Introduction

- llama.cpp: LLM inference project in C++
  - original goal: Run at the time state-of-the-art LLaMA model on an M1 MacBook
- whisper.cpp from same authors for audio transcription
- specialized on low-powered hardware, CPU inference

Why self-host

- Privacy
- Reliability
- Special Models
- Cost
- Offline Support

Whisper

- state-of-the-art transcription model by OpenAI
- use-case: transcribe long voice messages from contacts

Demo

download model

```
podman run -it --rm \
          -v $PWD/models:/models \
          ghcr.io/ggerganov/whisper.cpp:main "./models/downlo
```

transcribe using specific tag due to issue with main:

```
podman run -it --rm \
        -v $PWD/models:/models \
        -v $PWD/audios:/audios \
        ghcr.io/ggerganov/whisper.cpp:main-a9d06ce1518c5abd8c
```

LLamafile

- combines Cosmopolitan Libc and llama.cpp
- distribute LLMs as a single file, weights embedded
    - possible to use external weights
- file portable across Linux + Mac + Windows …
- webinterface
- OpenAI compatible API server
- supported by Mozilla

Demo

Just download available model and execute:

```
wget 'https://huggingface.co/Mozilla/llava-v1.5-7b-llamafile/
mv "llava-v1.5-7b-q4.llamafile?download=true" llava-v1.5-7b-q
chmod +x llava-v1.5-7b-q4.llamafile
./llava-v1.5-7b-q4.llamafile
```

Security

- Keep in mind: llamafiles are arbitrary executables!
- make sure you trust the source!
- Better use llamafile without embedded weights