Retrieval-Augmented Generation (RAG)

Overview

Retrieval-Augmented Generation (RAG) is an AI architecture that enhances large language models (LLMs) by combining text generation with external information retrieval. Instead of relying solely on pre-trained knowledge, RAG retrieves relevant documents from a knowledge source at query time and uses them to generate more accurate, up-to-date, and grounded responses.

How RAG Works

1. A user submits a query.
2. The query is converted into an embedding.
3. A retriever searches a knowledge base (e.g., vector database) for relevant documents.
4. Retrieved documents are passed to the language model as context.
5. The model generates a response grounded in the retrieved information.

Key Components

- **Embedding Model:** Converts text into numerical vectors.
- **Retriever:** Finds relevant documents based on similarity search.
- **Knowledge Store:** Vector database or document repository.
- **Generator (LLM):** Produces the final answer using retrieved context.

Benefits

- Reduces hallucinations by grounding responses in real data.
- Enables use of private or domain-specific knowledge.
- Keeps responses current without retraining the model.
- Improves transparency and explainability.

Common Use Cases

RAG is widely used in enterprise chatbots, question-answering systems, knowledge assistants, customer support automation, internal documentation search, and AI-powered research tools.