Running LLMs Locally on Your Computer

ClePy + Cleveland PyLadies

Large Language Models (LLMs) are powerful AI tools that can generate text, assist with coding, and provide insights across various domains

Prerequisites

https://github.com/eddie-cosma/clepy-ollama/blob/main/prerequisites.md

1. Install Python
2. Install Ollama
3. Download the tinyllama language model
4. Git clone github.com/eddie-cosma/clepy-ollama.git
5. Create a virtual environment and download dependencies

Goals

1. Explain benefits of self-hosted LLMs
2. Describe how to self-host LLMs using Ollama
3. Detail alternative methods of interaction with self-hosted LLMs
    a. Web API
    b. Python library

Benefits of running LLMs locally

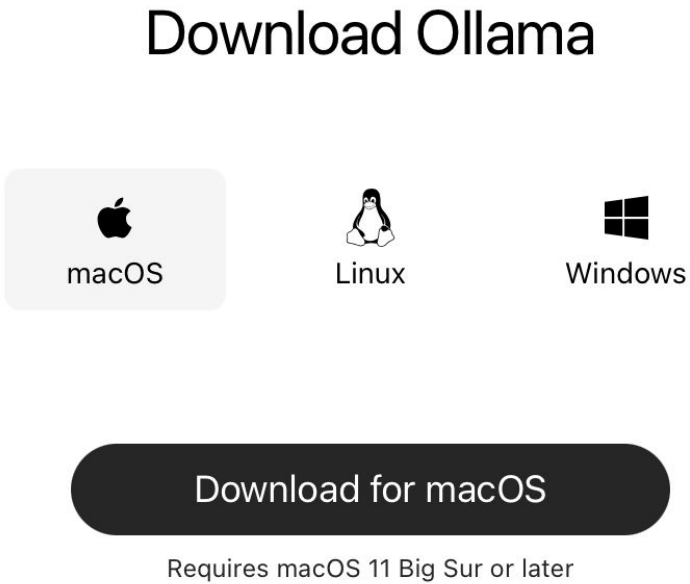
- **Privacy & Security** - Your data stays local
- **Offline Access** - No need for an internet connection
- **Cost Efficiency** - Avoid API fees and subscriptions
- **Customization** - Fine-tune models to your specific needs
- **Faster Response Times** - Local execution reduces latency

What is Ollama?

- Framework for running language models locally
- Open-source
- Includes API for managing models
- Includes access to pre-built models

Installing Ollama

1. Go to ollama.com/download
2. Select operating system
3. Click download button
4. Run the installer



Verify Installation

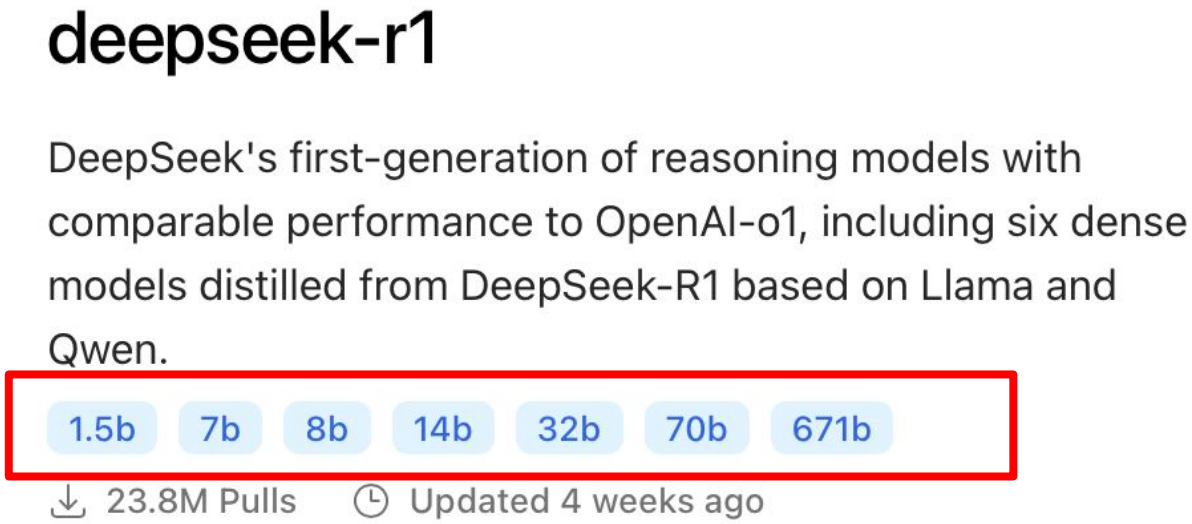
1. Open command prompt
2. Run `ollama --version`

```
eddiecosma@MacBookPro ~ % ollama --version
ollama version is 0.5.13
```

