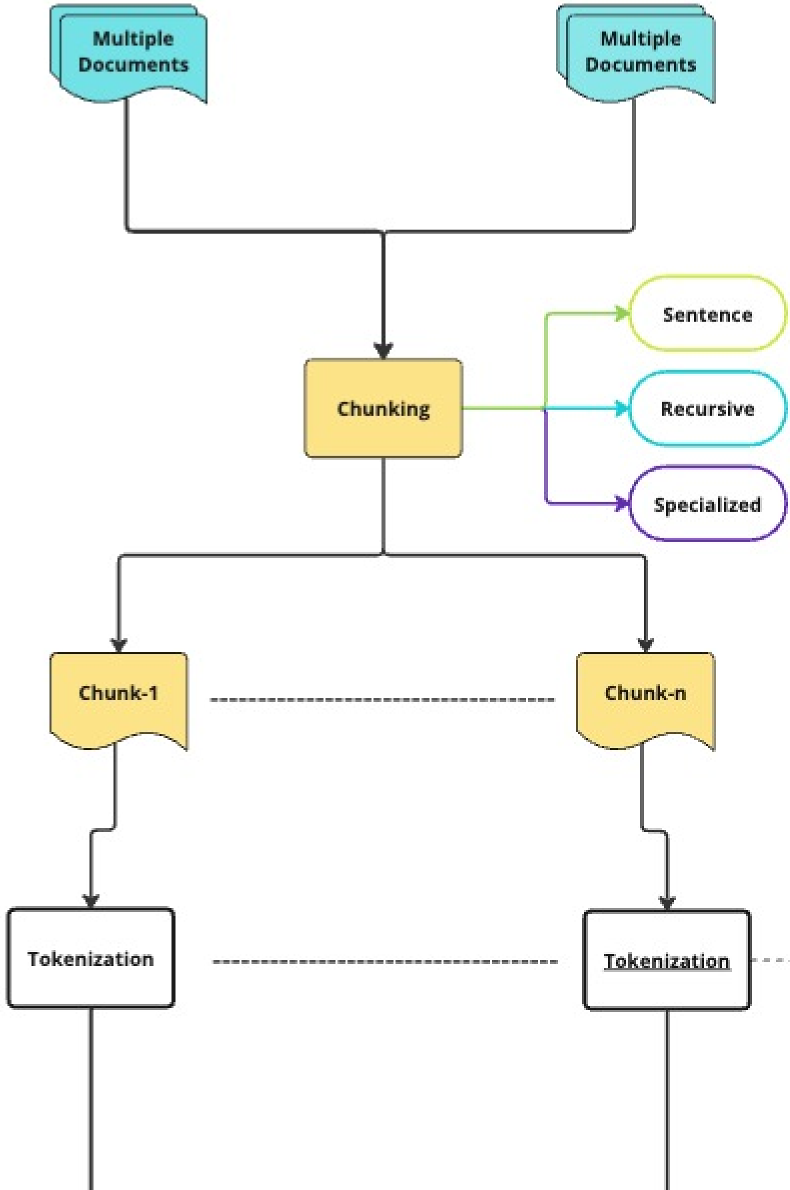
Stage 1 : Data Preparation

www.anyscale.com

Building RAG-based LLM Applications for Production

In this guide, we will learn how to develop and productionize a retrieval augmented generation (RAG) based LLM application, with a focus on scale and evaluation.



Multiple Documents

Multiple Documents

Chunking

Sentence

Recursive

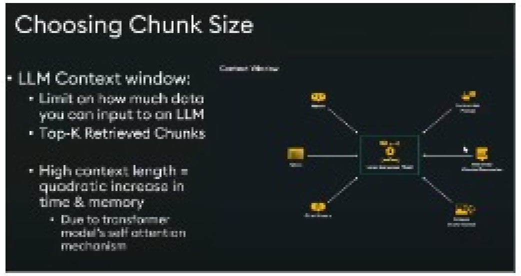
Specialized

Chunk-1

Chunk-n

Tokenization

Tokenization



Choosing Chunk Size

- LLM Context window:
  - Limit on how much data you can input to an LLM
  - Top-K Retrieved Chunks
  - High context length = quadratic increase in time & memory
    - Due to transformer model's self attention mechanism

