our multi-agent research system» · 2025 · anthropic.com/engineering/multi-agent-research-system

— ÚLTIMO PASO

Ahora elige con *criterio.*

La próxima vez que abras Claude, no elijas por costumbre. Mira qué tan difícil es la tarea que tienes al frente y elige el modelo que le corresponde.

Empieza barato, sube solo cuando el problema lo pida, y baja apenas puedas. Ese ritmo es la diferencia entre quemar tu plan y hacerlo rendir el triple.

— CHECKLIST DE DECISIÓN RÁPIDA

01 ¿Repetitivo y simple? → Sonnet, tu default.

02 ¿Necesita razonar o programar? → Opus.

03 ¿Arrancas un proyecto nuevo? → Fable, solo la partida.

04 ¿Puro volumen y velocidad? → Haiku.

05 ¿El modelo falló dos veces? → sube un nivel.

06 ¿Ya está la base y toca ejecutar? → baja un nivel.



— SOBRE EL AUTOR

Jorge *Riquelme.* Fundador · Otherway Labs

Ayudo a empresas y equipos a usar Claude todos los días — de la teoría a procesos que corren solos. Cada semana publico guías, MCPs y skills en **otherwaylabs.cl**.

— PRÓXIMOS PASOS

Tres formas *de seguir avanzando.*

— PARA EMPRESAS
Sesión gratuita

¿Hay procesos en tu empresa que *deberían correr solos*?

Agenda un diagnóstico y te muestro dónde la IA te ahorra horas cada semana.

Agenda en
otherwaylabs.cl →

— PARA EQUIPOS
Capacitación

Forma a tu equipo para que *use IA todos los días*.

Capacitaciones prácticas de Claude, con casos reales de tu propia operación.

Escríbeme en
otherwaylabs.cl →

— ACCESO ANTICIPADO
Grupo WhatsApp · Gratis

Recibe estos recursos *antes que nadie*.

Grupo exclusivo donde publico cada guía, MCP y skill antes de subirla a YouTube o Instagram. Sin esperar a que salgan los videos.

Únete en
chat.whatsapp.com →