Model Size Considerations

- Models can be described in terms of **number of parameters**
- Parameters are numeric values that define model behavior
- More parameters means better computation
- More parameters means bigger model size
  - Larger filesize
  - More RAM needed

Model Size Considerations



Download a Model

Run `ollama pull tinyllama`

```
eddiecosma@MacBookPro ~ % ollama pull tinyllama
pulling manifest
pulling 2af3b81862c6... 100%

637 MB
…
verifying sha256 digest
writing manifest
```

Run a Model Interactively

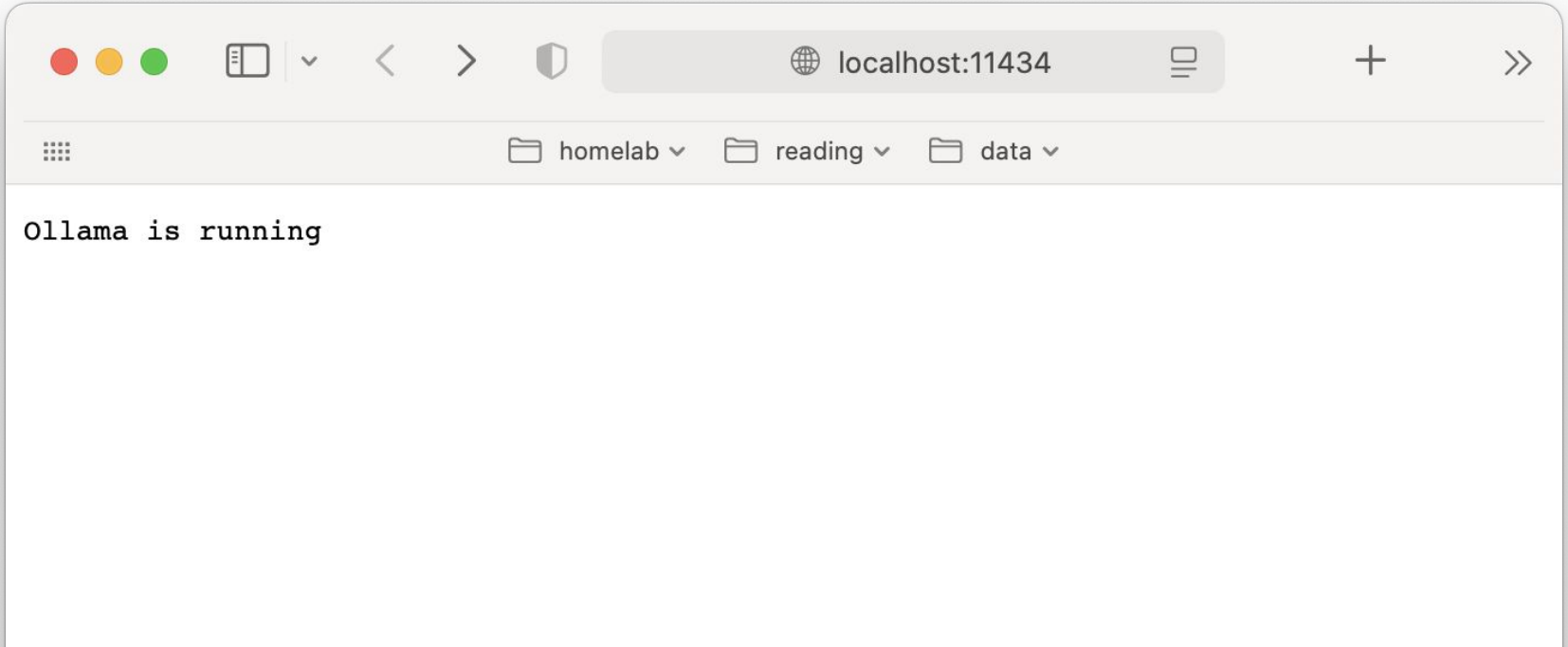
Run `ollama run tinyllama`

```
eddiecosma@MacBookPro ~ % ollama run tinyllama
>>> Send a message (/? for help)
```

To leave the prompt, type `/bye`

How else can we interact with this thing?

Using the Ollama Web API



I thought this was a Python Meetup...

Ollama Python Library

- Ollama can be accessed using a web API
- Ollama publishes a python library
- Library closely mirrors the web API
- Library allows for easy interaction using Python

Installing the Ollama Python Library

Install the library using `pip install ollama`

Activity

Run Jupyter Lab and open `ollama-python.ipynb`

Conclusion

- Ollama is a free, fast, and secure option for running LLMs locally
- You can easily interact with Ollama using python
  - Web API
  - Native ollama library

Join the Cleveland Tech Slack

https://cleveland-tech.vercel.app/