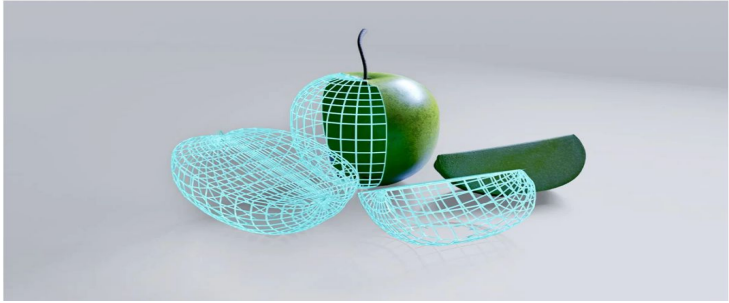
CWB- Module1

Introduction to GenAI Basics and Retrieval-Augmented Generation (RAG).

Introduction to GenAI



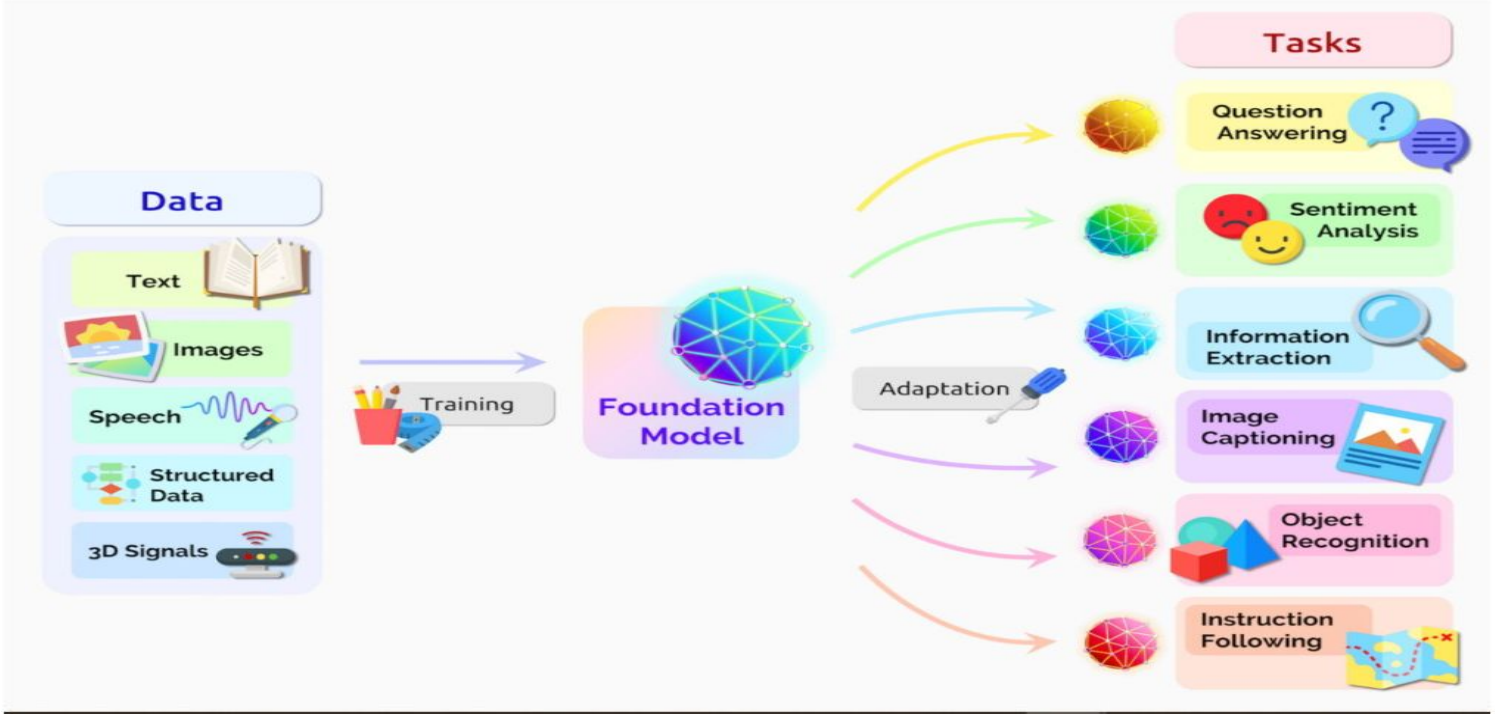
- Generative AI refers to the **use of AI to create new content,** like text, images, music, audio, and videos.

