Evaluation of RAG Pipelines

Building a PoC RAG pipeline is not overtly complex. LangChain and LlamaIndex have made it quite simple. Developing highly impressive Large Language Model (LLM) applications is achievable through brief training and verification on a limited set of examples. However, to enhance its robustness, thorough testing on a dataset that accurately mirrors the production distribution is imperative.

RAG is a great tool to address hallucinations in LLMs but...

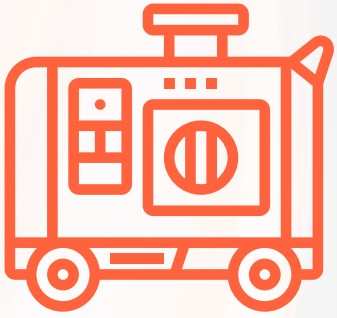
even RAGs can suffer from hallucinations

This can be because -

- The retriever fails to retrieve relevant context or retrieves irrelevant context
- The LLM, despite being provided the context, does not consider it
- The LLM instead of answering the query picks irrelevant information from the context

Two processes, therefore, to focus on from an evaluation perspective -

Search & Retrieval



Generation

- How good is the retrieval of the context from the Vector Database?
- Is it relevant to the query?
- How much noise (irrelevant information) is present?

- How good is the generated response?
- Is the response grounded in the provided context?
- Is the response relevant to the query?

Ragas (RAG Assessment)

Jithin James and Shahul ES from Exploding Gradients, in 2023, developed the Ragas framework to address these questions.

https://github.com/explodinggradients/ragas

Evaluation Data

To evaluate RAG pipelines, the following four data points are recommended

- A set of **Queries** or **Prompts** for evaluation
- **Retrieved Context** for each prompt
- Corresponding **Response** or **Answer** from LLM
- **Ground Truth** or known correct response

Evaluation Metrics

Evaluating Generation

Faithfulness — Is the **Response** faithful to the **Retrieved Context**?

Answer Relevance — Is the **Response** relevant to the **Prompt**?

Retrieval Evaluation

Context Relevance — Is the **Retrieved Context** relevant to the **Prompt**?

Context Recall — Is the **Retrieved Context** aligned to the **Ground Truth**?

Context Precision — is the **Retrieved Context** ordered correctly?

Overall Evaluation

Answer Semantic Similarity

is the **Response** semantically similar to the **Ground Truth**?

Answer Correctness

is the **Response** semantically and factually similar to the **Ground Truth**?

Evaluation Metrics

1 Faithfulness

Faithfulness is the measure of the extent to which the response is factually grounded in the retrieved context

Problem addressed : The LLM, despite being provided the context, does not consider it

or

Is the response grounded in the provided context?

Evaluated Process : Generation

Any measure of retrieval accuracy is out of scope

Score Range : (0,1) Higher score is better

Methodology

Faithfulness identifies the number of "claims" made in the response and calculates the proportion of those "claims" present in the context.

$$\textbf{Faithfulness} = \frac{\textit{Number of generated claims present in the context}}{\textit{Total number of claims made in the generated response}}$$

Illustrative Example

Query : Who won the 2023 ODI Cricket World Cup and when?

Context : The 2023 ODI Cricket World Cup concluded on 19 November 2023, with Australia winning the tournament.

Response 1 : High Faithfulness	**Response 2 : Low Faithfulness**
[Australia] won on [19 November 2023]	*[Australia] won on [15 October 2023]*

Evaluation Metrics

2 Answer Relevance

Answer Relevance is the measure of the extent to which the response is relevant to the query or the prompt

Problem addressed :The LLM instead of answering the query responds with irrelevant information

or

Is the response relevant to the query?

Evaluated Process : Generation

Any measure of retrieval accuracy is out of scope

Score Range : (0,1) Higher score is better

Methodology

For this metric, a response is generated for the initial query or prompt
To compute the score, the LLM is then prompted to generate questions for the generated response several times. The mean cosine similarity between these questions and the original one is then calculated. The concept is that if the answer correctly addresses the initial question, the LLM should generate questions from it that match the original question.

Answer Relevance = *Avg (Sc (Initial Query, LLM generated Query [i]))*

Illustrative Example

Query : Who won the 2023 ODI Cricket World Cup and when?

Response 1 : High Answer Relevance	**Response 2 : Low Answer Relevance**
India won on 19 November 2023	*Cricket world cup is held once every four years*

Note

Answer Relevance is **not a measure of truthfulness** but only of relevance. The response may or may not be factually accurate but may be relevant.

