Retrieval-Augmented Generation (RAG) is a technique that enhances large language models by providing them with external knowledge sources.

Instead of relying only on the model's training data, RAG retrieves relevant information from documents and passes it as context to the language model.

RAG Architecture:

1. Documents are split into smaller chunks.
2. Each chunk is converted into vector embeddings.
3. Embeddings are stored in a vector database.
4. A user query is embedded and matched against stored vectors.
5. The most relevant chunks are retrieved.
6. The language model generates an answer using the retrieved context.

Benefits of RAG:

- Reduces hallucinations
- Improves factual accuracy
- Enables domain-specific knowledge
- Keeps models up-to-date without retraining

RAG is widely used in enterprise search, knowledge bases, and AI assistants.