Retrieval Augmented Generation (RAG)

Agenda

- Retrieval Augmented Augmentation (RAG) Quiz
- Building Blocks of RAG
- Chunking
- Semantic Search
- Evaluation

Let’s begin the discussion by answering a few questions.

RAG Quiz

Which of the following represents the correct sequence of blocks for processing data in a retrieval-augmented generation (RAG) model?

A Embedding -> Vector DB -> Data Chunk -> Retriever -> LLM

B Data Chunk -> Embedding -> Vector DB -> Retriever -> LLM

C Vector DB -> Data Chunk -> Retriever -> Embedding -> LLM

D Data Chunk -> Embedding -> Retriever -> Vector DB

RAG Quiz

Which of the following represents the correct sequence of blocks for processing data in a retrieval-augmented generation (RAG) model?

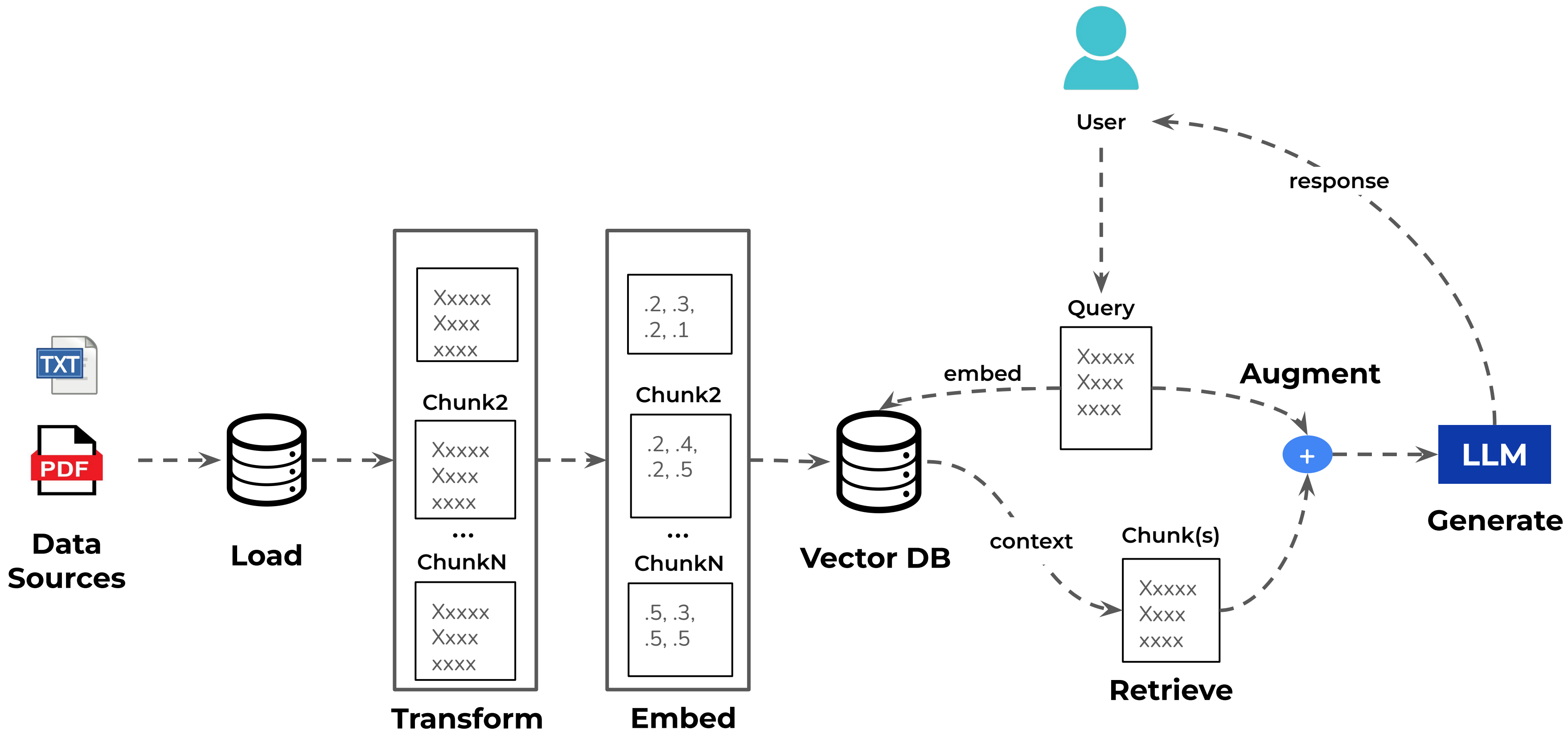
A Embedding -> Vector DB -> Data Chunk -> Retriever -> LLM

[✓ B Data Chunk -> Embedding -> Vector DB -> Retriever -> LLM]

C Vector DB -> Data Chunk -> Retriever -> Embedding -> LLM

D Data Chunk -> Embedding -> Retriever -> Vector DB

Retrieval Augmented Generation (RAG)



RAG Quiz

How does increasing the chunk overlap in a RAG system impact retrieval?

