

# Lab 6: RAG + Local LLM + Streamlit

CSI403 - Full Stack Development

Faculty of Information Technology

Sripatum University

Weight: 3.75%

# Objectives

- ✓ Setup Ollama
- ✓ Complete RAG pipeline
- ✓ Run Streamlit UI

**Repository:** <https://github.com/amornpan/Generic-RAG>

# Task 1: Install Ollama

- 1 Download from <https://ollama.ai>
- 2 Install and run

**Ollama** runs LLMs locally on your machine.  
No API keys or cloud services required!

## Task 2: Pull Model

```
ollama pull qwen2.5:7b
```

### Model Info:

- **qwen2.5:7b** - 7 billion parameters
- Good balance of quality and speed
- Requires 8GB RAM

## Task 3: Test LLM

```
ollama run qwen2.5:7b
```

Try asking:

```
>>> What is RAG?
```

Press Ctrl+D to exit.

## Task 4: Run API with LLM

```
cd Generic-RAG  
python api.py
```

The API now connects to:

- **OpenSearch** - for document retrieval
- **Ollama** - for LLM generation

## Task 5: Run Streamlit

```
streamlit run app.py
```

Opens at <http://localhost:8501>

## Task 6: Test Complete Flow

- 1 Open <http://localhost:8501>
- 2 Ask questions about your documents
- 3 Verify RAG responses include context

### **RAG Flow:**

Query → Search → Retrieve Context → LLM → Response

## Task 7: Modify Prompt

Edit the prompt template in `api.py`:

- Find the prompt template
- Customize instructions
- Test with different prompts

# Deliverables

Item	Check
Ollama running	<input type="checkbox"/>
LLM tested	<input type="checkbox"/>
Complete RAG working	<input type="checkbox"/>
Streamlit UI working	<input type="checkbox"/>

**Deadline: Sunday 23:59**