Run Your First Local LLM in 15 Minutes

January 27, 2025 · by Mark Bartlett

Download this guide as PDF

> **Quick Answer:** You'll install Ollama (free, one command), download an AI model, and have a working chatbot running entirely on your computer. No accounts, no API keys, no monthly fees. Works on Mac, Windows, and Linux—even on a laptop with 8GB of RAM.

📚 **More on this topic:** Ollama vs LM Studio · Ollama Troubleshooting · Best Models for Chat · VRAM Requirements

You've heard about ChatGPT, Claude, and all the other AI assistants. Maybe you've even used them. But here's the thing: every message you send goes to someone else's servers. Your questions, your ideas, your data—all processed in the cloud.

What if you could run the same kind of AI on your own computer? No internet required. No subscription fees. Complete privacy.

That's exactly what we're going to do. By the end of this guide, you'll have a fully functional AI chatbot running locally on your machine. And I promise—it's way easier than you think.

Why Run AI Locally?

Before we dive into the how, let's talk about the why. There are three big reasons people run AI locally:

Your Data Stays Yours

When you use ChatGPT or Claude, your conversations travel to data centers owned by OpenAI or Anthropic. With a local LLM (Large Language Model), everything stays on your machine. Your prompts never leave your computer. Nobody's training on your inputs. If you're working with sensitive information—personal journals, business ideas, code for a client—local AI keeps it private.

No Monthly Bills

Cloud AI services typically cost $20/month for premium features. Over a year, that's $240. Over five years, that's $1,200—plus whatever price increases come along.

A local LLM costs nothing to run after the initial setup. The software is free. The models are free. You just need a computer you already own.

Works Offline

Flying somewhere? Working from a cabin with no internet? Your local AI doesn't care. Once the model is downloaded, it runs entirely offline. No connection required.

What You'll Need

Minimum Hardware (It's Lower Than You Think)

Here's the good news: you don't need a fancy gaming PC to run local AI. Here's what actually works:

Bare minimum:

- 8GB RAM
- Any modern CPU (Intel or AMD from the last 6-7 years)
- 10GB free disk space
- macOS 11+, Windows 10+, or Ubuntu 18.04+

With 8GB of RAM, you can run smaller models like Llama 3.2 3B or Phi-3 Mini. These are genuinely useful—they can answer questions, help with writing, and handle basic coding tasks.

For a comfortable experience:

- 16GB RAM
- A dedicated GPU (nice to have, not required)
- 50GB free disk space (for trying different models)

With 16GB, you can run 7-8B parameter models like Llama 3.1 8B—these are significantly smarter and more capable.

Hardware Requirements by Model Size

Model Size	RAM Needed	Example Models	What It Can Do
3B	8GB	Llama 3.2 3B, Phi-3 Mini	Basic Q&A, simple writing

Model Size	RAM Needed	Example Models	What It Can Do
7-8B	16GB	Llama 3.1 8B, Mistral 7B	Good conversations, coding help
13B	32GB	Llama 2 13B	Better reasoning, longer context
70B+	64GB+	Llama 3.1 70B	Near-ChatGPT quality

Don't have a GPU? That's okay. Ollama works on CPU-only machines. Responses will be slower (maybe 3-6 words per second instead of 30+), but it absolutely works. Many people start this way.

→ Not sure what fits? Try our Planning Tool.

Step 1: Install Ollama

Ollama is the easiest way to run local AI. Think of it as a simple app that handles all the complicated stuff behind the scenes. You tell it which model you want, it downloads and runs it. That's it.

Mac Installation

1. Go to ollama.com
2. Click **Download** and select **Download for macOS**
3. Open the downloaded `.zip` file
4. Drag **Ollama** to your **Applications** folder
5. Open Ollama from Applications

You'll see a small llama icon appear in your menu bar. That means Ollama is running in the background and ready to go.

Windows Installation

1. Go to ollama.com
2. Click **Download** and select **Download for Windows**
3. Run the downloaded `.exe` installer
4. Follow the prompts (just click Next a few times)
5. Ollama installs and starts automatically

That's it. Ollama now runs in the background whenever your computer is on.

Linux Installation

Open your terminal and run this single command:

```
curl -fsSL https://ollama.com/install.sh | sh
```

The script downloads and installs everything automatically. When it finishes, Ollama is ready to use.

Verify It's Working

Open a terminal (Terminal on Mac/Linux, Command Prompt or PowerShell on Windows) and type:

```
ollama --version
```

You should see something like `ollama version 0.3.x`. If you see a version number, you're ready for the next step.

Don't see it? On Windows, try closing and reopening your terminal. On Mac, you might need to grant Ollama permission to install its command-line tool when prompted.

Step 2: Download Your First Model

Now for the fun part. We're going to download an AI model—the actual "brain" that generates responses.

Models come in different sizes. Bigger models are smarter but need more RAM. For your first model, we'll use **Llama 3.2 3B**—it's small enough to run on almost any computer but smart enough to be genuinely useful.

In your terminal, type:

```
ollama pull llama3.2
```

Press Enter and watch the magic happen:

```
pulling manifest
pulling dde5aa3fc5ff... 100% ▕████████████▏ 2.0 GB
pulling 966de95ca8a6... 100% ▕████████████▏ 1.4 KB
pulling fcc5a6bec9da... 100% ▕████████████▏ 7.7 KB
pulling a70ff7e570d9... 100% ▕████████████▏ 6.0 KB
pulling 56bb8bd477a5... 100% ▕████████████▏   96 B
pulling 34bb5ab01051... 100% ▕████████████▏  561 B
verifying sha256 digest
writing manifest
success
```

This downloads about 2GB of data—the model's "knowledge." On a decent internet connection, it takes 2-5 minutes.

