Information Retrieval and Synthesis Workflow with Gen AI

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution

This file is meant for personal use by m.sunita20@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Agenda

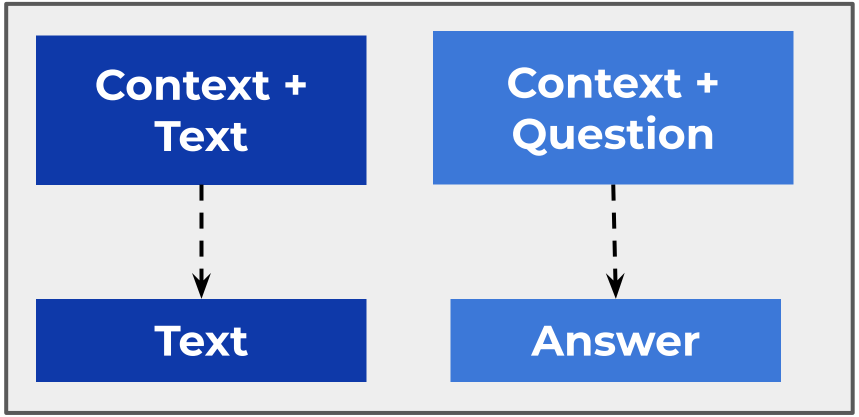
In this session, we will discuss :

- Overview of Retrieval Augmented Generation (RAG) and its Working
- Building Blocks of RAG
- Data Preparation Process with respect to RAG
- Devising and Evaluating Prompts with respect to RAG

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution

This file is meant for personal use by m.sunita20@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Retrieval Augmented Generation (RAG)



Context + Text

Text

Context + Question

Answer

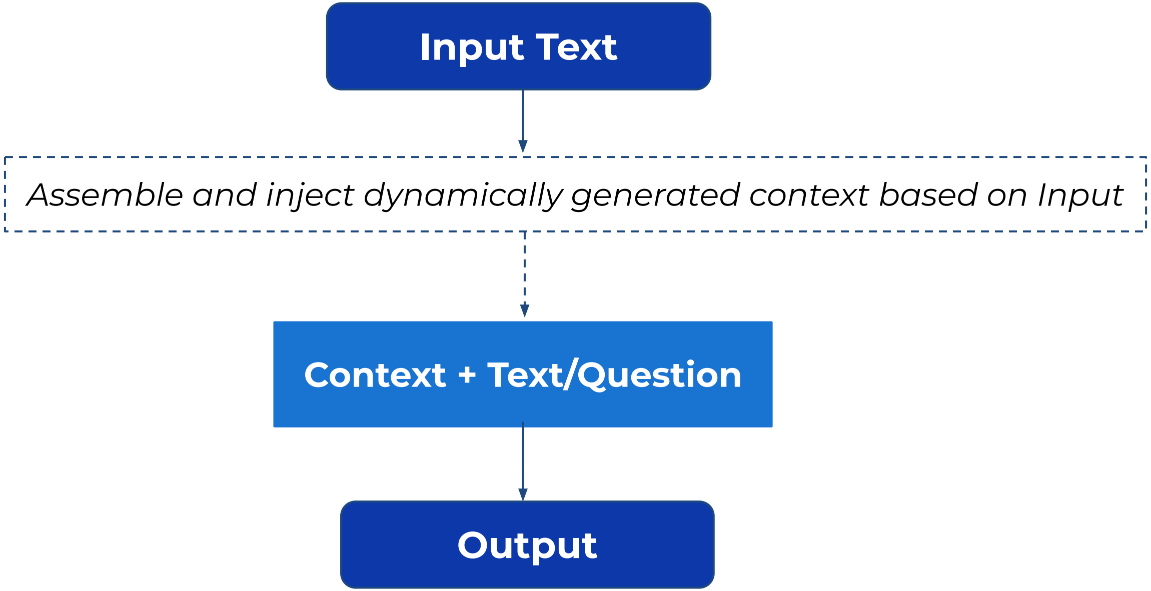
Contextualized

RAG

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution

This file is meant for personal use by m.sunita20@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Working of RAG