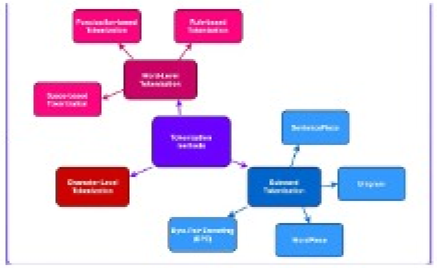
Context Window

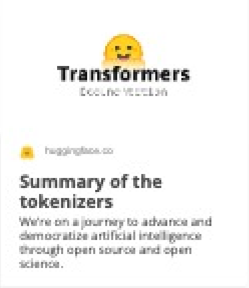
Chunking Best Practices for RAG Applications

YouTube · Updated X hours ago

LangChain Data Loaders, Tokenizers, Chunking, and...

YouTube · Updated X hours ago



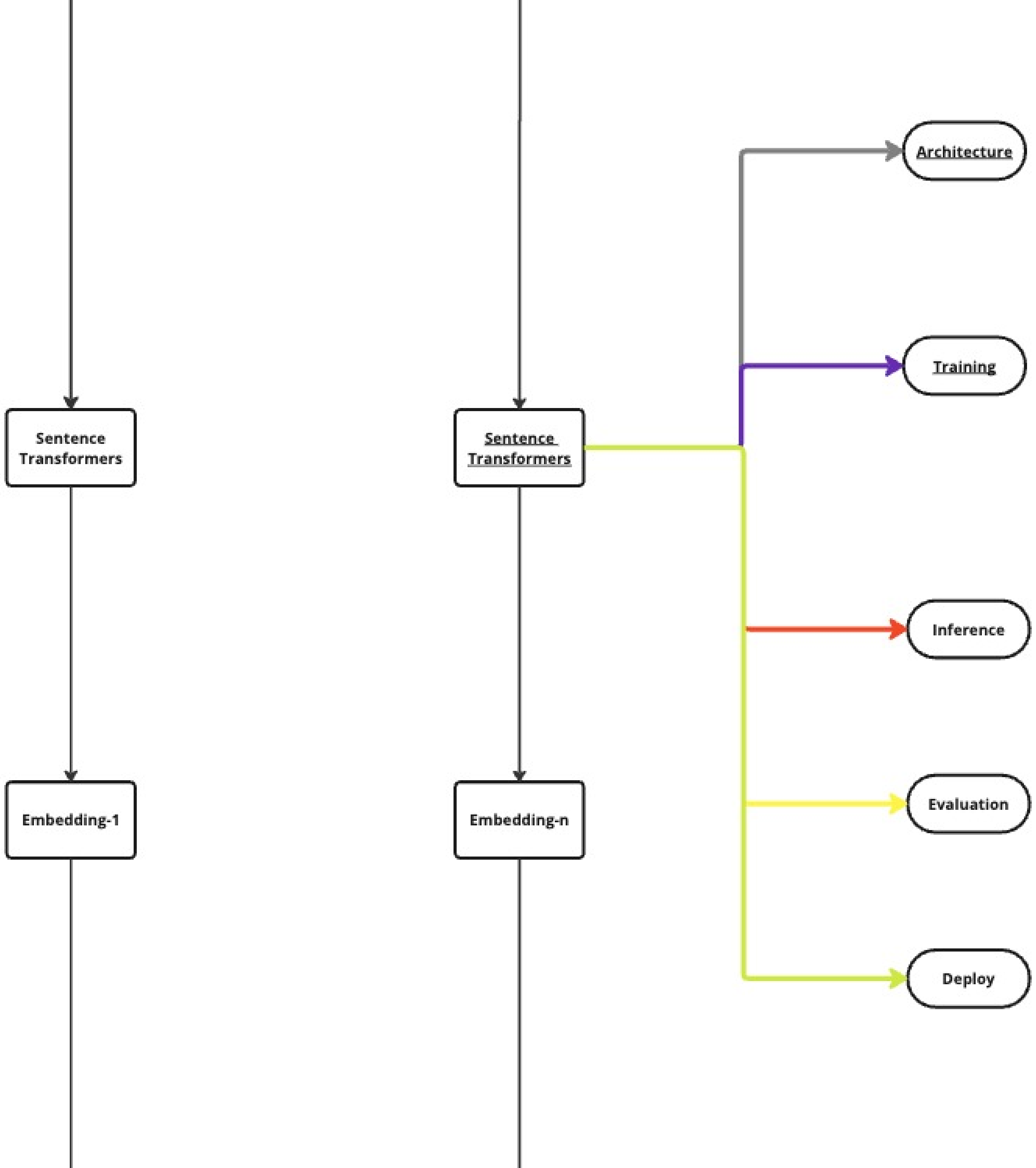


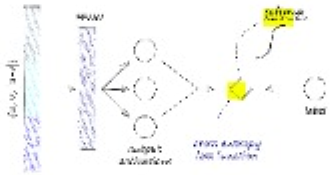
huggingface.co

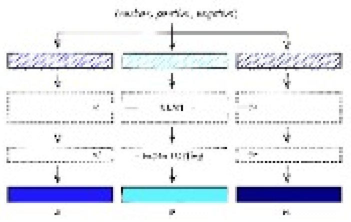
Summary of the tokenizers

