Data Science Collective

Member-only story

How to Run Large Language Models (LLMs) Locally: A Beginner's Guide to Offline AI

How to Set Up LLMs on Your Machine in Just a Few Minutes

Claudia Ng · Follow

Published in Data Science Collective · 4 min read · 3 days ago

If you've been following artificial intelligence (AI) advancements, you've likely noticed a thriving open-source community developing powerful large language models (LLMs). These models can **rival offerings from major players like OpenAI and Anthropic.**

The best part? You can run them **entirely offline**, keeping your data private while enjoying **unlimited AI access.**

In this guide, we'll explore:

- Why you should run LLMs locally
- The best tools for running LLMs offline
- How to choose the right AI model for your hardware

Image by StockSnap from Pixabay

Why Run AI Locally?

Running LLMs locally offer advantages over cloud-based solutions. Here's why I prefer it:

- 🛫 **No Internet Required** — Use AI **without WiFi**, whether you're on a long-haul flight, in a remote area, or during an OpenAI service outage.
- 🔒 **Enhanced Privacy & Data Control** — Local AI **keeps your data on your device**, preventing your interactions from being logged or used for future

model training.

- 💬 **Unlimited Usage** — Avoid **API rate limits, token restrictions, and paywalls**. Local models let you chat as much as you want — free, forever.

Every time you interact with a cloud-based AI, your prompts and responses may be stored as data for model training in the future. Running LLMs locally **puts you in control**.

How to Run LLMs Locally: Best Tools for Beginners & Advanced Users

Setting up local AI is easier than you might think — **some tools require zero coding!** Here are the best options based on your experience level:

1️⃣ LM Studio (Easiest, No Coding Required!)

LM Studio is the **fastest way to get started** with local LLMs. It provides a **user-friendly interface** to download models, chat with AI, and even upload documents for context.

💡 **Pro Tip:** LM Studio supports **context injection** — upload PDFs, CSVs, or DOCX files (up to 30MB each) to provide **background knowledge** to your AI assistant. This makes it **a local version of RAG (Retrieval-Augmented Generation)** — perfect for summarizing documents or extracting insights from reports!

2️⃣ Ollama (For Developers & Power Users)

Ollama is a **command-line tool** that makes downloading and running AI models seamless. It's slightly more technical but offers **flexibility and**

customization for those comfortable with the terminal.

3️⃣ vLLM (For Speed & Performance)

Developed by **UC Berkeley's Sky Computing Lab**, vLLM is optimized for **blazing-fast inference** and can handle **multiple concurrent requests** — ideal for those prioritizing speed.

4️⃣ Manual Installation (For AI Enthusiasts & Researchers)

If you prefer full control, you can manually download **GGUF models from Hugging Face** and use Python libraries like transformers to run them. **This option is best if you want to fine-tune models** for specific applications.

Choosing the Right AI Model for Your Needs

With many open-source LLMs available, picking the right one depends on your **hardware and use case**. Here are some well-regarded open-source AI models:

- DeepSeek R1
- Gemma 3 (My favorite!)
- DeepSeek V3
- QwQ 32B
- Llama 3.1

To find the best model for your needs, check out the Chatbot Arena LLM leaderboard — a ranking of AI models based on real-world user feedback.

How to Match Model Size to Your Computer

Once you've selected a model, the next step is to select a model size that aligns with your computer's capabilities. Researchers use "quantization", a method that reduces the precision of model parameters, to make larger models fit on less powerful devices.

If you're working on a high-precision reasoning task, prioritize accuracy with larger models. However, if you have limited compute power and want faster inference speeds, opt for a smaller, quantized version.

💡 **Key Factor: RAM (Memory) Matters!** LLMs require enough **RAM to load the model**. Here's a rough guide:

🖥️ **8GB RAM** → Small models (3B-7B) with aggressive quantization

🖥️ **16GB RAM** → Medium models (7B-13B) with moderate quantization

🖥️ **32GB+ RAM** → Large models (up to 30B) with higher precision

A **GPU (Graphics Processing Unit)** significantly improves performance. If you have a **CUDA-compatible NVIDIA GPU**, look for models **optimized for GPU inference**.

Tip: Start with a **smaller model** and gradually scale up based on performance!

