DIAGNÓSTICO // 01

El coste real de Claude Code

Anatomía financiera de un agente de IA:
Suscripción mensual vs. Consumo crudo por API.



API (Consumo) // Tokens/mes	→	SUBSCRIPCIÓN // Claude Pro (Plan)

Promedio: 12.000.000 // Coste estimado: € 1.932,23
[Token de Entrada + Token de Salida]
// Alto volumen, sin límite.
		/ Coste fijo: € 91,00 / mes
		// Uso prioritario, acceso extendido
		// Optimización de recursos.

Gemini Notebook

Evidencia Empírica: 15 Días de Auditoría

El Experimento

- Duración: 15 días activos
- Intensidad: 135 horas de uso
- Modelo: Opus 5.5
- Volumen: 20.710 peticiones

Coste Teórico en API

1.932,23 €

Equivale a una media de 14,31 € por hora de trabajo interactivo.

Coste Real Abonado

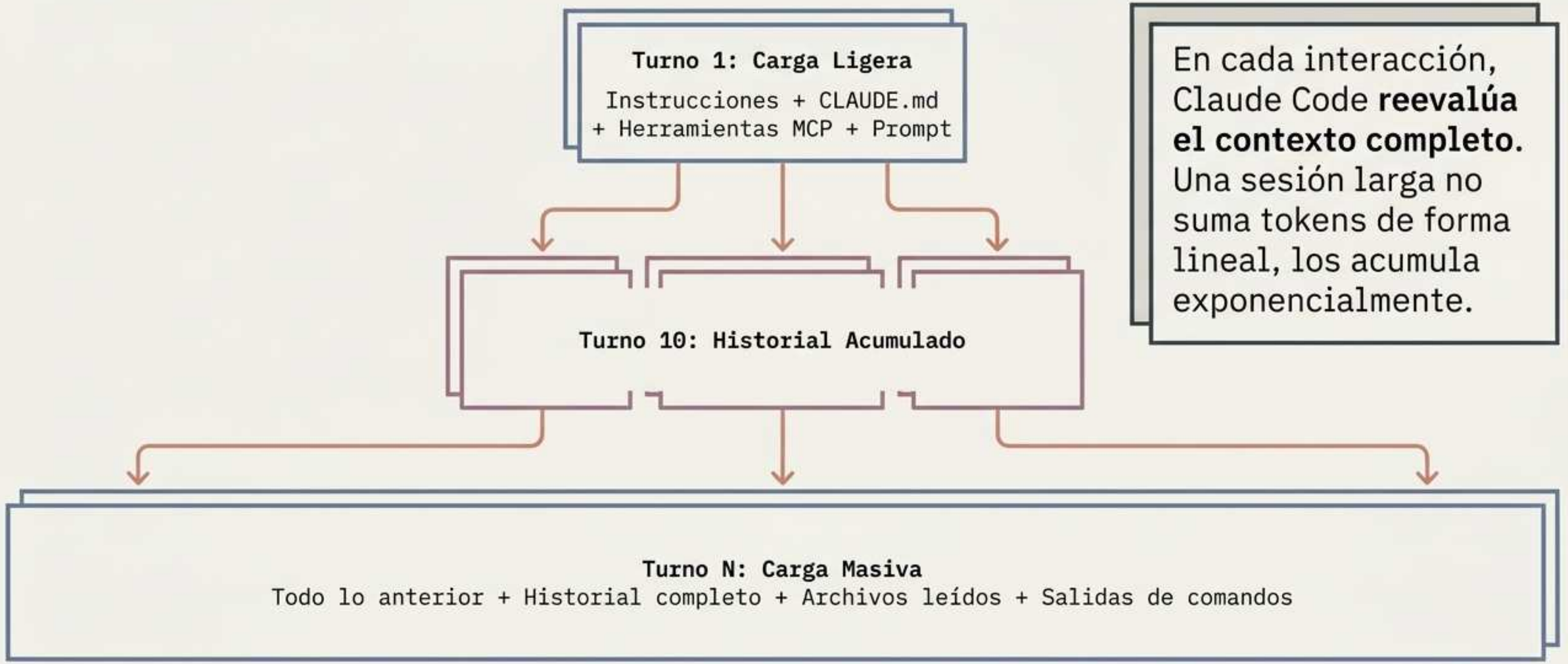
110,65 €

Plan Max (IVA 21% incluido).

