Hands-On: Deep Dive into RAG Evaluation Metrics - Retriever Metrics

Instructor

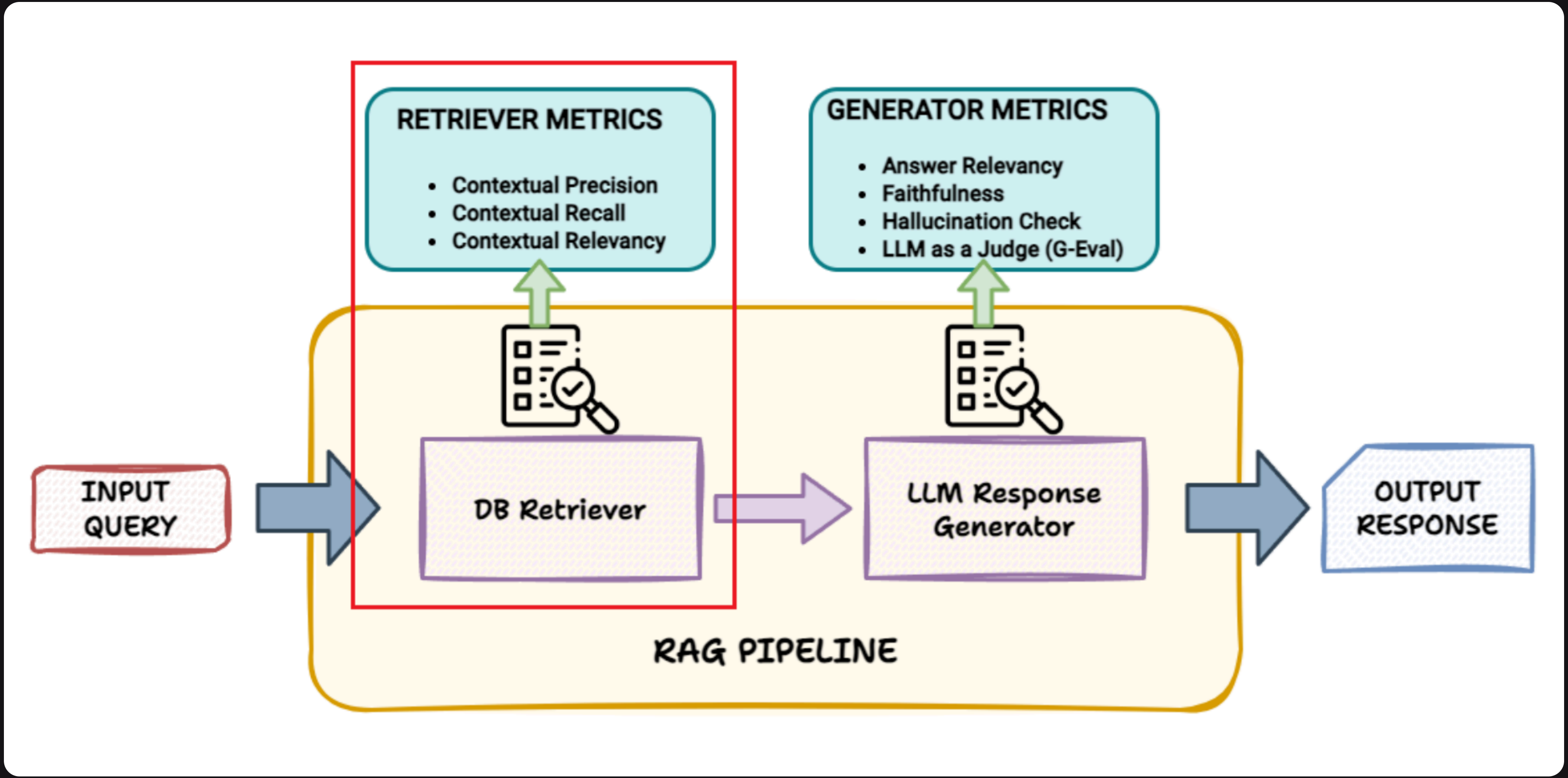
Dipanjan Sarkar

Head of Community & Principal AI Scientist at Analytics Vidhya

Google Developer Expert - ML & Cloud Champion Innovator

Published Author

Retriever Evaluation Metrics



Retriever Evaluation Metrics

- Contextual Precision
- Contextual Recall
- Contextual Relevancy

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 Precision

$$\text{Contextual Precision} = \frac{1}{\text{Number of Relevant Nodes}} \sum_{k=1}^{n} \left(\frac{\text{Number of Relevant Nodes Up to Position } k}{k} \times r_k \right)$$

