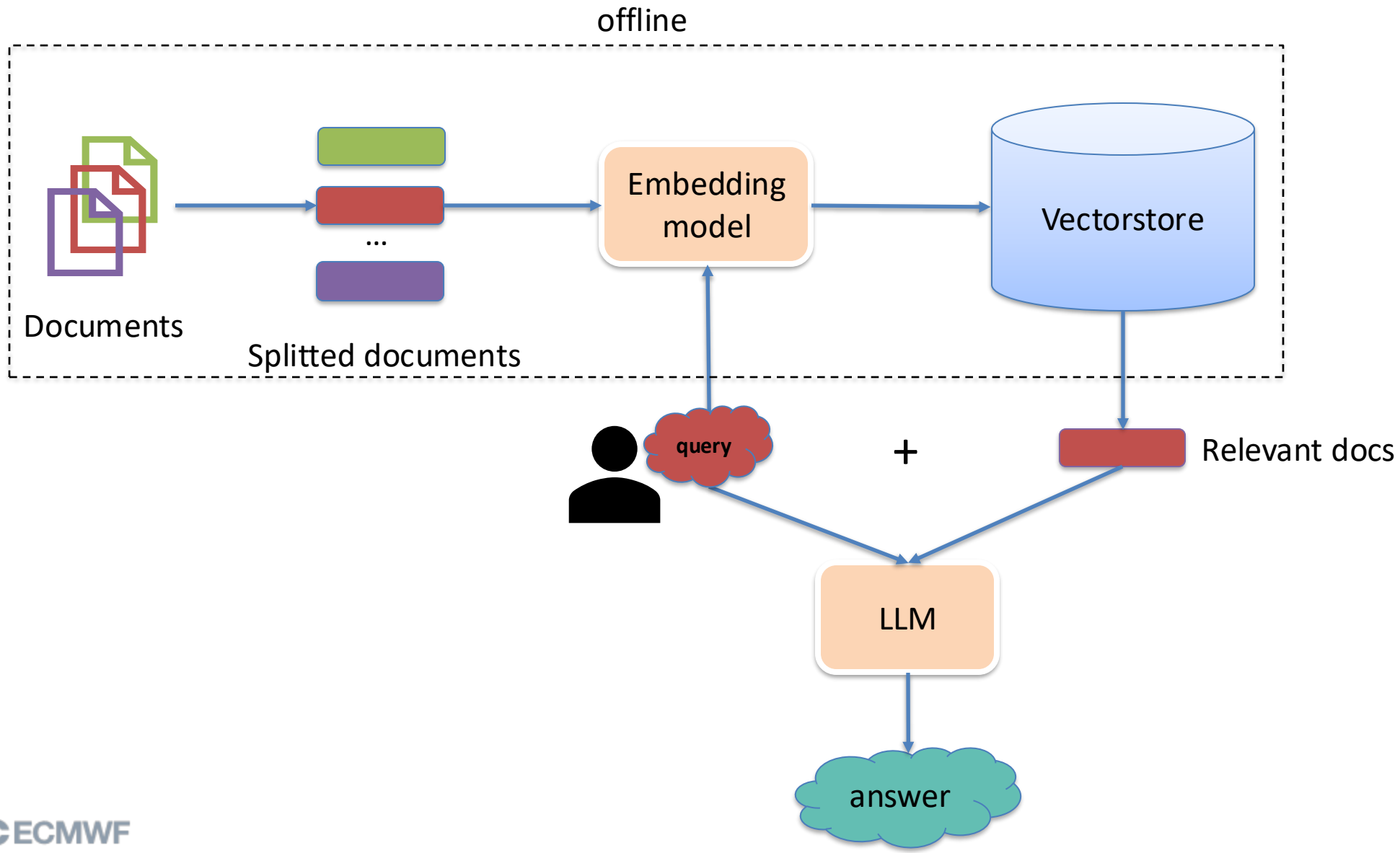
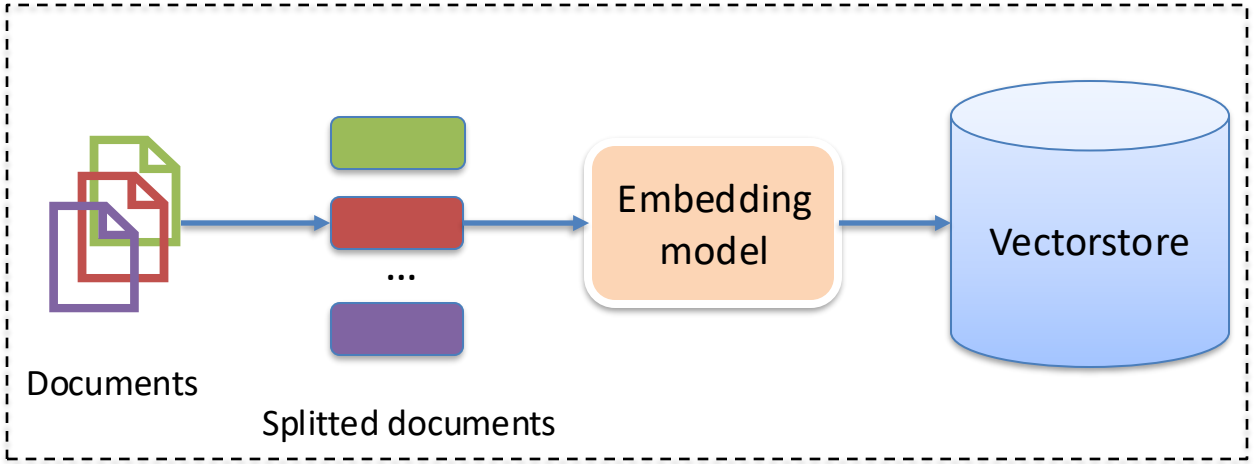
