Analytics Vidhya

Mastering RAG

RAG Terms that you should know

- Vector Embeddings
- Vector Database
- Dense Passage Retrieval
- BM25 (Best Matching 25)
- Tokenization
- Chunking
- Context Window
- Prompt Engineering
- Knowledge Augmentation
- Hybrid Search
- Re-Ranking
- Hallucination
- Fine-Tuning
- Knowledge Cutoff
- Multi-Modal RAG
- Contrastive Learning

Analytics Vidhya

RAG

A technique that combines information retrieval with text generation to improve the accuracy and relevance of responses by fetching external knowledge during inference.

Vector Embeddings

Numerical representations of text (or other data) in a high-dimensional space, allowing for semantic similarity searches in retrieval models.

Vector Database (Vector Store)

A specialized database optimized for storing and retrieving vector embeddings efficiently. Common examples:

- FAISS (Facebook AI Similarity Search)
- Chroma
- Weaviate
- Pinecone
- Milvus

Dense Passage Retrieval (DPR)

A neural-based retrieval method that encodes both queries and documents into embeddings for efficient similarity matching.

BM25 (Best Matching 25)

A traditional information retrieval algorithm based on term frequency and inverse document frequency (TF-IDF), used as a baseline for retrieval.

Tokenization

The process of breaking text into smaller units (tokens) before feeding them into a model, crucial for handling retrieval efficiency and generation accuracy.

Chunking

Splitting large text documents into smaller, retrievable chunks to improve retrieval relevance and reduce token usage in LLMs.

Knowledge Augmentation

The process of adding external retrieved information into the prompt before passing it to the LLM, enhancing response quality.

Prompt Engineering

Crafting specific query formats to optimize retrieval and generation performance in RAG pipelines

Context Window

The number of tokens an LLM can process in a single query-response cycle, which impacts the effectiveness of RAG-based retrieval.

Hybrid Search

A combination of BM25 (keyword-based retrieval) and Vector Search (semantic similarity-based retrieval) to achieve better relevance and diversity in retrieved documents.

Re-Ranking

A secondary filtering step where retrieved documents are scored and ranked to improve the selection of relevant information.

Hallucination

A phenomenon where LLMs generate false or misleading information due to missing or incorrect retrieval results.

LlamaIndex (Formerly GPT Index)

A framework designed for efficient indexing and querying of structured and unstructured data in RAG pipelines.

LangChain

A powerful framework for building LLM-powered applications, including RAG, through modular integration of retrievers, models, and workflows.

Fine-Tuning

Adjusting an LLM’s parameters using new training data.

Retrieval Latency

The time taken to fetch relevant documents from a vector store, affecting the speed and efficiency of RAG-based models.

Knowledge Cutoff

The last point in time when an LLM was trained on data. RAG overcomes this limitation by fetching the latest information.

Multi-Modal RAG

Extending RAG to support images, audio, video, and structured data alongside text retrieval.

Contrastive Learning

A technique to improve retrieval performance by training models to maximize the similarity of relevant document-query pairs and minimize irrelevant ones.