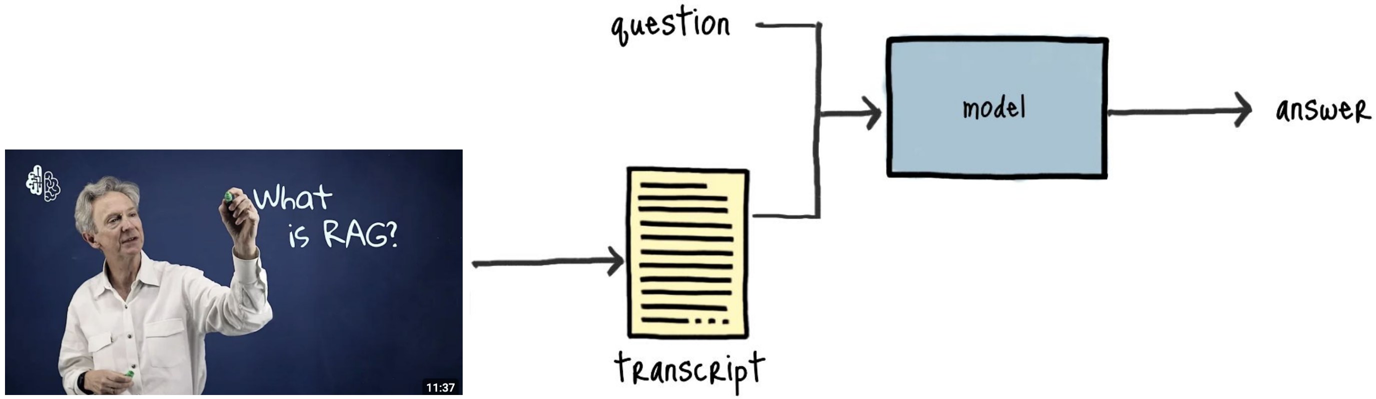
Retrieval Augmented Generation

- Apaar B.

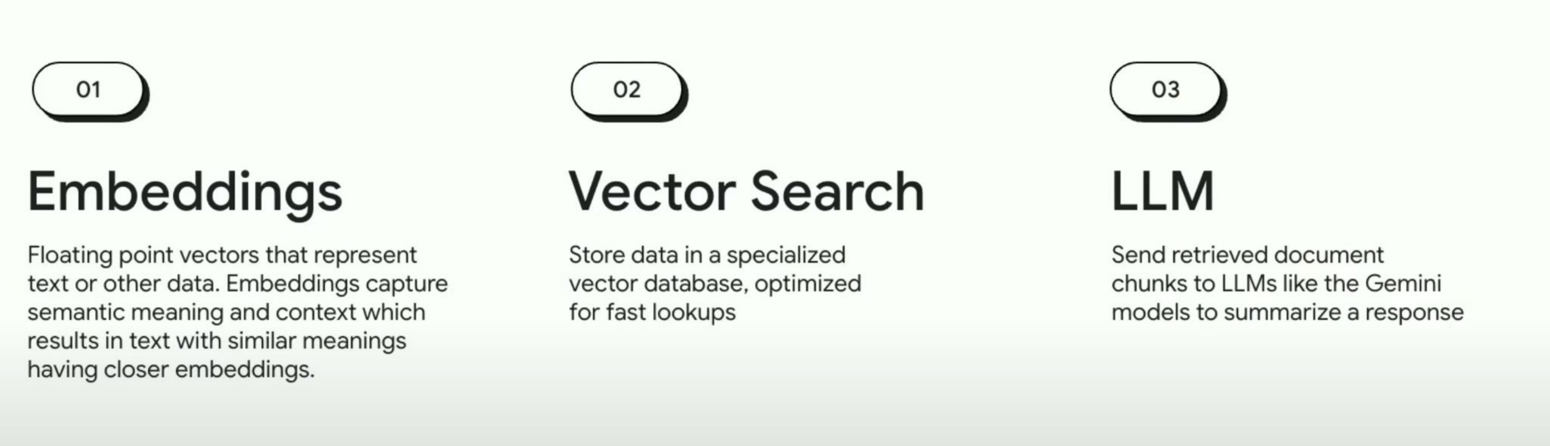
Retrieval Augmented Generation

Retrieval augmented generation (RAG), is a way to use external data or information to improve the accuracy of large language models (LLMs).

RAG doesn't train or fine-tune LLMs.



RAG Components



01

Embeddings

Floating point vectors that represent text or other data. Embeddings capture semantic meaning and context which results in text with similar meanings having closer embeddings.