Input Text

Assemble and inject dynamically generated context based on Input

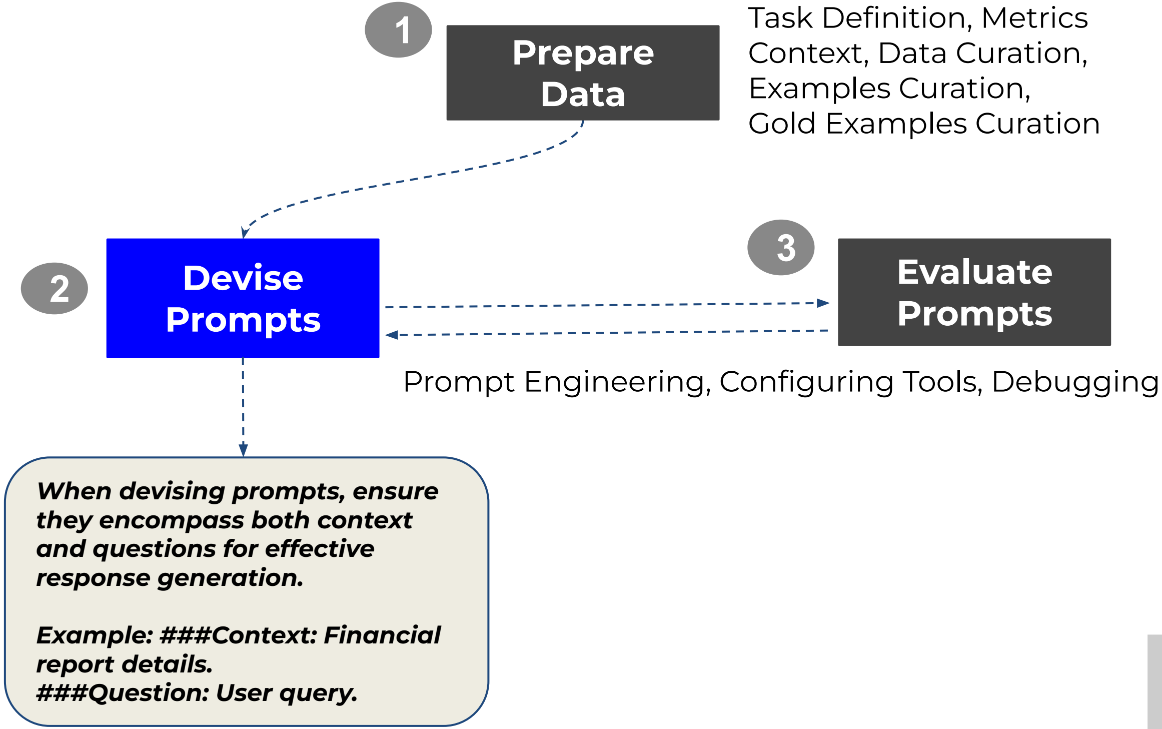
Context + Text/Question

Output

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution

This file is meant for personal use by m.sunita20@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Structure of RAG



1 Prepare Data

Task Definition, Metrics Context, Data Curation, Examples Curation, Gold Examples Curation

2 Devise Prompts

3 Evaluate Prompts

Prompt Engineering, Configuring Tools, Debugging

When devising prompts, ensure they encompass both context and questions for effective response generation.

Example: ###Context: Financial report details.
###Question: User query.

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution

This file is meant for personal use by m.sunita20@gmail.com only.

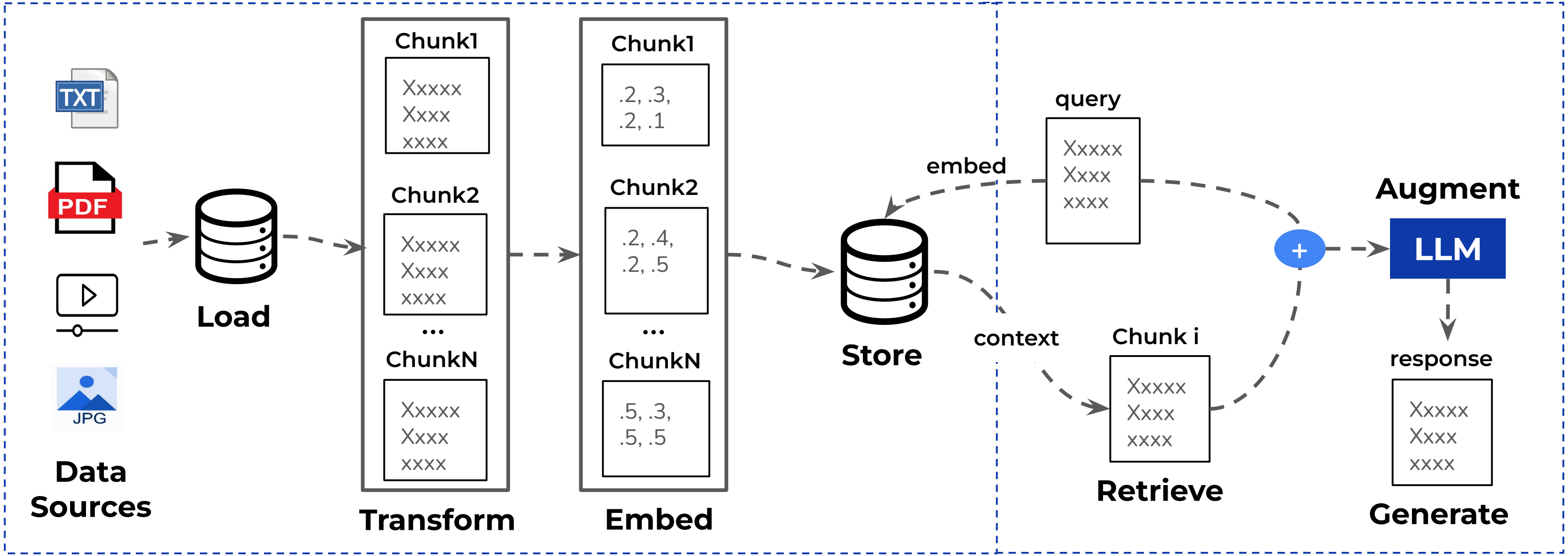
Sharing or publishing the contents in part or full is liable for legal action.