Explanation of Key Variables:

- **Number of Relevant Nodes:** Total count of nodes that are considered relevant in the retrieval context.
- **k:** The position of the node in the retrieval context (starting from 1).
- **n:** Total length of the retrieval context.
- **Number of Relevant Nodes Up to Position k:** Cumulative count of relevant nodes up to and including position k.
- **r_k:** Binary relevance for the k^{th} node, where:
  - r_k = 1 if the node is relevant
  - r_k = 0 if the node is not relevant

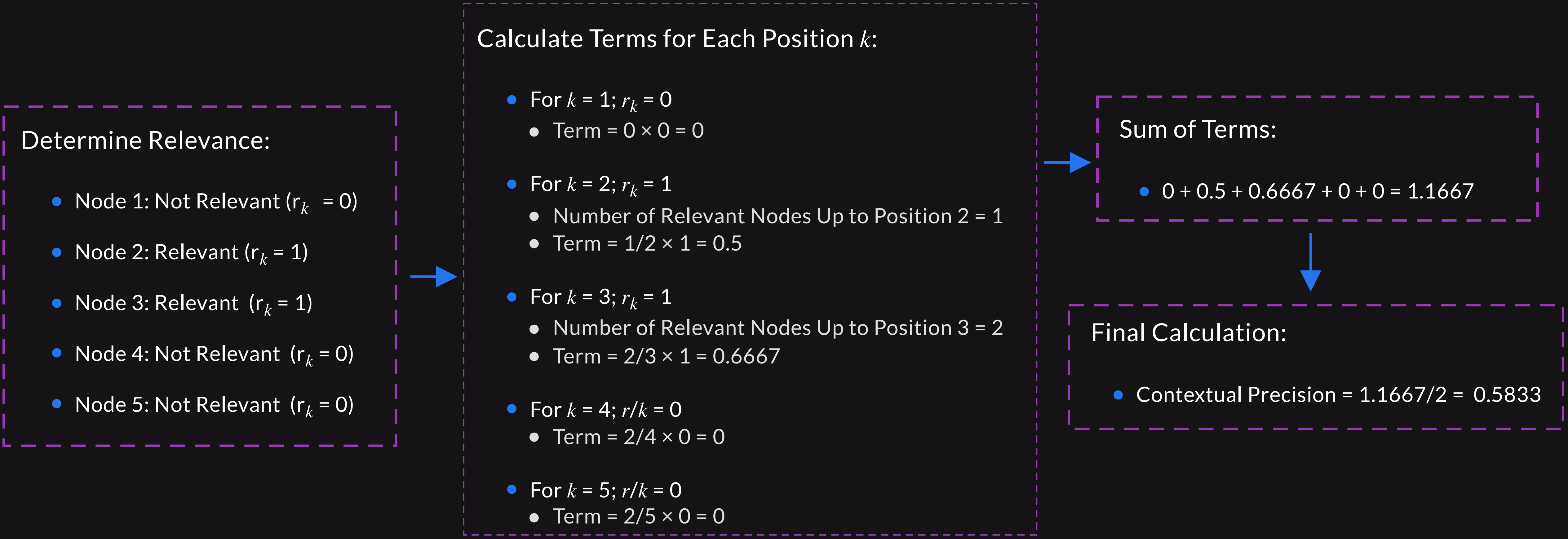
Context Precision

- ## Input Variables
  - "What is AI?"

