BIG school

GUÍA PRÁCTICA SOBRE IA LOCAL

QUÉ ES Y CÓMO USARLA

Cómo ejecutar Inteligencia Artificial en tu propio ordenador (y por qué deberías hacerlo)

Si ya tienes experiencia en programación, es muy probable que esto que te voy a describir te haya pasado alguna vez (y sino quédate porque de seguro te interesa, o te sirve como lección):

Estás trabajando en un proyecto serio de tu empresa, o en una idea propia que no quieres que nadie vea, y dudas si pegar ese trozo de código en ChatGPT para buscar un error.

Y haces bien. Cuando usas la IA en la nube, estás enviando tus datos, la lógica de tu negocio y el código a los servidores de otra empresa.

Para proyectos de casa, genial. **Para proyectos de clientes, es un riesgo de seguridad y de privacidad enorme**.

La solución a esto se llama IA Local.

En lugar de conectarte a internet para preguntarle a un modelo en la nube, e descargas el modelo a tu máquina y lo ejecutas en tu propio entorno. Sin intermediarios y sin ceder tus datos.

En esta guía te cuento por qué esto es el presente y cómo puedes empezar a montarlo en tu ordenador hoy mismo.

Vamos a ello.

1. Las 3 ventajas que lo cambian todo

¿Por qué complicarnos instalando cosas si el chat del navegador ya funciona? Muy sencillo:

Privacidad absoluta: Tu código nunca sale de tu ordenador. Si estás trabajando con datos sensibles o bajo contratos de confidencialidad, la IA local es la única forma 100% segura de programar asistido.

Cero costes extras: Se acabó preocuparse por la cuota mensual de tokens o las caídas de la API. Es gratis.

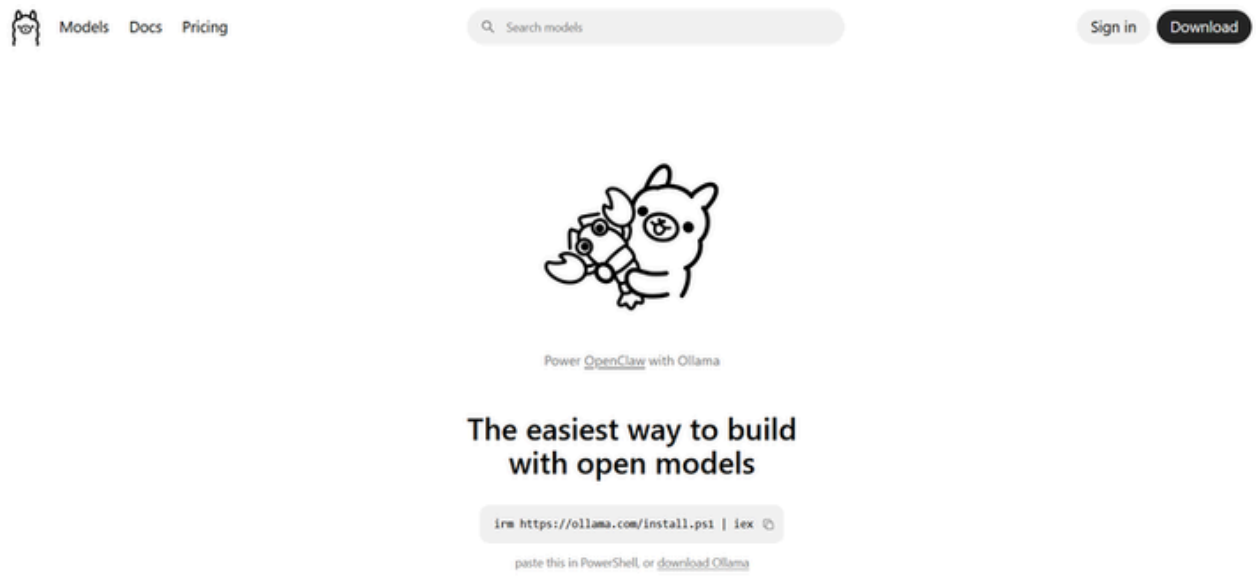
Modo offline: Funciona sin conexión a internet. Te vas a un tren o te quedas sin wifi, y sigues teniendo a tu copiloto listo para trabajar.

1. Las 3 ventajas que lo cambian todo

Ollama empaqueta modelos de lenguaje gigantes (como Llama 3 de Meta o modelos especializados en código) y los hace súper fáciles de ejecutar desde tu terminal.