Building Blocks of RAG

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution

This file is meant for personal use by m.sunita20@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Building Blocks of RAG



Step 1: Prepare Data

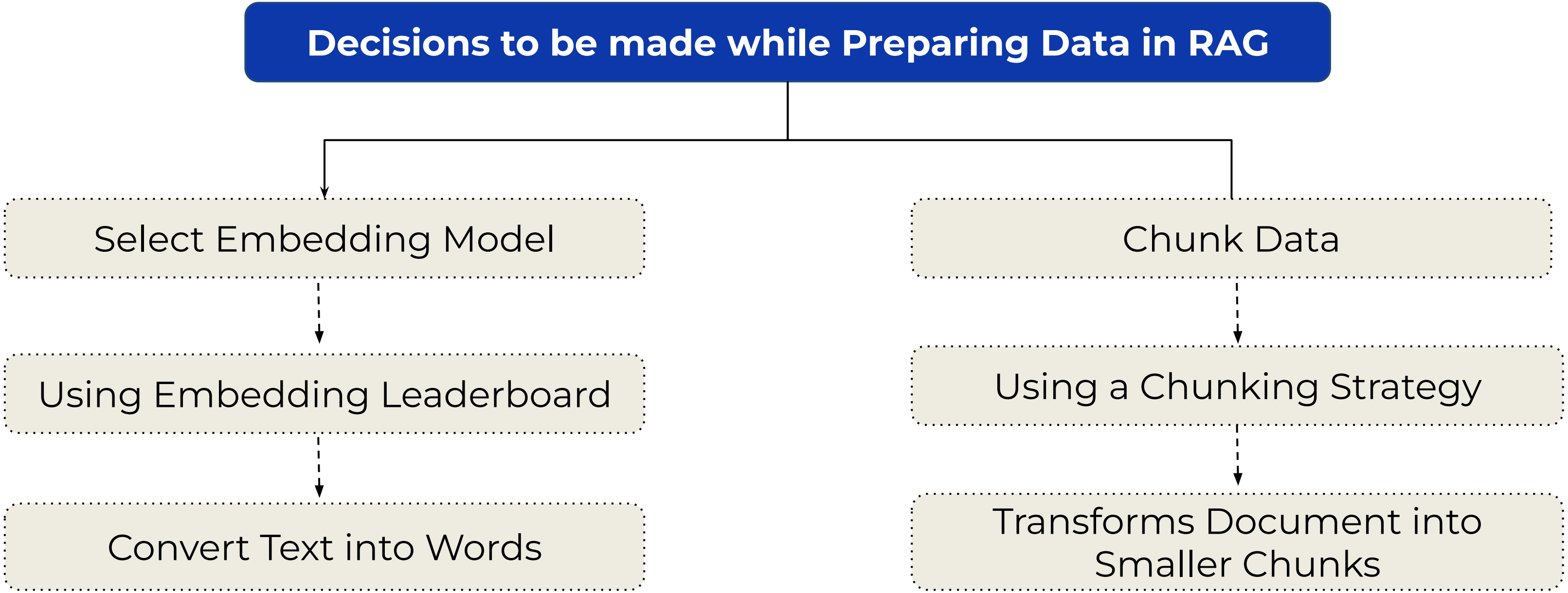
Step 2: Devise & Evaluate Prompts

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution

This file is meant for personal use by m.sunita20@gmail.com only.

Sharing or publishing the contents in part or full is liable for legal action.

Prepare Data in RAG



Decisions to be made while Preparing Data in RAG

Select Embedding Model

Using Embedding Leaderboard

Convert Text into Words

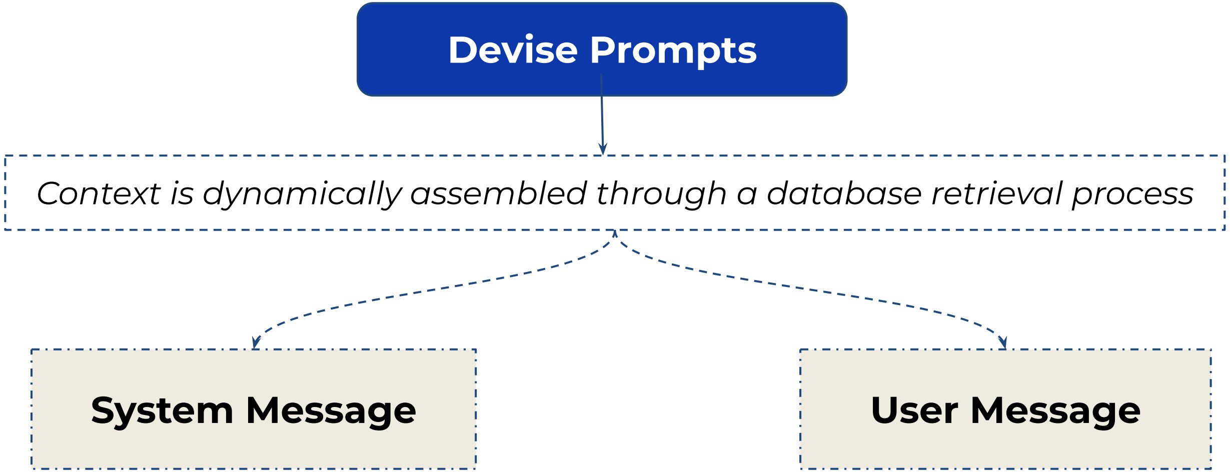
Chunk Data

Using a Chunking Strategy

Transforms Document into Smaller Chunks

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution
This file is meant for personal use by m.sunita20@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Devise Prompts in RAG