We're on a journey to advance and democratize artificial intelligence through open source and open science.

Stage 2 : Embedding







www.sbert.net

Pretrained Models - Sentence-Transformers documentation

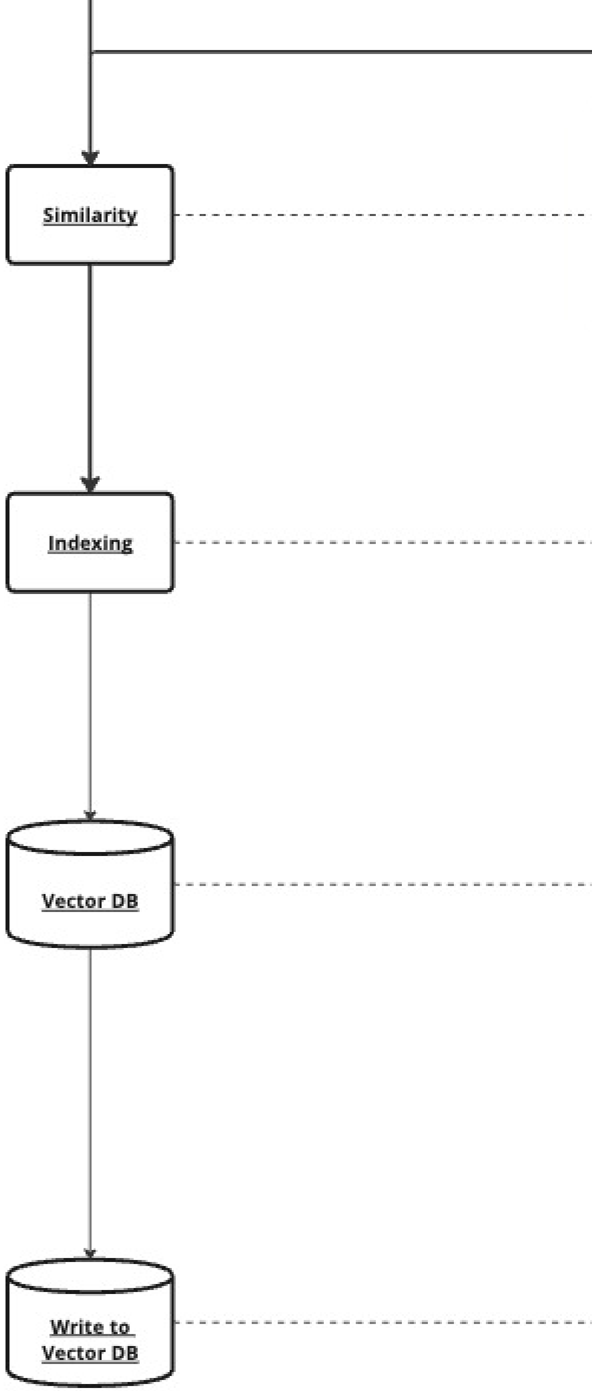
We provide various pre-trained models. Using these models is easy: All models are hosted on the HuggingFace Model Hub. The following table provides an overview of (selected) models. They have been extensively evaluated for their quality to embedded sent...

