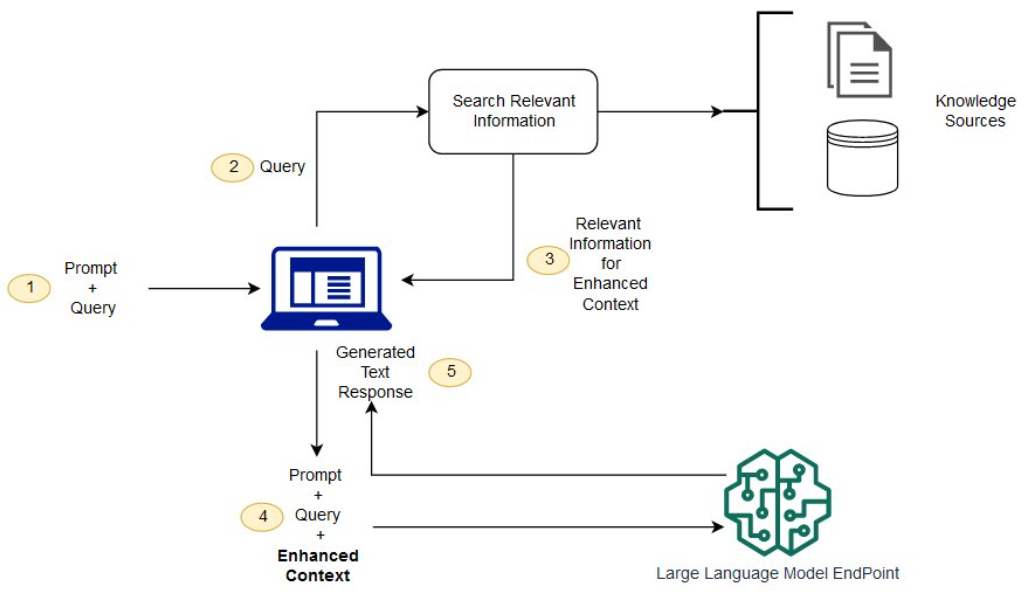
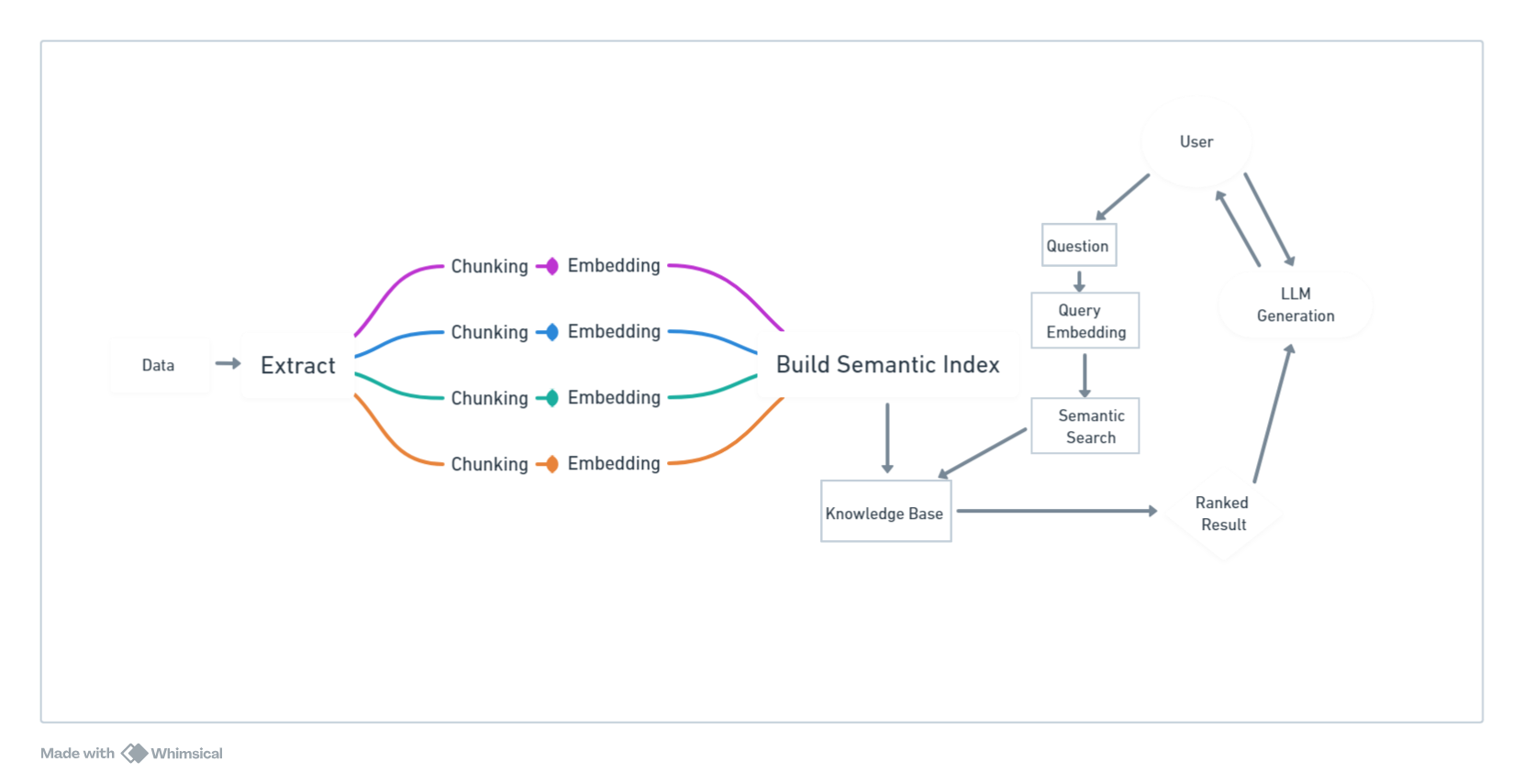
**RAG (Retrieval-Augmented Generation)**

