
CH Open
Source | Business | Community
Open Source AI
Workshops
8.-9. Mai 2025
BFH, Brückenstrasse 73
3005 Bern

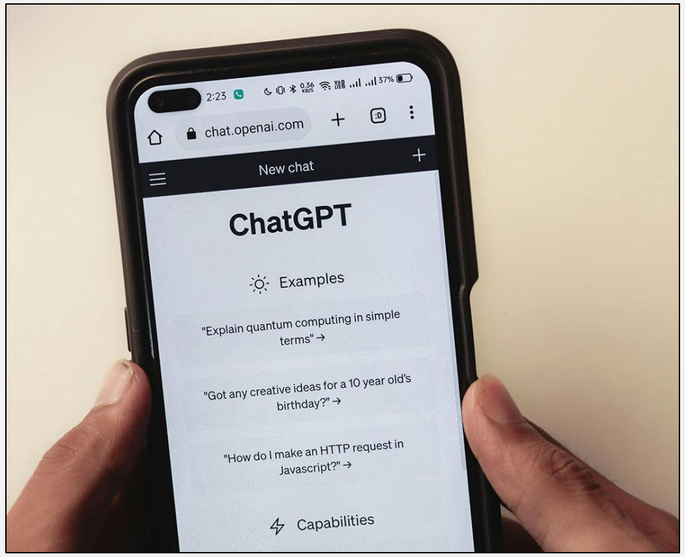
LLM-RAG from Scratch: Intelligent AI Retrieval Systems

Schedule:

08:30 – Coffee and croissants
09:00 – Start / Block 1
10:30 – Break
11:00 – Block 2
12:30 – Lunch break
13:30 – Block 3
15:00 – Break
15:30 – Block 4
17:00 – Closing

ornella.vaccarelli@hefr.ch

LLM-RAG from Scratch: Intelligent AI Retrieval Systems



ornella.vaccarelli@hefr.ch

LLM-RAG from Scratch: Intelligent AI Retrieval Systems

What Is a Large Language Model (LLM)?

A **Large Language Model** is a deep neural network trained on vast amounts of text data.
It learns to predict the next word in a sequence, allowing it to generate coherent, human-like text, perform question *answering*, *summarization*, *translation*, and more.

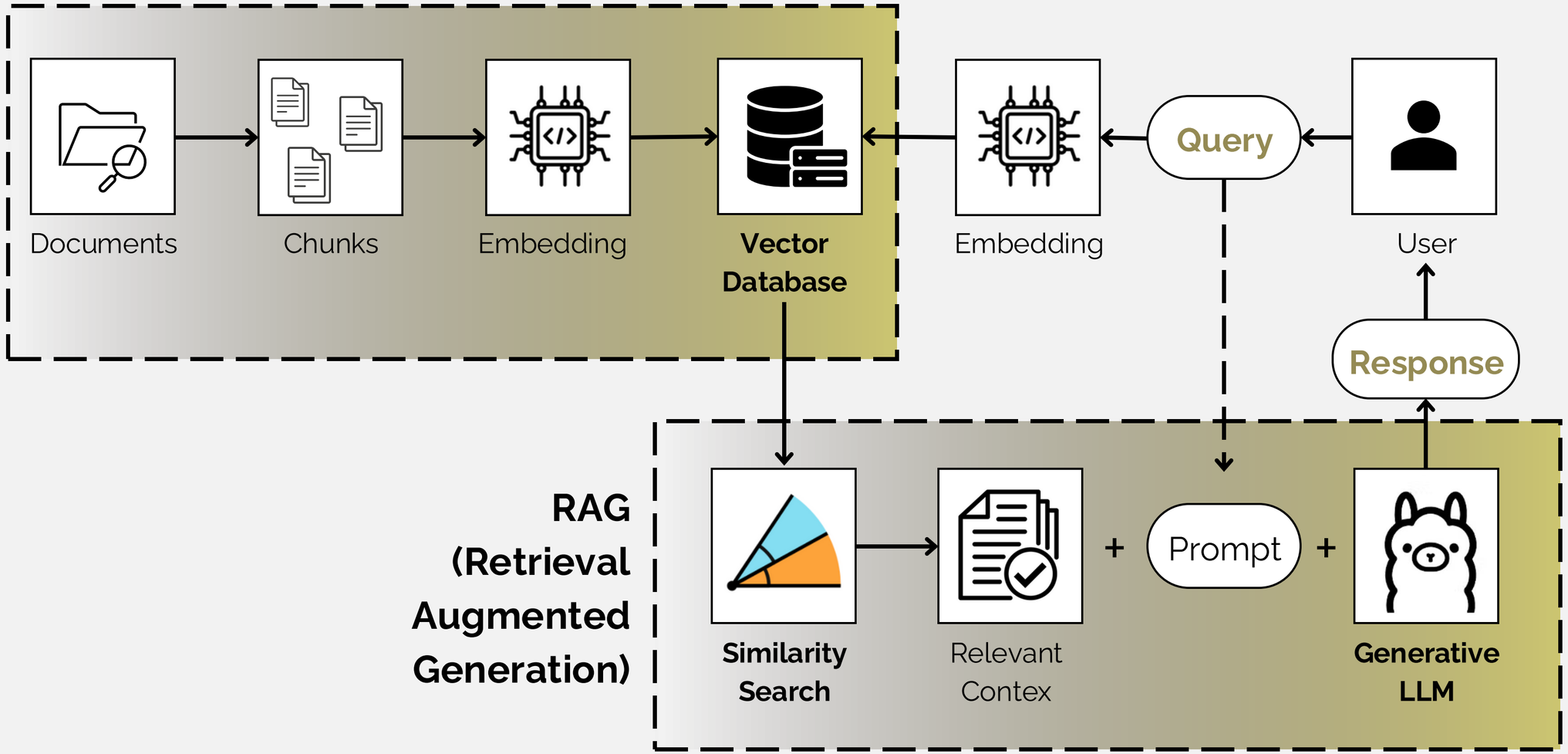
Limit of LLMs :

