Retrieval Augmented Generation - **RAG**

What is LLM?

Large Language Models (LLMs) are trained on vast volumes of data and use billions of parameters to generate original output for tasks like answering questions, translating languages, and completing sentences.

The model bases its responses on the textual content it has ingested during training

Challenges with LLM

Lack of Specific data - Language models are limited to providing generic answers based on their training data. If users were to ask domain specific questions, a traditional LLM may not be able to provide accurate answers.

Hallucination - The model bases its responses on the textual content it has ingested during training – and there's no telling exactly what went into that data, or how the model recombines it to generate novel text.

Generic Responses - Language models often provide generic responses that aren't tailored to specific contexts. This can be a major drawback in a customer support scenario since individual user preferences are usually required to facilitate a personalized customer experience.

What is RAG?

Retrieval-Augmented Generation (RAG) is the process of optimizing the output of a large language model, so it references an authoritative knowledge base outside of its training data sources before generating a response.

How RAG works?

Step 1: Data collection- You must first gather all the data that is needed for your application.

Step 2: Data chunking- Data chunking is the process of breaking your data down into smaller, more manageable pieces.

Step 3: Document embeddings- Now that the source data has been broken down into smaller parts, it needs to be converted into a vector representation. This involves transforming text data into embeddings, which are numeric representations that capture the semantic meaning behind text.

How RAG works?

Step 4: Handling user queries - When a user query enters the system, it must also be converted into an embedding or vector representation. The same model must be used for both the document and query embedding to ensure uniformity between the two.

Step 5: Generating responses with an LLM - The retrieved text chunks, along with the initial user query, are fed into a language model.

How RAG works?