Limitations of Running LLMs Locally

While local AI offers **privacy and unlimited use**, there are trade-offs:

❌ **No Internet Access** — Unlike ChatGPT, local models can't browse the web or perform live fact-checking.

PC.

To compensate for these limitations, I like to regularly check for **newer, optimized models** trained on recent data!

Final Thoughts: Reclaiming Your AI Experience

Running AI **offline** puts you in control — no paywalls, no data collection, no internet needed. Here's my set-up:

📌 **My Go-To Tool:** LM Studio (**best balance of ease and functionality!**)

📌 **Best Model Right Now:** Gemma 3

📌**Key Benefit:** Work distraction-free by turning off WiFi while using AI

Open-source LLMs are so powerful. There are options for different tasks, and you no longer have to rely on big tech — you can run powerful AI models **on your own terms.**

Want to build your AI skills?

👉 Join the **AI Weekender** for ideas on more AI projects you can build in a weekend!

Llm AI Python Genai Data Science

Responses (2)

Alex Mylnikov

What are your thoughts?

Raphael Schols
2 days ago (edited)

Ohh nice, I'll give this a try! Thanks for the article. Have you built anything with local models that you use frequently? I'd also be interested in the limitations you've seen, especially compared to online LLMs.

1 1 reply Reply

Biyesse
3 days ago

Nice article 👍

1 1 reply Reply

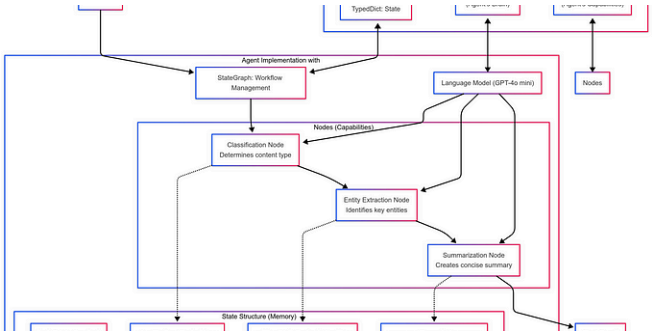
More from Claudia Ng and Data Science Collective

In TDS Archive by Claudia Ng

The Most Expensive Data Science Mistake I've Witnessed in My...

Why true success in machine learning goes beyond optimizing a single metric

Nov 28, 2024 1.3K 29



In Data Science Collective by Paolo Perrone

The Complete Guide to Building Your First AI Agent with...

Three months into building my first commercial AI agent, everything collapsed...

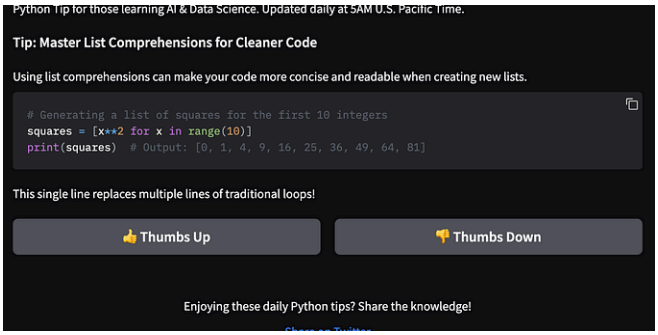
Mar 11 2.9K 61

In Data Science Collective by Buse Şenol

Model Context Protocol (MCP): An End-To-End Tutorial With Hands-...

What is MCP? How to create an MPC Server that brings news from a web site with Claude...

Mar 18 1.2K 15



In TDS Archive by Claudia Ng

How I Built My First AI-Powered Web App in 20 Minutes

A beginner's guide to building your AI-driven web application without front-end...

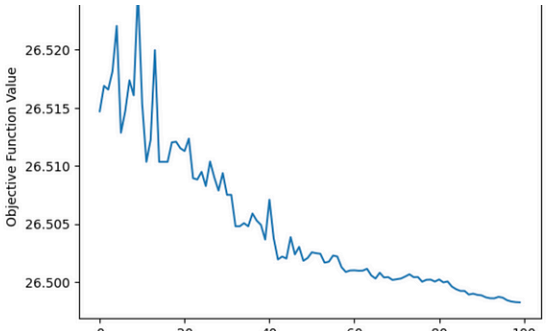
Feb 1 96 2

See all from Claudia Ng

See all from Data Science Collective

Recommended from Medium

Dataset	Sampling prop.	Epochs	Disk size
CommonCrawl	67.0%	1.10	3.3 TB
C4	15.0%	1.06	783 GB
Github	4.5%	0.64	328 GB
Wikipedia	4.5%	2.45	83 GB
Books	4.5%	2.23	85 GB
ArXiv	2.5%	1.06	92 GB
StackExchange	2.0%	1.03	78 GB



In Data Science Collective by tangbasky

Understanding Llama-1: A Personal Perspective

Llama 1 can be regarded as the first open-source large language model that can...