// El uso crudo de la API habría multiplicado por 21x la cuota mensual del plan Max en solo dos semanas.

Gemini Notebook

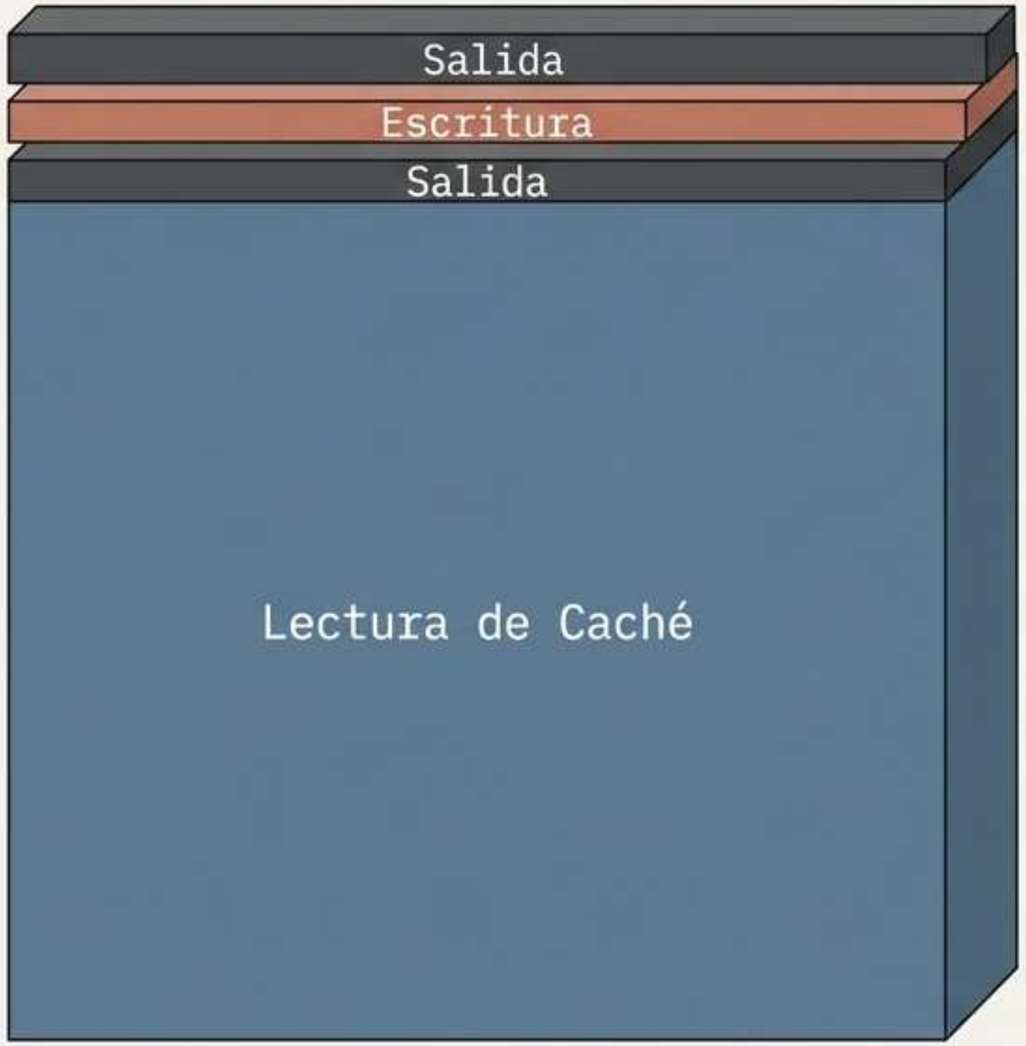
La paradoja del agente: pagas por la memoria, no por el mensaje



En cada interacción, Claude Code **reevalúa el contexto completo.** Una sesión larga no suma tokens de forma lineal, los acumula exponencialmente.

