Analytics Vidhya

Introduction to RAG Evaluation Metrics

Instructor

Dipanjan Sarkar

Head of Community & Principal AI Scientist at Analytics Vidhya

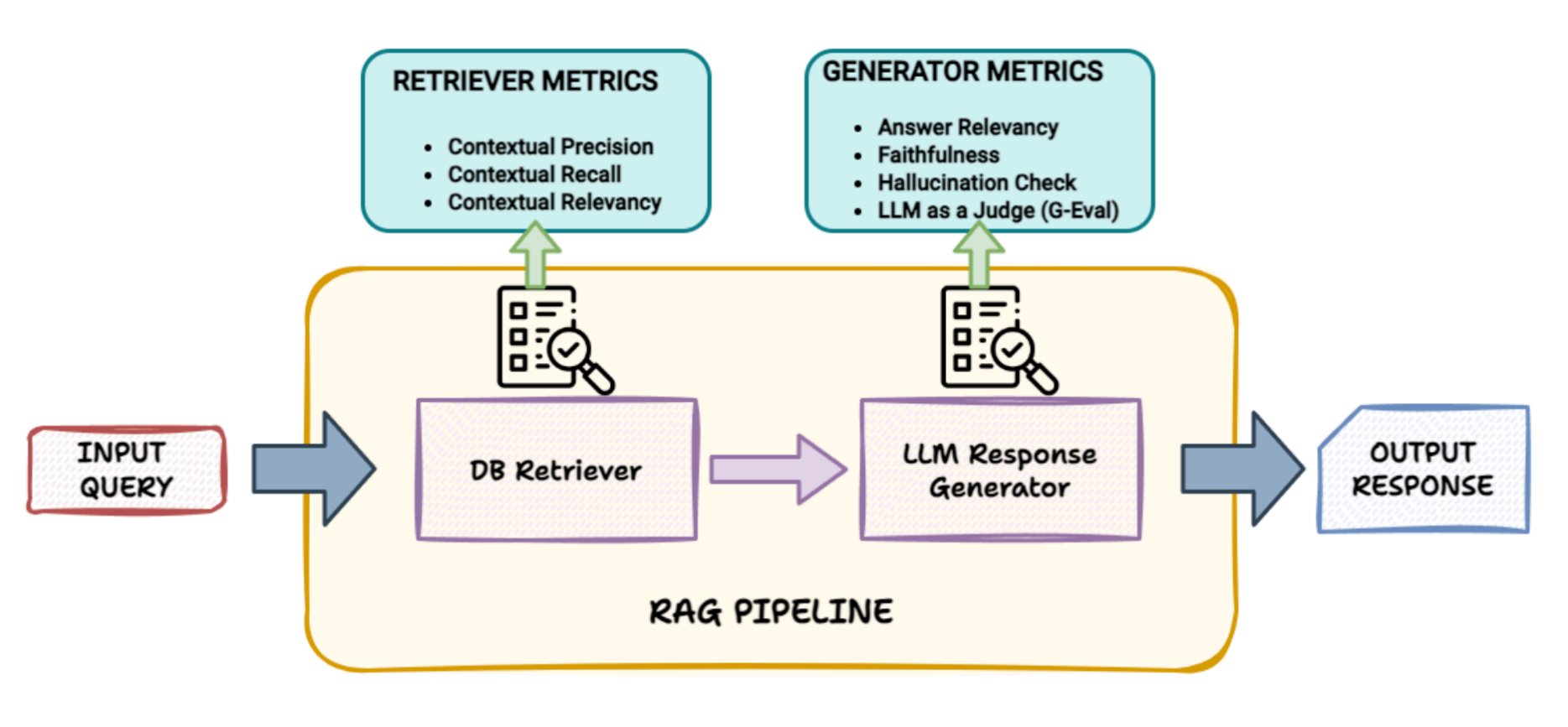
Google Developer Expert - ML & Cloud Champion Innovator