👉 **Paso a paso para empezar:**

1. Descarga **Ollama** gratis desde su web oficial.



2. Abre tu terminal.
3. Escribe este comando: `ollama run llama3` (o el modelo de código que prefieras).

```
ollama run llama3.2-vision
```

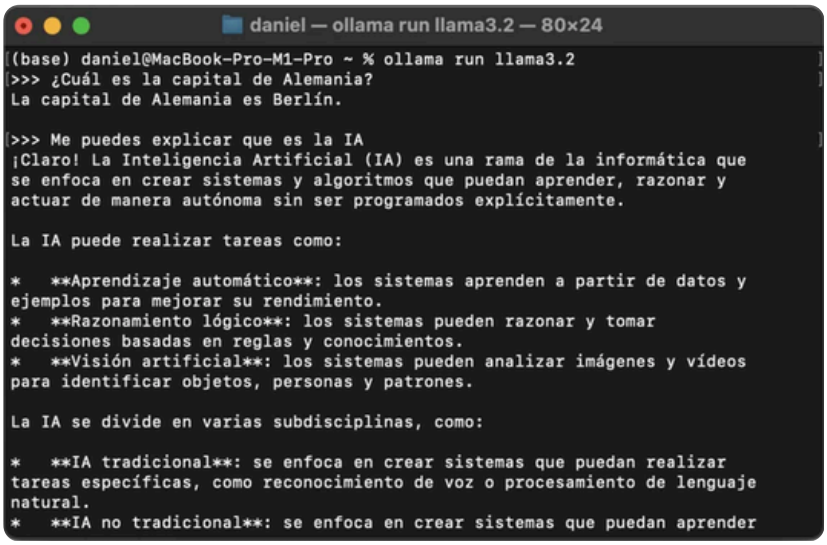
Para utilizar el modelo 90B, de mayor tamaño:

```
ollama run llama3.2-vision:90b
```

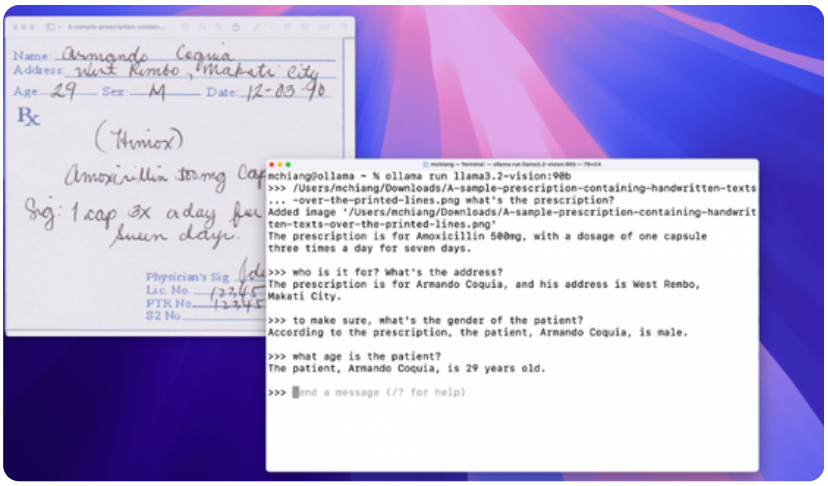
Ejemplos de uso:

Aqui tienes algunos ejemplos prácticos de todos los usos que puedes darle ¡Son infinitos!

Puedes hacerle preguntas simples cómo:

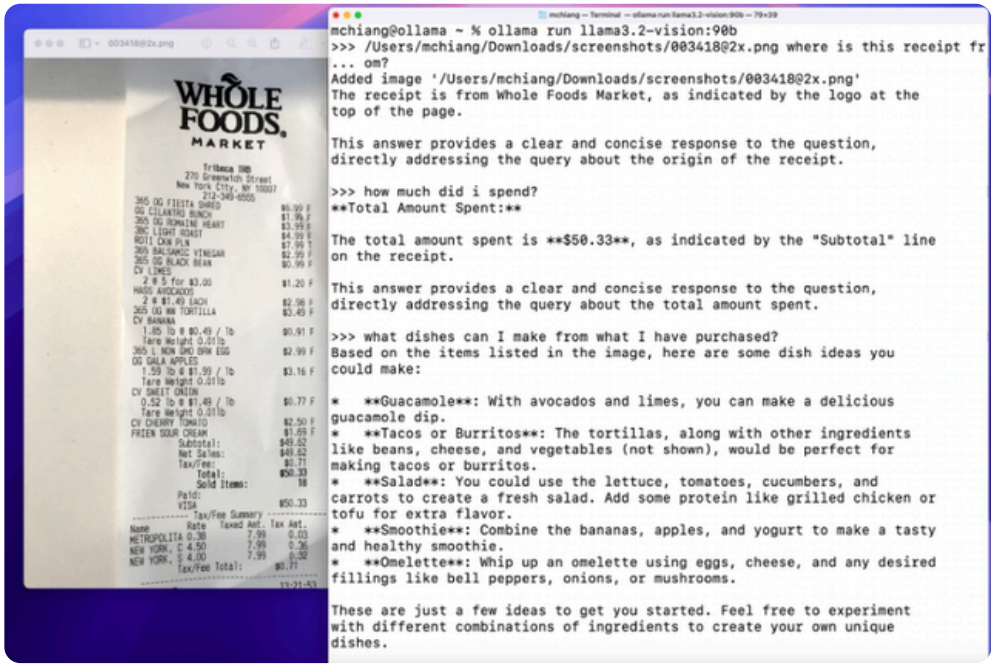


O pedirle que te interprete una receta médica (con lo dificil que es...)