* Process of Augmenting the output of Large Language Models(LLMs).
* LLMs are trained on vast data and use billions of parameters to generate original output.
* LLMs perform the task like question answering, language translation, completing sentence etc.
* RAG extends the capabilities of LLMs to specific domains or organizations internal knowledge base all this without retraining the model.
* It is a cost-effective approach to remain relevant, accurate and useful in various context.

Below is the High Level and Low Level RAG architecture:





**Why RAG?**

* The drawback of LLMs responses are very unpredictable and also LLM training data is static and has cut-off date on the knowledge it has.
* It gives false information when it does not have the answer.
* Presents out-of-date or generic information when the user expects specific response.
* It can give inaccurate information due to terminology confusion.
* RAG is the approach to tackle few of the challenges. It redirects the LLM to retrieve the information from authoritative pre-determined knowledge sources. Organizations have the control over generated output.

**How does Retrieval-Augmented Generation work?**

Without RAG, the LLM takes the user input and creates a response based on information it was trained on—or what it already knows.

With RAG, an information retrieval component is introduced that utilizes the user input to first pull information from a new data source.

The user query and the relevant information are both given to the LLM. The LLM uses the new knowledge and its training data to create better responses.

Following is the process of RAG:

**Create external data**

The new data outside of the LLM's original training data set is called external data. It can come from multiple data sources, such as a APIs, databases, or document repositories. The data may exist in various formats like files, database records, or long-form text. Another AI technique, called embedding language models, converts data into numerical representations and stores it in a vector database. This process creates a knowledge library that the generative AI models can understand.

**Retrieve relevant information**

The next step is to perform a relevancy search. The user query is converted to a vector representation and matched with the vector databases. For example, consider a smart chatbot that can answer human resource questions for an organization. If an employee searches, *"How much annual leave do I have?"* the system will retrieve annual leave policy documents alongside the individual employee's past leave record. These specific documents will be returned because they are highly-relevant to what the employee has input. The relevancy was calculated and established using mathematical vector calculations and representations.

**Augment the LLM prompt**

Next, the RAG model augments the user input (or prompts) by adding the relevant retrieved data in context. This step uses prompt engineering techniques to communicate effectively with the LLM. The augmented prompt allows the large language models to generate an accurate answer to user queries.

**Update external data**

The next question may be—what if the external data becomes stale? To maintain current information for retrieval, asynchronously update the documents and update embedding representation of the documents. You can do this through automated real-time processes or periodic batch processing.

**Benefits of RAG:**

* Cost-effective implementation
* Current Information
* Enhanced User Trust
* More Developer Control

Knowledge Resources are the vector database which store the additional data in vector form.

This is known as vector DB.

* Data is stored as arrays of numbers representing various data types like text, images etc.
* It is optimized for handling high dimensional space data.
* Efficiently does similarity search using metrics like cosine similarity or Euclidean distance.
* Indexing technique for retrieving fast data.
* Embeddings from models like Word2Vec and Bert can be stored for similarity search.

**Application of Vector Databases:**

* Search Engines
* Recommendation Systems
* NLP
* Computer Vision
* Anamoly Detection

**There are four major databases widely used in industries:**

* Chroma DB
* Weaviate DB
* FAISS DB
* Pinecone DB

**Chroma DB**

* [Chroma DB](https://docs.trychroma.com/) is an open-source vector store used for storing and retrieving vector embeddings.
* Its main use is to save embeddings along with metadata to be used later by large language models. Additionally, it can also be used for semantic search engines over text data.

**Weaviate DB**

Weaviate is an open-source vector database. It allows you to store data objects and vector embeddings and scale seamlessly into billions of data objects.

Key features of weaviate are:

* Speed: It can search ten nearest neighbors from millions of data objects in few milliseconds.
* Flexibility: With Weaviate, either vectorize data during import or upload your own, leveraging modules that integrate with platforms like OpenAI, Cohere, HuggingFace, and more.
* Production-ready. From prototypes to large-scale production, Weaviate emphasizes scalability, replication, and security.
* Beyond search: Apart from fast vector searches, Weaviate offers recommendations, summarizations, and neural search framework integrations.

**FAISS DB:**

It is an open source library for efficient similarity search and clustering of dense vectors.

**PineCone DB:**

Pinecone is a managed vector database platform that has been purpose-built to tackle the unique challenges associated with high-dimensional data. Equipped with cutting-edge indexing and search capabilities, Pinecone empowers data engineers and data scientists to construct and implement large-scale machine learning applications that effectively process and analyze high-dimensional data.

* Fully managed service
* Highly scalable
* Real-time data ingestion
* Low-latency search
* Integration with LangChain

**Conclusion:**

The most popular is the weaviate and Pinecone in the industry due to its feature.