Published Author

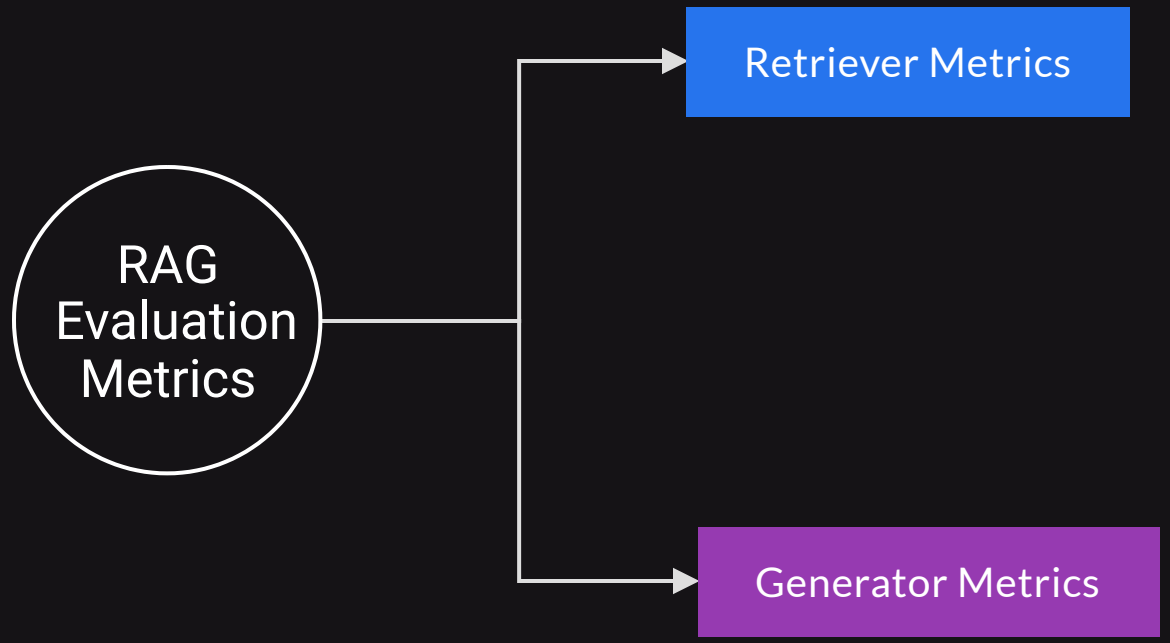
Introduction

RAG Evaluation Points & Metrics

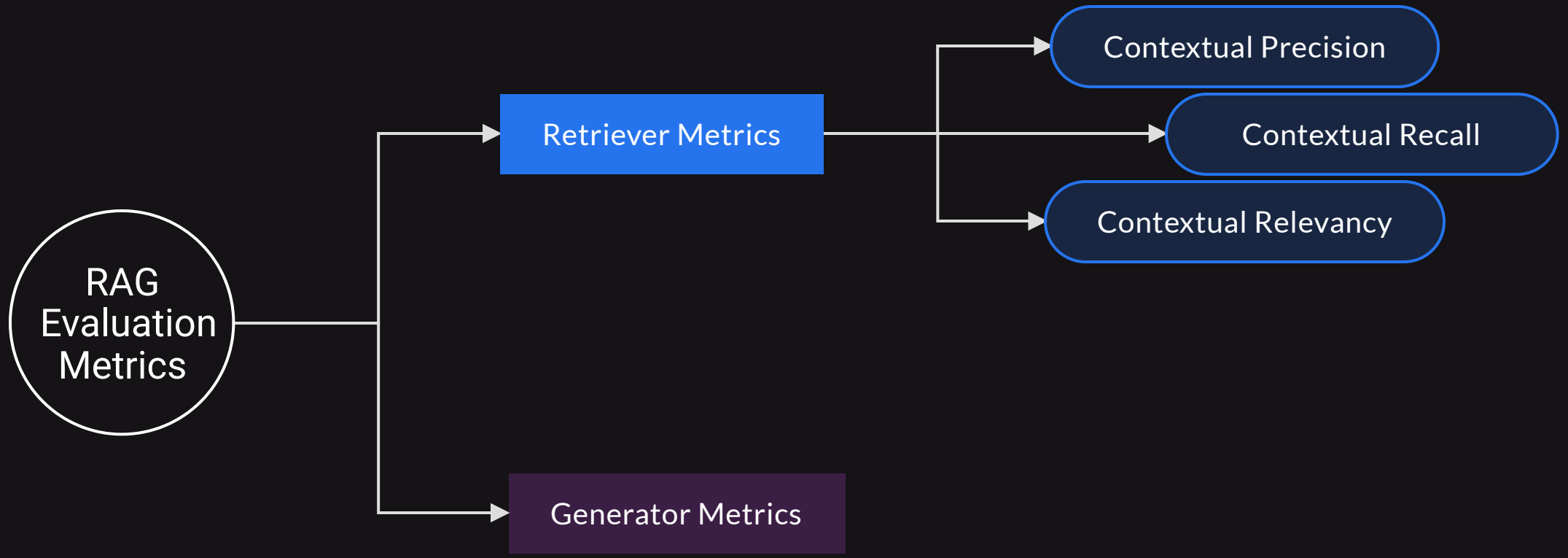
RAG Evaluation Points & Metrics



Classification: RAG Evaluation Metrics



Classification: RAG Evaluation Metrics



Context Precision

- Measures, whether **retrieved context** document chunks (nodes) that are relevant to the given **input query,** are ranked higher than irrelevant ones