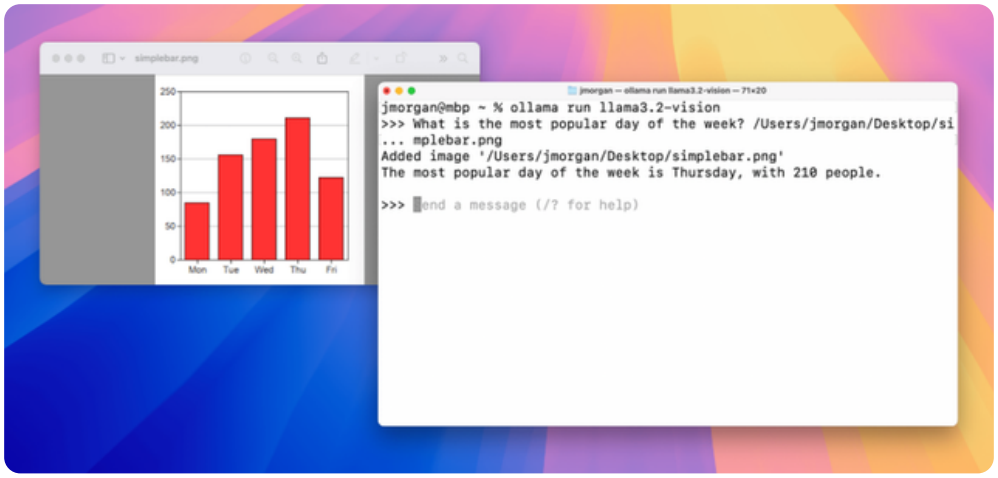


O pedirle que te interprete una receta médica (con lo dificil que es...)

También podrías solicitarle que te controle un ticket de compra y saque conclusiones:

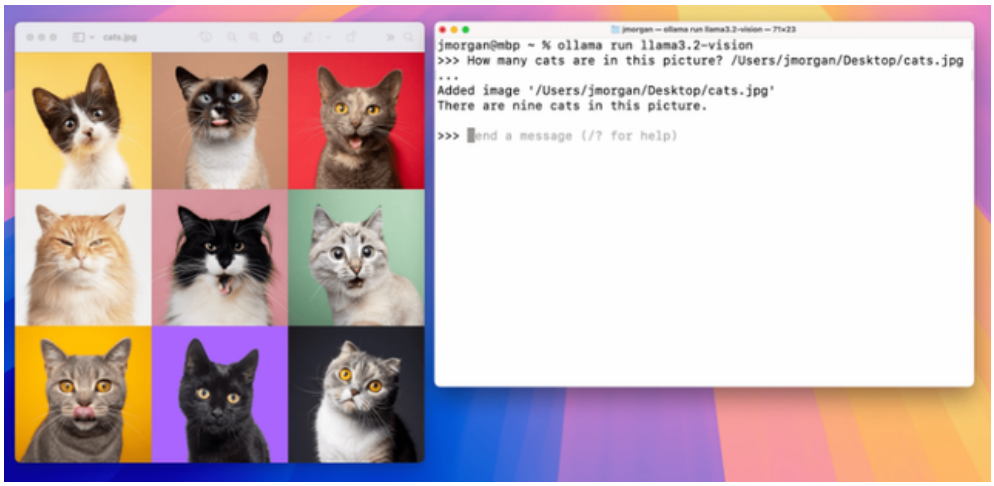


O incluso que te ayude a analizar un gráfico y extraiga datos de él:



O que interprete o identifique los elementos en una imagen:

O que interprete o identifique los elementos en una imagen:

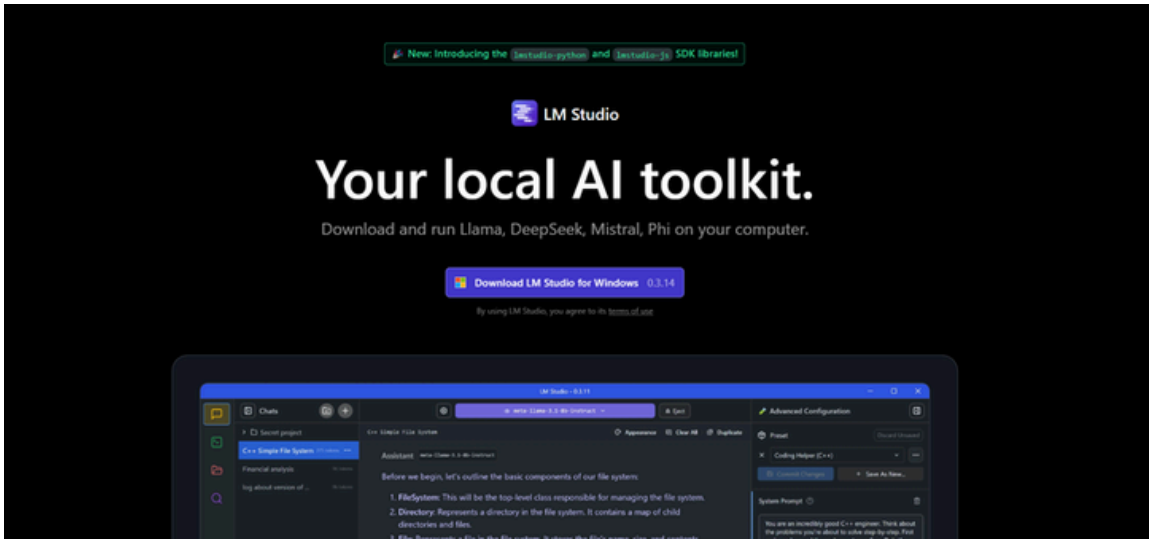


Estos son solo algunos ejemplos, pero las posibilidades son enormes.

La primera vez tardará un poco en descargar el "cerebro" del modelo a tu disco duro. Una vez termine, tendrás un chat interactivo directamente en tu terminal.

Le preguntas, te responde. Todo en tu máquina. Brutal.

3. LM Studio: Para los que prefieren interfaz visual



Si la terminal te da pereza y quieres una experiencia calcada a la de ChatGPT, tienes LM Studio.

Es una aplicación de escritorio que te permite buscar modelos de código abierto, descargarlos con un solo clic y chatear con ellos en una interfaz súper limpia.

Pero antes de emocionarnos, hablemos de las capacidades de tu máquina.

👉 Paso 1: ¿Tu ordenador aguanta esto?

LM Studio está muy bien optimizado, pero necesita algo de "gasolina":

RAM: Con 16GB vas sobrado para mover modelos decentes. Si tienes 8GB, funciona, pero te tocará usar modelos más pequeñitos.

Almacenamiento: Haz hueco en el disco, los modelos pesan entre 2 y 20GB.

Potencia: Si tienes tarjeta gráfica dedicada (NVIDIA/AMD) irá como un tiro. ¿Tienes un Mac con chip M1/M2/M3? Estás salvado, vuelan.

No necesitas el ordenador de la NASA. Un equipo normalito de hace unos años con 16GB de RAM lo mueve como un campeón. Cero estrés.

** **TRUCO:** LM Studio te indica cuánta memoria RAM va a consumir el modelo antes de descargarlo. Fíjate bien en los colores y en la barra de memoria para asegurarte de que tu ordenador tiene potencia gráfica suficiente para correrlo sin quedarse "congelado".

👉 Paso 2: Instalación para vagos

Entras en lmstudio.ai y te bajas el instalador (hay para Mac, Windows y Linux).
Lo ejecutas. El clásico "siguiente, siguiente, finalizar" de toda la vida.

Al abrirlo te saldrá un panel lleno de cosas. Puede abrumar un poco al principio, pero vamos a lo fácil.