- ## Retrieved Context:
  - **Node 1:** "Machine Learning is the study of algorithms which learn with more data."
  - **Node 2:** "AI is known as Artificial Intelligence."
  - **Node 3:** "Artificial intelligence refers to machines mimicking human intelligence, like problem-solving and learning. AI includes applications like virtual assistants, robotics, and autonomous vehicles. It's evolving rapidly with advancements in machine learning and deep learning."
  - **Node 4:** "NLP is a branch of AI that enables computers to understand, interpret, and generate human language. Techniques include tokenization, stemming, and sentiment analysis. Applications range from chatbots to language translation services."
  - **Node 5:** "Machine learning is a field of artificial intelligence focused on enabling systems to learn patterns from data. Algorithms analyze past data to make predictions or classify information. Popular applications include recommendation systems and image recognition."

Context Precision

$$\text{Contextual Precision} = \frac{1}{\text{Number of Relevant Nodes}} \sum_{k=1}^{n} \left(\frac{\text{Number of Relevant Nodes Up to Position } k}{k} \times r_k \right)$$



Determine Relevance:

- Node 1: Not Relevant (r_k = 0)
- Node 2: Relevant (r_k = 1)
- Node 3: Relevant (r_k = 1)
- Node 4: Not Relevant (r_k = 0)
- Node 5: Not Relevant (r_k = 0)

Calculate Terms for Each Position k:

- For k = 1; r_k = 0
  - Term = 0 × 0 = 0
- For k = 2; r_k = 1
  - Number of Relevant Nodes Up to Position 2 = 1
  - Term = 1/2 × 1 = 0.5
- For k = 3; r_k = 1
  - Number of Relevant Nodes Up to Position 3 = 2
  - Term = 2/3 × 1 = 0.6667
- For k = 4; r/k = 0
  - Term = 2/4 × 0 = 0
- For k = 5; r/k = 0
  - Term = 2/5 × 0 = 0

Sum of Terms:

- 0 + 0.5 + 0.6667 + 0 + 0 = 1.1667

Final Calculation:

- Contextual Precision = 1.1667/2 = 0.5833

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 which can capture all relevant context information from your Vector DB

Context Recall

$$\text{Contextual Recall} = \frac{\text{Number of Attributable Statements}}{\text{Total Number of Statements}}$$

Explanation of Key Variables:

- **Number of Attributable Statements:**
  - The count of statements in the expected output that can be attributed to nodes in the retrieved context.
- **Total Number of Statements:**
  - The total count of statements in the expected output.

How it Works:

- **Attributable Statements:**
  - Using an LLM as a judge, it identifies statements in the expected output that can be supported by information within the retrieved context.
- **Expected Output:**
  - The ideal or ground truth answer for the input query. This is used instead of the actual output to assess the retrieval system's performance quality.

Context Recall

Input Variables

1. Input Query:
   - "What is AI?"
2. Expected Output:
   - "AI, also known as Artificial Intelligence, is used to build complex systems for applications like virtual assistants, robotics, and autonomous vehicles."
3. Retrieved Context:
   - Node 1: "NVIDIA makes chips for AI."
   - Node 2: "AI is an acronym for Artificial Intelligence."

Context Recall

$$\text{Contextual Recall} = \frac{\text{Number of Attributable Statements}}{\text{Total Number of Statements}}$$

Identify Statements in the Expected Output:

- **Expected Output:** "AI, also known as Artificial Intelligence, is used to build complex systems for applications like virtual assistants, robotics, and autonomous vehicles."
- This output contains 2 statements:
  - **Statement 1:** "AI, also known as Artificial Intelligence..."
  - **Statement 2:** "...is used to build complex systems for applications like virtual assistants, robotics, and autonomous vehicles."