- Higher Context Precision score represents a better retrieval system which can correctly rank relevant nodes higher

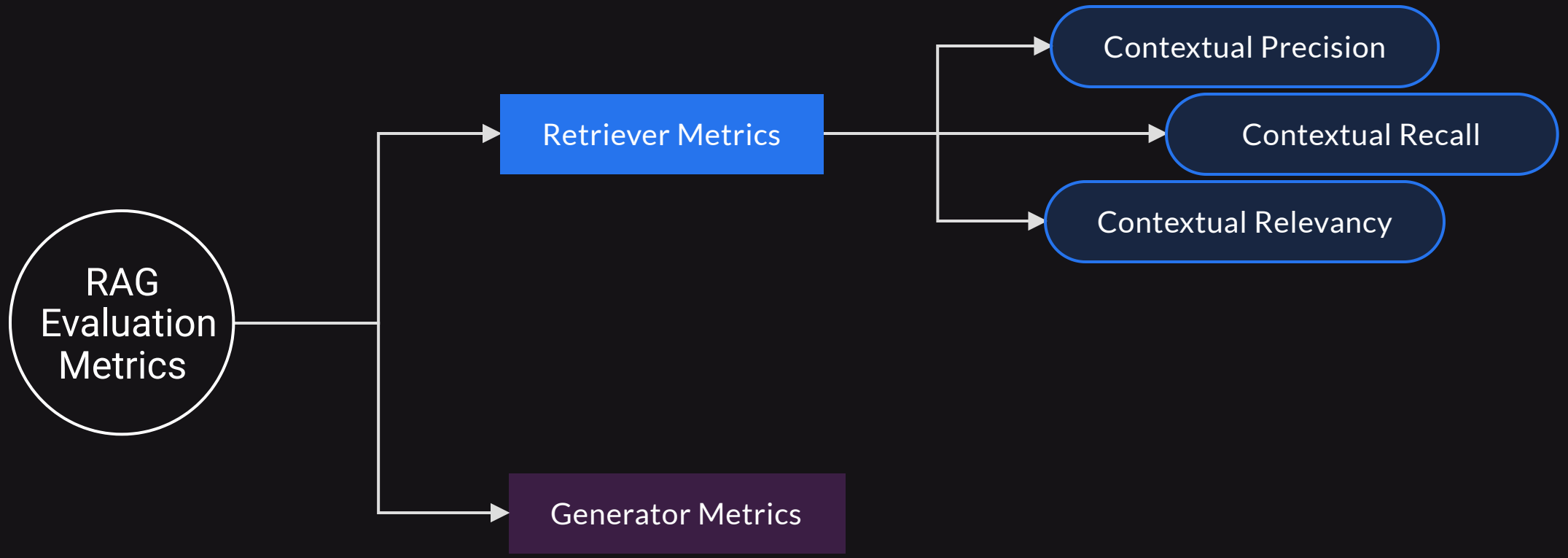
Context Recall

- Measures the extent of which of the **retrieved context** document chunks (nodes) aligns with the **expected response answer** (ground truth reference).
- Higher Context Recall score represents a better retrieval system that can capture all relevant context information from your Vector DB