02

Vector Search

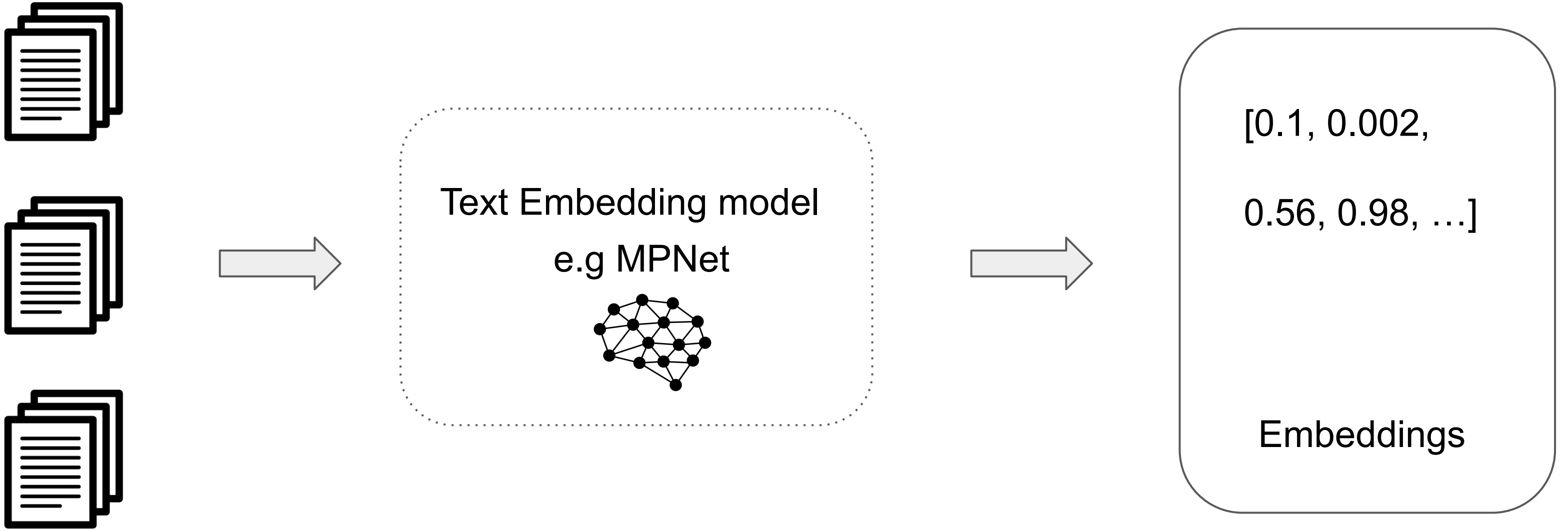
Store data in a specialized vector database, optimized for fast lookups

03

LLM

Send retrieved document chunks to LLMs like the Gemini models to summarize a response

