Run Your First Local LLM in 15 Minutes

January 27, 2026 · Updated: March 31, 2026

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:** Qwen 3.5 9B Setup Guide · 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?

Three 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 (or 8GB of GPU VRAM), you can run Qwen 3.5 9B at Q4 quantization. This is a genuinely capable model — it has built-in vision (describe images, read screenshots), 262K context, and beats models three times its size on reasoning benchmarks. Two years ago, you needed 16GB for anything useful. The 9B class changed that.

For a comfortable experience:

- 16GB RAM (or a GPU with 12-16GB VRAM)
- A dedicated GPU (nice to have, not required)
- 50GB free disk space (for trying different models)

With 16GB, you can run the 9B at higher quality (Q8) or step up to larger models like Qwen 3.5 27B.

Hardware Requirements by Model Size

Model Size	RAM/VRAM Needed	Example Models	What It Can Do
3-4B	4-6GB	Qwen 3.5 4B, Llama 3.2 3B	Basic Q&A, simple tasks
9B	8GB	Qwen 3.5 9B (Q4_K_M)	Strong chat, coding, vision — the starter model
27-35B	16-24GB	Qwen 3.5 27B, Qwen 3.5 35B-A3B (MoE)	Near-frontier quality for most tasks
70B+	48GB+	Llama 3.3 70B, Qwen 3.5 72B	Best local quality, needs serious hardware

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.

Mac users with 32GB+ RAM: Ollama 0.19 (released March 31, 2026) includes an MLX preview that uses Apple's machine learning framework for faster inference on Apple Silicon. Prefill speed

roughly doubles compared to the previous version on M5-series chips. This is automatic — just update Ollama and it uses MLX when available.

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 0.18.x` or `0.19.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.

The model to start with in 2026 is **Qwen 3.5 9B**. It fits on 8GB GPUs, has built-in vision (can describe images and read screenshots), supports 262K context tokens, and beats models three times its size on reasoning benchmarks. It's the best small model available right now by a wide margin.

In your terminal, type:

```
ollama pull qwen3.5:9b
```

This downloads about 6GB of data. On a decent internet connection, it takes 5-10 minutes.

Only have 4-6GB of VRAM or running on CPU with 8GB RAM? Start with the smaller version:

```
ollama pull qwen3.5:4b
```

This one is about 3GB. Less capable, but runs on almost anything.

Have 16GB+ VRAM? Go bigger:

```
ollama pull qwen3.5:27b
```

This 27B model is about 17GB and noticeably smarter. If you have exactly 16GB VRAM, try the MoE variant instead — it only activates 3B parameters per token despite having 35B total:

```
ollama pull qwen3.5:35b-a3b
```

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 qwen3.5:9b
```

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:

What model should I use?

The right model depends on your hardware. Here's the quick answer for each VRAM tier:

Your VRAM/RAM	Model	Command	Download Size	Why
8GB	Qwen 3.5 9B Q4	`ollama run qwen3.5:9b`	~6GB	Best quality at this tier, built-in vision
12GB	Qwen 3.5 9B Q8	`ollama run qwen3.5:9b-q8_0`	~10GB	Higher quality quant of the same model
16GB	Qwen 3.5 35B-A3B Q4	`ollama run qwen3.5:35b-a3b`	~8GB	MoE model, only activates 3B per token — fast and smart
24GB	Qwen 3.5 27B Q4	`ollama run qwen3.5:27b`	~17GB	Best local model for most tasks

📖 **Our full Qwen 3.5 9B setup guide is here — covers quantization tables, vision features, and thinking mode.**

Other models worth trying

Once you're comfortable with Qwen 3.5, experiment with other models. Each has different strengths:

```
# Good alternative to Qwen 3.5 — different "personality"
ollama pull llama3.3

# Strong at coding
ollama pull qwen3.5:9b-coder

# Smaller, runs on very limited hardware
ollama pull qwen3.5:4b

# Good for reasoning tasks
ollama pull gemma3:12b
```

Model	Size	Best For	VRAM/RAM Needed
qwen3.5:9b	~6GB	General use, vision, coding — the default	8GB
qwen3.5:27b	~17GB	Best quality for most tasks	24GB
qwen3.5:35b-a3b	~8GB	MoE model, great quality/speed ratio	16GB
llama3.3	~4.7GB	Alternative chat personality	8GB
gemma3:12b	~8GB	Reasoning, explanations	12GB
qwen3.5:4b	~3GB	Very limited hardware	4-6GB

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)

Prefer clicking over typing? **LM Studio** (currently at v0.4.8) is a free app with a full graphical interface — browse models, download with one click, chat in a polished window. The 0.4 release

added a server mode with continuous batching, and the recent 0.4.7 update added an Anthropic-compatible API so you can use LM Studio models with Claude Code and similar tools.

Many people use LM Studio for casual chatting and Ollama for scripts and automation. For a detailed comparison, 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, Continue)
- Run a ChatGPT-like web UI with Open WebUI
- Talk to your LLM with voice
- Build your own chatbot or automation workflows

The API runs at `http://localhost:11434` and is compatible with the OpenAI format, so most existing tools work out of the box. Ollama also supports structured output (JSON schema) for building apps that need reliable formatting.

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 qwen3.5:4b`
- 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 qwen3.5:4b`
- 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 (qwen3.5:4b)
- 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:

- **Read the Qwen 3.5 9B deep dive** — quantization tables, vision features, thinking mode
- **Try bigger models** as you get more comfortable (or upgrade your hardware)
- **Install Open WebUI** for a ChatGPT-like web interface
- **Add voice chat** — talk to your LLM instead of typing
- **Connect Ollama to your workflow** — code editors, note apps, automation tools

What required a data center five years ago now runs on a laptop. What needed 16GB of VRAM a year ago fits in 8GB. The barrier keeps dropping.

Your AI, your hardware, your rules.

Related Guides

- Qwen 3.5 9B Setup Guide: The Best 8GB Model
- Ollama vs LM Studio: Which Should You Use?
- How Much VRAM Do You Need for Local LLMs?
- GPU Buying Guide for Local AI
- Voice Chat with Local LLMs

Get notified when we publish new guides.

Subscribe — free, no spam

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