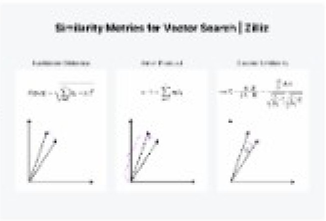
www.deepset.ai

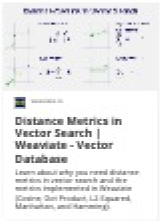
Evaluating RAG Part I: How to Evaluate Document Retrieval

A guide to the evaluation of components in retrieval augmented generation

Stage 3 : Add / Update Record in Vector DB




Similarity Metrics for Vector Search | Zilliz


weaviate.io

Distance Metrics in Vector Search | Weaviate - Vector Database

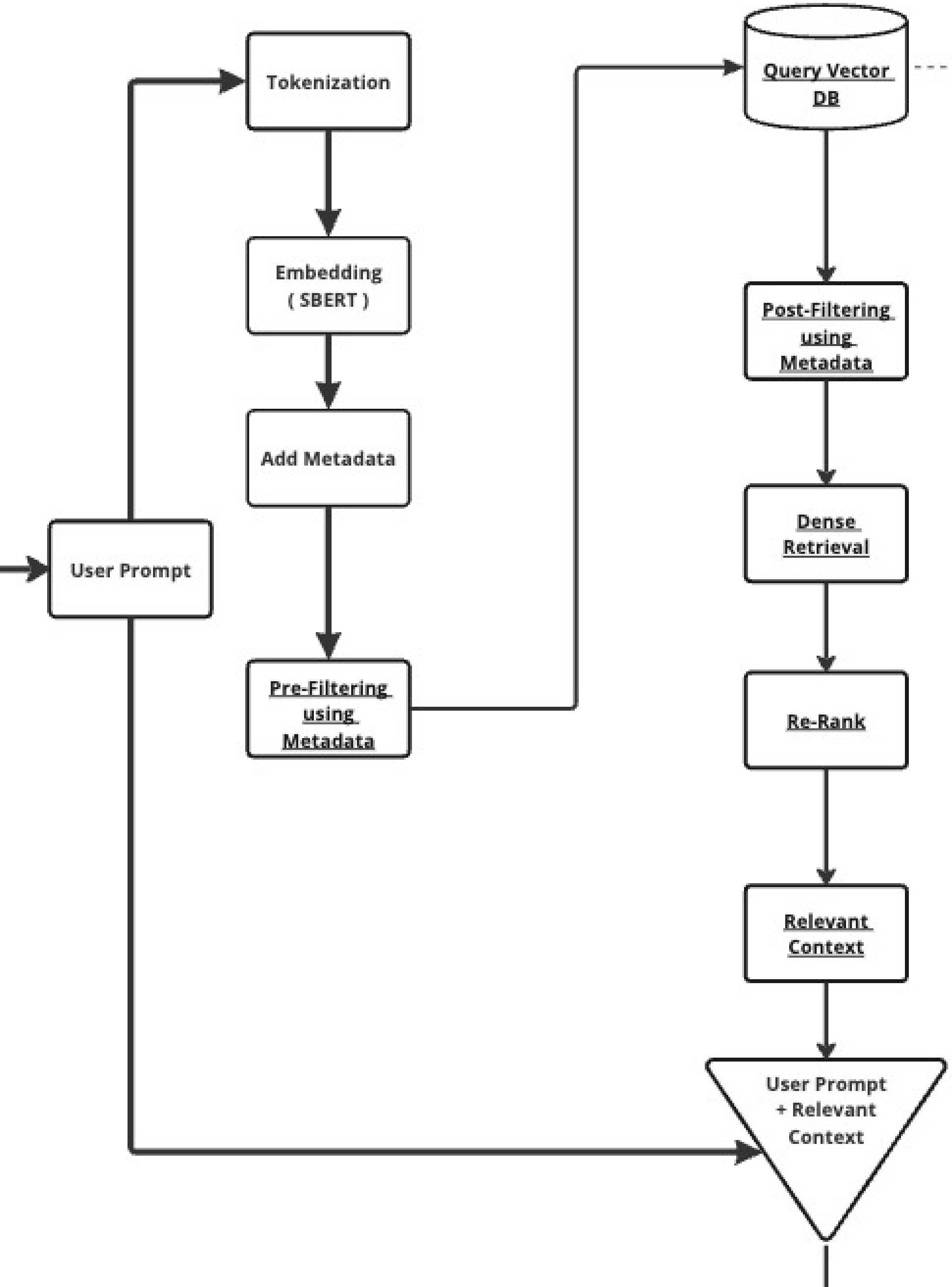
Learn about why you need distance metrics in vector search and the metrics implemented in Weaviate (Cosine, Dot Product, L2-Squared, Manhattan, and Hamming).

