# Week 1: Introduction to RAG + Local LLM

## Full Stack RAG with Local LLM

Semester 2/2568

# Agenda

1. Course Overview
2. What is RAG?
3. Why Local LLM?
4. Technology Stack
5. Assessment
6. Setup

# Course Information

| Item | Detail |
|---|---|
| Course Code | CSI403 |
| Course Name | Full Stack RAG with Local LLM |
| Credits | 3 (2-3-5) |
| Semester | 2/2568 |

**"Build Your Own AI, Not Just API Calls"**

# What Makes This Course Different?

**Traditional Course:**

- OpenAI API Key
- Pay per Token $$$
- Cloud Dependency
- Data sent to Cloud

**This Course:**

- Local Ollama LLM
- Free & Unlimited
- Self-Hosted
- Data Stays Local

# The Problem with LLMs

**Large Language Models have limitations:**

- Knowledge cutoff date
- No access to your private documents
- Can hallucinate (make up facts)
- Generic responses

**Solution: RAG (Retrieval-Augmented Generation)**

## Retrieval-Augmented Generation

**RAG = Search your documents + Use LLM to generate answers**

1. **Retrieval:** Find relevant documents
2. **Augmented:** Add context to prompt
3. **Generation:** LLM creates answer

# Cost Comparison

**Cost per 1 Million Tokens**

| Service | Cost | Type |
|---|---|---|
| OpenAI GPT-4 | $30.00 | Cloud API |
| OpenAI GPT-3.5 | $2.00 | Cloud API |
| **Ollama (Local)** | **$0.00** | **Self-Hosted** |

**Local LLM = Free Forever!**

# Benefits of Local LLM

**Cost & Freedom:**

- No API fees
- Unlimited usage
- No rate limits

**Privacy & Control:**

- Data stays local
- Offline capable
- Full customization

# Tech Stack (100% Self-Hosted)

| Component | Technology | Type |
|---|---|---|
| Frontend | Streamlit | Local |
| Backend | FastAPI | Local |
| Embedding | HuggingFace (bge-m3) | Local |
| Vector DB | OpenSearch | Local |
| **LLM** | **Ollama (qwen2.5:7b)** | **Local** |

**No API Key Required!**

# Assessment Overview

| Category | Total |
|---|---|
| Attendance | 10% |
| Quiz (4 times) | 20% |
| Lab (8 times) | 30% |
| Project | 40% |
| **Total** | **100%** |

# Software to Install

**Please install before Week 2:**

1. **Miniconda** - Python 3.10+
2. **Git** - git-scm.com
3. **VS Code** - code.visualstudio.com
4. **Docker Desktop** - docker.com
5. **Ollama** - ollama.ai

# Questions?

Welcome to Full Stack RAG with Local LLM! "Build Your Own AI, Not Just API Calls"