Devise Prompts

Context is dynamically assembled through a database retrieval process

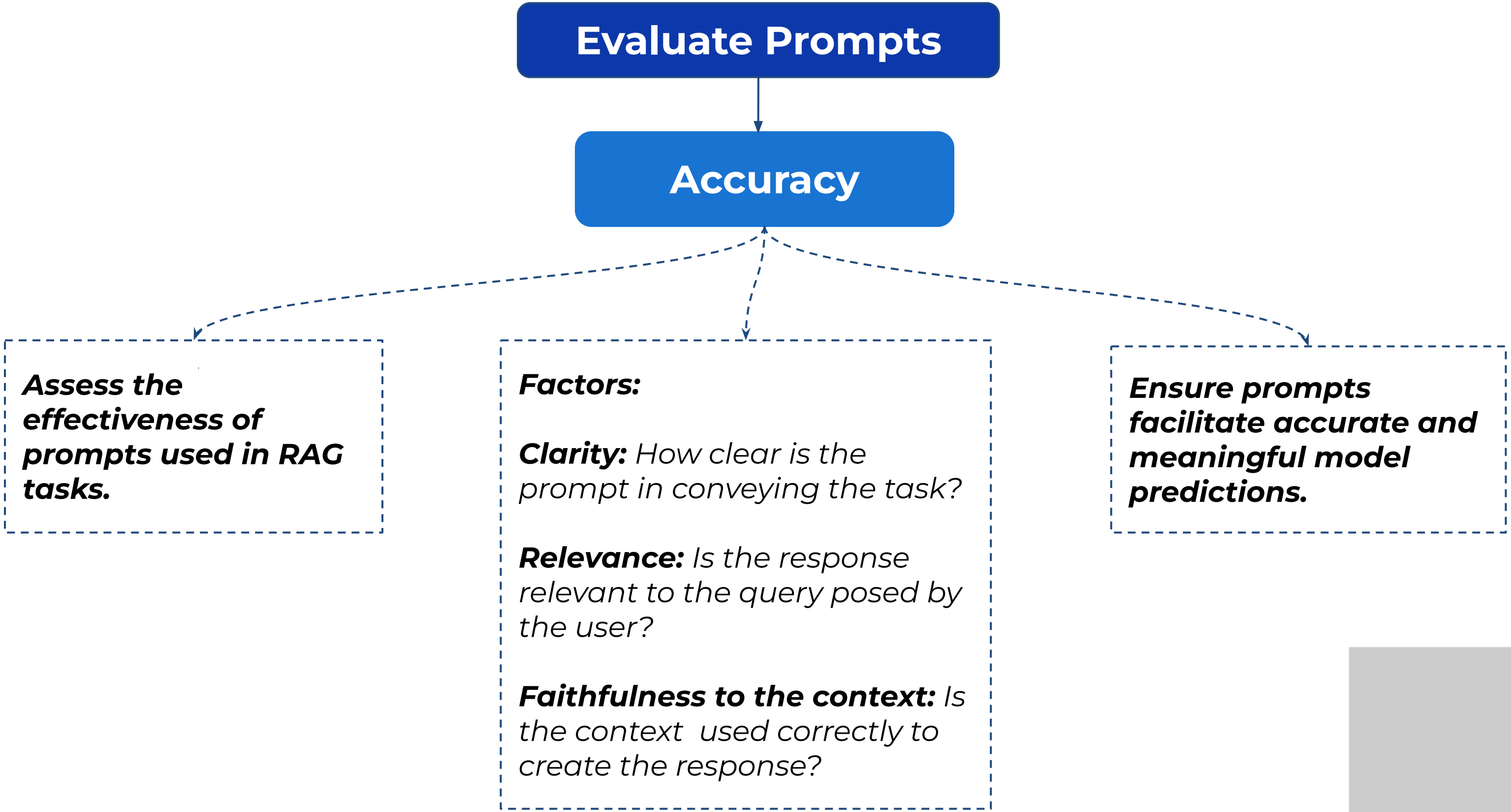
System Message

User Message

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution

This file is meant for personal use by m.sunita20@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Evaluate Prompts in RAG



Evaluate Prompts

Accuracy

Assess the effectiveness of prompts used in RAG tasks.

Factors:

Clarity: *How clear is the prompt in conveying the task?*

Relevance: *Is the response relevant to the query posed by the user?*

Faithfulness to the context: *Is the context used correctly to create the response?*

Ensure prompts facilitate accurate and meaningful model predictions.

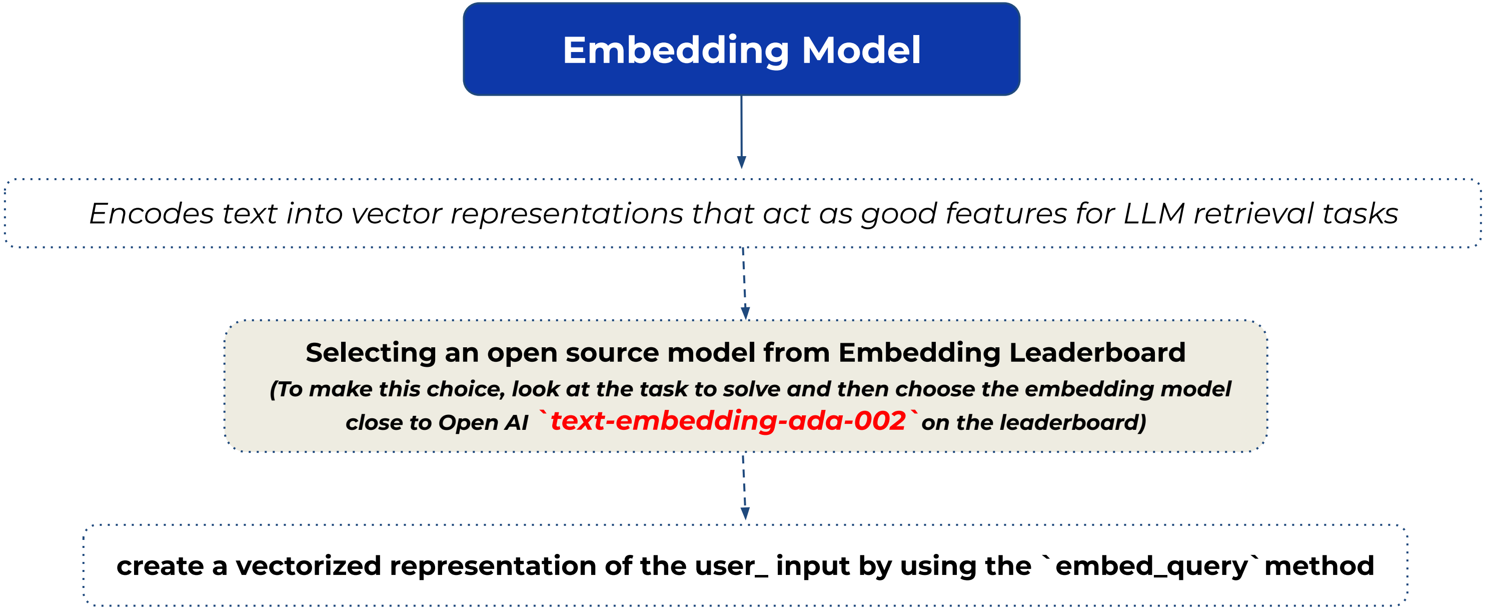
This file is meant for personal use by m.sunita20@gmail.com only.
Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution
Sharing or publishing the contents in part or full is liable for legal action.

