INSTALA CLAUDE CODE LOCAL GRATIS

GUÍA PASO A PASO: SIN LÍMITES USO, SIN APIS Y GRATIS

HÉCTOR JAYAT

Instala Claude Code Local Gratis

Guía completa

Por Héctor Jayat

Guía paso a paso para usar Claude Code gratis en local con Ollama, sin API, sin nube y con modelos open source

Puntos Destacados

- Puedes usar Claude Code en local conectándolo a Ollama, eliminando costes de API y dependencia de la nube
- El sistema funciona usando modelos open source como Qwen o Gemma ejecutados directamente en tu máquina
- El rendimiento y la calidad dependen del hardware, especialmente de la RAM disponible

Tabla de Contenido

Qué Significa Realmente Usar Claude Code en Local

En las últimas semanas ha empezado a circular una alternativa muy interesante: usar **Claude Code en local, sin pagar API y sin depender de servidores externos**.
Y sí, se puede hacer. Pero no exactamente como muchos lo están contando.

En esta guía te explicamos paso a paso cómo hacerlo de verdad, qué está pasando por debajo y qué puedes esperar en la práctica.

Antes de entrar en la guía, hay algo importante que debemos dejar claro porque aquí es donde mucha gente se confunde.

Claude Code, como producto oficial de Anthropic, no está diseñado para ejecutarse completamente en local con sus propios modelos. Por defecto, siempre se conecta a sus servidores.

Entonces, ¿qué estamos haciendo aquí?

Lo que hacemos es aprovechar que Claude Code permite conectarse a endpoints compatibles con su API. Y aquí es donde entra en juego Ollama, que actúa como un servidor local capaz de ejecutar modelos open source en tu propia máquina.

En otras palabras:

- Claude Code → interfaz inteligente que actúa como agente
- Ollama → servidor local que responde como si fuera una API
- Modelo open source → el verdadero "cerebro"

Esto cambia completamente el paradigma: dejamos de pagar por uso y pasamos a depender de nuestro hardware.

Paso 1 Instalar Ollama el Motor Local

El primer paso es instalar Ollama, que será el componente encargado de ejecutar modelos de lenguaje en local.

La instalación es muy sencilla y está disponible para macOS, Windows y Linux. Solo tienes que descargarlo desde su web oficial e instalarlo como cualquier otra aplicación.

Una vez instalado, Ollama se ejecuta automáticamente en segundo plano. No necesitas abrir ninguna interfaz gráfica ni hacer configuraciones complejas. Esto es importante porque convierte tu ordenador en un pequeño servidor de IA sin que tengas que preocuparte por ello.

Desde nuestra experiencia, este es uno de los puntos fuertes de Ollama: elimina completamente la fricción técnica que antes implicaba trabajar con modelos locales.

Paso 2 Comprobar Que Ollama Está Activo

Después de instalarlo, necesitamos asegurarnos de que todo está funcionando correctamente.

Puedes hacerlo de dos formas muy simples:
Abriendo en el navegador:

http://localhost:11434

O desde terminal:
ollama list

Si todo está correcto, Ollama responderá aunque todavía no tengas ningún modelo descargado.

Este detalle es importante porque ese endpoint (localhost:11434) será el que usaremos más adelante para conectar Claude Code a tu máquina.

En esencia, aquí estamos verificando que ya tienes tu "API local" lista.

Paso 3 Descargar un Modelo Para Programar

Aquí es donde realmente empieza lo interesante. Necesitamos descargar un modelo que actúe como sustituto de Claude.

Dependiendo de la potencia de tu equipo, puedes elegir entre varias opciones:

Para equipos potentes:

- qwen3-coder:30b

Para equipos medios:

- qwen2.5-coder:7b

Para equipos más modestos:

- gemma:2b

La descarga se hace desde terminal con un solo comando:
ollama pull NOMBRE_DEL_MODELO

Por ejemplo:
ollama pull qwen2.5-coder:7b

Este proceso puede tardar bastante dependiendo del tamaño del modelo y tu conexión.
Aquí hay algo que conviene entender bien: cuanto más grande el modelo, mejor será la calidad de las respuestas, pero

también mayor será el consumo de RAM y CPU (o GPU si tienes).

Nosotros hemos probado varios y la diferencia es bastante clara. Los modelos pequeños funcionan, pero los grandes son los que realmente se sienten como un copiloto serio.

