what is rag?

RAG in AI stands for Retrieval-Augmented Generation. It is a technique that combines the power of retrieval-based systems with generative models like GPT. In a RAG system, when a user asks a question, the model first retrieves relevant documents or pieces of information from a large knowledge base (like Wikipedia or a private database). Then, it uses this retrieved information to generate a more accurate and informed answer. This approach helps improve the quality of responses, especially for questions that require up-to-date or specific knowledge not fully contained in the model’s training data.

what is flow?

In AI, flow usually refers to the movement of data and control through different stages of an AI system or pipeline. It describes how information is processed step by step — from input (like user data) through various components (like preprocessing, model inference, and post-processing) to the final output.

For example, in a machine learning pipeline, the flow might include:

1. Collecting and cleaning data

2. Feeding data into a model

3. Getting predictions

4. Displaying or using the results

Flow is important in AI because it helps organize complex processes clearly and efficiently, making it easier to design, debug, and scale systems. Tools like TensorFlow, Apache Airflow, or Microsoft Power Automate help manage or visualize these flows in AI projects.  
  
what is vector db?

A vector database is a special type of database designed to store, search, and manage vector embeddings — which are numerical representations of data like text, images, or audio. In AI, models often convert complex data into these vectors so that similar items can be compared using math.

For example, when you search using natural language, the system can convert your query into a vector and find the most similar vectors in the database — enabling things like semantic search, recommendation systems, or retrieval-augmented generation (RAG).

Popular vector databases include Pinecone, FAISS, Weaviate, and Milvus. They are optimized for fast similarity search using techniques like approximate nearest neighbor (ANN) algorithms.