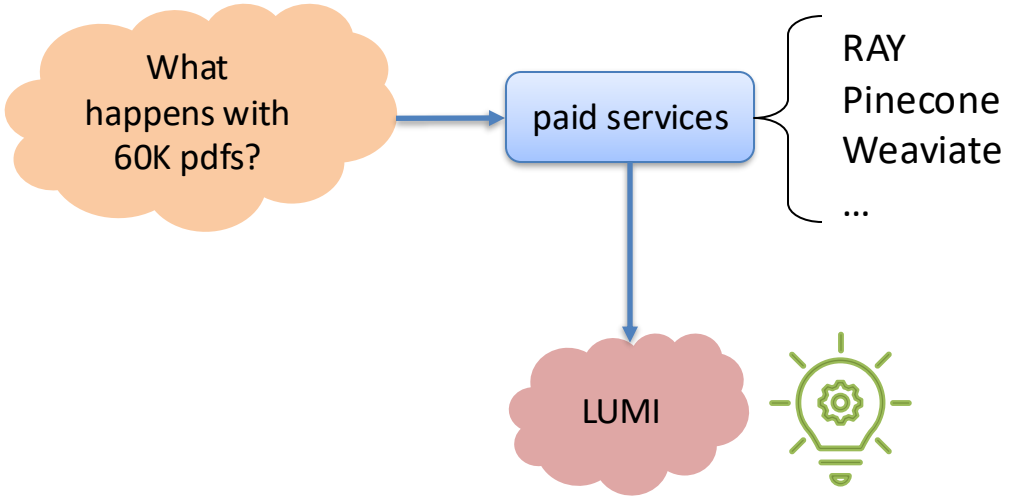
RAG PIPELINE



