Corre AI en tu Mac. Sin internet. Sin pagar nada.

CATEGORÍA · CONECTORES, MCP Y HERRAMIENTAS

Guia paso a paso para instalar Ollama y correr Gemma 4 de Google en tu computadora. El modelo vive en tu disco. Nadie lo lee. Ni Google, ni OpenAI, ni nadie.

QUÉ VAS A ENCONTRAR

01 Instalar Ollama en Mac en menos de 5 minutos
02 Descargar Gemma 4 (9.6 GB, gratis, directo del registro oficial)
03 Chatear con el modelo desde la terminal
04 Verificar que funciona sin conexion a internet
05 Cuándo usar AI local vs AI en la nube

FICHA DEL RECURSO

Qué es, para qué sirve y cómo se implementa

FUNCIÓN

Guia paso a paso para instalar Ollama y correr Gemma 4 de Google en tu computadora. El modelo vive en tu disco. Nadie lo lee. Ni Google, ni OpenAI, ni nadie.

PARA QUÉ SIRVE

Cada vez que pegas un contrato, un reporte interno, o los datos de un cliente en ChatGPT o Claude, esa informacion sale de tu computadora y va a servidores en la nube. Puede usarse para entrenar modelos futuros. Puede estar sujeta a politicas de privacidad que cambian. Correr AI local resuelve esto. El modelo vive en tu disco. Ninguna peticion sale de tu Mac. No necesitas internet para usarlo. No tienes cuenta. No hay terminos de servicio que aceptar. El costo: 9.6 GB de espacio en disco y 10 minutos de instalacion. $0 0 datos COSTO MENSUAL SALEN DE TU MAC

DÓNDE APLICARLO

Cubre: Instalar Ollama en Mac en menos de 5 minutos; Descargar Gemma 4 (9.6 GB, gratis, directo del registro oficial); Chatear con el modelo desde la terminal; Verificar que funciona sin conexion a internet; Cuándo usar AI local vs AI en la nube.

CÓMO IMPLEMENTARLO

1. Por Que Importa: Todo lo que le dices a ChatGPT, OpenAI lo guarda.
2. Antes De Empezar: Necesitas una Mac y 10 minutos.
3. Paso 01 · Instalar Ollama: Un comando. Eso es todo.
4. Paso 02 · Descargar El Modelo: Gemma 4. El modelo de Google que nadie menciona.
5. Paso 03 · Chatear Con El Modelo: El >>> es tu nueva interfaz.
6. Paso 04 · La Prueba Final: Apaga el WiFi.
7. Modelos Alternativos: Si tu Mac no aguanta 9.6 GB, usa esto.

POR QUE IMPORTA

Todo lo que le dices a ChatGPT, OpenAI lo guarda.

Cada vez que pegas un contrato, un reporte interno, o los datos de un cliente en ChatGPT o Claude, esa informacion sale de tu computadora y va a servidores en la nube. Puede usarse para entrenar modelos futuros. Puede estar sujeta a politicas de privacidad que cambian. Correr AI local resuelve esto. El modelo vive en tu disco. Ninguna peticion sale de tu Mac. No necesitas internet para usarlo. No tienes cuenta. No hay terminos de servicio que aceptar. El costo: 9.6 GB de espacio en disco y 10 minutos de instalacion. $0 0 datos COSTO MENSUAL SALEN DE TU MAC

ANTES DE EMPEZAR

Necesitas una Mac y 10 minutos.

Ollama funciona en Mac (Intel y Apple Silicon), Linux y Windows. Esta guia esta optimizada para Mac. Si tienes Apple Silicon (M1, M2, M3, M4) el modelo va a correr notablemente mas rapido. No necesitas conocimientos tecnicos. No necesitas saber programar. Si puedes abrir una terminal y escribir un comando, puedes hacer esto.

REQUISITOS

Mac con macOS 12 o superior
10 GB de espacio libre en disco
4 GB de RAM minimo (8 GB recomendado)
Conexion a internet solo para la descarga inicial

PASO 01 · INSTALAR OLLAMA

Un comando. Eso es todo.

Ollama es el motor que permite correr modelos de lenguaje grandes en tu computadora. Es open source, gratuito y tiene mas de 2 millones de descargas. No es experimental. Es el estandar para AI local. Hay dos formas de instalarlo: descargando la app desde el sitio oficial, o con un comando en la terminal. Cualquiera de las dos funciona igual. Despues de instalar, Ollama corre silenciosamente en segundo plano. Aparece como un icono en la barra de menu de tu Mac. No consume recursos hasta que llamas a un modelo.

OPCION A · INSTALAR CON TERMINAL

curl -fsSL https://ollama.ai/install.sh | sh
OPCION B · Descarga directa
Ve a ollama.com y descarga el instalador para Mac.
Arrastralo a tu carpeta Aplicaciones como cualquier app.

PASO 02 · DESCARGAR EL MODELO

Gemma 4. El modelo de Google que nadie menciona.