Data Preparation Process

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution

This file is meant for personal use by m.sunita20@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Select Embedding Model



Embedding Model

Encodes text into vector representations that act as good features for LLM retrieval tasks

Selecting an open source model from Embedding Leaderboard

(To make this choice, look at the task to solve and then choose the embedding model close to Open AI `text-embedding-ada-002` on the leaderboard)

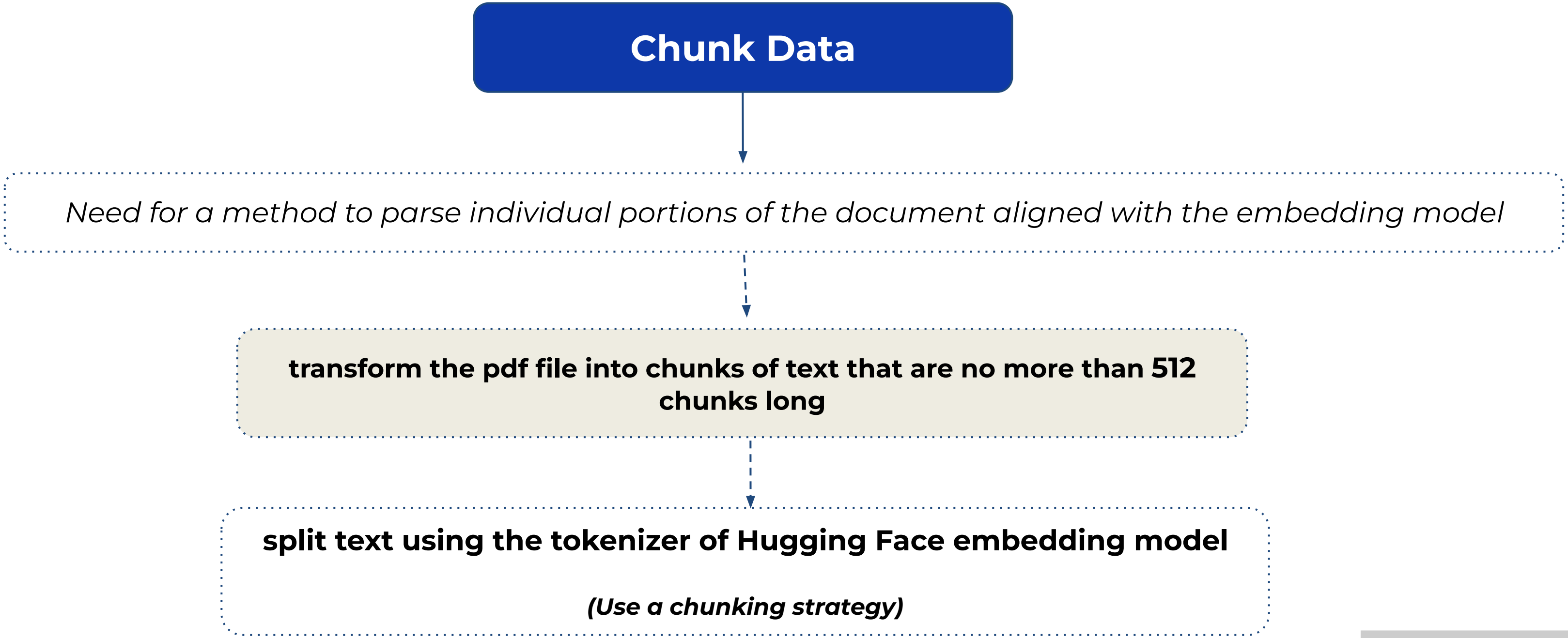
create a vectorized representation of the user_ input by using the `embed_query` method

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution

This file is meant for personal use by m.sunita20@gmail.com only.

Sharing or publishing the contents in part or full is liable for legal action.

