Ollama Setup Guide — Run Local LLMs on Your Laptop

For Students and Workshop Participants

1■■ What is Ollama?

Ollama is an open-source platform that allows you to run large language models (LLMs) locally — such as Llama 3, Mistral, and Gemma. It simplifies downloading, managing, and running AI models without cloud access. Perfect for Agentic AI and hands-on learning.

2■■ System Requirements

OS	Minimum Requirements
Windows 10/11	8 GB RAM (16 GB recommended)
macOS 12+ (M1/M2/M3)	8 GB RAM
Linux (Ubuntu/Debian)	8 GB RAM, Python 3.9+ optional

3■■ Installation Instructions

Windows:
1. Visit https://ollama.com/download
2. Download and run the installer
3. Open Command Prompt and verify with:
ollama --version

macOS:
1. Install via Homebrew:
brew install ollama/tap/ollama
2. Start service:
ollama serve

Linux:
1. Run installer:
curl -fsSL https://ollama.com/install.sh | sh
2. Verify with:
ollama --version

4■■ Running Your First Model

Run Llama 3.1 locally:

ollama run llama3.1

When the prompt appears, type something like:
Write a haiku about AI students.

5■■ Managing Models

Command	Description
ollama list	View installed models
ollama pull <model>	Download a model
ollama run <model>	Run interactively
ollama rm <model>	Remove model

6■■ Using Ollama in Python

pip install ollama
Example code:

```
import ollama
response = ollama.chat(model='llama3.1', messages=[{'role': 'user', 'content': 'Explain Agentic AI.'}])
print(response['message']['content'])
```

7■■ Troubleshooting Tips

- Restart terminal if 'command not found'
- Pull smaller models if memory is low
- Run ollama serve if service is inactive

8■■ Workshop Extension

Connect Ollama with LangChain or CrewAI to build autonomous agents that can research, plan, and summarize automatically.

■ **Tip:** Try smaller models like 'mistral' or 'phi3' for laptops with 8GB RAM.