Evaluation Metrics

3 Context Relevance

Context Relevance is the measure of the extent to which the retrieved context is relevant to the query or the prompt

Problem addressed :The retriever fails to retrieve relevant context

or

Is the retrieved context relevant to the query?

Evaluated Process : Retrieval

Indifferent to the final generated response

Score Range : (0,1) Higher score is better

Methodology

The retrieved context should contain information only relevant to the query or the prompt. For context relevance, a metric 'S' is estimated. 'S' is the number of sentences in the retrieved context that are relevant for responding to the query or the prompt.

$$\text{Context Relevance} = \frac{S}{\text{Total number of sentences in the retrieved context}}$$

S = *(number of relevant sentences from the context)*

Illustrative Example

Query : Who won the 2023 ODI Cricket World Cup and when?

Context 1 : High Context Relevance	**Context 2 : Low Context Relevance**
The 2023 Cricket World Cup, concluded on 19 November 2023, with Australia winning the tournament. The tournament took place in ten different stadiums, in ten cities across the country. The final took place between India and Australia at Narendra Modi Stadium	*The 2023 Cricket World Cup was the 13th edition of the Cricket World Cup. It was the first Cricket World Cup which India hosted solely. The tournament took place in ten different stadiums.In the first semi-final India beat New Zealand, and in the second semi-final Australia beat South Africa.*

Evaluation Metrics

Ground Truth

Ground truth is information that is known to be real or true. In RAG, or Generative AI domain in general, Ground Truth is a prepared set of **Prompt-Response examples**. It is akin to *labelled data* in Supervised Learning parlance.

Calculation of certain metrics necessitates the availability of Ground Truth data

4 Context Recall

Context recall measures the extent to which the retrieved context aligns with the "provided" answer or Ground Truth

Problem addressed :The retriever fails to retrieve accurate context

or

Is the retrieved context good enough to provide the response?

Evaluated Process : Retrieval

Indifferent to the final generated response

Score Range : (0,1) Higher score is better

Methodology

To estimate context recall from the ground truth answer, each sentence in the ground truth answer is analyzed to determine whether it can be attributed to the retrieved context or not. Ideally, all sentences in the ground truth answer should be attributable to the retrieved context.

$$\text{Context Recall} = \frac{\textit{Number of Ground Truth sentences in the context}}{\textit{Total number of sentences in the Ground Truth}}$$

Illustrative Example

Query : Who won the 2023 ODI Cricket World Cup and when?

Ground Truth : Australia won the world cup on 19 November, 2023.

Context 1 : High Context Recall	Context 2 : Low Context Recall
The 2023 Cricket World Cup, concluded on 19 November 2023, with Australia winning the tournament.	*The 2023 Cricket World Cup was the 13th edition of the Cricket World Cup. It was the first Cricket World Cup which India hosted solely.*

Evaluation Metrics

5 Context Precision

Context Precision is a metric that evaluates whether all of the ground-truth relevant items present in the contexts are ranked higher or not.

Problem addressed :The retriever fails to rank retrieve context correctly

or

Is the higher ranked retrieved context better to provide the response?

Evaluated Process : Retrieval

Indifferent to the final generated response

Score Range : (0,1) Higher score is better

Methodology

Context Precision is a metric that evaluates whether all of the ground-truth relevant items present in the all retrieved context documents are ranked higher or not. Ideally all the relevant chunks must appear at the top

$$\text{Context Precision @ k} = \frac{\textit{Sum(Precision@k)}}{\textit{Total number of relevant documents in the top results}}$$

$$\text{Precision @ k} = \frac{\textit{True Positives @ k}}{\textit{(True Positives @ k + False Positives @ k)}}$$

Precision @ k

Precision@k is a metric used in information retrieval and recommendation systems to evaluate the accuracy of the top k items retrieved or recommended. It measures the proportion of relevant items among the top k items.

Evaluation Metrics

6 Answer semantic similarity

Answer semantic similarity evaluates whether the generated response is similar to the “provided” response or Ground Truth.

Problem addressed : The generated response is incorrect

or

Does the pipeline generate the right response?

Evaluated Process : Retrieval & Generation

Score Range : (0,1) Higher score is better

Methodology

Answer semantic similarity score is calculated by measuring the semantic similarity between the generated response and the ground truth response.

Answer Semantic Similarity = *Similarity (Generated Response, Ground Truth Response)*

7 Answer Correctness

Answer correctness evaluates whether the generated response is semantically and factually similar to the “provided” response or Ground Truth.

Problem addressed : The generated response is incorrect

or

Does the pipeline generate the right response?