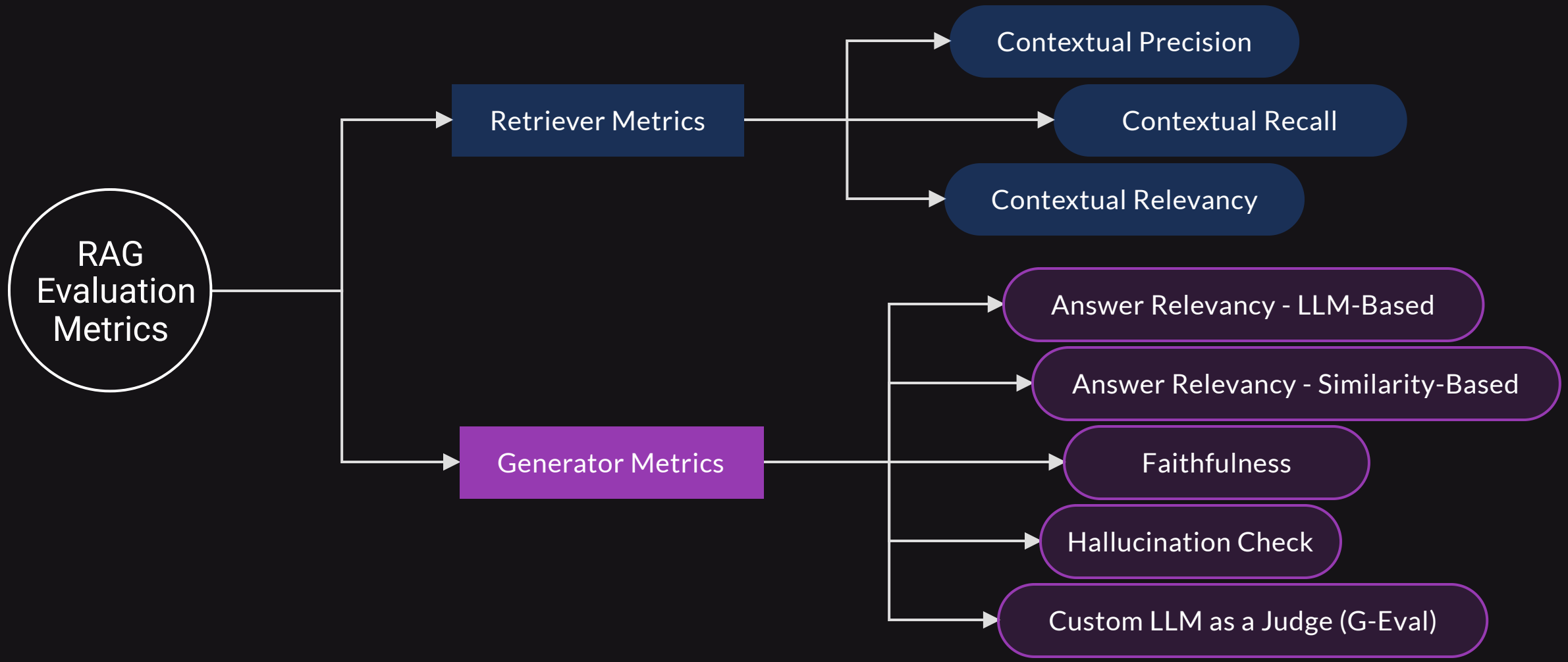
Context Relevancy

- Measures the relevancy of the information in the **retrieved context** document chunks (nodes) to the given **input query**
- Higher Context Relevance score represents a better retrieval system which can retrieve more semantically relevant nodes for queries

Classification: RAG Evaluation Metrics



Classification: RAG Evaluation Metrics



Answer Relevancy - LLM-Based

- Measures the relevancy of the information in the **generated response** to the provided **input query** using LLM as a Judge
- Higher Answer Relevance shows the LLM Generator is able to generate better quality relevant responses for queries

Answer Relevancy - Similarity-Based

- Measures the relevancy of the information in the **generated response** to the provided **input query** using semantic similarity between LLM generated queries from the response and the input query
- Higher Answer Relevance shows the LLM Generator is able to generate better quality relevant responses for queries

Faithfulness

- Measures if the information in the **generated response** factually aligns with the contents of the **retrieved context** document chunks (nodes)
- Higher Faithfulness means the generated response is more grounded with regard to the retrieved context reducing contradictions

Hallucination Check

- Measures the proportion of contradictory statements by comparing the **generated response** to the **expected context** document chunks (ground truth reference).
- Lower the hallucination score, lower the proportion of contradictory statements making the response more grounded and relevant

Custom LLM as a Judge (G-Eval)

- G-Eval is a framework that uses LLMs with chain-of-thoughts (CoT) to evaluate LLM responses and retrieved contexts based on **ANY** custom criteria.
- You can decide the judging criteria and give detailed evaluation steps using prompt instructions

Thank You

Analytics Vidhya