Assess Attributability for Each Statement:

- **Statement 1:** "AI, also known as Artificial Intelligence..."
  - This can be attributed to **Node 2** in the retrieved context: "AI is an acronym for Artificial Intelligence."
  - **Verdict:** Yes (Attributable)
- **Statement 2:** "...is used to build complex systems for applications like virtual assistants, robotics, and autonomous vehicles."
  - No nodes in the retrieved context support this statement.
  - **Verdict:** No (Not Attributable)

Calculate Contextual Recall:

- **Number of Attributable Statements:** 1 (Statement 1)
- **Total Number of Statements:** 2
- **Contextual Recall:** 1/2 = 0.5

Analytics Vidhya

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

Context Relevancy

$$\text{Contextual Relevancy} = \frac{\text{Number of Relevant Statements}}{\text{Total Number of Statements}}$$

Explanation of Key Variables:

- **Number of Relevant Statements:**
  - The count of statements in the retrieval context that are relevant to the input query.
- **Total Number of Statements:**
  - The total count of statements in the retrieval context.

How it Works:

- **Relevant Statements**:
  - The system, using an LLM as a judge, identifies statements in the retrieval context that are directly relevant to answering the input query.
- **Input Query**:
  - The specific question or topic the retrieval context aims to address.

A higher **Contextual Relevancy** score indicates that a greater portion of the information in the retrieved context is directly relevant to the input query.

Context Relevancy

Input Variables

1. ### Input Query:

- "What is AI?"

2. ### Retrieved Context:

- **Node 1:** "NVIDIA makes chips for AI."
- **Node 2:** "Google and Microsoft are battling out the market share for AI Chatbots."
- **Node 3:** "Artificial intelligence refers to machines mimicking human intelligence, like problem-solving and learning. AI includes applications like virtual assistants, robotics, and autonomous vehicles. It's evolving rapidly with advancements in machine learning and deep learning."
- **Node 4:** "NLP is a branch of AI that enables computers to understand, interpret, and generate human language. Techniques include tokenization, stemming, and sentiment analysis. Applications range from chatbots to language translation services."
- **Node 5:** "Machine learning is a field of artificial intelligence focused on enabling systems to learn patterns from data. Algorithms analyze past data to make predictions or classify information. Popular applications include recommendation systems and image recognition."

Context Relevancy

1) Identify Statements and Their Relevance in the Retrieved Context

- Node 1:
  - Statements: ["NVIDIA makes chips for AI."]
  - Verdicts: [Not Relevant]
- Node 2:
  - Statements: ["Google and Microsoft are battling out the market share for AI Chatbots."]
  - Verdicts: [Not Relevant]
- Node 3:
  - Statements: ["Artificial intelligence refers to machines mimicking human intelligence, like problem solving and learning.", "AI includes applications like virtual assistants, robotics, and autonomous vehicles.", "It's evolving rapidly with advancements in machine learning and deep learning."]
  - Verdicts: [Relevant, Relevant, Relevant]
- Node 4:
  - Statements: ["NLP is a branch of AI that enables computers to understand, interpret, and generate human language.", "Techniques include tokenization, stemming, and sentiment analysis.", "Applications range from chatbots to language translation services."]
  - Verdicts: [Relevant, Relevant, Relevant]
- Node 5:
  - Statements: ["Machine learning is a field of artificial intelligence focused on enabling systems to learn patterns from data.", "Algorithms analyze past data to make predictions or classify information.", "Popular applications include recommendation systems and image recognition."]
  - Verdicts: [Relevant, Relevant, Relevant]

Context Relevancy

2) Count Relevant Statements

- Number of Relevant Statements:
  - 9 (all statements in Nodes 3, 4, and 5)
- Total Number of Statements:
  - 11 (including irrelevant statements in Nodes 1 and 2)

Context Relevancy

3) Calculate Contextual Relevancy

$$\text{Contextual Relevancy} = \frac{\text{Number of Relevant Statements}}{\text{Total Number of Statements}}$$

Contextual Relevancy = 9/11 =0.8182

Thank You

Analytics Vidhya