A It reduces the number of retrieved chunks, improving efficiency

B It increases redundancy, but helps maintain continuity in retrieved content

C It eliminates the need for chunking altogether

D It prevents the model from retrieving any irrelevant information

RAG Quiz

How does increasing the chunk overlap in a RAG system impact retrieval?

A It reduces the number of retrieved chunks, improving efficiency

B [✓ It increases redundancy, but helps maintain continuity in retrieved content]

C It eliminates the need for chunking altogether

D It prevents the model from retrieving any irrelevant information

Chunking

The process of **breaking a document** (or text corpus) **into smaller, manageable pieces** (chunks) before storing them

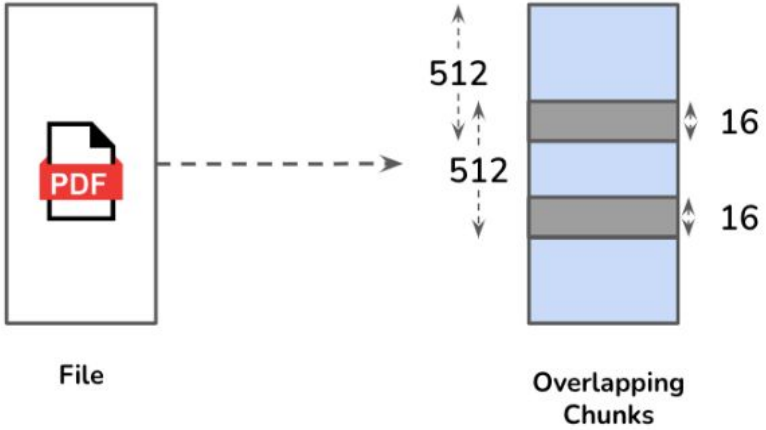
Chunks are then **indexed and used to retrieve relevant information efficiently** when generating responses

512 => number of tokens in a chunk

16 => number of tokens overlapping between any two consecutive chunks

Increasing the chunk overlap parameter increases redundancy as there will be larger overlaps

But this helps maintain continuity in retrieved content as information might be spread across multiple chunks



RAG Quiz

Why do we need to create embeddings for our data in a Retrieval-Augmented Generation (RAG) system?

- A To improve keyword-based searching
- B To enable semantic similarity search
- C To reduce the amount of stored data
- D To make data more structured

RAG Quiz

Why do we need to create embeddings for our data in a Retrieval-Augmented Generation (RAG) system?

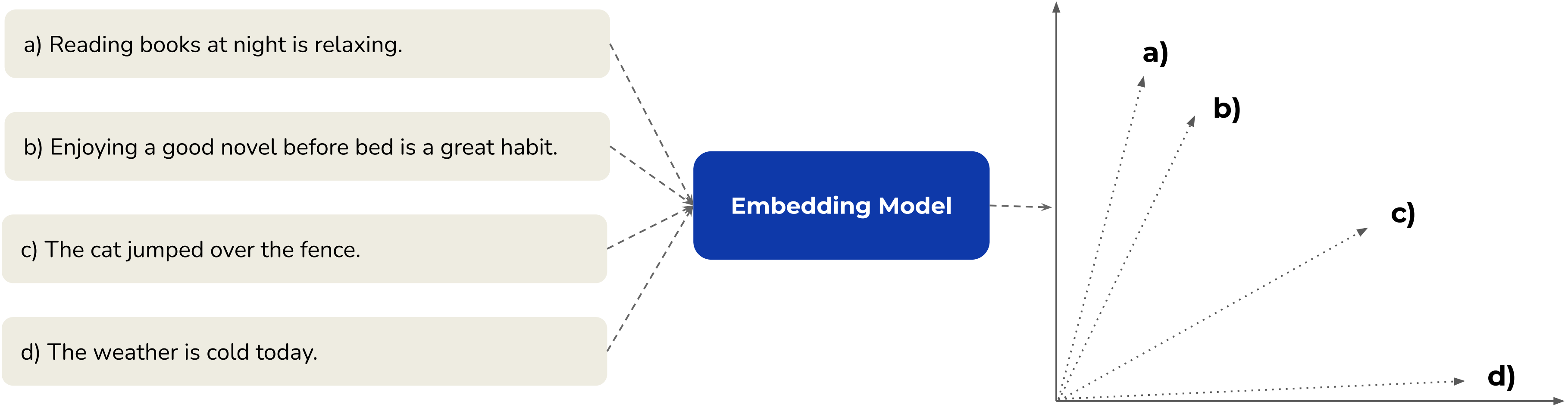
A To improve keyword-based searching

B [✓ To enable semantic similarity search]