- Knowledge cutoff → outdated or incomplete answers
- Tendency to “hallucinate” facts

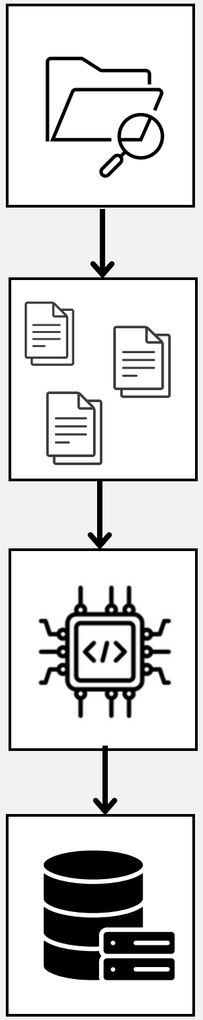
RAG :

- **Retrieval**: find relevant documents or passages
- **Augmentation**: inject retrieved context into prompt
- **Generation**: LLM produces fact-grounded, context-aware output

Technical Framework - Architecture



VECTOR DATABASE



PDF Extractor — Extract text from PDF

Text Segmentation — Divide the text in chunks

Embedding Model — Transform the chunks in vectors

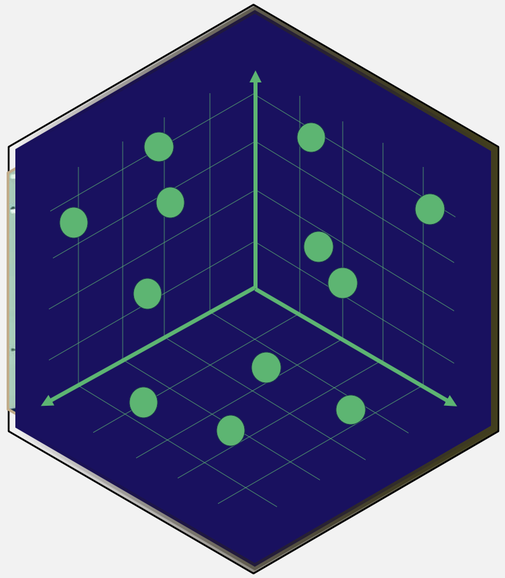
Vector Database — Store the vectors in a database