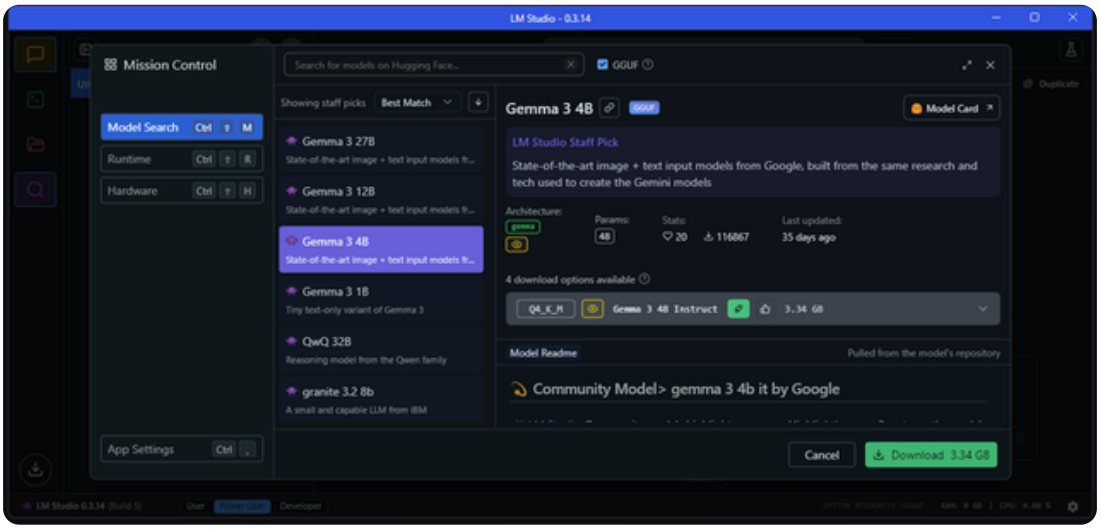
👉 Paso 3: Descarga tu primer modelo

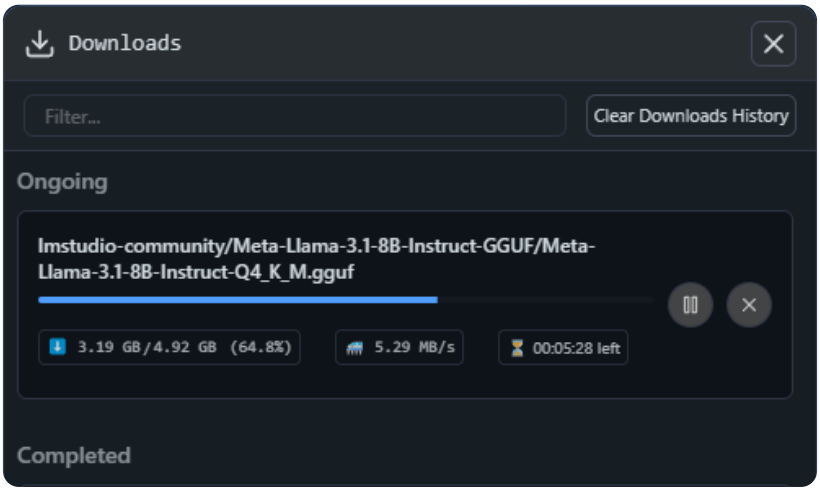
Esto bebe directamente de Hugging Face (el mayor repositorio de IA open source).

1. Vete a la lupa (pestaña Descubrir).
2. Busca algo como `"Meta-Llama-3.1-8B-Instruct"`

3. Haz clic en "Descargar" junto al modelo que hayas elegido.
4. Y ahora, vete a prepararte un café. Estos modelos pesan lo suyo (unos 5GB tranquilamente) y todo dependerá de la velocidad de tu Wi-Fi. Te puede tardar unos 10 minutitos. Puedes ir vigilando cómo avanza la cosa desde la pestaña de Descargas.

TRUCO DE SENIOR: Elige las versiones "quantizadas" (las que **acaban en Q4_K_M**). Están comprimidas para pesar menos (4-8GB) y responden rapidísimo sin perder casi calidad.





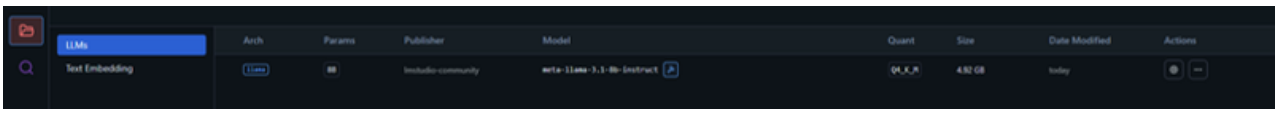
AVISO: Si te lo marca en rojo, ni lo intentes o congelarás la máquina.

👉 Paso 4: Cargar y chatear

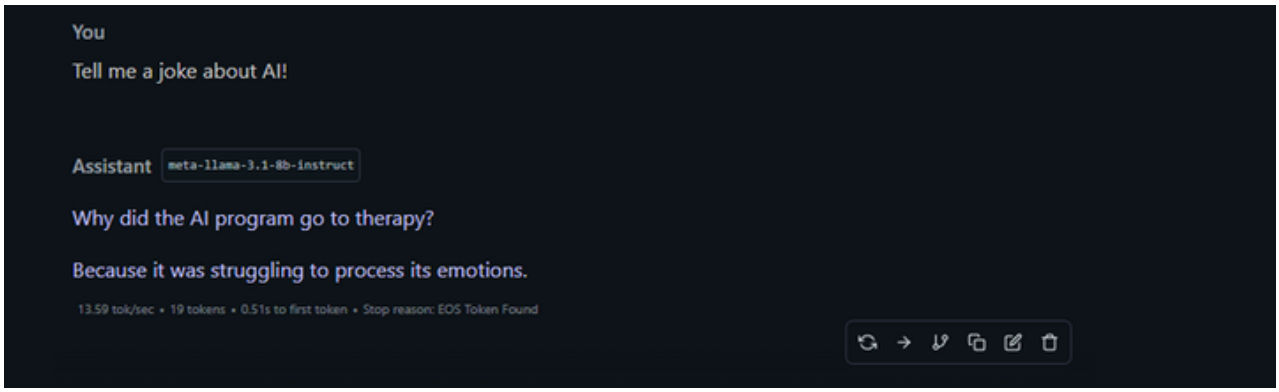
Vete a la pestaña "**Mis Modelos**" 📁, seleccionas el que has bajado y le das a **Cargar**. Dale unos segundos para que se meta en la memoria RAM.

Vete a la pestaña de "**Chat**" 💬, lo eliges en el desplegable superior y listo.

Vete a la pestaña "**Mis Modelos**" 📁, seleccionas el que has bajado y le das a **Cargar**. Dale unos segundos para que se meta en la memoria RAM.