Evaluated Process : Retrieval & Generation

Score Range : (0,1) Higher score is better

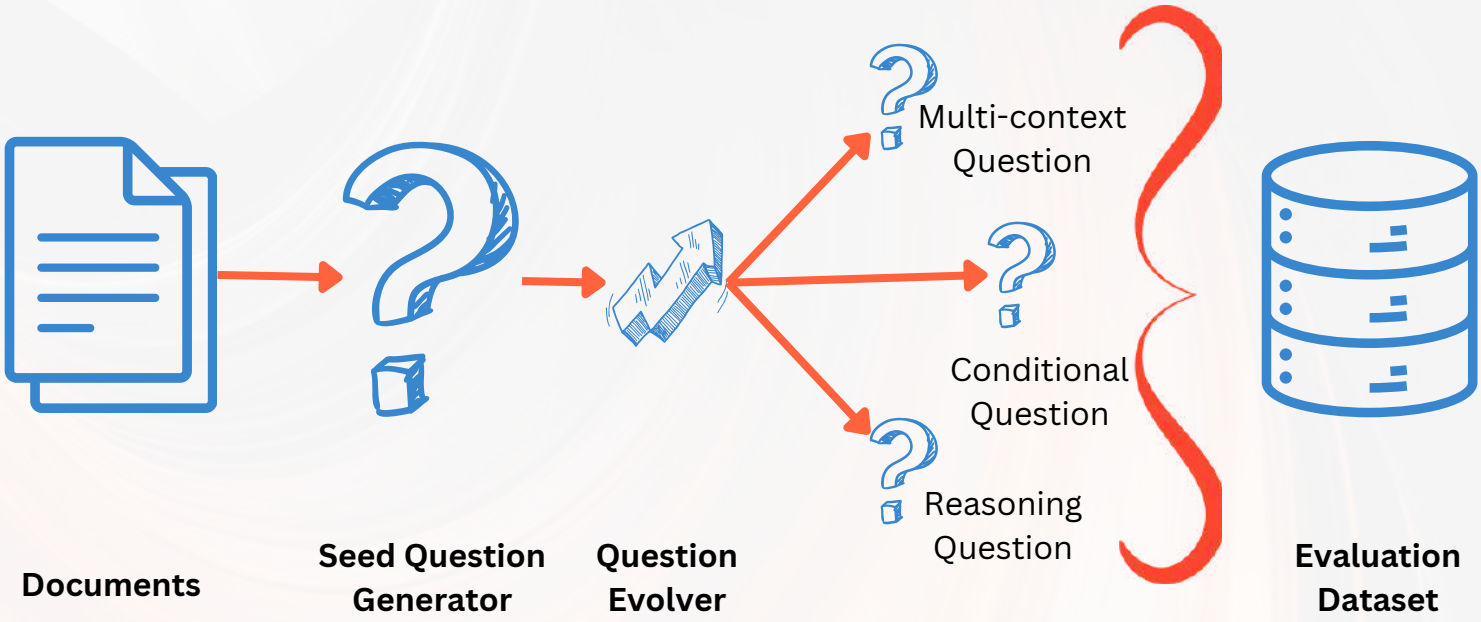
Methodology

Answer correctness score is calculated by measuring the semantic and the factual similarity between the generated response and the ground truth response.

Synthetic Test Data Generation

Generating hundreds of QA (Question-Context-Answer) samples from documents manually can be a time-consuming and labor-intensive task. Moreover, questions created by humans may face challenges in achieving the necessary level of complexity for a comprehensive evaluation, potentially affecting the overall quality of the assessment.

Synthetic Data Generation uses Large Language Models to generate a variety of Questions/Prompts and Responses/Answers from the Documents (Context). It can greatly reduce developer time.



Synthetic Data Generation Pipeline

	question	context	answer	question_type	episode_done
0	What technique improves the performance of lar...	- "We explore how generating a chain of though...	The technique that improves the performance of...	simple	True
1	What phenomenon is discussed in the paper rega...	- This paper instead discusses an unpredictabl...	The phenomenon discussed in the paper is the e...	reasoning	True
2	What is the purpose of chain-of-thought (CoT) ...	- Providing these steps for prompting demonstr...	The purpose of chain-of-thought (CoT) promptin...	simple	True
3	What is the performance of the largest fine-tu...	On the MathQA-Python dataset, the largest fine...	The performance of the largest fine-tuned mode...	simple	True
4	What is the accuracy increase of Zero-shot-CoT...	Experimental results demonstrate that our Zero...	The accuracy increase of Zero-shot-CoT on Mult...	reasoning	True