BUILD VECTOR STORE

- Langchain





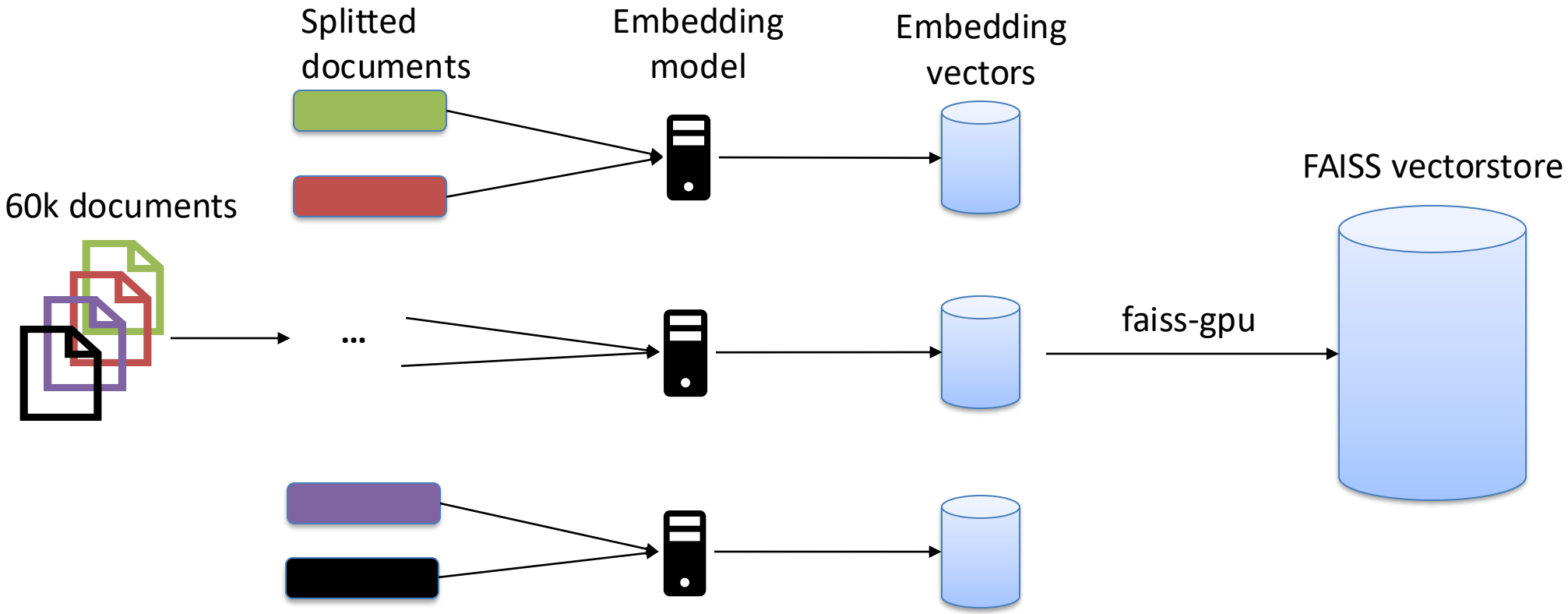
$$\boldsymbol{documents} = PyMuPDFLoader(pdf\_path).load()$$

$$\boldsymbol{text\_splitter} = RecursiveCharacterTextSplitter(chunk\_size = 1000, chunk\_overlap = 200)$$
$$\boldsymbol{split\_documents} = text\_splitter.split\_documents(documents)$$

$$\boldsymbol{vectordb} = Chroma.from\_documents(split\_documents, embedding = OpenAIEmbeddings(), persist\_directory = "./db")$$

BUILD VECTOR STORE FROM SCRATCH

- PyTorch Distributed Data Parallel (DDP)



