# Evaluation Metric Proposal for RAG's Generation

Retrieval-augmented generation (RAG) is a powerful technique that enhances large language models (LLMs) by incorporating external knowledge sources. This approach allows LLMs to generate more informative and factually grounded responses, especially in specialized domains or when dealing with private data 1. Evaluating the performance of RAG systems is crucial to ensure their effectiveness and reliability. This article presents a comprehensive evaluation metric proposal for RAG's generation, encompassing both stand-alone metrics and RAG-based generation metrics.

## Stand-alone Evaluation Metrics for Text Generation

Stand-alone metrics assess the quality of generated text without explicitly considering the retrieval component of RAG. These metrics are widely used in various natural language generation (NLG) tasks, including machine translation, summarization, and dialogue generation. Here are some prominent stand-alone metrics:

* **BLEU (Bilingual Evaluation Understudy):** This precision-focused metric calculates the n-gram overlap between the generated text and a reference text 1. It is widely used in machine translation but also applicable to other NLG tasks. BLEU scores range from 0 to 1, with higher scores indicating greater similarity between the generated and reference texts. While BLEU is a popular metric, it has limitations in capturing semantic similarity and is sensitive to sentence length. For example, two sentences with different word order but similar meaning might receive a low BLEU score.
* **ROUGE (Recall-Oriented Understudy for Gisting Evaluation):** This recall-focused metric measures the overlap of n-grams, word sequences, or word pairs between the generated text and a set of human-generated reference texts 1. It is commonly used for evaluating summarization and machine translation systems. There are different types of ROUGE, including:
  + **ROUGE-N:** Measures the overlap of n-grams (e.g., ROUGE-1 for unigrams, ROUGE-2 for bigrams) 4.
  + **ROUGE-L:** Calculates the longest common subsequence (LCS) between the generated and reference texts 4.
  + **ROUGE-S:** Focuses on skip-bigrams, which are any pair of words in their sentence order 4.
* **METEOR (Metric for Evaluation of Translation with Explicit ORdering):** This metric combines precision and recall by aligning unigrams between the generated and expected text based on exact match, stemming, synonyms, and paraphrases 1. It provides a more balanced evaluation of model performance than BLEU or ROUGE alone. METEOR also considers word order and semantic similarity, making it more robust to variations in phrasing.
* **BERTScore:** This metric leverages contextual embeddings from BERT to compute a similarity score for each token in the candidate sentence with each token in the reference sentence 7. It has shown strong correlation with human judgments in machine translation and image captioning tasks. Unlike BLEU and ROUGE, which rely on exact matches, BERTScore can capture semantic similarity between words with different surface forms.
* **Perplexity:** This metric measures how well a probability distribution predicts a sample 2. In the context of language models, it reflects the model's ability to predict the next word in a sequence. Lower perplexity values indicate better performance, as they suggest that the model assigns higher probabilities to the observed sequence of words.
* **Mover Distance:** This metric evaluates semantic similarity by considering the "movement" required to transform one sentence into another in a semantic space 9. It uses text embeddings to represent sentences as points in a high-dimensional space and calculates the optimal "distance" to move the points of one sentence to align with the points of another sentence. This approach provides a more nuanced measure of semantic similarity than simple word overlap.
* **Reference-free Metrics:** These metrics evaluate the quality of generated text without relying on a reference text 5. They are particularly useful when ground truth data is scarce or difficult to obtain. Some examples include:
  + **Entailment-based metrics:** These metrics assess whether the generated text logically follows from a given premise or context. They can be used to detect factual inconsistencies or contradictions.
  + **Factuality metrics:** These metrics evaluate the factual accuracy of the generated text by checking it against external knowledge sources or fact-checking databases.

### Analysis of Stand-alone Metrics

The choice of stand-alone metric depends on the specific NLG task and the desired evaluation criteria. BLEU and ROUGE are widely used due to their simplicity and ease of calculation, but they have limitations in capturing semantic similarity and are sensitive to sentence length. METEOR offers a more balanced evaluation by considering both precision and recall, as well as word order and semantic similarity. BERTScore provides a more sophisticated measure of semantic similarity by leveraging contextual embeddings. Perplexity is useful for evaluating language models and their ability to predict the next word in a sequence. Mover distance offers another perspective on semantic similarity by considering the "movement" required to transform one sentence into another in a semantic space. Reference-free metrics are valuable when ground truth data is limited or unavailable.