kdb.ai

Methods of Similarity - KDB.AI: The Smarter Database for AI.

The vector database that extends the knowledge of Generative AI applications with contextual search at scale.

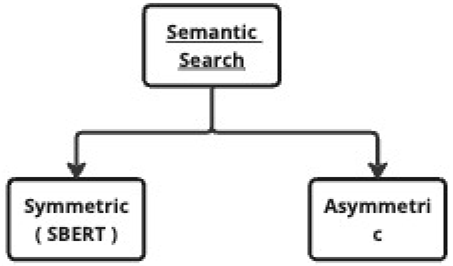
Stage 4 : Query Processing + Semantic Search



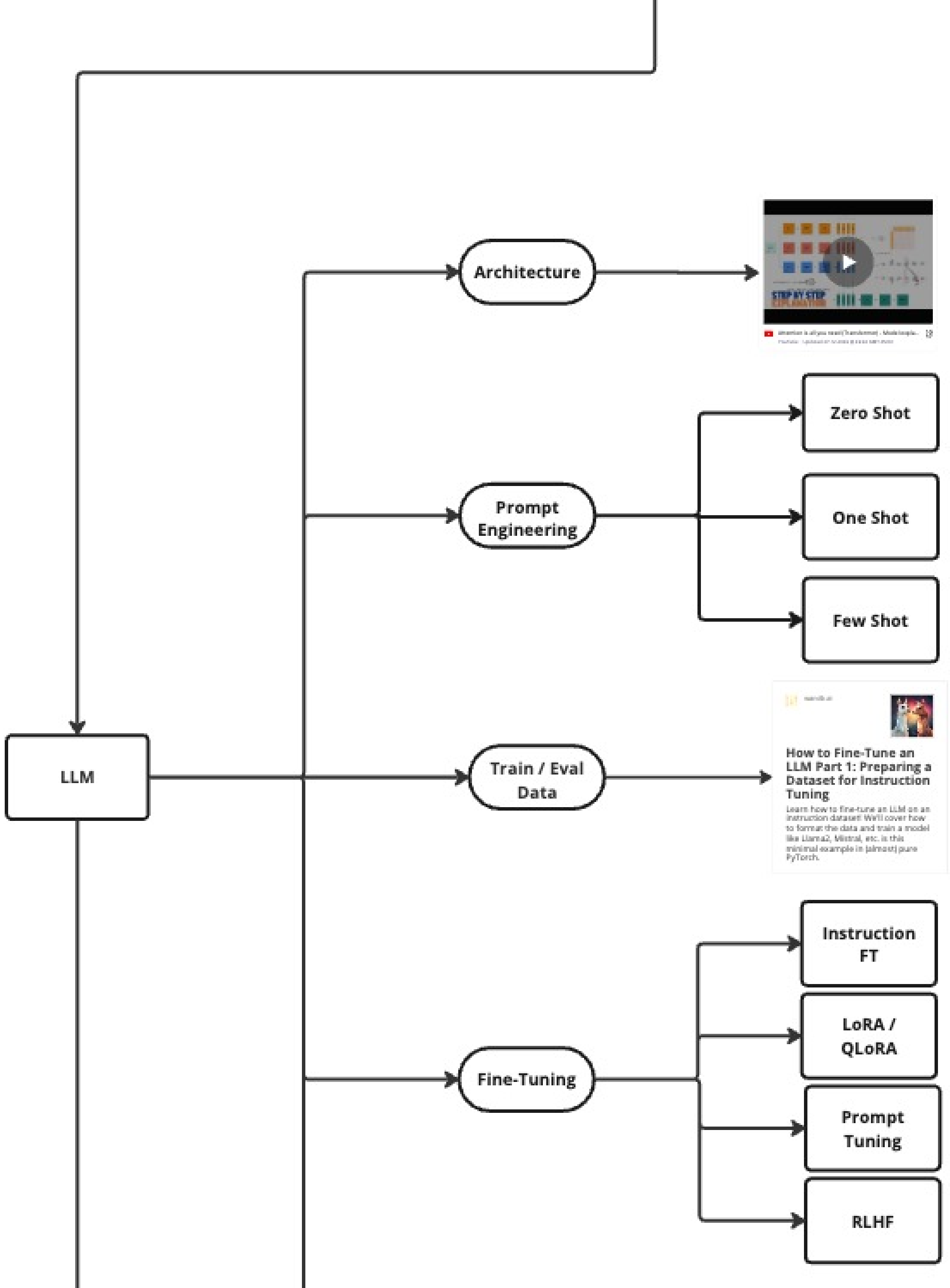
www.sbert.net

Semantic Search - Sentence-Transformers documentation

Semantic search seeks to improve search accuracy by understanding the content of the search query. In contrast to traditional search engines which only find documents based on lexical matches, semantic search can also find synonyms. The idea behind sema...