Paso 4 Conectar Claude Code a Tu Servidor Local

Este es el paso clave de toda la guía.

Por defecto, Claude Code se conecta a los servidores de Anthropic. Pero nosotros vamos a redirigirlo a nuestro servidor local.

Esto se hace configurando la variable de entorno:
export ANTHROPIC_BASE_URL=http://localhost:11434

En algunos casos también puedes necesitar una API key dummy, pero normalmente no es necesario si todo está bien configurado.

Lo que estamos haciendo aquí es redirigir Claude Code para que, en lugar de consultar a Anthropic, consulte a Ollama.
Y como Ollama está sirviendo un modelo open source, todo el procesamiento ocurre en tu máquina.

Este pequeño cambio es lo que convierte toda la experiencia en:

- **Gratis**
- **Local**
- **Sin límites de uso**

Paso 5 Ejecutar Claude Code en Tu Proyecto

Una vez hecho esto, ya puedes empezar a usar Claude Code como siempre.

Entra en tu proyecto:
cd mi-proyecto

Y lanza Claude Code.
A partir de aquí, el comportamiento es prácticamente el mismo:

- **Analiza archivos automáticamente**
- **Propone cambios**
- **Edita código**
- **Ejecuta tareas**

La diferencia es que ahora todo ocurre en local.

Desde nuestra experiencia, este momento es bastante sorprendente porque sientes que tienes un agente de desarrollo completamente autónomo funcionando sin depender de internet.

Paso 6 Usarlo Como Copiloto De Desarrollo Real

Aquí es donde realmente se ve el potencial.
Puedes pedirle cosas como:

- Crear una web desde cero
- Refactorizar código existente
- Explicar funciones complejas
- Añadir tests
- Detectar errores

Lo interesante no es solo que responda, sino que actúa directamente sobre tu proyecto.

Lee archivos, los modifica y ejecuta acciones reales. Esto lo acerca mucho más a un agente que a un simple chatbot.

Y todo esto sin enviar tu código a ningún servidor externo.

Requisitos Reales de Hardware

Aquí es donde mucha gente se lleva una sorpresa.
Ejecutar modelos en local no es gratis en términos de recursos.
Para modelos pequeños (2B–7B):

- Mínimo 16 GB de RAM
- Recomendado 32 GB

Para modelos grandes (30B):

- Mínimo 32 GB
- Ideal 64 GB

Si no tienes este hardware, la experiencia será lenta o directamente frustrante.

Nosotros lo hemos probado en diferentes configuraciones y la diferencia es enorme. Con poca RAM, el sistema funciona, pero no es práctico para trabajar en serio.

Limitaciones que debes conocer

Aunque esta solución es muy potente, no es perfecta.

Primero, la calidad del modelo no es la misma que Claude original. Los modelos open source han mejorado muchísimo, pero todavía hay diferencias en tareas complejas.

Segundo, el rendimiento depende completamente de tu máquina. No hay escalabilidad como en la nube.

Y tercero, algunas funcionalidades avanzadas de Claude Code pueden no comportarse igual con modelos locales.

Aun así, para muchos casos de uso, especialmente desarrollo diario, el resultado es más que suficiente.

Nuestra experiencia usando este setup

Después de probar esta configuración durante varios días, hay algo que tenemos claro.

No sustituye completamente a Claude en la nube, pero sí abre una alternativa muy potente.

Para tareas como:

- Prototipado rápido
- Edición de código
- Automatización básica

Funciona sorprendentemente bien.

Además, el hecho de trabajar sin límites de uso cambia completamente la forma en la que interactúas con la herramienta. Dejas de pensar en costes y empiezas a experimentar más.

Eso, para nosotros, es uno de los mayores cambios.

Conclusión final

Usar Claude Code en local con Ollama no es una función oficial “mágica”, sino una combinación inteligente de herramientas que aprovecha la compatibilidad de APIs para crear algo nuevo.

No estás ejecutando Claude como tal, pero sí estás usando su interfaz y su forma de trabajar con modelos que corren en tu propio ordenador.

Esto elimina costes, mejora la privacidad y te da control total, a cambio de depender de tu hardware y aceptar ciertas limitaciones en calidad.