Gemma 4 es el modelo open source de Google publicado en abril 2025. En benchmarks independientes supera a versiones anteriores de GPT-4 en razonamiento, matematicas y codigo. Y es completamente gratuito. Pesa 9.6 GB. La descarga tarda entre 5 y 20 minutos segun tu conexion. Se hace una sola vez. Despues el modelo vive en tu disco y no necesitas internet nunca mas. Hay modelos mas pequenos si tienes menos de 10 GB disponibles. Al final de esta guia hay una tabla de alternativas.

COMANDO DE DESCARGA

ollama run gemma4
Ollama descarga el modelo automaticamente.
Vas a ver las barras de progreso en terminal.
Cuando termina: 'success' y el prompt >>>

PASO 03 · CHATEAR CON EL MODELO

El >>> es tu nueva interfaz.

Cuando ves el prompt >>> en la terminal, el modelo esta cargado y listo. Escribe cualquier mensaje y presiona Enter. El modelo responde palabra por palabra directamente en tu pantalla. No hay interfaz grafica por default. La terminal es la interfaz. Esto parece tecnico al principio pero tiene una ventaja: ves exactamente lo que esta pasando, sin distracciones. Para cerrar el chat escribe /bye. Para volver a abrirlo, corre ollama run gemma4 de nuevo. El modelo ya esta descargado, carga en segundos.

COMANDOS BASICOS

ollama run gemma4 # abrir chat
ollama list # ver modelos instalados
ollama rm gemma4 # borrar modelo
/bye # cerrar chat (dentro del >>>)
Ctrl + C # forzar salida si se congela

PASO 04 · LA PRUEBA FINAL

Apaga el WiFi.

El modelo sigue funcionando. Esta es la diferencia que cambia todo. Con el modelo cargado en terminal, desactiva tu WiFi desde la barra de menu de tu Mac. Luego escribe cualquier pregunta en el >>>. El modelo responde igual. Sin internet. Sin cuenta. Sin datos saliendo de tu computadora. El calculo ocurre completamente en tu procesador y memoria RAM. Esto es lo que lo hace util para informacion sensible: contratos, datos de clientes, reportes internos, analisis financieros. Nada de eso toca un servidor externo.

PARA QUE USARLO SIN INTERNET

Revisar contratos con datos de clientes
Analizar reportes financieros internos
Trabajar en vuelos o lugares sin conexion
Cualquier tarea donde la privacidad sea critica

MODELOS ALTERNATIVOS

Si tu Mac no aguanta 9.6 GB, usa esto.

Gemma 4 con 9.6 GB necesita al menos 8 GB de RAM disponible para correr con fluidez. Si tu Mac tiene 8 GB en total (con otras apps abiertas), puede quedarse sin memoria. La solucion es usar un modelo mas pequeno. Pierdes algo de calidad en respuestas largas, pero para tareas cotidianas la diferencia es minima. Cierra Chrome antes de correr cualquier modelo local.

TABLA DE MODELOS POR TAMAÑO

gemma4 9.6 GB RAM recomendada: 16 GB
llama3.2 2.0 GB RAM recomendada: 8 GB
mistral 4.1 GB RAM recomendada: 8 GB
phi4-mini 2.5 GB RAM recomendada: 8 GB
Comando para cualquiera: ollama run [nombre]

LOCAL VS NUBE

No es uno u otro.

Es saber cuándo usar cada uno. AI local no reemplaza a Claude o ChatGPT para todo. Los modelos en la nube tienen mas contexto, conectividad, y actualizaciones constantes. Son mejores para tareas creativas complejas, investigacion en tiempo real, y flujos que necesitan integraciones. AI local gana cuando la privacidad importa, cuando no hay internet, o cuando quieres cero costo por uso intensivo. Un contrato de 50 paginas para revisar clausula por clausula: local. Una campana de marketing con investigacion de mercado: nube.

REGLA SIMPLE

DATOS SENSIBLES usar local
SIN INTERNET usar local
ALTO VOLUMEN / $0 usar local
CREATIVIDAD COMPLEJA usar nube
INVESTIGACION REAL usar nube
INTEGRACIONES usar nube

CHECKLIST DE INSTALACION

Paso a paso en orden.

INSTALACION (10 MINUTOS)

Descargar Ollama desde ollama.com o con el comando curl
Abrir Terminal (Cmd + Espacio, escribir Terminal)
Correr: ollama run gemma4
Esperar que descargue los 9.6 GB (barras de progreso)
Ver el mensaje 'success' y el prompt >>>
VERIFICACION
Escribir una pregunta en el >>> y confirmar que responde
Apagar el WiFi y hacer otra pregunta
Confirmar que responde igual sin internet
Escribir /bye para cerrar el chat
SI ALGO FALLA
Cerrar Chrome y otras apps pesadas antes de correr el modelo
Verificar que tienes 10 GB libres en disco
Si se queda sin RAM: usar llama3.2 en vez de gemma4

SIGUIENTE PASO

Ya tienes AI local. Ahora aprende a usarla bien.

Esto es el primer paso. La diferencia entre instalar un modelo y sacarle leverage real esta en saber que pedirle, como estructurar las tareas, y cuando combinarlo con AI en la nube. Cada semana publico casos

MÁS GUÍAS GRATIS EN 377CREATIVE.COM/RECURSOS