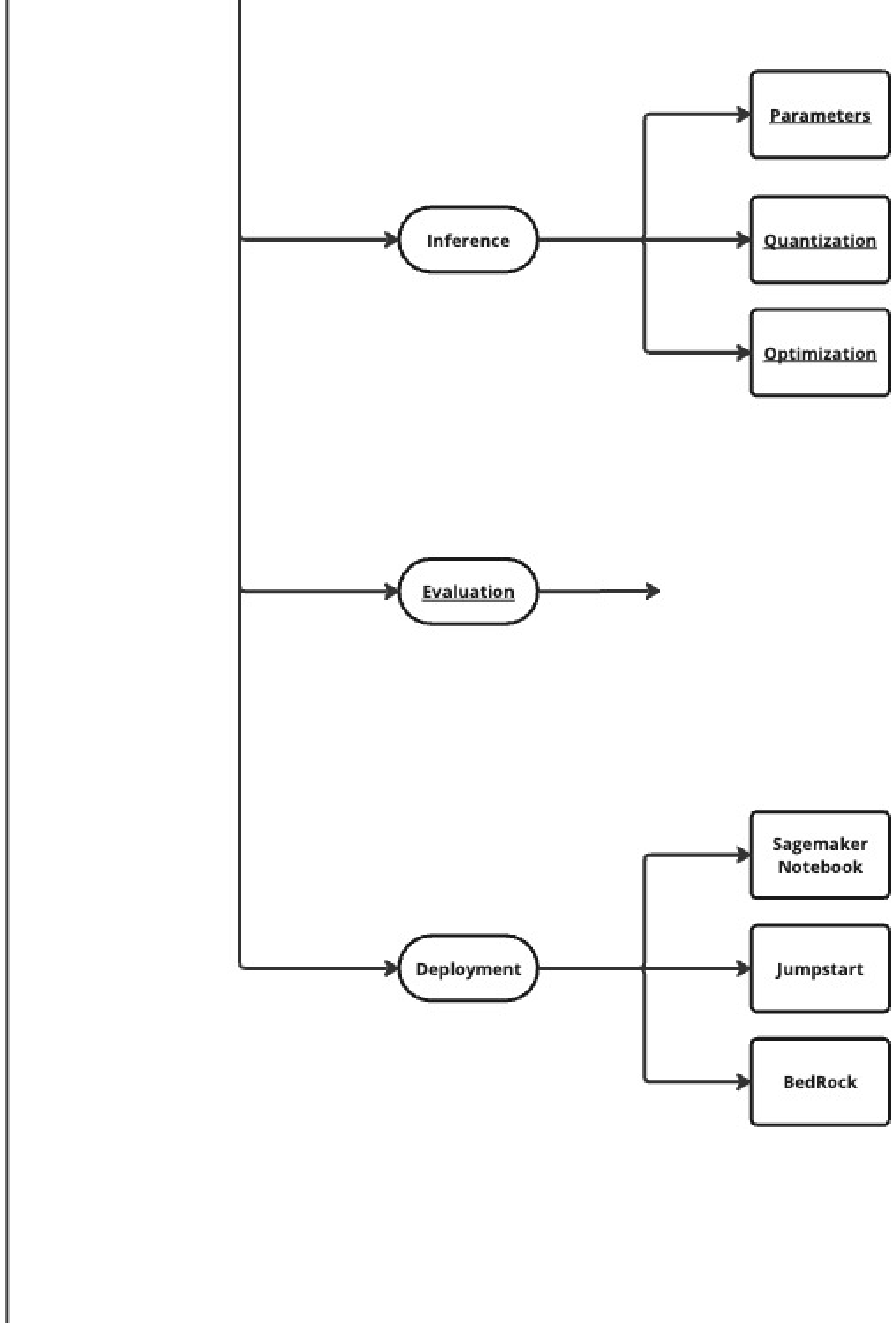


Stage 5: Generation LLM



How to Fine-Tune an LLM Part 1: Preparing a Dataset for Instruction Tuning

Learn how to fine-tune an LLM on an instruction dataset! We'll cover how to format the data and train a model like Llama2, Mistral, etc. is this minimal example in (almost) pure PyTorch.



LLM Chronicles #7: How To Evaluate LLMs?