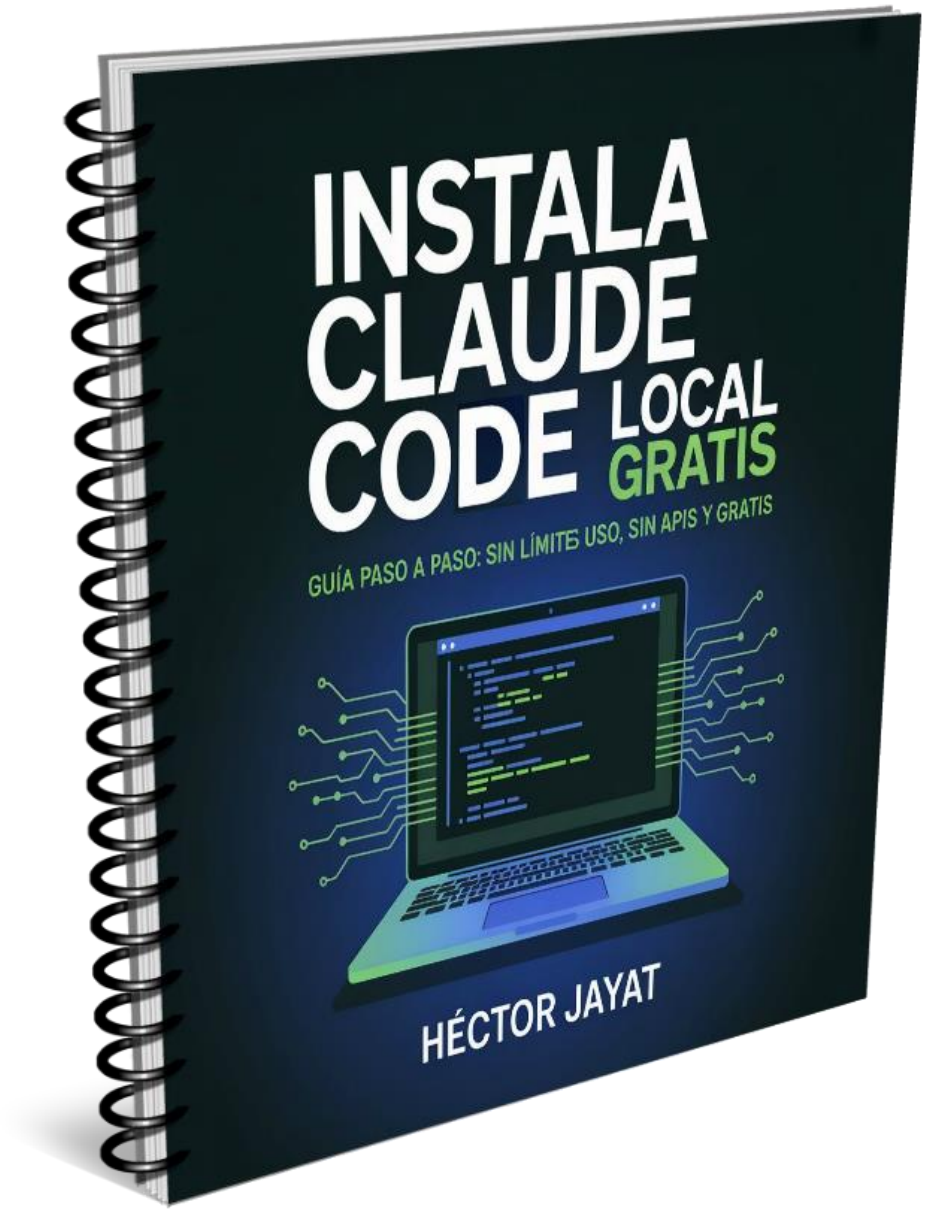
Si tienes un equipo suficientemente potente, es una de las formas más interesantes que hemos visto últimamente de convertir tu máquina en un entorno de desarrollo asistido por IA completamente autónomo.

Y lo más importante: es solo el principio de hacia dónde está evolucionando este tipo de herramientas.

Para más información puedes ver este video corto:

[(1) Tu propio Claude Code gratis en local - YouTube]

Fuente de Información: **Intelarter**


INSTALA
CLAUDE
CODE LOCAL
GRATIS
GUÍA PASO A PASO: SIN LÍMITES USO, SIN APIS Y GRATIS
HÉCTOR JAYAT