Escribe algo. Pídele un script en Python. **Se siente exactamente igual que ChatGPT, pero con la tranquilidad de que si desconectas el router, sigue funcionando**, y tus palabras permanecen privadas.



4. ¿Y cómo conecto esto a mi editor de código?

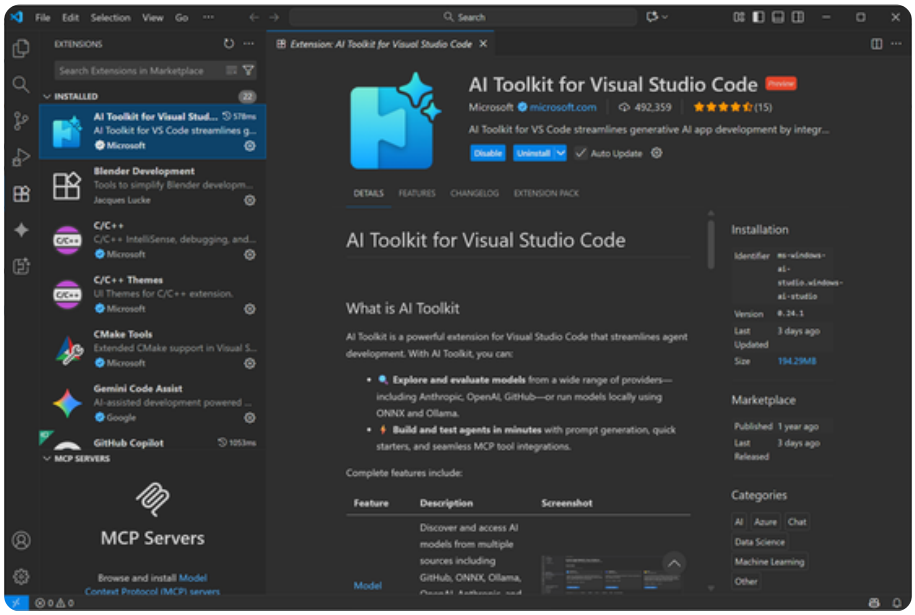
Aquí es donde ocurre la verdadera magia. Porque ya hemos hablado de que estar haciendo "Alt+Tab" todo el rato para copiar y pegar código de una pestaña a tu editor, no es programar de verdad.

Tanto **Ollama** como **LM Studio** te permiten levantar un servidor local.

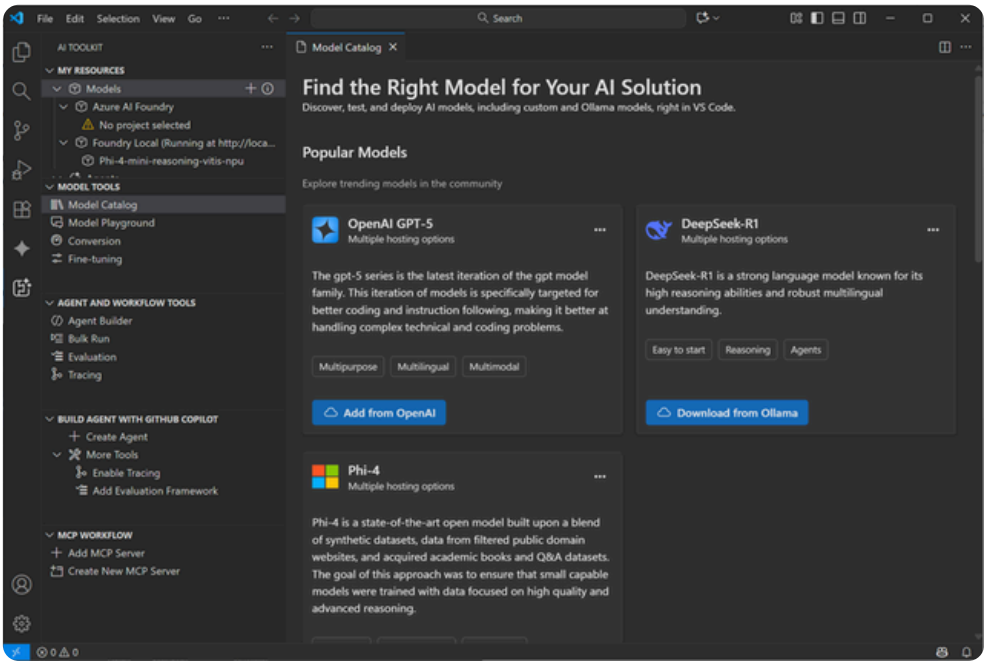
Esto significa que puedes irte a tu editor (como **VS Code**) e instalar extensiones de copiloto, y decirles: *"Ey, no uses la API de OpenAI, conéctate al modelo que tengo corriendo en mi propio ordenador"*.