Gemini Notebook

El salvavidas arquitectónico: Prompt Caching



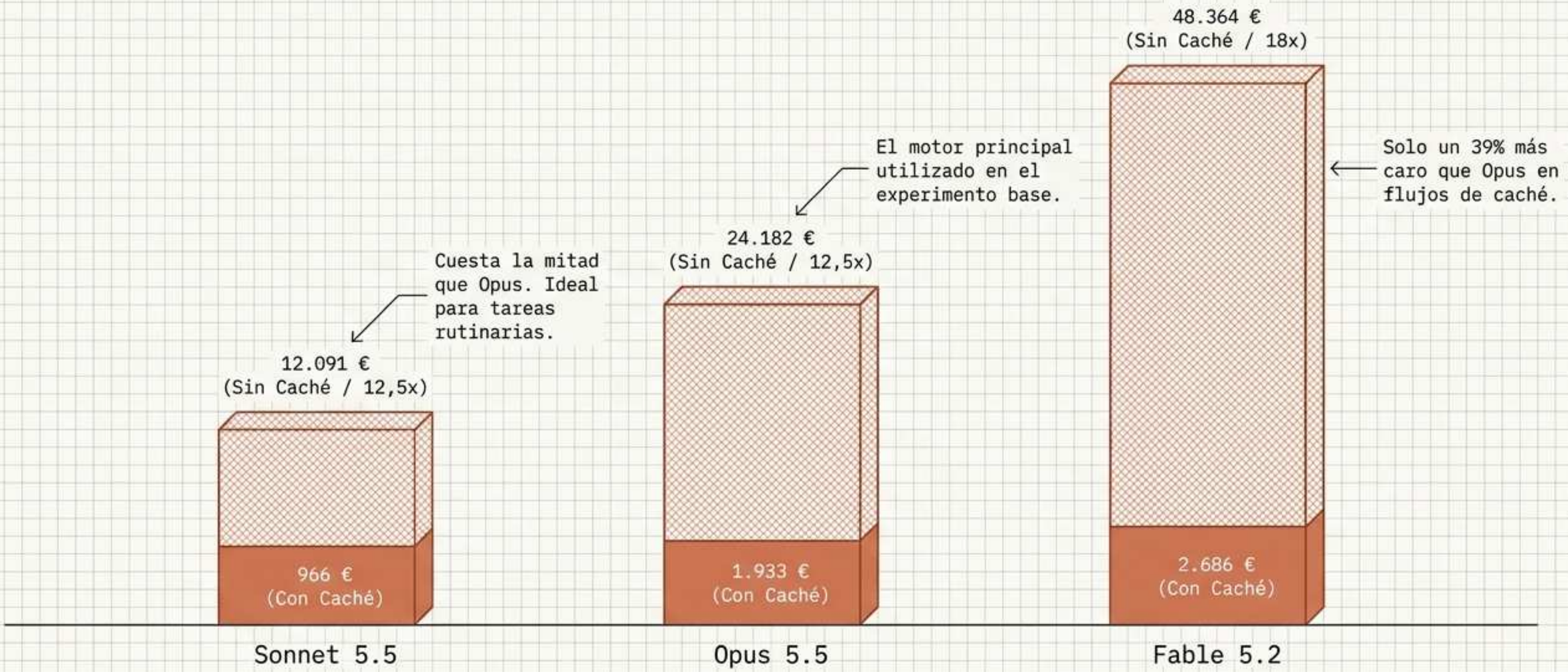
Lectura de Caché	6.446 millones de tokens	60,6% del coste
Escritura en Caché	71,6 millones de tokens	21,7% del coste
Salida del Modelo	18,6 millones de tokens	17,6% del coste
Entrada sin Caché	<0,1 millones	0,01% del coste

Casi dos tercios del coste provienen de releer el pasado. Sin un sistema de caché que cobre esta lectura al 5% de su valor nominal, la IA interactiva sería inviable.

Gemini Notebook

Simulación de modelos en un escenario Sin Caché

La viabilidad de tu agente depende de la arquitectura de la caché; apagarla multiplica tu factura hasta por 18x.



Gemini Notebook

Matriz de Diagnóstico de Suscripciones (España, 21% IVA incl.)