## RAG-based Generation Metrics

RAG-based generation metrics specifically evaluate the quality of generated text in the context of the retrieved information. These metrics consider the interplay between the retrieval and generation components of RAG. Here are some essential RAG-based generation metrics:

* **Faithfulness:** This metric measures the factual consistency of the generated answer against the given context 5. It penalizes claims made in the answer that cannot be deduced from the retrieved context. To calculate faithfulness, one can extract claims from the generated answer and verify them against the retrieved context using techniques like natural language inference or fact verification 6.
* **Answer Relevancy:** This metric assesses the degree to which a response directly addresses and is appropriate for a given question or context 5. It penalizes the presence of redundant information or incomplete answers. Answer relevancy can be measured by comparing the generated answer to the user's query or by assessing whether the answer fulfills the user's information needs.
* **Context Relevance:** This metric measures how relevant the retrieved contexts are to the question 5. It penalizes the presence of redundant or irrelevant information in the retrieved context. Context relevance can be assessed by comparing the retrieved context to the user's query or by evaluating the semantic similarity between the context and the question.
* **Context Recall:** This metric measures the recall of the retrieved context using the annotated answer as ground truth 5. It assesses how well the retrieved context covers the information needed to answer the question. Context recall can be calculated by comparing the retrieved context to a set of relevant documents or by evaluating the proportion of information from the ground truth answer that is present in the retrieved context.
* **Context Precision:** This metric measures the precision of the retrieved context by considering the proportion of relevant chunks among the retrieved ones 13. It focuses on the accuracy of the retrieval process. Context precision can be calculated by manually assessing the relevance of each retrieved chunk or by using an LLM to judge the relevance of the chunks.
* **PII Filter:** This filter ensures that the generated response does not include any personally identifiable information (PII), such as names, addresses, or phone numbers 1. PII filters are crucial for protecting user privacy and complying with data protection regulations.
* **HAP Filter:** This filter monitors the generated response for hate speech, abuse, and profanity (HAP) 1. HAP filters are essential for ensuring responsible use of RAG and preventing the generation of harmful or offensive content.

### Analysis of RAG-based Generation Metrics

RAG-based generation metrics provide a more comprehensive evaluation of RAG systems by considering the interplay between retrieval and generation. Faithfulness ensures that the generated response is grounded in the retrieved context and avoids hallucinations. Answer relevancy assesses the relevance of the response to the user's query. Context relevance measures the quality of the retrieved context itself. Context recall and context precision evaluate the effectiveness of the retrieval process in terms of coverage and accuracy. PII and HAP filters are crucial for ensuring responsible use of RAG in production systems.

## Evaluating Multi-turn Conversations

RAG is increasingly used in conversational AI applications, where the system needs to maintain context and coherence across multiple turns. Evaluating the performance of RAG in multi-turn conversations requires specific metrics that capture these aspects. One such metric is:

* **R2R (Response to Response):** This metric evaluates the relevance and coherence of a response in relation to the previous turn in a conversation 19. It assesses whether the response appropriately addresses the previous turn and maintains the flow of the conversation.

## Industry Applications and Evaluation Metrics

Different industries are leveraging RAG for various applications, and they employ specific evaluation metrics based on their needs and priorities. Here are some examples:

* **Question Answering:** In question answering systems, metrics like faithfulness, answer relevancy, and context relevance are crucial to ensure that the system provides accurate and relevant answers to user queries 11. For example, a legal question answering system might prioritize faithfulness to ensure that the answers are legally sound and supported by relevant legal documents.
* **Customer Support:** For customer support applications, metrics like answer correctness, response time, and customer satisfaction are essential to evaluate the effectiveness and efficiency of the RAG system 20. For instance, a customer support chatbot for a telecommunications company might prioritize answer correctness and response time to quickly and accurately resolve customer issues.
* **Content Generation:** In content generation tasks, metrics like fluency, coherence, and diversity are important to assess the quality and creativity of the generated text 12. For example, a marketing content generation system might prioritize fluency and coherence to create engaging and persuasive marketing materials.
* **Financial Industry:** In the financial industry, RAG can be used for tasks like financial analysis, risk assessment, and fraud detection. Metrics like factual accuracy, risk mitigation, and compliance with regulations are crucial in this domain 20.
* **Healthcare Industry:** In the healthcare industry, RAG can be applied to tasks like medical diagnosis, treatment recommendation, and patient education. Metrics like patient privacy, safety, and accuracy of medical information are paramount in this domain 20.

## Challenges of Evaluating RAG Systems

Evaluating RAG systems presents unique challenges due to the complex interplay between retrieval and generation. Some key challenges include:

* **Need for High-Quality Ground Truth Data:** Many evaluation metrics, such as context recall and faithfulness, rely on ground truth data for comparison. Creating high-quality ground truth data can be time-consuming and expensive, especially for specialized domains or complex tasks 14.
* **Computational Cost of Some Metrics:** Certain metrics, such as those involving LLM judges or complex semantic similarity calculations, can be computationally expensive 15. This can make it challenging to evaluate RAG systems at scale or in real-time applications.
* **Difficulty of Evaluating Subjective Aspects:** Some aspects of RAG generation, such as creativity, humor, or persuasiveness, can be subjective and difficult to evaluate with automated metrics. Human evaluation might be necessary in these cases, but it can be time-consuming and prone to bias.
* **Domain Specificity:** Evaluating RAG systems in specialized domains often requires domain-specific knowledge and expertise. Generic metrics might not be sufficient to capture the nuances of the domain or the specific requirements of the task.

