AI and Python -
Ollama for Local LLM AI Usage

This Still Isn’t AI…

- AI is a buzzword for investors…
- You need a correct Frame of Reference to find solutions…
- Turning Objects into text…

AI is HARD, API’s are NOT

- The smartest minds of a generation accessed with 10 lines of code…

What is Ollama

- Framework for running LLM’s on your local system
- Run from CLI
- REST API and cURL
- And Python Module
- Model Quantization
    - Compression so it runs on your system
    - Personal llama2 experience…
    - https://symbl.ai/developers/blog/a-guide-to-quantization-in-llms/

Models

- Models are trained for types of tasks.
- Smaller models run on worse machines
- https://ollama.com/library
- License -
  - https://ollama.com/library/llama3.1
- https://ollama.com/library/samantha-mistral

Installing

- https://ollama.com/

Running from CLI

- ollama
- pull
- list
- run
- /bye

Python

- python3 -m pip install ollama
- Ollama must be running
- Ollama configures to run at startup
- The model has to be downloaded, but does not need to be running

Text - Prompting

- Roles
- Example = ollama-python.py

Image Recognition - LLaVA

- ollama pull llava
- https://llava-vl.github.io/
- jpg or png
- Example = ollama-image.py

cURL

```
curl --location 'http://localhost:11434/api/generate' \
--data '{
    "model": "llama2:7b",
    "prompt": "what is a pbj sandwich",
    "stream": false
}'
```

Labs

- lab-chat.py
  - Talk to your AI
- lab-chat-memory.py
  - Give your AI a memory
- lab-image.py
  - Create a Title, Description and CSV formatted Tags for image