Week 7: Local LLM + RAG + Streamlit

Full Stack RAG with Local LLM

Semester 2/2568

Agenda

1. Ollama
2. RAG Pipeline
3. Streamlit
4. Lab 6

What is Ollama?

Local LLM Runtime

- Run LLMs locally
- No API key
- Free unlimited
- Data stays local

We use: qwen2.5:7b

Ollama Commands

```
# Pull model
ollama pull qwen2.5:7b

# Run
ollama run qwen2.5:7b

# API at http://localhost:11434
```

Call from Python

```
import requests

response = requests.post(
    "http://localhost:11434/api/generate",
    json={
        "model": "qwen2.5:7b",
        "prompt": "What␣is␣RAG?",
        "stream": False
    }
)
print(response.json()["response"])
```

RAG Flow

Question → Embed → Search → Context → LLM → Answer

1. Embed question (bge-m3)
2. Search OpenSearch
3. Build context
4. Send to Ollama
5. Return answer

Streamlit UI

```
import streamlit as st

st.title("RAG␣Q&A")

if prompt := st.chat_input("Ask..."):
    st.chat_message("user").write(prompt)
    response = call_api(prompt)
    st.chat_message("assistant").write(response)
```

Lab 6: Complete RAG (3.75%)

Tasks:

1. Setup Ollama
2. Test LLM
3. Run complete RAG
4. Run Streamlit UI

Questions? Complete RAG System!