LLM-RAG from Scratch: Build Your Own Open-Source AI Chatbot

📅 **Wednesday, 10 September 2025**
📍 **ETH Zurich, Room F91**

This hands-on workshop offers a deep dive into the development of intelligent AI systems using the Retrieval-Augmented Generation (RAG) framework and open-source tools. Participants will learn how to build a fully functional AI chatbot that can search and understand large collections of documents to generate accurate, context-aware responses.

Throughout the session, you'll be guided through the complete RAG pipeline: from data ingestion and embedding generation to vector database integration and connection to an open-source LLM. Using frameworks such as LangChain, Hugging Face Transformers, and Ollama, you will set up each component yourself. The workshop also covers practical aspects like data handling and strategies to ensure data privacy and security during deployment.

Whether you're a developer looking to integrate LLMs into real-world applications or simply curious about how open-source AI chatbots work under the hood, this session will give you the skills and insights to build and deploy your own system.

By the end of the workshop, you'll walk away with a working open-source chatbot and a clear understanding of how to customize and expand it for your own use cases.

During this Workshop, You Will:

- **Ingest & Preprocess Data**
  Convert PDFs, Markdown, or text files into clean, chunked passages ready for embedding.
- **Generate & Store Embeddings**
  Use Hugging Face Transformers to produce vector embeddings and load them into a vector database.
- **Implement Retrieval & Generation**
  Orchestrate your RAG flow with LangChain: retrieve the top-k passages for a query, then feed them into an open-source LLM (e.g., Ollama or Llama) to generate context-aware answers.
- **Secure & Package Your Pipeline**
  Wrap your components in Docker (or Docker Compose), so you can run the entire pipeline in one reproducible environment.

By the end of the day, you'll have a fully functional Open-Source AI chatbot!

Requirements

- Basic Skills: Familiarity with Python scripting and command-line usage. No advanced ML or DevOps experience needed.
- What to Bring: Your own laptop (admin rights to install Docker) and a stable Wi-Fi connection.
- Tools & Frameworks: We'll use open-source libraries for each step:
  Data & Embeddings: Python + Hugging Face Transformers
  Pipeline Orchestration: LangChain
  LLM Serving: Ollama (or another local LLM)
  Deployment: Docker

All installations and setup will be performed live: just bring your laptop and enthusiasm!