Module 3 of 10

RAG evaluation

Grid University

This module aims to teach you the essential methods for assessing the performance of Retrieval Augmented Generation systems. You will learn how to evaluate the RAG pipeline as a whole, as well as its individual components (retrieval and generation), and understand the key metrics used to quantify their effectiveness.

Key learning areas:

- **Metrics for Overall RAG System Evaluation:** Explore end-to-end metrics that assess the final output of the RAG system.
- **Metrics for the Components:** Dive into metrics specifically for evaluating how well the system fetches relevant documents and the quality of the LLM's output based on the retrieved context.
- **Main Tools and Frameworks for RAG Evaluation:** Get familiar with specialised tools that automate and streamline the evaluation process.
- **Dataset Creation:** Learn techniques for creating robust evaluation datasets, including question-answer pairs and relevant contexts.

By the end of this module, you will be able to choose appropriate metrics and tools to accurately measure the performance of the RAG system, allowing you to effectively compare the accuracy of different RAG implementations or improvements you make.

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

How to assess the overall RAG system

In these materials you'll learn how to assess the overall RAG system and its individual components (retrieval and generation), covering key metrics like answer correctness, relevance, faithfulness, precision, and recall to quantify performance and overall evaluation pipeline setup.

RAG Evaluation
This notebook demonstrates how you can evaluate your RAG, by building a synthetic evaluation dataset and using LLM-as-a-judge to compute the accuracy of your system.

READ

Deep Dive Into Evaluating RAG Outputs
In this notebook, we'll show you how to evaluate the output of a RAG system. The high-level RAG flow is depicted in the diagram below.

READ

How to create synthetic datasets for RAG evaluation

The following resources explain how to create synthetic datasets for RAG evaluation. You'll learn methods to generate synthetic queries and question-answer pairs using LLMs, focusing on strategies to mitigate bias and enhance diversity (e.g., varied question types, user personas).

Generating Synthetic Dataset for RAG
Unfortunately, in the life of a Machine Learning Engineer, there's often a lack of labeled data or very little of it.

READ

Tackling Generated Datasets Diversity
In the previous chapter, we discussed the potential of using LLM for synthetic dataset generation to further finetune a local Retriever model.

READ

RAG evaluation tools

The following articles will introduce you to popular RAG evaluation tools, including RAGAS and DeepEval that assist in building robust evaluation benchmarks.

RAGAS framework
The purpose of this guide is to illustrate a simple workflow for testing and evaluating a RAG system with ragas.

READ

DepEval framework
GitHub

READ

Next: Final Assessment

In Progress 50%

Retrieval-Augmented Generator

- Retrieval-Augmented Generation (RAG)
- Final Assessment

Course Tasks 0/1

Course Evaluation