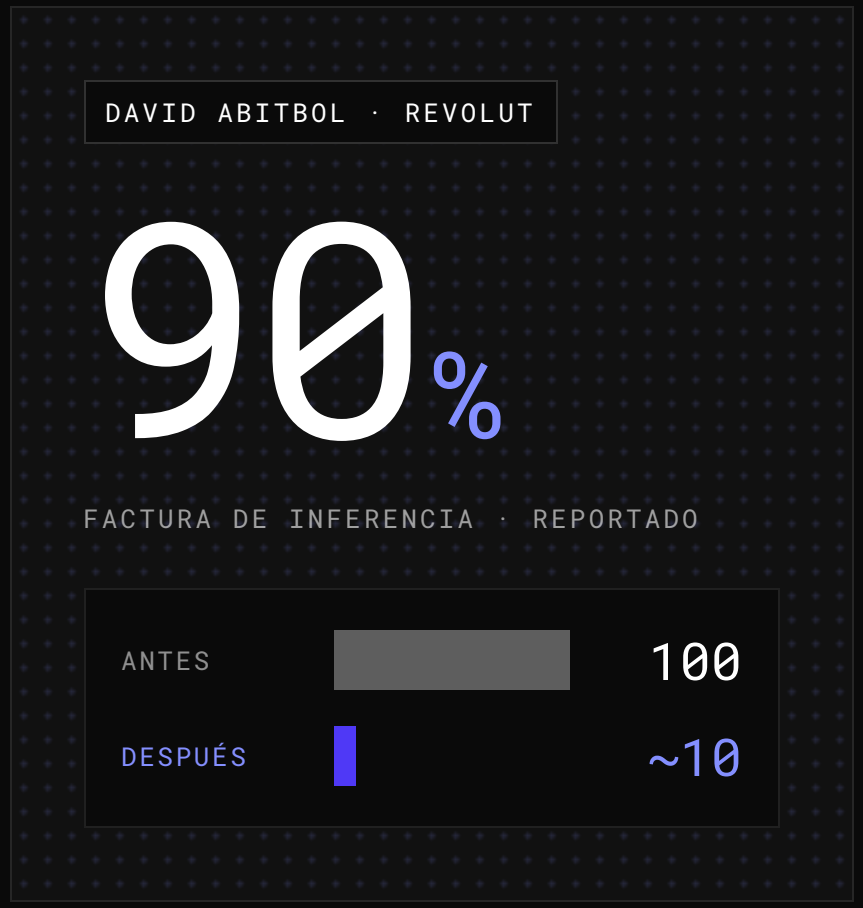
Helmcode

// CASO REVOLUT · MODELOS ABIERTOS

Cómo Revolut redujo **un 90 % su factura de inferencia** para programación con IA.

Fuente: David Abitbol (Revolut), AI Engineer NYC, 13 de octubre de 2026.



Los costes limitaban el uso de la IA.
Cambiaron el modelo por defecto.

01 · HERRAMIENTAS

Asistentes y agentes de coding

→

02 · GASTO

1.000 a 10.000 $

AL MES POR INGENIERO, CASOS CITADOS

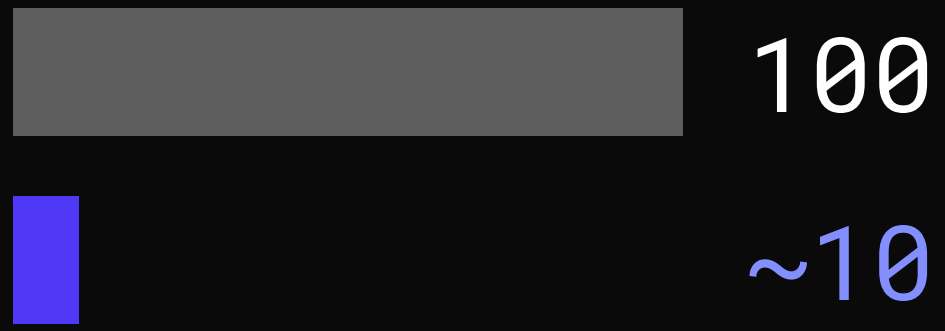
→

03 · RESPUESTA

Límites de gasto

ANTES	APIs propietarias por defecto	→	Límites de uso
DESPUÉS	**GLM, modelo abierto por defecto**	+	Propietarios cuando hacen falta

Una décima parte de la factura anterior.



ÍNDICE · FACTURA ANTERIOR = 100 · FACTURA REPORTADA, NO COSTE TOTAL

2.000
ingenieros en el despliegue descrito

GLM
modelo abierto por defecto para coding

56 B200
capacidad dedicada, servida con Together AI

Fuente: David Abitbol (Revolut), AI Engineer NYC, 13 de octubre de 2026.

Tu empresa no necesita la infraestructura de Revolut.

56 NVIDIA B200

REVOLUT

Una carga de trabajo real

TU EMPRESA

TU EQUIPO O TU CONSULTORA: INTEGRA

HELMCODE: SIRVE LA INFERENCIA

Tu aplicación → API compatible con OpenAI → Inferencia gestionada → Modelos abiertos, uno por tarea

SIN OPERAR GPUS

PROCESAMIENTO EN LA UE

SIN RETENCIÓN DE PROMPTS

Antes de migrar, se mide con una carga real.

01

Elige

Una operación con volumen: soporte, RAG, documentos o código.

02

Línea base

Coste, calidad y latencia de hoy, con casos reales.

03

Compara

Los mismos casos con modelos abiertos, en las mismas condiciones.

04

Decide

Si el coste por tarea aceptada no baja, no se migra.

COSTE POR TAREA ACEPTADA

`coste completo ÷ tareas aceptadas`

QUÉ APORTAS

- Volumen de uso y factura actual de esa carga.
- Qué cuenta como tarea aceptada.
- Una muestra de casos, anonimizada.

QUÉ RECIBES

- Calidad, latencia y coste, con sus fuentes.
- Una recomendación: mantener, combinar o PoC.
- También por escrito si no compensa.

METODOLOGÍA COMPLETA

helmcode.com/es/guias/optimizacion-costes-ia

Nueve modelos abiertos. Una sola API.

LENGUAJE · EL MODELO DE REVOLUT **glm5.3** Coding agéntico de horizonte largo	LENGUAJE **glm5.3-flash** Agentes de código, contexto de 1M	LENGUAJE **deepseek-v4-flash** Razonamiento y agentes
LENGUAJE **qwen3.6** RAG de alto volumen y clasificación	LENGUAJE **gemma4** Asistentes y documentos	EMBEDDINGS **qwen3-embedding** Recuperación en más de 100 idiomas
RERANKING **rerank** Ordena los resultados de RAG	VOZ A TEXTO **whisper** Llamadas, reuniones y audio	TEXTO A VOZ **kokoro** Agentes de voz en tiempo real

Prueba un modelo **abierto** con una carga real de tu empresa.

METODOLOGÍA COMPLETA

helmcode.com/es/guias/optimizacion-costes-ia

CALCULA TU ESCENARIO

helmcode.com/es/roi-calculator

TU EQUIPO O TU CONSULTORA INTEGRA · HELMCODE SIRVE LA INFERENCIA