Pro	Max	Ultra
22,13 €/mes (Anual: 18,44 €/mes)	110,65 €/mes	221,30 €/mes
• Capacidad base (1x) • Incluye Opus y Sonnet	• Capacidad 5x • Añade modelo Fable • Prioridad de red • 91 €/mes en créditos API	• Capacidad 10x • Elimina ventana de 5h • 183 €/mes en créditos API

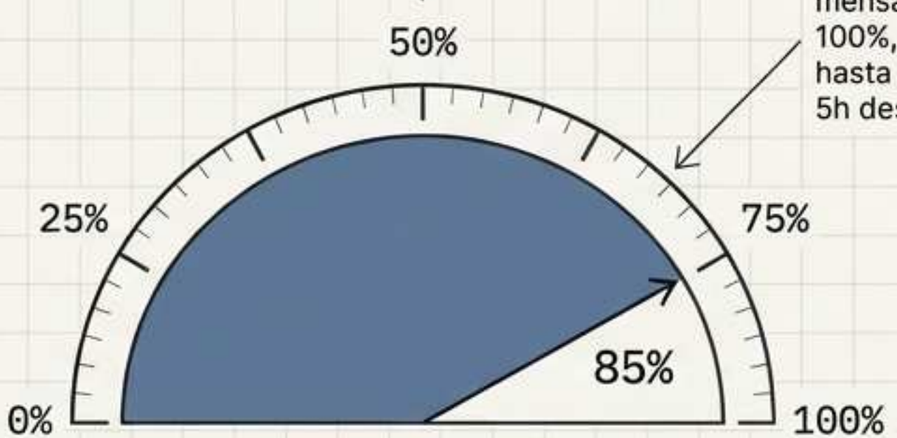
Aviso de Créditos: Los créditos de API incluidos en Max/Ultra son para desarrollos externos. No sirven para ampliar los límites del agente Claude Code cuando se agota la cuota interactiva.

Gemini Notebook

Anatomía de los límites ocultos

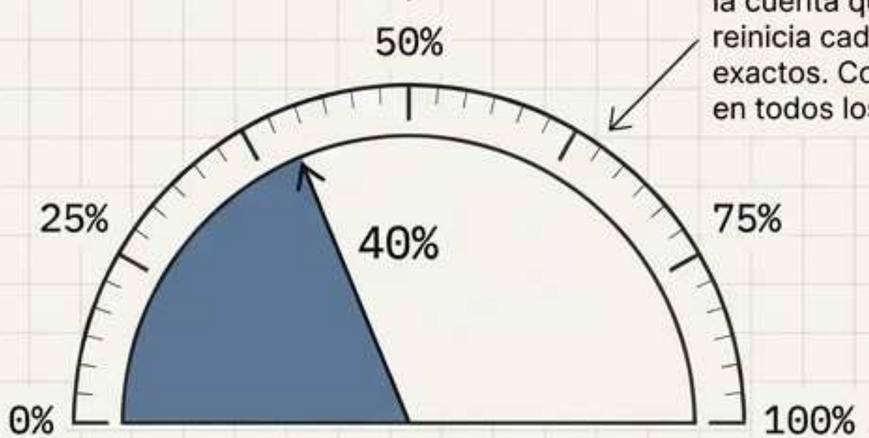
Origen del Tráfico

claude.ai (Web) + Claude Code (Terminal)



Se activa con el primer mensaje. Al llegar al 100%, hay bloqueo hasta que transcurran 5h desde el inicio.

La Ventana de 5 Horas



Contador global de la cuenta que se reinicia cada 7 días exactos. Compartido en todos los planes.

El Límite Semanal

REGLA CRUCIAL: claude.ai_web_chat + claude_code_cli = 1_shared_limit.
El uso en tu navegador consume tu presupuesto de terminal.

Gemini Notebook

Traducción de límites a horas reales de ingeniería

Métrica base: Consumo real intensivo de 14,31 €/hora

Plan Pro	Valor API: ~255 €/sem | Soporta ~18 horas/semana



Veredicto: La ventana de 5h se agota tras ~3 horas continuas. Insuficiente para jornadas completas.

Plan Max	Valor API: ~1.273 €/sem | Soporta ~89 horas/semana



Veredicto: Rara vez se agota en una jornada ininterrumpida. El estándar para profesionales.