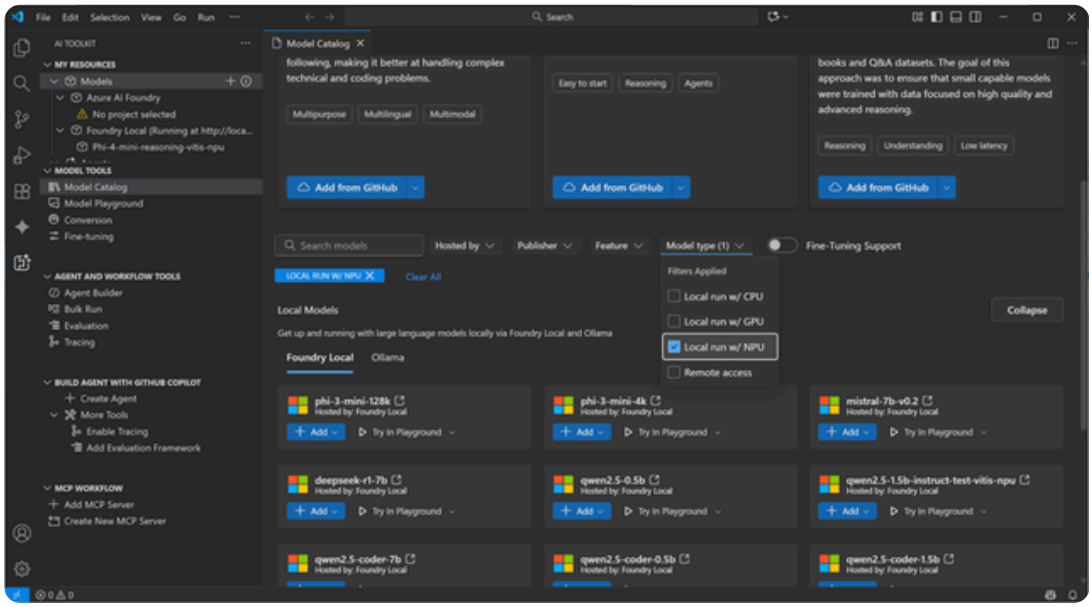
Estos son los pasos a seguir para lograr esto (tomando **VS Code** como editor de código):

1. Instala la extensión desde la pestaña "**Extensiones**" de VS Code, abre la extensión (el icono de la parte inferior izquierda en mis capturas de pantalla) y abre el catálogo de modelos:



Si te desplazas hacia abajo, podrás filtrar por varias categorías, como "local" o "en la nube", y, si eliges **"local", podrás seleccionar entre CPU, GPU o NPU**:

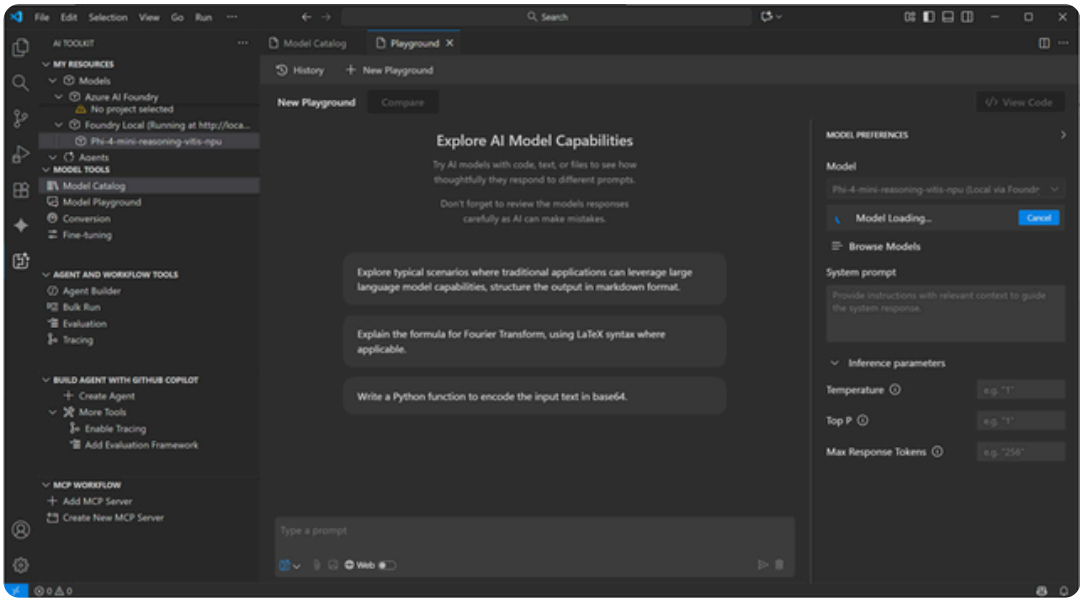




Ahora solo tienes que hacer clic en “Añadir” debajo del modelo que quieras probar.