Open LLM Leaderboard

pakodas.substack.com

LLM Chronicles #7: How To Evaluate LLMs? | Open LLM Leaderboard

Let's explore the internals of automated LLM evaluation

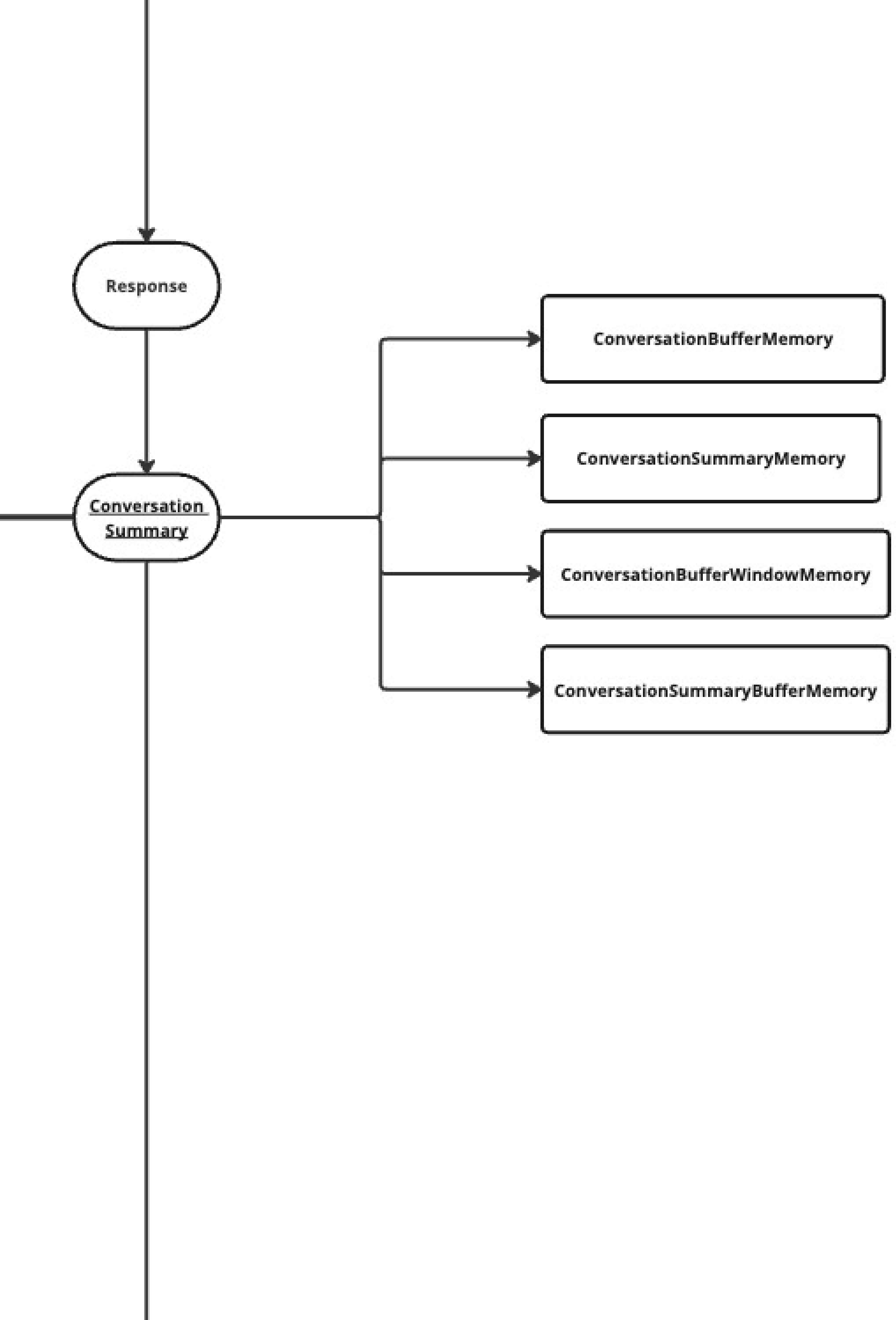
medium.com

#2. Deep Dive on Evaluation of Large Language Models (LLMs)

An effective Large Language Model (LLM) evaluation workflow would involve creating evaluating tasks / data that highly correlates to the...

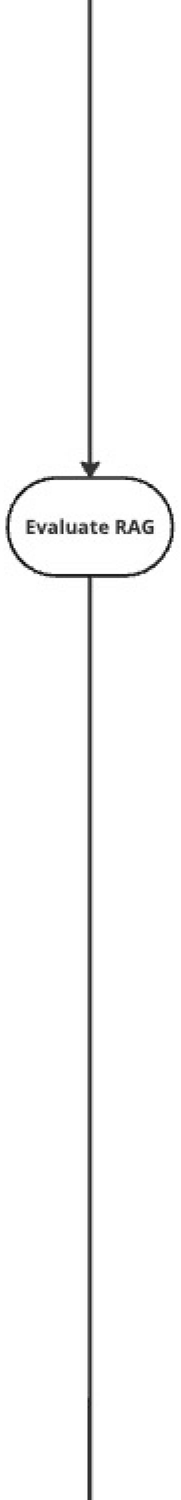
Stage 5: Generation LLM

Stage 6: Conversation Summary