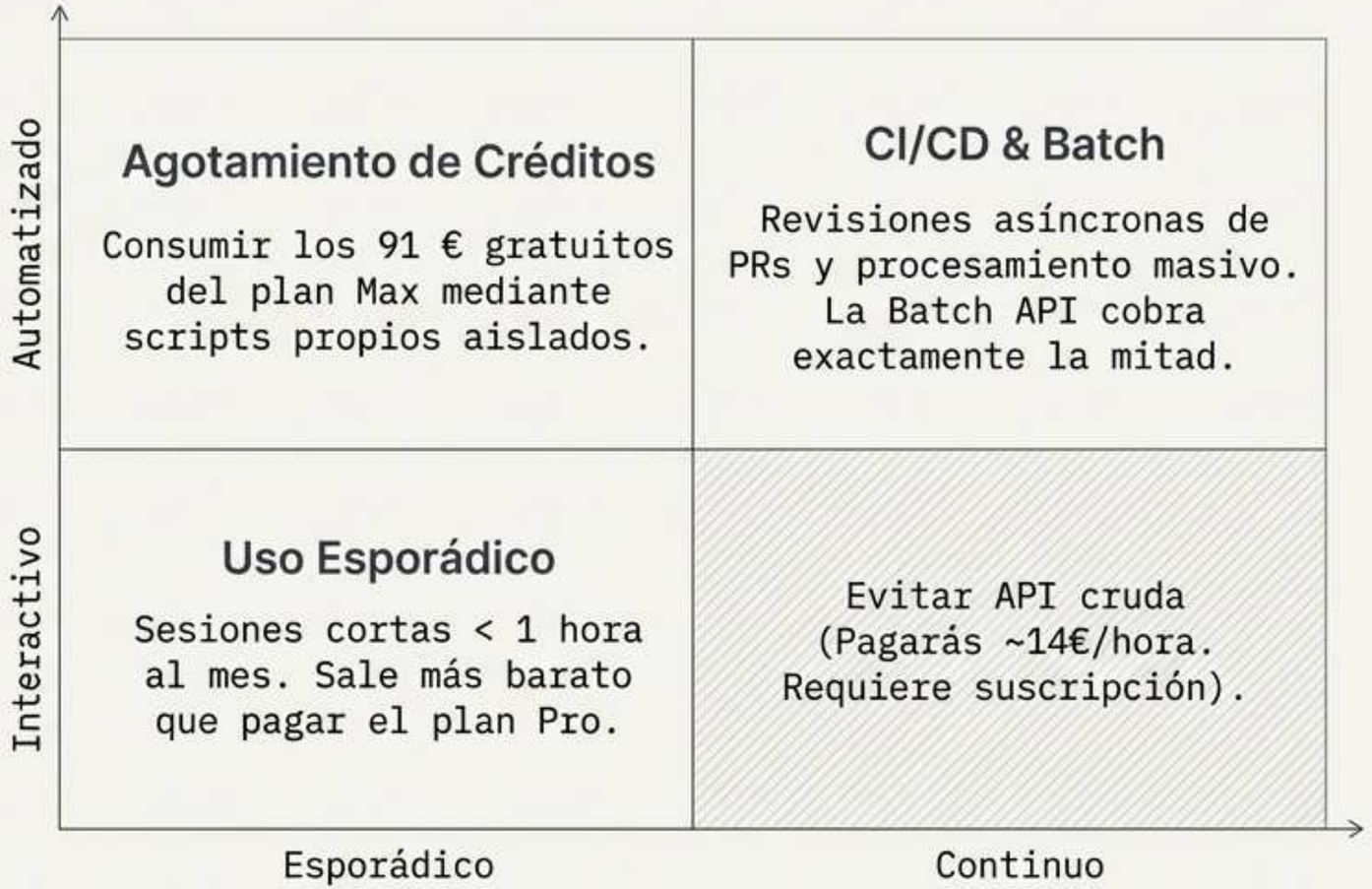
Plan Ultra	Valor API: ~2.546 €/sem | Soporta ~178 horas/semana



Veredicto: Sin cortes intradiarios. Solo necesario para flujos de múltiples agentes paralelos.