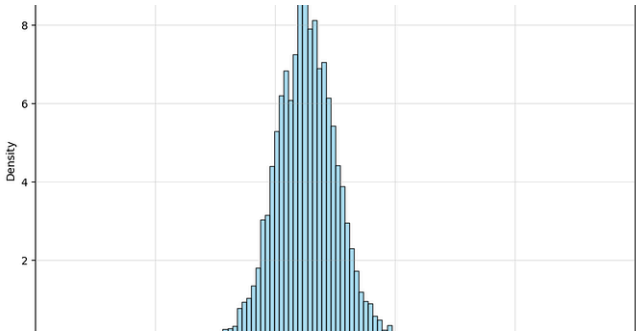
3d ago 4

Xin Cheng

Quantum Machine Learning for MNIST classification

Harnessing Qubits to Read Digits: MNIST Meets Quantum Circuits

3d ago 1

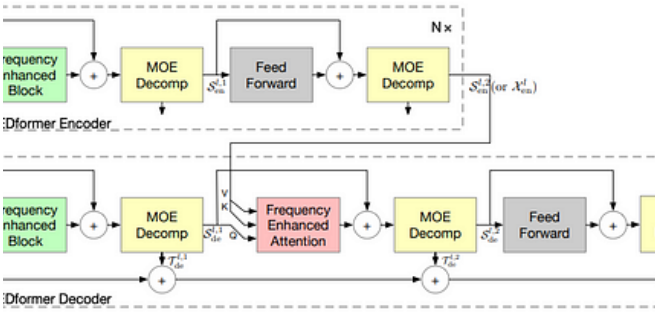


RationalPursuit

Is This Coin Fair?

Imagine you and your best friend have flipped a coin to decide who buys lunch every day fo...

Mar 24 70 2

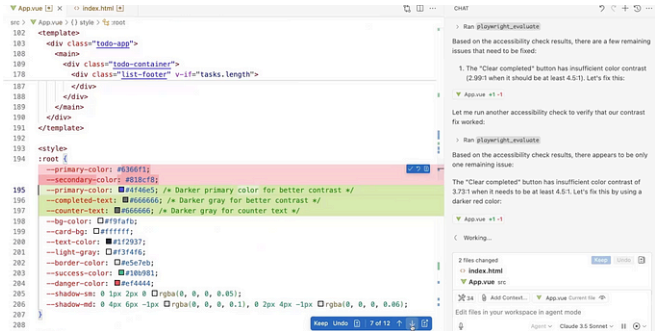


Dong-Keon Kim

FEDformer: Unleashing the Power of Frequency in Time Series...

A Deep Dive into Frequency Enhanced Decomposed Transformers for Long-Term...

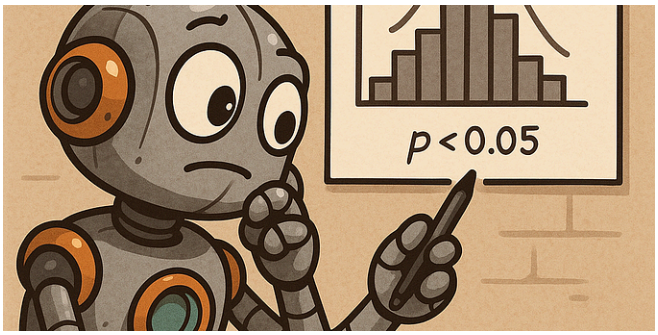
5d ago 18



In Coding Beauty by Tari Ibaba

VS Code's new AI agent mode is an absolute game changer

Wow this is insane.



In Towards AI by Robert Martin-Short

Data-Driven LLM Evaluation with Statistical Testing

Helping iterative projects move in the right direction.

6d ago 779 29

3d ago 47 1

See more recommendations