setup-local-llm

- Why run LLM locally?
  - ◇ privacy
  - ◇ own data
  - ◇ customization
  - ◇ cost

- Why use Llama.cpp instead of alternatives?
  - ◇ Ollama
    - ■ is a blackbox (sometimes)
  - ◇ Llama Edge
    - ■ isn't fully baked
  - ◇ VLLM
    - ■ dependency purgatory!
  - ◇ S-LANG
    - ■ dependency purgatory!

- Download LLM model
  - ◇ https://huggingface.co/models?sort=trending&search=qwen+gguf
  - ◇ https://huggingface.co/Qwen/Qwen1.5-0.5B-Chat-GGUF
    - ■ go to file and versions.
    - ■ should be about 298 MB

- Setup Llama.cpp Server in a Docker container
  - ◇ sudo docker run -v /home/hd/ml/models:/models -p

8000:8000 ghcr.io/ggml-org/llama.cpp:server -m /models/qwen1_5-0_5b-chat-q2_k.gguf --port 8000 --host 0.0.0.0 -n 512 --api-key apple

- Access Local LLM Via:
  - ◇ Browser
    - ■ http://0.0.0.0:8000/#/
  - ◇ Command line

```
curl -XPOST -H 'Authorization: Bearer apple' -H "Content-type: application/json" -d '{
 "model": "/models/qwen1_5-0_5b-chat-q2_k.gguf",
 "messages": [
  {
   "role": "system",
   "content": "You are a helpful coding assistant."
  },
  {
   "role": "user",
   "content": "How do I check if a Python object is an instance of a class?"
  }
 ]
}' 'http://0.0.0.0:8000/v1/chat/completions'
```

  - ◇ Python
    - ■ nano main.py with contents below

```
from openai import OpenAI

client = OpenAI()

completion = client.chat.completions.create(
  model="/models/qwen1_5-0_5b-chat-q2_k.gguf",
  messages=[
    {"role": "developer", "content": "You are a coding assistant."},
    {
      "role": "user",
      "content": "How do I check if a Python object is an instance of a class?",
    },
  ],
)

print(completion.choices[0].message.content)
```

    - ■ install dependencies

```
pip install openai
```

    - ■ export environment variables

```
export OPENAI_API_KEY="apple"
export OPENAI_BASE_URL="http://0.0.0.0:8000/"
```

- run

python main.py

- **Optimization Tips**

◇ use {"max_tokens": 200}

◇ use these args when running

- --top_k 40 --top_p 0.9 --temp 0.5 --repeat_last_n 64 --repeat_penalty 1.3 -t 16