Gemini Notebook

Cuándo romper la regla: El caso de uso para la API cruda

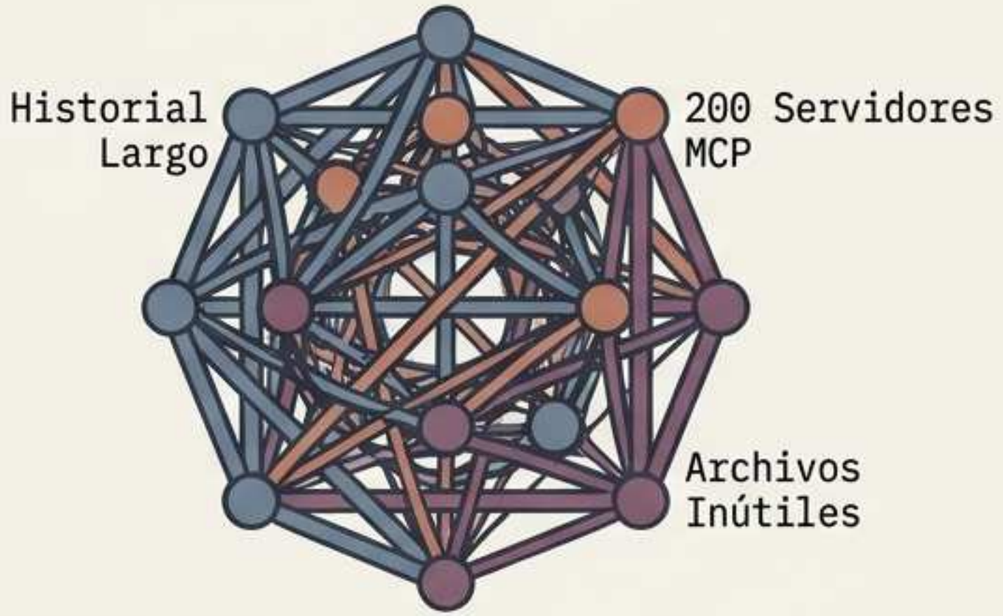


sudo alert: Peligro con ANTHROPIC_API_KEY. Si esta variable está expuesta en tu entorno, Claude Code facturará a tu tarjeta saltándose el plan de pago. Verifica con /usage.

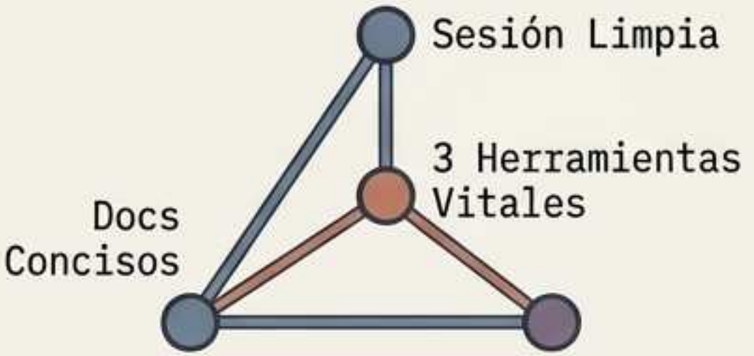
Gemini Notebook

Playbook Estrategia #1: Poda agresiva de contexto

Carga de Contexto Acumulativa



Carga Optimizada



Comando /clear

Limpia la sesión al cambiar de tarea. En auditorías reales, previene el 83% del gasto innecesario de límite.

Micro-Sesiones

Evita sesiones activas donde el turno 100 relee los 99 anteriores.

Archivos CLAUDE.md

Cada línea sobrante se cobra cientos de veces por sesión. Mantén la documentación dividida y concisa.

Dieta MCP

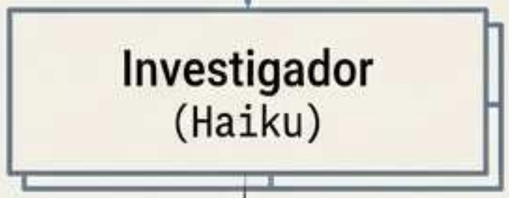
Cargar cientos de herramientas añade un catálogo completo a todas y cada una de las peticiones. Carga solo lo vital.

Gemini Notebook

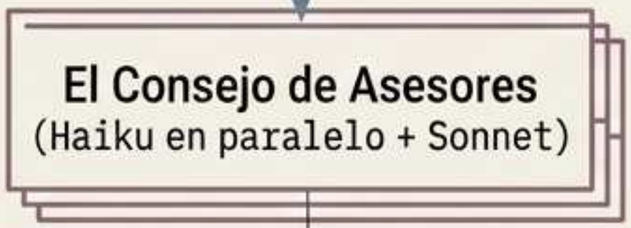
Playbook Estrategia #2: Topología de Subagentes



Toma decisiones arquitectónicas.
No ejecuta tareas de alto volumen para aislar la basura del contexto.



Delegación de operaciones locuaces.
Lectura masiva de logs a bajo coste.

