Introduction to Retrieval-Augmented Generation (RAG) for Study Assistants

Team Syntax Syndicate
KIU Fall 2025

December 2025

1 Abstract

This lecture note explores the architecture and benefits of Retrieval-Augmented Generation (RAG). We discuss how connecting Large Language Models (LLMs) to external knowledge sources reduces hallucinations and provides verifiable citations, which is critical for educational tools like Cognify.

2 The Problem of LLM Hallucinations

Large Language Models, while powerful, suffer from a phenomenon known as **Hallucination**. This occurs when the model generates factually incorrect information with high confidence. In a study environment, hallucinations are dangerous because they can lead to student misinformation.

3 What is RAG?

Retrieval-Augmented Generation is a technique that combines a generative model with a retrieval system. Instead of relying solely on its training data, the model follows a three-step process:

1. **Retrieval:** The system searches a document database for relevant text snippets.
2. **Augmentation:** The snippets are added to the user's prompt.
3. **Generation:** The model generates an answer based strictly on the provided snippets.

4 Core Terminology

Understanding the following terms is essential for building AI applications:

- **Vector Embedding:** A numerical representation of text that captures semantic meaning.

- **Cosine Similarity:** A mathematical formula used to find how related two pieces of text are.
- **Context Window:** The maximum amount of text an AI can ”read” at one time.
- **Prompt Engineering:** The art of crafting instructions to get better results from an LLM.

5 Evaluating RAG Performance

To ensure the system is working, developers use the **Golden Set** method. This involves creating a set of ”ground truth” questions and answers to test the system’s accuracy, latency, and reliability.

6 Conclusion

RAG represents a major shift in AI development. By grounding models in real data, we can create trustworthy assistants that help students master complex subjects without the risk of false information.