## Research Papers on Evaluation Metrics for RAG

Several research papers have explored evaluation metrics for RAG systems. These papers provide valuable insights into the challenges and considerations for evaluating RAG's performance. Some notable papers include:

* "Evaluation of RAG Metrics for Question Answering in the Telecom Domain" 21 This paper analyzes the challenges of evaluating RAG-based question answering in a specialized domain and proposes modifications to existing evaluation frameworks.
* "Evaluating RAG Applications with RAGAs" 13 This paper introduces RAGAs, a framework for evaluating RAG applications with a focus on component-level metrics and LLM-generated data.
* "Retrieval-Augmented Generation Assessment" 13 This paper presents a framework for quantifying RAG performance without human annotation, using synthetic data and LLM judges.

## Open-Source Libraries or Tools for Evaluating RAG

Several open-source libraries and tools provide implementations of evaluation metrics for RAG. These resources facilitate the evaluation process and offer standardized ways to assess RAG systems. Some popular options include:

* **Ragas:** This framework offers a comprehensive set of evaluation metrics for RAG, including faithfulness, answer relevancy, context recall, and context precision. It supports both Langchain and Llama-Index16.
* **DeepEval:** This framework provides a wide range of evaluation metrics for both RAG and fine-tuning use cases. It offers over 14 evaluation metrics, including G-Eval, summarization, hallucination, faithfulness, and contextual relevancy6.
* **Phoenix:** This tool, developed by Arize AI, focuses on AI observability and evaluation. It provides metrics like faithfulness, relevance, and semantic similarity to assess retrieval and answer quality18.
* **Evidently AI:** This open-source Python library includes tools for evaluating RAG systems, such as scoring context relevance, running ranking metrics, and evaluating generation quality with or without ground truth25.

## Proposed Set of Evaluation Metrics for RAG's Generation

Based on the available resources, the following set of evaluation metrics is proposed for RAG's generation:

| **Metric Category** | **Metric Name** | **Description** |
| --- | --- | --- |
| Stand-alone Metrics | BLEU | Measures n-gram overlap between generated and reference texts. |
|  | ROUGE | Measures the overlap of n-grams, word sequences, or word pairs between generated and reference texts. |
|  | METEOR | Combines precision and recall, considering word order and semantic similarity. |
|  | BERTScore | Leverages contextual embeddings to compute semantic similarity. |
|  | Perplexity | Measures how well a language model predicts the next word in a sequence. |
| RAG-based Metrics | Faithfulness | Measures the factual consistency of the generated answer against the retrieved context. |
|  | Answer Relevancy | Assesses the relevance of the response to the user's query. |
|  | Context Relevance | Measures the relevance of the retrieved context to the question. |
|  | Context Recall | Evaluates the coverage of relevant information in the retrieved context. |
|  | Context Precision | Measures the accuracy of the retrieval process. |
|  | PII Filter | Ensures that the generated response does not include PII. |
|  | HAP Filter | Monitors the generated response for hate speech, abuse, and profanity. |
| Multi-turn Conversation Metrics | R2R | Evaluates the relevance and coherence of a response in relation to the previous turn in a conversation. |

This combination of metrics provides a comprehensive evaluation of RAG's generation capabilities, considering both the quality of the generated text and its alignment with the retrieved context. It also includes filters to ensure responsible use and addresses the specific challenges of evaluating multi-turn conversations.

## Conclusion

Evaluating RAG systems requires a multifaceted approach that considers various aspects of performance. By incorporating both stand-alone and RAG-based generation metrics, we can gain a holistic understanding of the strengths and weaknesses of RAG systems. The proposed set of evaluation metrics provides a solid foundation for assessing RAG's generation capabilities and guiding further improvements. Future advancements in RAG evaluation might involve developing new metrics that capture more nuanced aspects of performance, such as creativity or user satisfaction, and creating more sophisticated evaluation frameworks that can handle the increasing complexity of RAG systems.