C To reduce the amount of stored data

D To make data more structured

Sentence Embeddings



a) Reading books at night is relaxing.

b) Enjoying a good novel before bed is a great habit.

c) The cat jumped over the fence.

d) The weather is cold today.

Similar sentences are closer in the embedding space; **dissimilar sentences are further apart**

This enables **semantic search to retrieve sentences** (or chunks) from the data source that are **“similar” to the user query**

RAG Quiz

How does RAG enhance the capabilities of LLMs in natural language processing tasks ?

A By limiting the access to external knowledge sources for generating responses

B By reducing the need for fine-tuning LLMs on specific tasks

C By integrating external knowledge sources to provide contextually rich and accurate responses

D By reducing the model's size

RAG Quiz

How does RAG enhance the capabilities of LLMs in natural language processing tasks ?

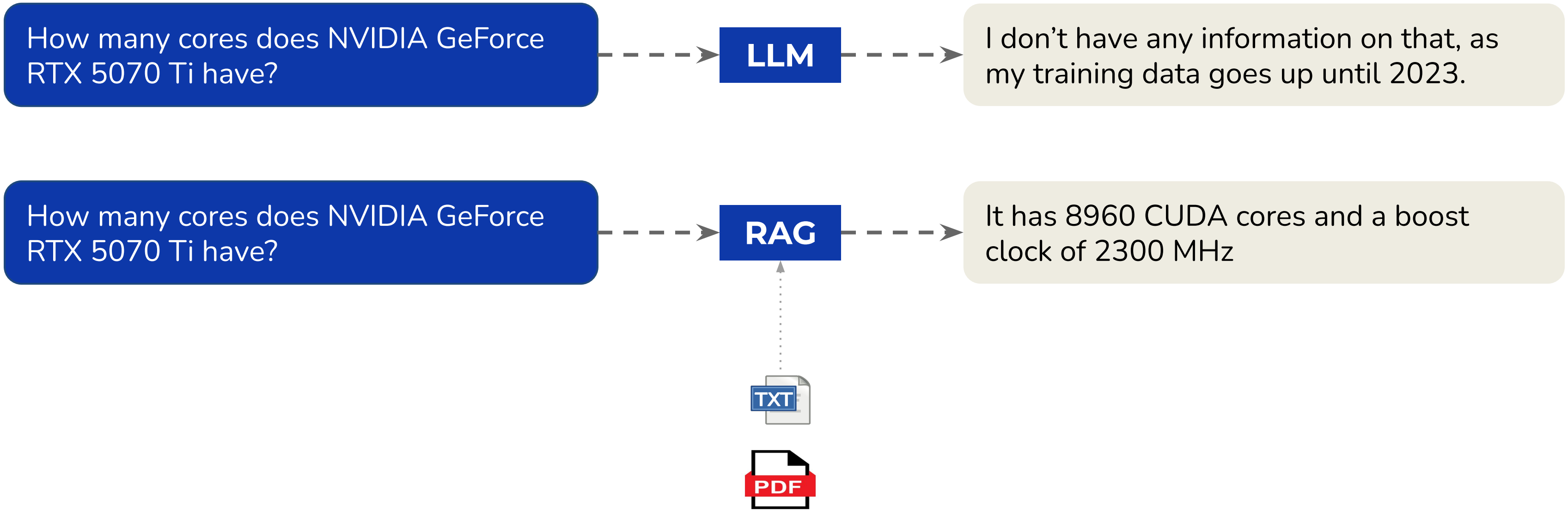
A By limiting the access to external knowledge sources for generating responses

B By reducing the need for fine-tuning LLMs on specific tasks

C By integrating external knowledge sources to provide contextually rich and accurate responses

D By reducing the model's size

Advantage of RAG



How many cores does NVIDIA GeForce RTX 5070 Ti have?

LLM

I don’t have any information on that, as my training data goes up until 2023.