Synthetic Data Generated Using Ragas

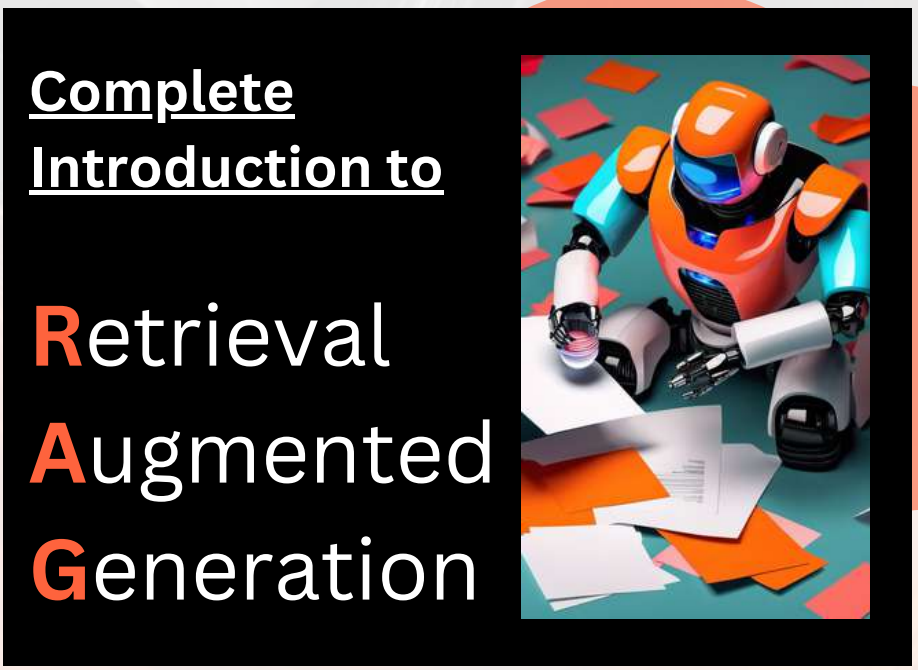
Ragas Documentation

Yarnit

Generate blogs, emails, social media posts, advertisements at the speed of your thought using Generative AI

Sign Up

www.yarnit.app

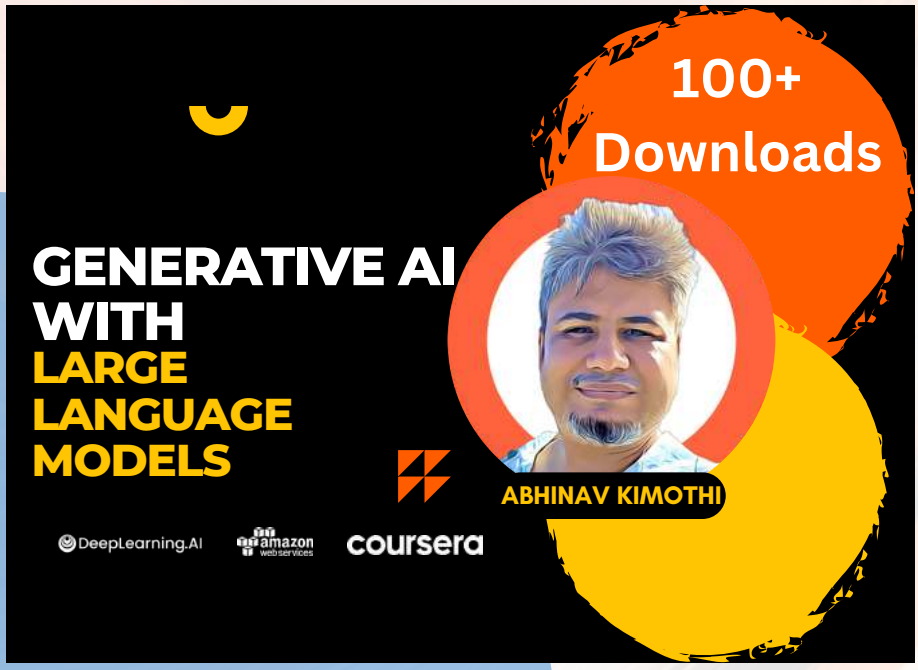


Coming Soon

Introduction to RAG architecture and builiding RAG Apps with code examples using LangChain & LlamaIndex

Subscribe

Subscribe for Updates



LLM Notes

Detailed Notes from Generative AI with Large Language Models Course by Deeplearning.ai and AWS.

Download

Free version available

Follow Abhinav Kimothi for more posts like this