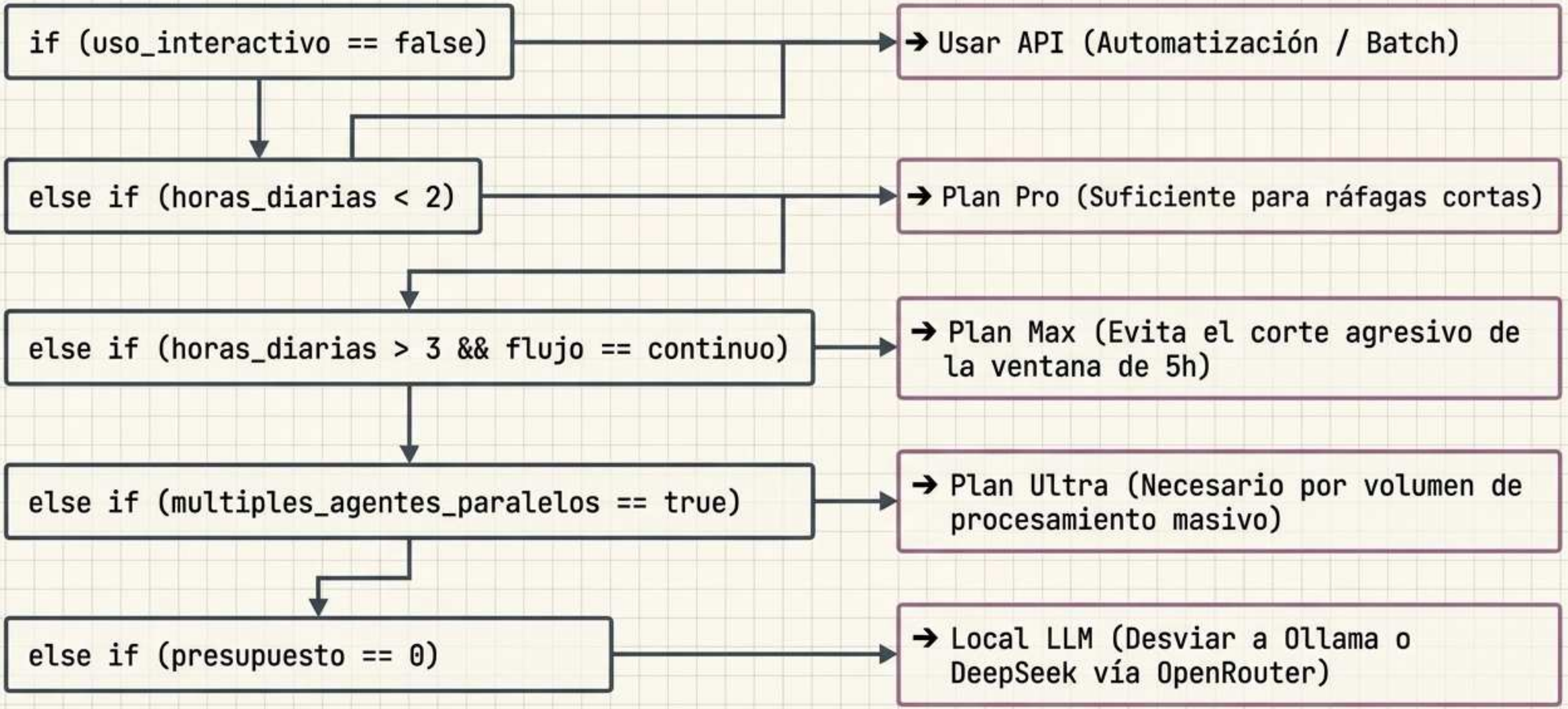


Múltiples subagentes revisan decisiones complejas y las resumen
por apenas 0,11 € el ciclo, sin quemar límite premium.

La ingeniería de agentes consiste en explotar modelos asimétricos. No uses el modelo más pesado para la tarea más ligera.

Gemini Notebook

El Algoritmo de Decisión (Síntesis)



Gemini Notebook

Toolkit del Desarrollador (Dependencias y Auditoría)

`ccusage [Open Source]`

Herramienta que lee el directorio local `~/.claude/projects` para calcular el coste exacto por día, sesión y modelo directamente en tu terminal.

`Configuración de Línea de Estado`

Scripts para inyectar el coste en tiempo real o el porcentaje de uso de la cuota (sin abrir /usage) en tu interfaz `CLI`.

`Documentación de Anthropic`

Guías oficiales sobre Prompt Caching y metodologías de reducción de tokens para agentes locales.

```
// Monitoriza tu uso, poda tu contexto, protege tu límite.
```

Gemini Notebook