How many cores does NVIDIA GeForce RTX 5070 Ti have?

RAG

It has 8960 CUDA cores and a boost clock of 2300 MHz

NVIDIA GeForce RTX Graphics Cards Specification Document

RAG Quiz

In a Retrieval-Augmented Generation (RAG) system, which factor primarily ensures that the response correctly utilizes the provided information?

A Clarity in conveying the task

B Relevance of the response to the query

C Faithfulness to the context

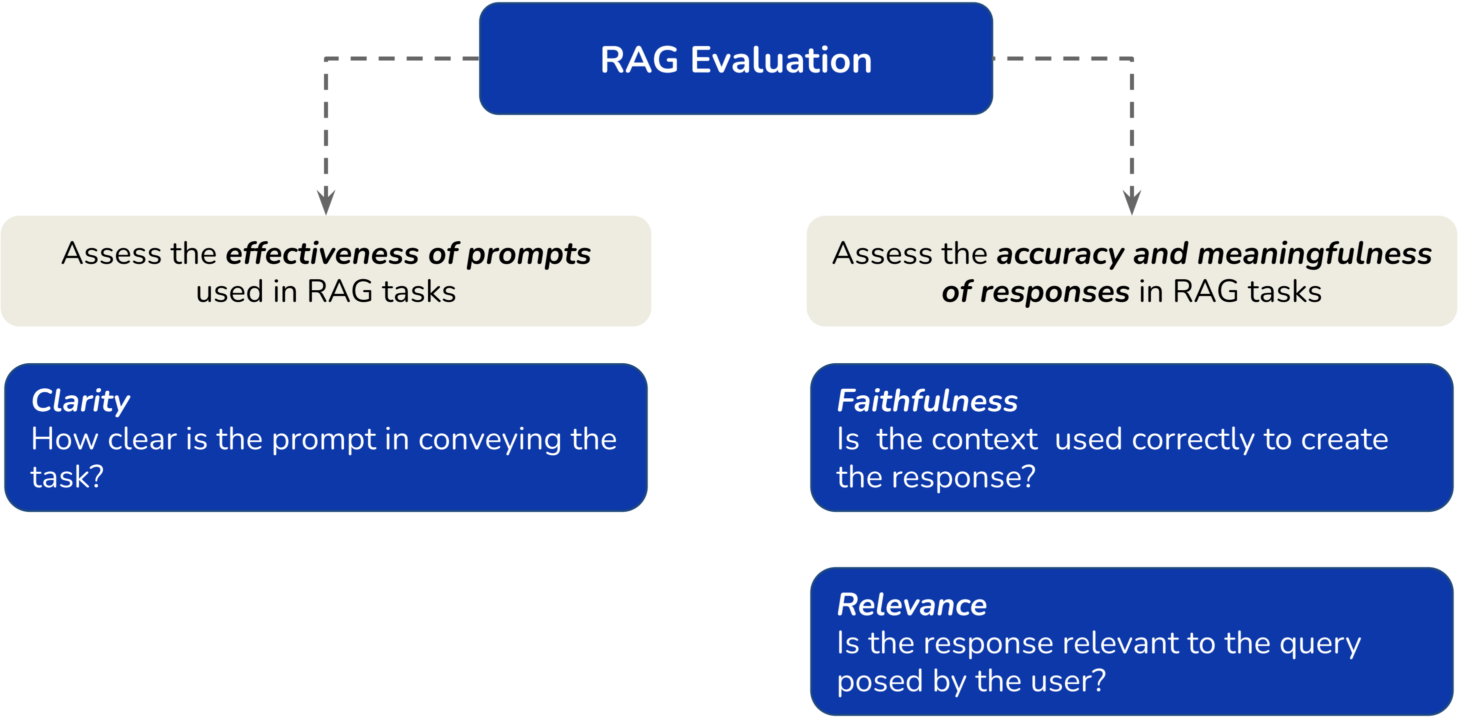
D Length of the response

RAG Quiz

In a Retrieval-Augmented Generation (RAG) system, which factor primarily ensures that the response correctly utilizes the provided information?

- A Clarity in conveying the task
- B Relevance of the response to the query
- [✓ C Faithfulness to the context]
- D Length of the response

Evaluation



RAG Evaluation

Assess the ***effectiveness of prompts*** used in RAG tasks

Clarity
How clear is the prompt in conveying the task?

Assess the ***accuracy and meaningfulness of responses*** in RAG tasks

Faithfulness
Is the context used correctly to create the response?

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

Great Learning

Power Ahead!