Training Day 15 Report

Overview

Retrieval-Augmented Generation (RAG) is a hybrid architecture that combines traditional information retrieval methods with generative capabilities of language models. It is designed to overcome the limitation of language models relying only on their pre-trained knowledge, by allowing them to access external data sources during inference.

Architecture Components

RAG models integrate two main systems:

- **Retriever:** Locates relevant documents from an external corpus using semantic search techniques.
- **Generator:** A language model (e.g., GPT, BERT) that generates output based on both the query and the retrieved context.

How RAG Works

1. User inputs a query.
2. The retriever searches an indexed document store for top-k relevant passages.
3. Retrieved passages are appended to the query and sent to the language model.
4. The model generates a grounded and context-aware response.

Advantages of RAG

- Reduces hallucination in LLM outputs.
- Dynamically integrates new knowledge without retraining.
- Ideal for domains where accuracy and references matter (e.g., healthcare, law, research).

Challenges and Considerations

- Requires a high-quality document corpus and embeddings.
- Retrieval quality heavily impacts generation accuracy.
- Needs chunking, indexing, and efficient vector similarity mechanisms.

Conclusion

RAG systems offer a promising solution to combine the power of large language models with the flexibility of real-time knowledge access. They are becoming essential in AI applications that demand high factual correctness, source grounding, and domain adaptability.