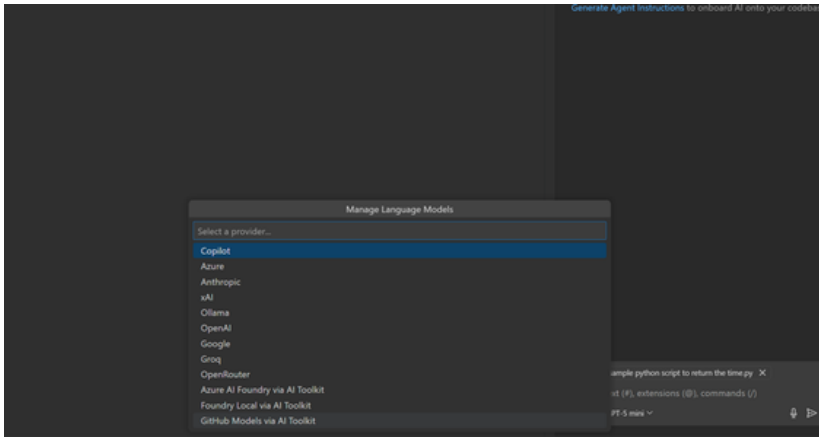
En este caso, verás que se ha añadido **Phi-4 mini reasoning** simplemente porque es un modelo pequeño que funciona en esa NPU, pero quizá te resulte más útil algo como **Qwen Coder** para tareas de programación.

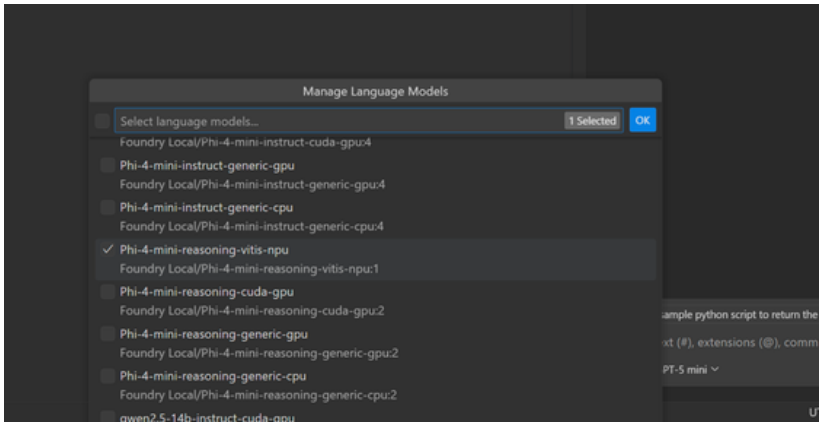
Una vez descargado el modelo, ve a la sección “**Modelos**” en la parte superior, dentro de “**MIS RECURSOS**”, y haz doble clic en el que quieras ejecutar. En nada verás como se cargará el modelo:

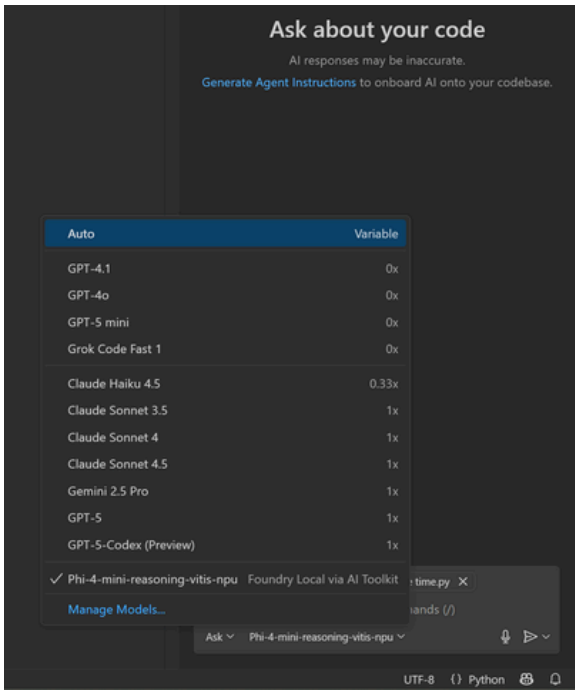


Ahora, vuelve a tu código, luego abre Copilot, haz clic en el modelo y, a continuación, haz clic en "Gestionar modelos".

En este caso, por ejemplo, he seleccionado Foundy Local a través de AI Toolkit:







¡Eso es todo! A partir de ese momento, tienes un asistente revisando tu código, sugiriendo autocompletados y ayudándote a construir tu proyecto... de forma privada, segura y sin pagar a terceros.

Espero que te haya sido útil y que lo pruebes hoy mismo.

Descárgate una de estas herramientas y comprueba cómo rinde tu máquina.

Aprender a elegir el modelo adecuado y dónde ejecutarlo es la diferencia entre ser un programador que tira prompts a ciegas, y uno que tiene el control técnico de su entorno.

mouredev

BIG school

No te pierdas la Sesión 1

Activa la campanita para que YouTube te avise cuando empiece el directo

ACTIVA LA CAMPANITA