Module 6 of 10

RAG Advanced approaches

Grid University

This module aims to equip learners with knowledge of sophisticated techniques to significantly enhance Retrieval Augmented Generation (RAG) pipelines. The objective is to understand various advanced approaches that improve RAG's latency, accuracy, and robustness, enabling you to optimise RAG systems for specific business needs.

Key learning areas:

- **Optimising Retrieval:** Explore methods like hybrid search, re-ranking, ColBERT, and HyDE to improve the relevance and quality of information fetched for the LLM.
- **Enhancing Generation & Robustness:** Delve into techniques such as self-reflective RAG and query reformulation that enable the RAG system to intelligently refine its output and adapt to complex queries.
- **Boosting Performance & Efficiency:** Understand approaches like semantic caching and other strategies focused on reducing latency and minimising computational costs, directly impacting user experience and operational efficiency.

By the end of this module, you will be able to apply advanced techniques to optimise your RAG system for specific needs, significantly improving its overall performance in terms of latency, accuracy, and robustness.

To complete this module, you need to finish the course listed below and review the reading materials.

Advanced Retrieval Techniques for RAG Pipelines

This content focuses on advanced techniques to significantly enhance the retrieval phase of your RAG pipeline. You'll explore methods like hybrid search, re-ranking, and advanced models such as ColBERT and HyDE (Hypothetical Document Embeddings). The goal is to understand how these strategies directly improve the relevance and quality of the information provided to the Large Language Model, leading to more accurate and useful generated responses.

Hybrid search + reranking
For an AI model to be useful in specific contexts, it often needs access to background knowledge.
READ

ColBERT approach
Hybrid search combines dense and sparse retrieval to deliver precise and comprehensive results.
READ

HyDE approach
HyDE uses a Language Learning Model, like ChatGPT, to create a theoretical document when responding to a query, as opposed to using the query and its computed vector to directly seek in the vector database.
READ

Optimizing Queries for Accurate RAG Results

These reading materials cover query rewriting techniques to optimize user queries for better RAG retrieval. You'll learn how to improve ambiguous or short queries using methods like query expansion, decomposition, and conversational rewriting to ensure the system fetches the most relevant information.

Retrieval
Overwire + check Multi-Query, Decomposition and Step Back approaches.
READ

Advanced RAG Techniques
Self-reflective RAG, Corrective RAG, query reformulation approaches.
READ

Optimizing RAG with Semantic Caching

In this part of the course you'll explore semantic caching, a powerful technique to optimize RAG performance by reducing latency and cost. You'll learn how it stores and retrieves results based on query meaning (using embeddings), avoiding redundant LLM calls and retrieval operations, thus improving user experience and cutting expenses:

Implementing semantic cache to improve a RAG system with FAISS
In this notebook, we will explore a typical RAG solution where we will utilize an open-source model and the vector database Chroma DB.
READ

Additional reading:

Advanced RAG: Fine-Tune Embeddings from HuggingFace for RAG
Fine-tuning embeddings approach.
READ

Better RAG 3: The text is your friend
Metadata extraction.
READ

Agentic RAG:
turbocharge your RAG with query reformulation and self-query!
READ

RAG is dead, long live agentic retrieval
RAG has come a long way since the days of naive chunk retrieval; now agentic strategies are table stakes.
READ

Optimizing LLM Accuracy
Maximize correctness and consistent behavior when working with LLMs.
READ

Building Performant RAG Applications for Production
Prototyping a RAG application is easy, but making it performant, robust, and scalable to a large knowledge corpus is hard.
READ

A Cheat Sheet and Some Recipes For Building Advanced RAG
It's the start of a new year and perhaps you're looking to break into the RAG scene by building your very first RAG system.
READ