- Generative AI is **powered by foundation models** that can multi-task and perform out-of-the-box tasks, including summarization, Q&A, classification, and more.

Foundation Models

- Models that are trained on a broad set of unlabeled data that can be used for different tasks, with minimal fine-tuning.
- Foundation Models like OpenAI GPT-3, BERT, or DALL-E 2, etc.
- They are called **Foundation Models because they serve as the foundation for many applications of AI model.**

The domain-specific foundation model can be used for many tasks as opposed to the previous technologies that required building models from scratch in each use case.


Data
Text
Images
Speech
Structured Data
3D Signals
Training
Foundation Model
Adaptation
Tasks
Question Answering
Sentiment Analysis
Information Extraction
Image Captioning
Object Recognition
Instruction Following

What are Large Language Models?

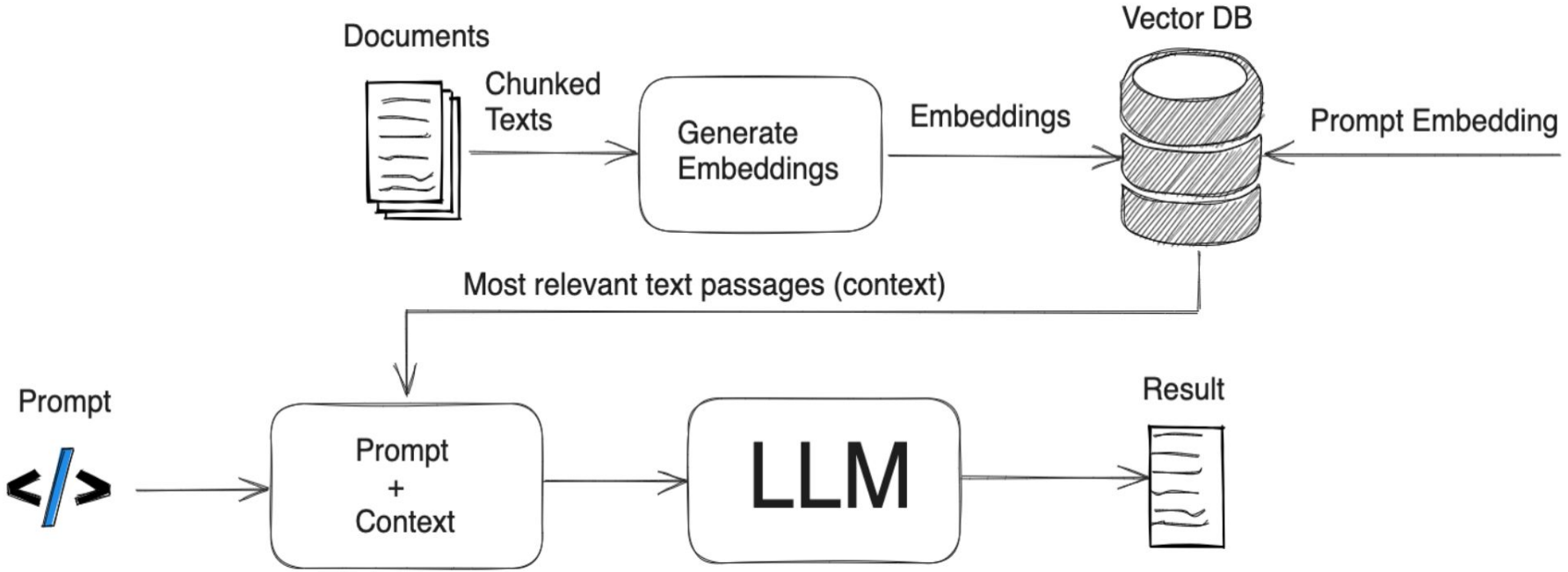
An LLM, **such as ChatGPT**, is an instance or a specific use case of a foundational model. It is a large language model based on a foundational architecture but is fine-tuned and specialized for generating human-like text in conversational interactions.

LLMs like ChatGPT are designed for chatbots, virtual assistants, or text-based dialog systems. They have undergone additional training to make them more coherent, context-aware, and suitable for natural language conversations.

Introduction to RAG

RAG, or Retrieval Augmented Generation, is a technique that combines the capabilities of a pre-trained large language model with an external data source.

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. 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


Documents
Chunked
Texts
Generate
Embeddings
Embeddings
Vector DB
Prompt Embedding
Most relevant text passages (context)
Prompt
Prompt
+
Context
LLM
Result

Applications of RAG

Text summarization

Personalized recommendations

Business intelligence