Chunk Data



Chunk Data

Need for a method to parse individual portions of the document aligned with the embedding model

transform the pdf file into chunks of text that are no more than 512 chunks long

split text using the tokenizer of Hugging Face embedding model

(Use a chunking strategy)

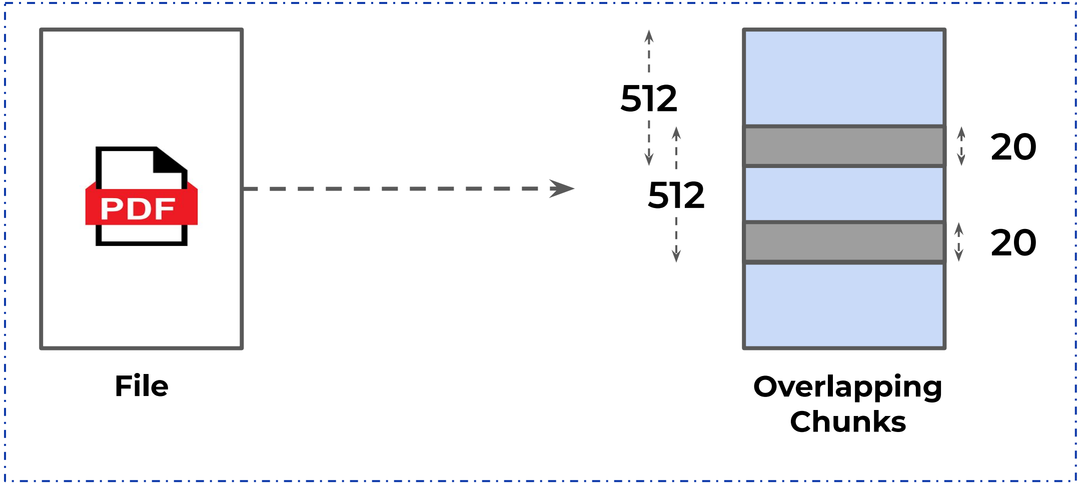
Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution

This file is meant for personal use by m.sunita20@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Chunking Strategy: Example

chunk_overlap = 20

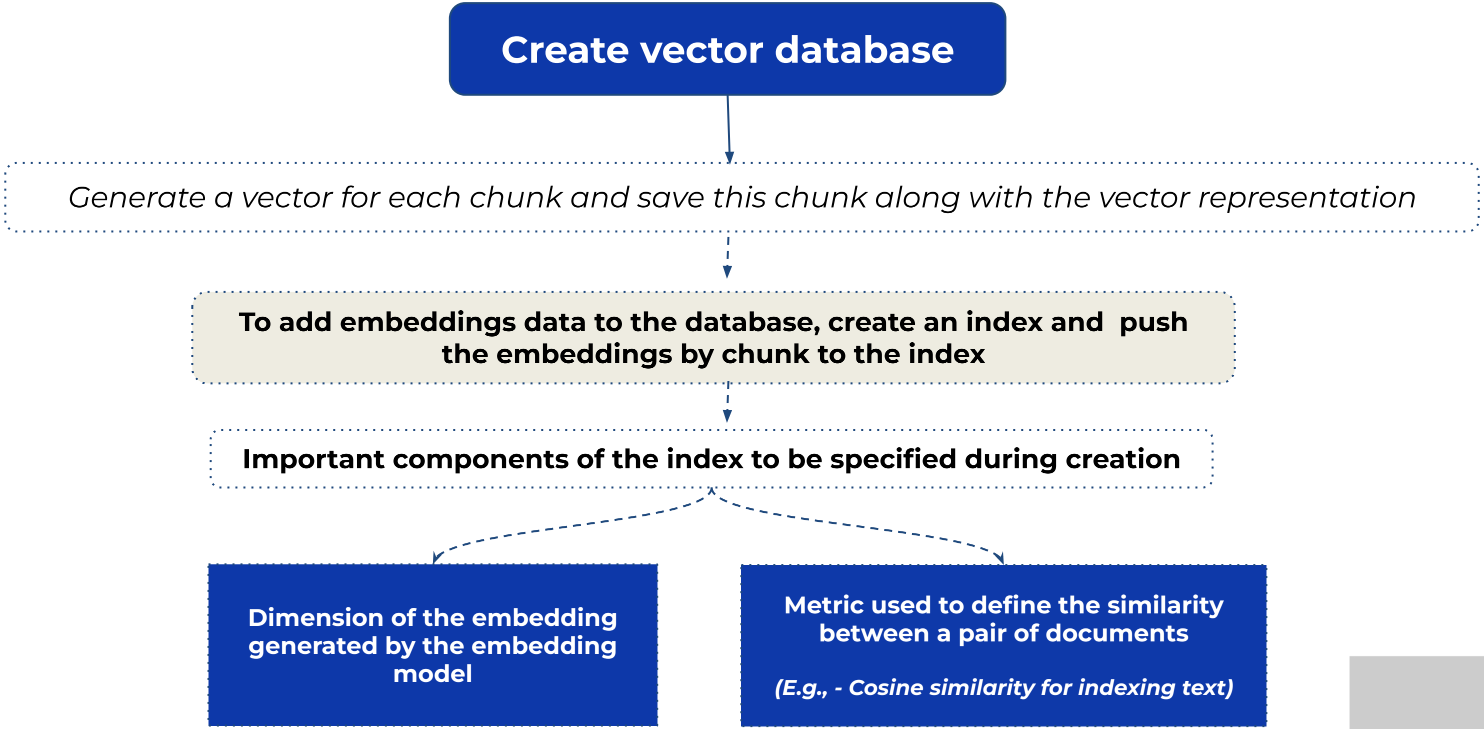
ensures that the chunks are related to each other (i.e., there is some continuity between the chunks)



Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution

This file is meant for personal use by m.sunita20@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Create Vector Database



Create vector database

Generate a vector for each chunk and save this chunk along with the vector representation

To add embeddings data to the database, create an index and push the embeddings by chunk to the index

Important components of the index to be specified during creation

Dimension of the embedding generated by the embedding model

Metric used to define the similarity between a pair of documents

(E.g., - Cosine similarity for indexing text)

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution

This file is meant for personal use by m.sunita20@gmail.com only.