1. Embeddings

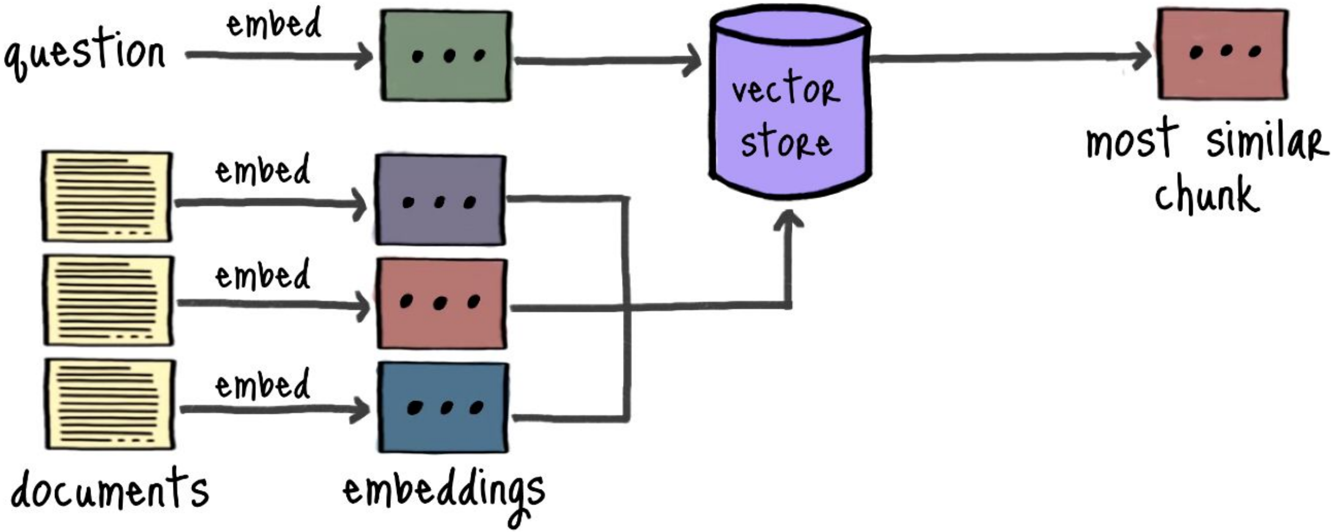


Text Embedding model
e.g MPNet

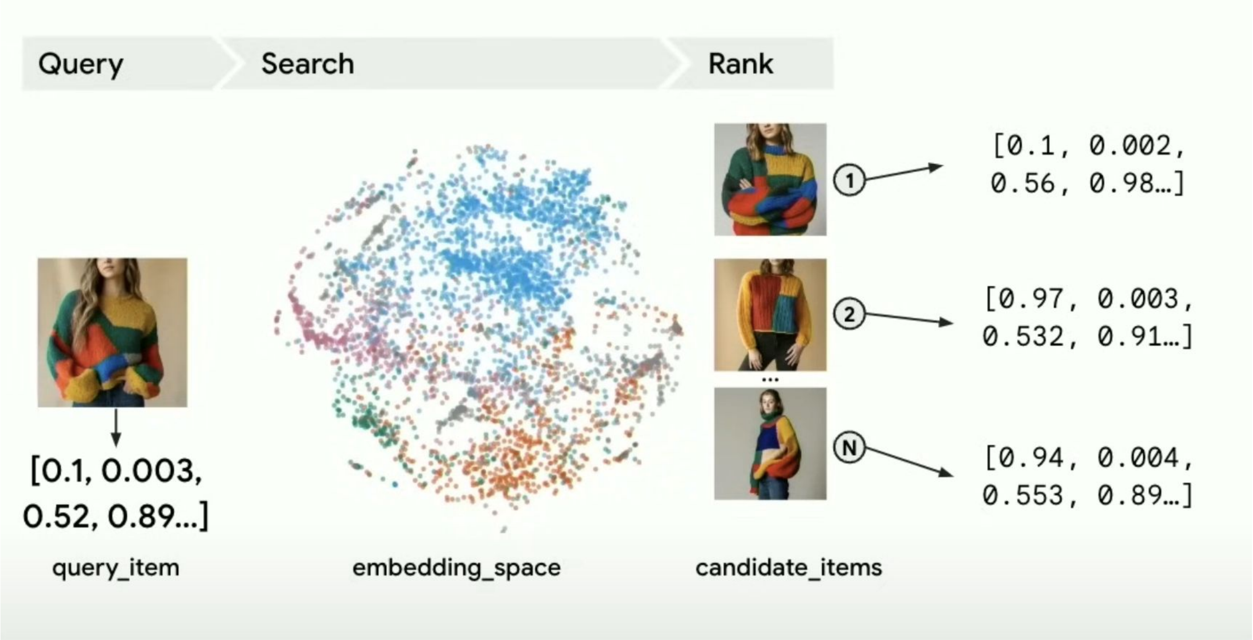
[0.1, 0.002,
0.56, 0.98, …]