Open Source AI workshop Chunk 1

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

LLM-RAG from Scratch: Crafting Intelligent AI Retrieval Systems

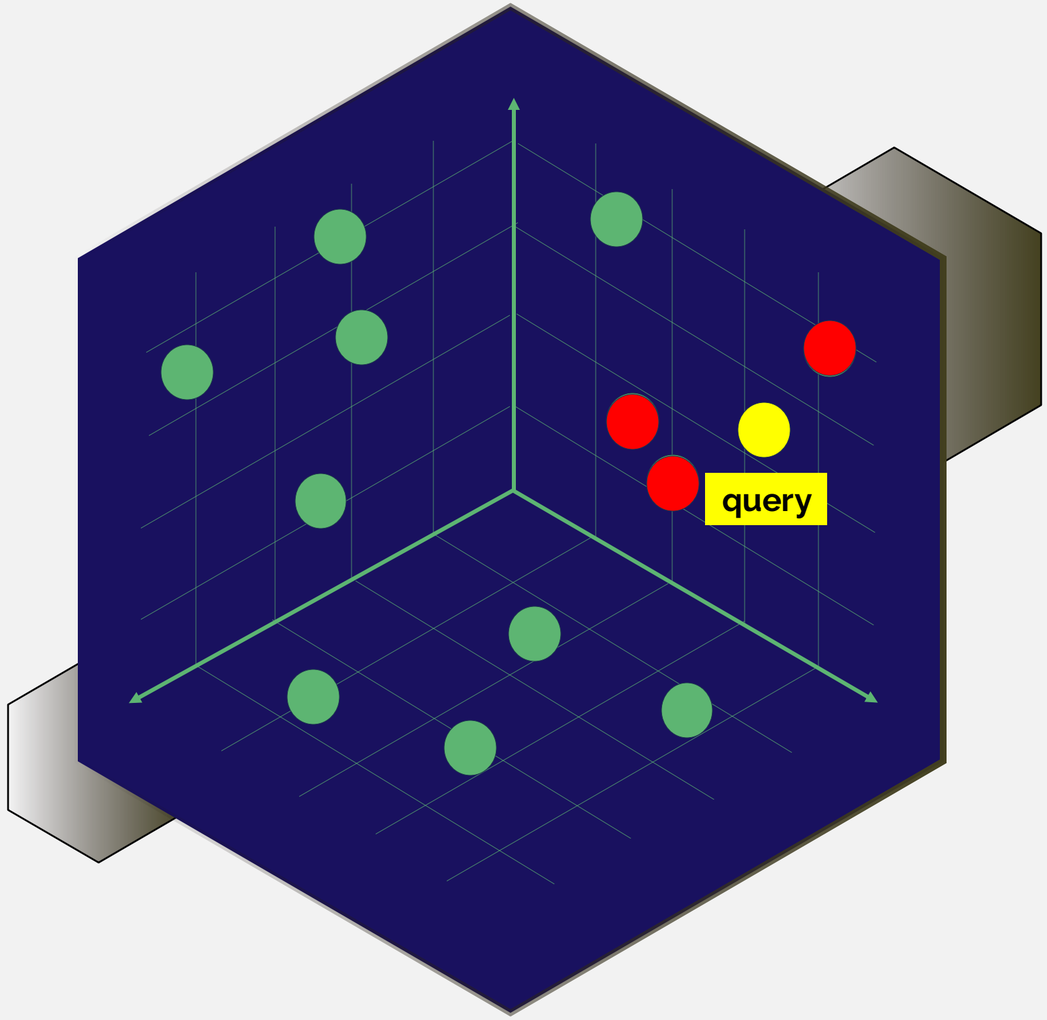
Embark on a hands-on journey to build your very own Retrieval-Augmented Generation (LLM-RAG) system from scratch using open source AI models and cutting-edge tools. In this workshop, you'll learn how to integrate large language models with retrieval pipelines and dive into vast repositories of data,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



SIMILARITY SEARCH

Query : also transform into a vector.
We search all the chunks "similar" to the query in order to properly answer.

- **Metric:** Cosine similarity $\mathrm{Cos\_sim}(\mathbf{q}, \mathbf{v}_i) = \frac{\mathbf{q} \cdot \mathbf{v}_i}{\|\mathbf{q}\| \|\mathbf{v}_i\|}$



LLM

LLM open-source:

- Full control of data.
- Local hosting (via Colab).

LLM that we will test:

- Mistral-7B: (small and performant model, multilingual, and ideal for starting)

Workshop

GitHub: https://github.com/ovaccarelli/LLM-RAG

```
LLM-RAG/
 ├ notebooks/
 │    ├ llm_rag_Open_Source_AI_Workshop_1.ipynb
 │    ├ llm_rag_Open_Source_AI_Workshop_2.ipynb
 │    ├ llm_rag_Open_Source_AI_Workshop_3.ipynb
 │    ├ llm_rag_Open_Source_AI_Workshop_4.ipynb
 │    ├ llm_rag_Open_Source_AI_Workshop_final_to_complete.ipynb
 │    └ llm_rag_Open_Source_AI_Workshop_final.ipynb
 ├ data/
 │    ├ sample_pdf/      ← test PDF(s) for extraction
 │    ├ PDFs/            ← the main PDF(s) of our LLM-RAG
 │    └ vectorstores/    ← saved FAISS indices
 ├ README.md
```