AICamp

AISG Day 5

Ollama &
Local LLMs

David Tang

What is Ollama?

Local LLMs

Open-source

Ease of use

https://ollama.ai

Contents

(1) Info deck
(2) Ollama spin-up
(3) Usecase demo

Why tho?

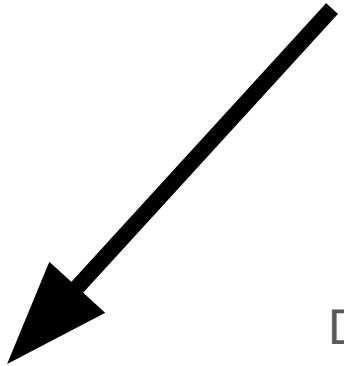
High privacy is crucial (sensitive data)

Offline access is essential

Have the necessary hardware

Cost of API usage is a concern

Prefer more control over the model



Meta AI

Llama 3.1
Llama 3.2
Llama 3.3

Multimodal
Multilingual
Open source
Deploy anywhere

Models

https://github.com/ollama/ollama?tab=readme-ov-file#model-library

Performance

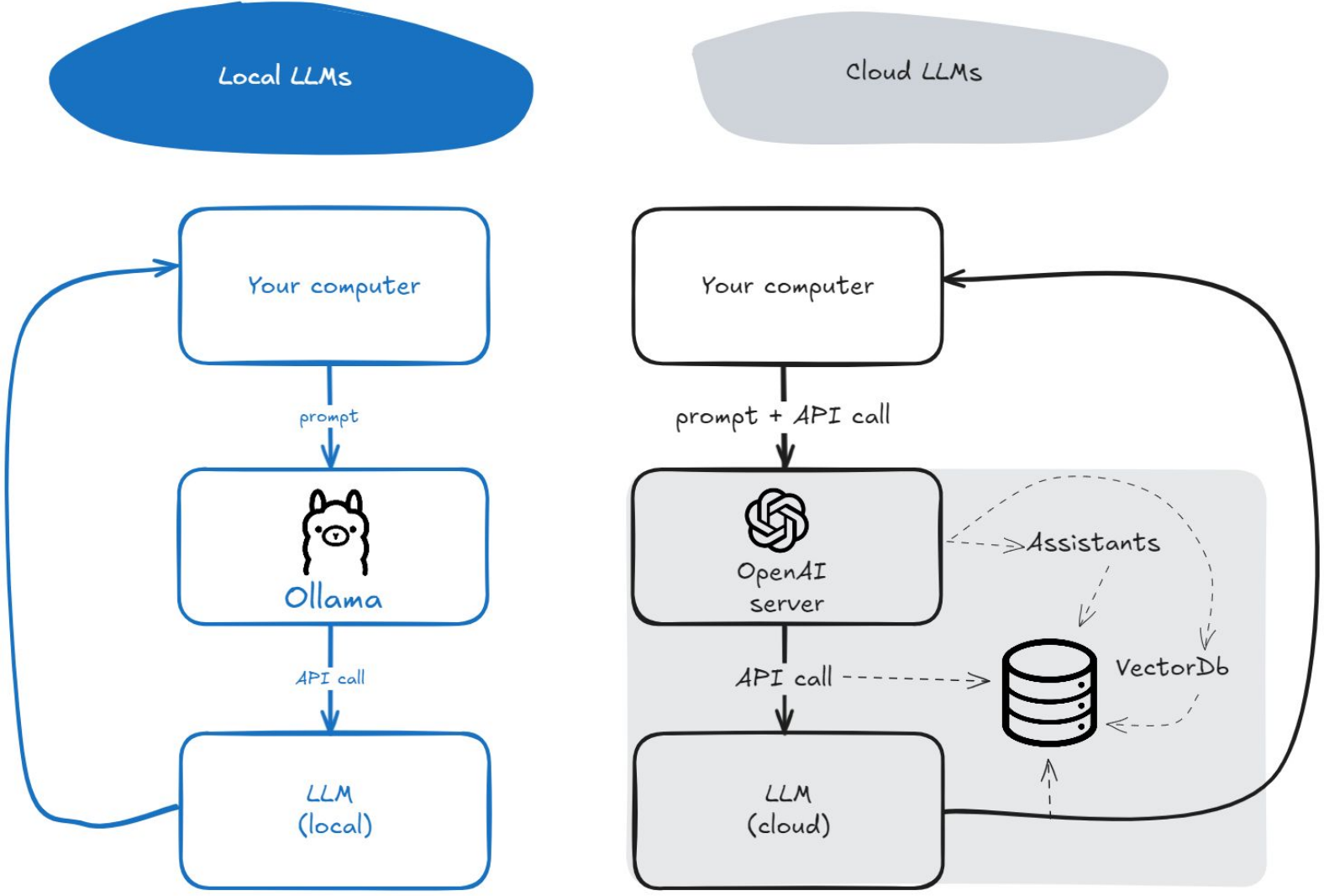
Quantization reduces the size of a machine learning model by reducing the *precision of the numbers*.

Benefits include **reduced model size**, **faster inference** and lower memory bandwidth.

No network latency.

Floating-Point Precision and Quantization

- FP32: 32-bit floating point, highest precision, largest size.
- FP16: 16-bit floating point, half the size of FP32, good balance of size and performance.
- 8-bit (INT8): 8-bit integer, smaller and faster than floating point, potential for accuracy loss.
- 4-bit: Even smaller and faster than 8-bit, but greater potential for accuracy loss.


Local LLMs
Cloud LLMs
Your computer
prompt
Ollama
API call
LLM
(local)
Your computer
prompt + API call
OpenAI
server
Assistants
API call
VectorDb
LLM
(cloud)

Let's connect!

www.linkedin.com/in/drdavidtang

 @davidtang.ai