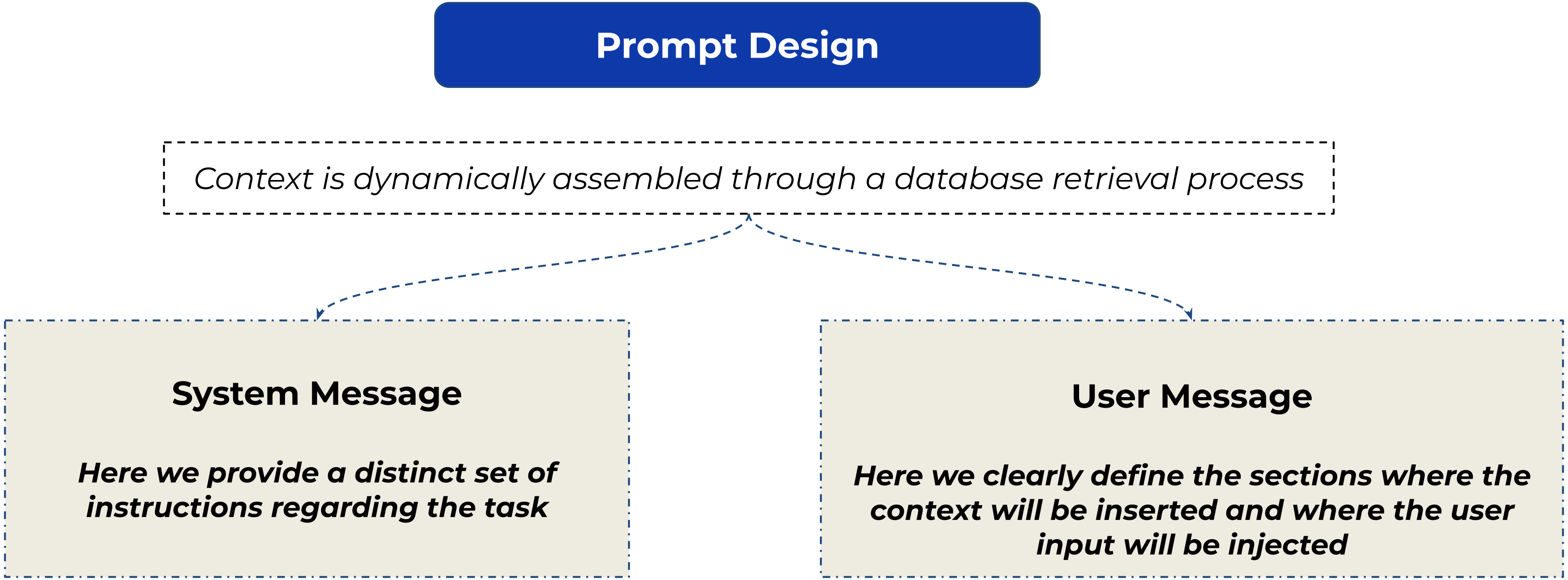
Sharing or publishing the contents in part or full is liable for legal action.

Devising and Evaluating Prompts

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution

This file is meant for personal use by m.sunita20@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Devising Prompts



