Title

Introduction to Retrieval-Augmented Generation (RAG)

---

1. What is Retrieval-Augmented Generation?

Retrieval-Augmented Generation (RAG) is an AI architecture that combines information retrieval with text generation.

Instead of relying only on a language model's internal knowledge, RAG retrieves relevant documents from an external knowledge source and uses them to generate grounded answers.

---

2. Why is RAG Important?

Large language models may generate incorrect or fabricated information, a phenomenon known as hallucination.

RAG reduces hallucinations by grounding responses in retrieved documents. If the relevant information is not found in the retrieved data, a properly designed RAG system should avoid answering.

---

3. Core Components of a RAG System

A typical RAG system consists of the following components:

1. **Document Store** A collection of external documents such as PDFs, reports, or knowledge bases.
2. **Text Chunking** Documents are split into smaller chunks so they can be embedded efficiently.
3. **Embedding Model** Text chunks are converted into numerical vectors that represent semantic meaning.
4. **Vector Database** Vector embeddings are stored in a database that supports similarity search.
5. **Retriever** Given a user query, the retriever finds the most relevant document chunks.
6. **Generator** A language model generates an answer using only the retrieved context.

---

4. Role of Embeddings in RAG

Embeddings convert text into dense numerical vectors. Texts with similar meanings have vectors that are close to each other in vector space.

In RAG, embeddings are used to match user queries with relevant document chunks using similarity search.

---

5. Vector Databases and FAISS

FAISS (Facebook AI Similarity Search) is a library for efficient similarity search and clustering of dense vectors.

FAISS enables fast retrieval of relevant document chunks even on CPU-only systems.

---

6. How RAG Reduces Hallucinations

RAG reduces hallucinations by enforcing two constraints:

1. Answers must be generated from retrieved documents.
2. If no relevant documents are retrieved, the system should refuse to answer.

This makes RAG systems more reliable and trustworthy than standalone language models.

---

7. Limitations of RAG

While RAG improves factual grounding, it still depends on:

- The quality of source documents
- The quality of embeddings
- The design of prompts and retrieval logic

Poorly designed RAG systems can still produce incorrect answers.

---

8. Summary

Retrieval-Augmented Generation is a practical approach for building reliable AI systems.

By combining document retrieval with controlled generation, RAG enables AI models to provide accurate and explainable answers grounded in external knowledge.