$sampler = DistributedSampler(dataset)$

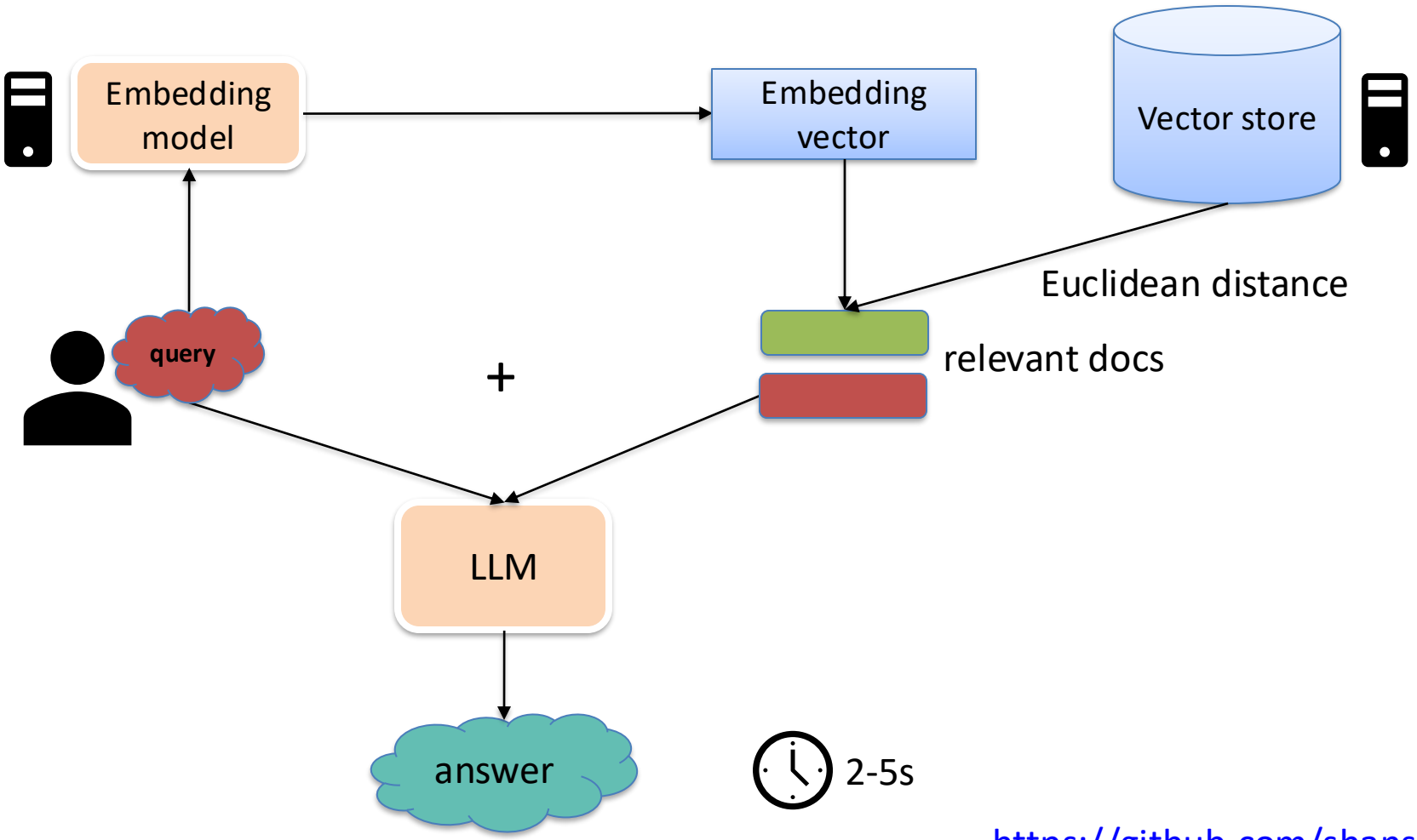
$model = DDP(model, device\_ids = [local\_rank])$

~1B tokens
15-45 minutes

$gpu\_index = faiss.index\_cpu\_to\_all\_gpus(cpu\_index)$

3.3 million vectors
0.5s

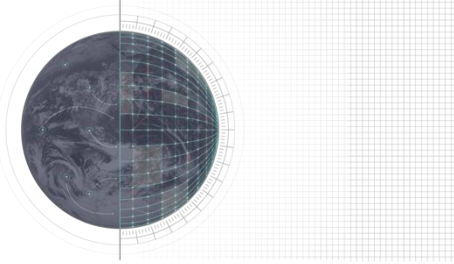
RETRIEVAL & GENERATION



SCAN ME

https://github.com/shanshanwangcsc/RAG-60K/tree/main

FAISS VS CHROMA



	Vector store building		Vector store retrieval			
	Insert speed	capacity	retrieve speed	retrieve quality	Metadata filtering	
Chroma	**	**	**	**	****	
Faiss	*****	*****	*****	*****	**	