Embeddings

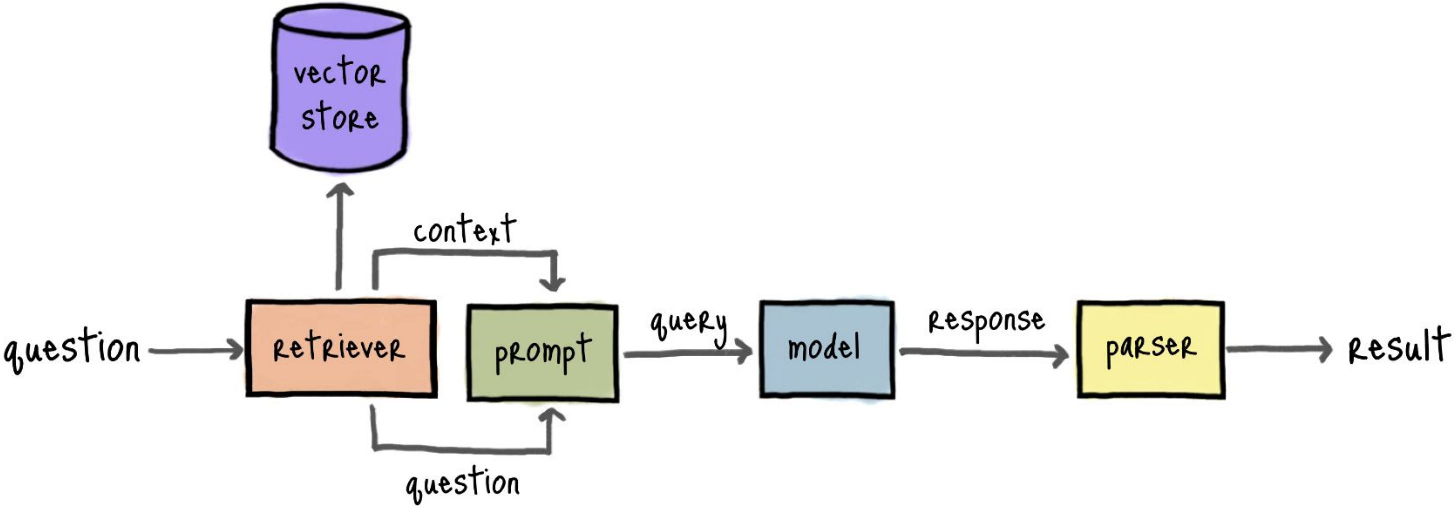
2. Vector Search



2. Vector Search



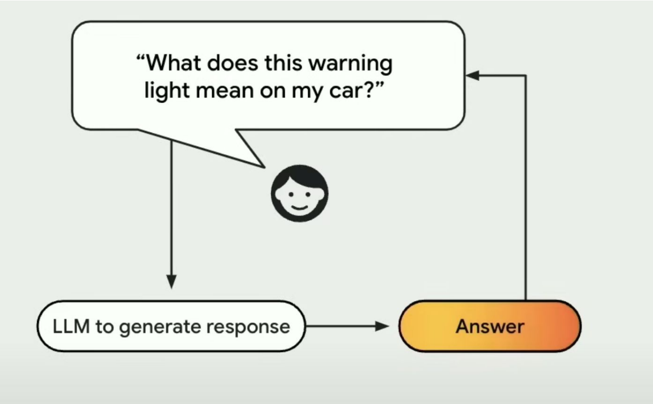
3. LLMs



1. Chunks retrieved from vector search are fed into the LLM

2. This augments the existing LLM's knowledge with the information it wasn't trained on

3. The LLM generates a response that weaves together retrieved chunks + pretrained knowledge

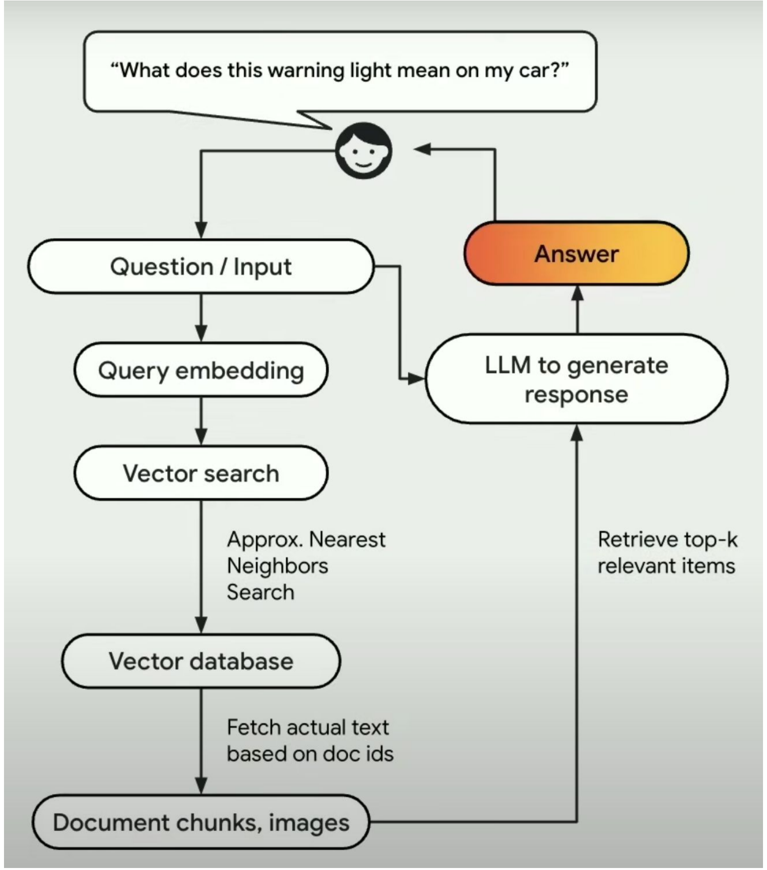


"What does this warning light mean on my car?"

LLM to generate response

Answer

Standalone LLM



"What does this warning light mean on my car?"

Question / Input

Query embedding

Vector search

Approx. Nearest Neighbors Search

Vector database

Fetch actual text based on doc ids

Document chunks, images

Retrieve top-k relevant items

LLM to generate response

Answer

RAG