Have 16GB+ RAM? Try the larger, smarter model:

```
ollama pull llama3.1:8b
```

This one is about 4.7GB and noticeably more capable.

Step 3: Have Your First Conversation

Here's the moment you've been waiting for. Let's talk to your AI.

Type this command:

```
ollama run llama3.2
```

After a moment, you'll see a prompt:

```
>>>
```

That's it. You're in. Type anything and press Enter:

```
>>> What is the capital of France?
```

```
The capital of France is Paris.

>>>
```

Congratulations—you just ran AI entirely on your own computer!

Prompts to Try

Here are some things to ask your new AI assistant:

Ask a question:

```
>>> Explain photosynthesis like I'm 10 years old
```

Get help writing:

```
>>> Write a professional email declining a meeting invitation
```

Learn something:

```
>>> What are three interesting facts about octopuses?
```

Get coding help:

```
>>> Write a Python function that checks if a number is prime
```

How to Exit

When you're done chatting, type `/bye` or press `Ctrl+D`:

```
>>> /bye
```

You're back to your normal terminal. The model unloads after a few minutes of inactivity to free up memory.

What to Try Next

You've got local AI running. Here's how to explore further:

Other Models Worth Trying

Once you're comfortable, experiment with different models. Each has its own personality and strengths:

```
# A great all-around model (needs 16GB RAM)
ollama pull llama3.1:8b

# Excellent for coding tasks
ollama pull codellama

# Fast and capable
ollama pull mistral

# Very small, runs on anything
ollama pull phi3:mini

# Good at reasoning and explaining
ollama pull gemma2
```

Model	Size	Best For	RAM Needed
llama3.2	2GB	Quick answers, basic tasks	8GB
llama3.1:8b	4.7GB	General assistant, writing	16GB
codellama	3.8GB	Programming help	16GB
mistral	4.1GB	Fast responses, good quality	16GB
phi3:mini	2.2GB	Limited hardware	8GB
gemma2	5.4GB	Explanations, reasoning	16GB

To see what you've downloaded:

```
ollama list
```

To remove a model you don't want:

```
ollama rm model-name
```

Try a Visual Interface (LM Studio)

Love what Ollama does but prefer clicking over typing? Check out **LM Studio**. It's a free app with a full graphical interface—browse models, download with one click, and chat in a nice window.

You can even run both. Many people use LM Studio for casual chatting and Ollama for scripts and automation. For a detailed comparison of both tools, read our Ollama vs LM Studio guide.

Download it at lmstudio.ai.

Connect to Other Apps

Ollama runs a local server that other apps can talk to. This means you can:

- Use AI in your code editor (VS Code, Cursor)
- Build your own chatbot
- Connect to note-taking apps
- Create automation workflows

The API runs at `http://localhost:11434` and is compatible with the OpenAI format, so many existing tools work out of the box.

Troubleshooting Common Issues

Running into problems? Here are the most common issues and how to fix them:

"Model not found" or download fails

Problem: You typed `ollama pull` but got an error.

Solutions:

- Check your internet connection
- Make sure you typed the model name correctly (they're case-sensitive)
- Try a different model: `ollama pull phi3:mini`

- Check available models at ollama.com/library

Slow responses

Problem: The AI takes forever to respond, typing out words very slowly.

Why it happens: Your model is running on CPU instead of GPU, or the model is too large for your RAM.

Solutions:

- Try a smaller model: `ollama run phi3:mini`
- Close other applications to free up RAM
- If you have an NVIDIA GPU, make sure drivers are up to date
- CPU-only is just slower—3-6 words/second is normal without a GPU

Out of memory errors

Problem: You see “not enough memory” or the model won’t load.

Why it happens: The model needs more RAM than you have available.

Solutions:

- Try a smaller model (phi3:mini or llama3.2)
- Close Chrome and other memory-hungry apps
- Restart your computer to clear memory
- Check the model size before downloading—don’t try 70B models on 16GB RAM

Ollama won’t start

Problem: The `ollama` command isn’t recognized, or the app won’t open.

Solutions:

- **Mac:** Open Ollama from Applications first, then try the terminal command
- **Windows:** Close and reopen your terminal after installation
- **Linux:** Run `sudo systemctl start ollama` to start the service
- Try reinstalling from ollama.com

Model takes too long to load

Problem: You run a model and it sits there for minutes before responding.

Why it happens: Large models take time to load into memory, especially on HDDs.

Solution: First response is always slowest. Subsequent responses will be much faster since the model stays loaded. If it's consistently slow, try a smaller model.

You Did It!

Take a second to appreciate what you just accomplished. You installed software, downloaded an AI model, and had a conversation with it—all running locally on your own hardware.

No monthly fees. No data leaving your computer. No corporate oversight.

This is just the beginning. From here, you can:

- **Try bigger models** as you get more comfortable (or upgrade your hardware)
- **Explore LM Studio** for a visual experience
- **Connect Ollama to your workflow**—code editors, note apps, automation tools
- **Fine-tune models** for specific tasks (that's an advanced topic for another day)

The local AI ecosystem is growing fast. New models drop almost weekly, each one more capable than the last. What required a data center five years ago now runs on a laptop.

Welcome to the club. Your AI, your hardware, your rules.

Related Guides

- Ollama vs LM Studio: Which Should You Use?
- How Much VRAM Do You Need for Local LLMs?
- GPU Buying Guide for Local AI

Source: https://insiderllm.com/guides/run-first-local-llm/