Prompt Design

Context is dynamically assembled through a database retrieval process

System Message

Here we provide a distinct set of instructions regarding the task

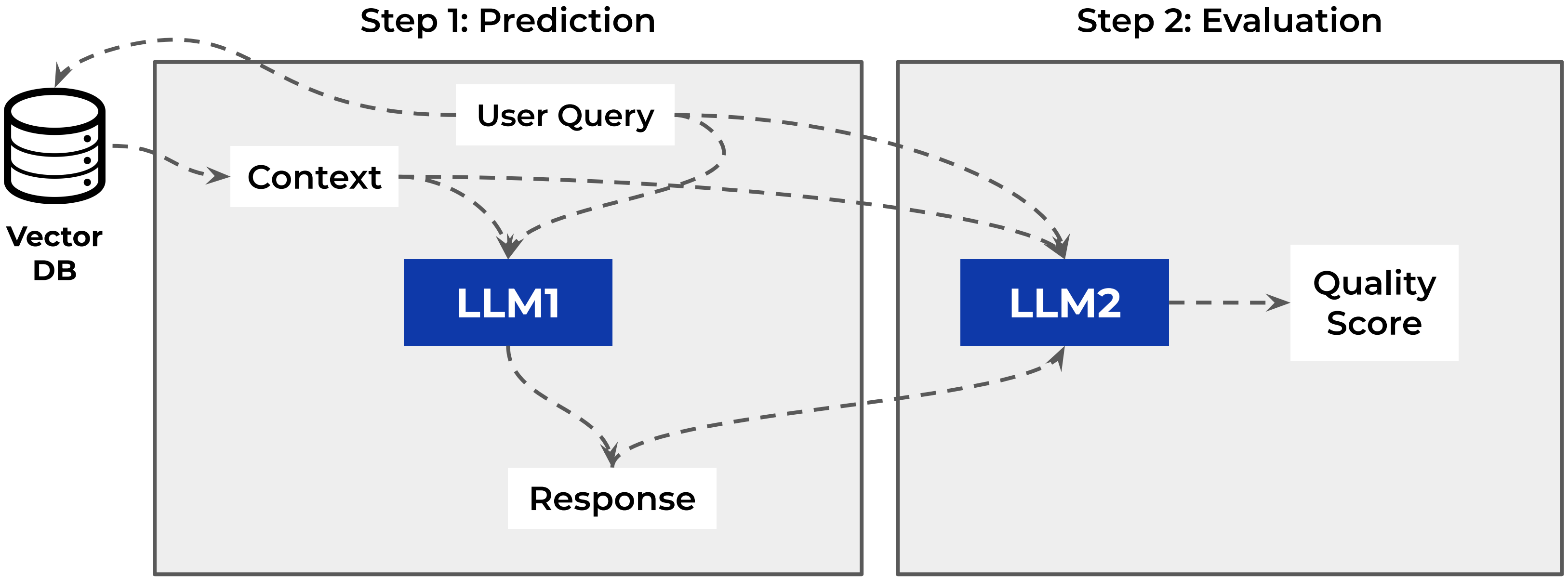
User Message

Here we clearly define the sections where the context will be inserted and where the user input will be injected

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution

This file is meant for personal use by m.sunita20@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

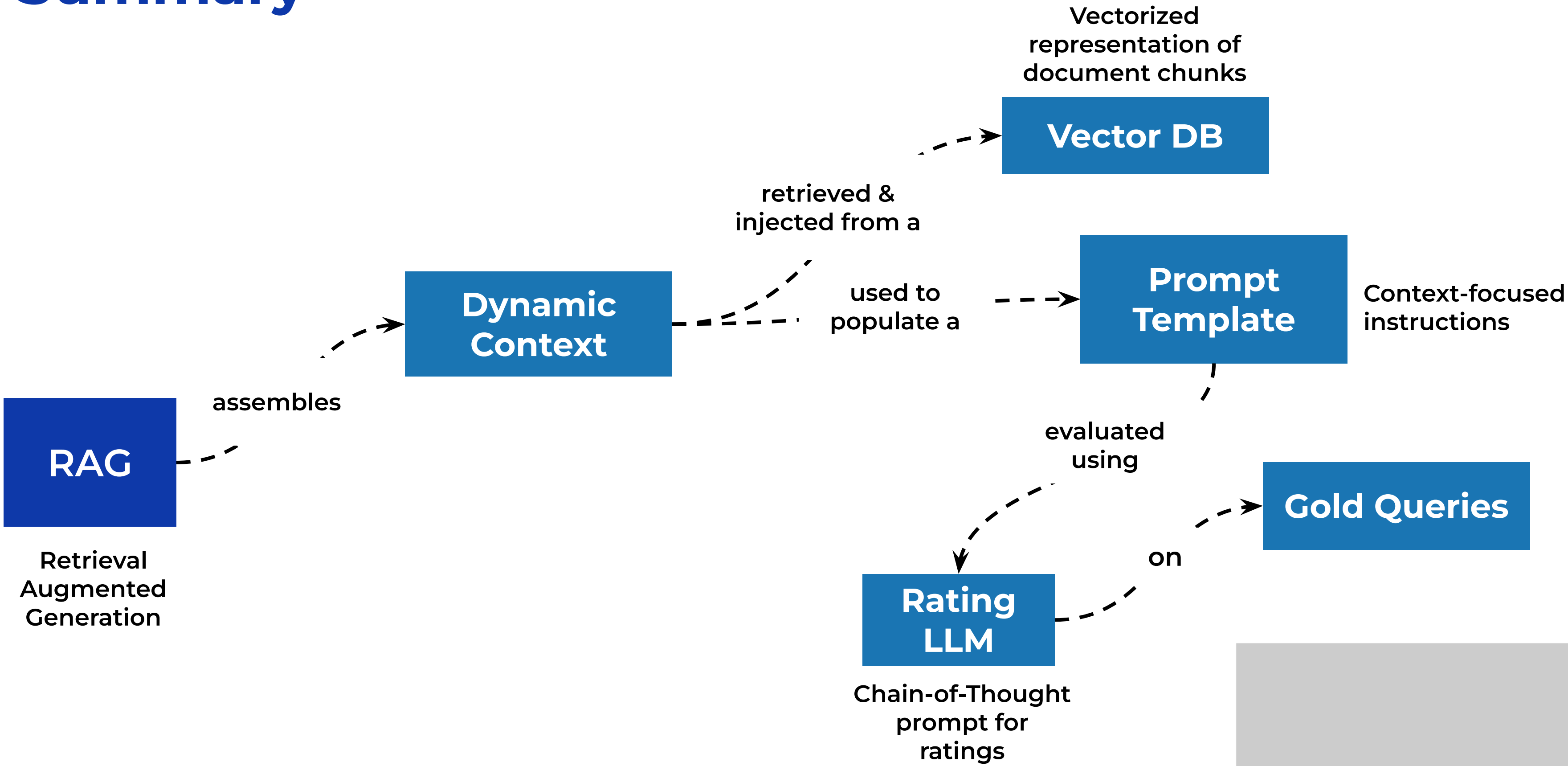
Evaluation Process in RAG



Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution

This file is meant for personal use by m.sunita20@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Summary



RAG

Retrieval Augmented Generation

assembles

Dynamic Context

retrieved & injected from a

Vector DB

Vectorized representation of document chunks

used to populate a

Prompt Template

Context-focused instructions

evaluated using

Rating LLM

Chain-of-Thought prompt for ratings

on

Gold Queries

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution

This file is meant for personal use by m.sunita20@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.