Open Source AI workshop

May 8, 2025, 9:00 – Bern University of Applied Science (BFH)

LLM-RAG from Scratch: Crafting Intelligent AI Retrieval Systems

Embark on a hands-on journey to build your very own Retrieval-Augmented Generation (LLM-RAG) system from scratch using open source AI models and cutting-edge tools. In this workshop, you'll learn how to integrate large language models with retrieval pipelines and dive into vast repositories of data, extracting the precise information you need to generate clear, accurate, and contextually rich responses.

During this Workshop, You Will:

- **Master the Fundamentals:** Learn how the RAG framework combines retrieval and generation to produce responses that are both context-aware and highly accurate.
- **Build Your Own RAG System:** Follow a step-by-step guide to develop a complete LLM-RAG pipeline—from setting up a retrieval database and generating document embeddings to integrating these components with an open source LLM.
- **Gain Hands-On Experience:** Work through practical exercises and real-world examples that demonstrate how to efficiently search extensive documentation and leverage that data to inform precise responses.
- **Ensure Data Security:** Discover effective strategies for deploying open source LLMs with a strong emphasis on maintaining data privacy and ensuring secure, robust operations.

By the end of the day, you will have the skills and knowledge necessary to deploy a fully operational LLM-RAG solution for your personalized chatbot.

Requirements

- **Target Audience:** This workshop is designed for any individual with basic programming skills who is interested in learning more about AI and how to apply it in practical projects.
- **What to Bring:** Bring your own laptop with a reliable internet connection (WiFi).
- **Prerequisites:** A basic understanding of Python is required. No advanced machine learning or AI expertise is necessary.

Tools and Frameworks

We will work with a range of open source tools and frameworks for every stage of the RAG pipeline, including libraries for data retrieval, language model orchestration, and embedding generation, such as **LangChain, Hugging Face Transformers, Ollama**, and various tools for document indexing and vector search.