#### Works cited

1. What Are the Evaluation Metrics for RAGs? | Baeldung on Computer Science, accessed March 12, 2025, <https://www.baeldung.com/cs/retrieval-augmented-generation-evaluate-metrics-performance>

2. How to evaluate Text Generation Models ? Metrics for Automatic Evaluation of NLP Models, accessed March 12, 2025, <https://towardsdatascience.com/how-to-evaluate-text-generation-models-metrics-for-automatic-evaluation-of-nlp-models-e1c251b04ec1/>

3. Evaluation metrics for text generation | PyTorch - DataCamp, accessed March 12, 2025, <https://campus.datacamp.com/courses/deep-learning-for-text-with-pytorch/text-generation-with-pytorch?ex=10>

4. Automated metrics for evaluating the quality of text generation | DigitalOcean, accessed March 12, 2025, <https://www.digitalocean.com/community/tutorials/automated-metrics-for-evaluating-generated-text>

5. A list of metrics for evaluating LLM-generated content - Microsoft Learn, accessed March 12, 2025, <https://learn.microsoft.com/en-us/ai/playbook/technology-guidance/generative-ai/working-with-llms/evaluation/list-of-eval-metrics>

6. RAG Evaluation Metrics Explained: A Complete Guide | by Mohamed EL HARCHAOUI, accessed March 12, 2025, <https://medium.com/@med.el.harchaoui/rag-evaluation-metrics-explained-a-complete-guide-dbd7a3b571a8>

7. RAG evaluation metrics: A journey through metrics - Elasticsearch Labs, accessed March 12, 2025, <https://www.elastic.co/search-labs/blog/evaluating-rag-metrics>

8. A Practical Framework for Evaluating Text Generation LLMs | by Daniel Bukowski | Medium, accessed March 12, 2025, <https://medium.com/@bukowski.daniel/a-practical-framework-for-evaluating-text-generation-llms-4016ffa93736>

9. How do we evaluate the outputs of text generation models? - Data Science Stack Exchange, accessed March 12, 2025, <https://datascience.stackexchange.com/questions/118739/how-do-we-evaluate-the-outputs-of-text-generation-models>

10. Best Practices in RAG Evaluation: A Comprehensive Guide - Qdrant, accessed March 12, 2025, <https://qdrant.tech/blog/rag-evaluation-guide/>

11. Evaluation Metrics for Retrieval-Augmented Generation (RAG) Systems - GeeksforGeeks, accessed March 12, 2025, <https://www.geeksforgeeks.org/evaluation-metrics-for-retrieval-augmented-generation-rag-systems/>

12. RAG Evaluation Metrics Starter Kit - Arize AI, accessed March 12, 2025, <https://arize.com/blog-course/rag-evaluation/>

13. Evaluating RAG Applications with RAGAs - Towards Data Science, accessed March 12, 2025, <https://towardsdatascience.com/evaluating-rag-applications-with-ragas-81d67b0ee31a/>

14. Evaluating the evaluators: know your RAG metrics - Tweag, accessed March 12, 2025, <https://tweag.io/blog/2025-02-27-rag-evaluation/>

15. Evaluate the reliability of Retrieval Augmented Generation applications using Amazon Bedrock | AWS Machine Learning Blog, accessed March 12, 2025, <https://aws.amazon.com/blogs/machine-learning/evaluate-the-reliability-of-retrieval-augmented-generation-applications-using-amazon-bedrock/>

16. RAG Evaluation Survey: Framework, Metrics, and Methods | EvalScope - Read the Docs, accessed March 12, 2025, <https://evalscope.readthedocs.io/en/latest/blog/RAG/RAG_Evaluation.html>

17. Mastering RAG Evaluation: Best Practices & Tools for 2025 - Orq.ai, accessed March 12, 2025, <https://orq.ai/blog/rag-evaluation>

18. Top 10 Open Source RAG Evaluation Frameworks You Must Try | by Sebastian Petrus, accessed March 12, 2025, <https://sebastian-petrus.medium.com/top-10-open-source-rag-evaluation-frameworks-you-should-try-4fb7cee9d18a>

19. List of available metrics - Ragas, accessed March 12, 2025, <https://docs.ragas.io/en/stable/concepts/metrics/available_metrics/>

20. Top Metrics to Monitor and Improve RAG Performance - Galileo AI, accessed March 12, 2025, <https://www.galileo.ai/blog/top-metrics-to-monitor-and-improve-rag-performance>

21. Evaluation of RAG Metrics for Question Answering in the Telecom Domain - arXiv, accessed March 12, 2025, <https://arxiv.org/abs/2407.12873>

22. Ragas, accessed March 12, 2025, <https://docs.ragas.io/>

23. ‼️ Top 5 Open-Source LLM Evaluation Frameworks in 2025 - DEV Community, accessed March 12, 2025, <https://dev.to/guybuildingai/-top-5-open-source-llm-evaluation-frameworks-in-2024-98m>

24. RAG Evaluation | DeepEval - The Open-Source LLM Evaluation Framework, accessed March 12, 2025, <https://docs.confident-ai.com/guides/guides-rag-evaluation>

25. Open-source RAG evaluation and testing - Evidently 0.6.3, accessed March 12, 2025, <https://www.evidentlyai.com/blog/open-source-rag-evaluation-tool>