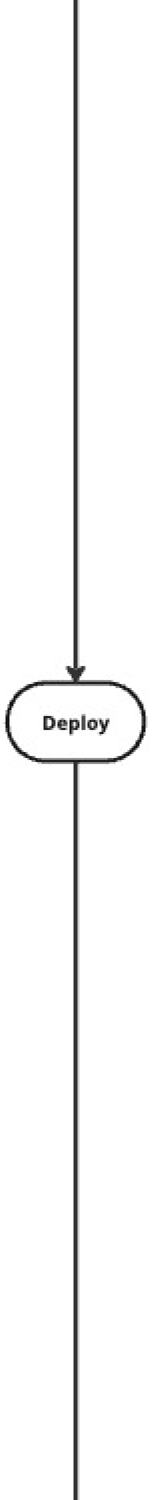
Response
Conversation Summary
ConversationBufferMemory
ConversationSummaryMemory
ConversationBufferWindowMemory
ConversationSummaryBufferMemory

Stage 7: Evaluation



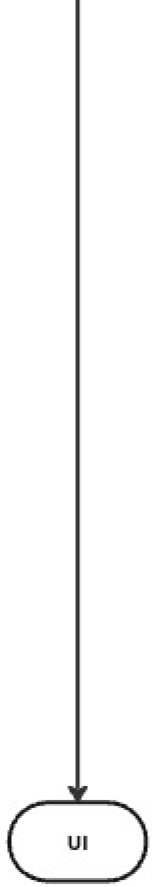
Evaluate RAG

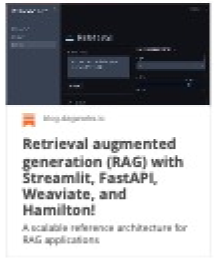
Stage 8 : Deployment



Deploy

Stage 10: Front-End





blog.dagworks.io

Retrieval augmented generation (RAG) with Streamlit, FastAPI, Weaviate, and Hamilton!

A scalable reference architecture for RAG applications