**Retriever Evaluation Metrics**

1. **Context Precision**:
   * **Definition**: Measures how many of the retrieved contexts (documents or passages) are relevant compared to the total retrieved.
   * **Formula**: Context Precision=Relevant Retrieved ContextsTotal Retrieved Contexts\text{Context Precision} = \frac{\text{Relevant Retrieved Contexts}}{\text{Total Retrieved Contexts}}Context Precision=Total Retrieved ContextsRelevant Retrieved Contexts​
   * **Why it's important**: Ensures that the retriever fetches high-quality, relevant information.
2. **Context Recall**:
   * **Definition**: Measures how many of the relevant contexts in the entire knowledge base were retrieved.
   * **Formula**: Context Recall=Relevant Retrieved ContextsTotal Relevant Contexts in the Knowledge Base\text{Context Recall} = \frac{\text{Relevant Retrieved Contexts}}{\text{Total Relevant Contexts in the Knowledge Base}}Context Recall=Total Relevant Contexts in the Knowledge BaseRelevant Retrieved Contexts​
   * **Why it's important**: Ensures the retriever doesn't miss important information that could improve the generation.

**Generator Evaluation Metrics**

1. **Faithfulness**:
   * **Definition**: Measures whether the generated response accurately reflects the retrieved context without introducing hallucinations or incorrect information.
   * **Why it's important**: Ensures the output is grounded in the retrieved knowledge, crucial for trust in real-world applications.
   * **Evaluation Methods**:
     + Manual evaluation by comparing the output to the retrieved context.
     + Automatic tools like **FactCC** or **BERTScore** can help but may not fully capture nuanced errors.
2. **Answer Relevancy**:
   * **Definition**: Measures how relevant the generated response is to the input query.
   * **Why it's important**: Ensures the response addresses the user's question or need.
   * **Evaluation Methods**:
     + Semantic similarity metrics like **ROUGE**, **BLEU**, or **METEOR**.
     + Human evaluation for subjective tasks.
3. **AspectCritique**:
   * **Definition**: Assesses specific dimensions of the response, such as coherence, fluency, informativeness, or empathy (depending on the use case).
   * **Why it's important**: Breaks down the evaluation into granular aspects to identify strengths and weaknesses.
   * **Evaluation Methods**:
     + Human evaluation or rule-based scoring (for specific aspects).
     + Embedding-based similarity methods like **BERTScore** for coherence.

**Applicability to RAG**

These metrics align well with RAG systems because:

1. **Retriever Metrics** ensure the foundation (retrieval stage) provides the necessary high-quality context.
2. **Generator Metrics** ensure the system effectively uses the retrieved information to produce useful and trustworthy outputs.