OPEN ACCESS JOURNALS
Indiana Journal of Multidisciplinary Research
Abbreviated Tittle- Ind J Multi Res
ISSN *(Online)*- 2583-3820
Journal Homepage Link- https://indianapublications.com/Journals/IJMR
DOI: https://doi.org/10.5281/zenodo.21607583

INDIANA PUBLICATIONS
PRODUCTIVE AND QUALITY RESEARCH

Research Article

Volume-06|Issue-04|2026

Evaluation Metrics for Retrieval-Augmented Generation (RAG) Systems: A Framework Proposal

Ullas C L[1], Punitha M[2], Reshma B[3], Shashaank K[4], Shashank V Wali[5], Tanush A K[6]

[1,4,5,6]Ullas C L, Student, Department of Information Science and Engineering, JSS Academy of Technical Education, Bengaluru, Karnataka, India.
[2,3]Punitha M, Assistant Professor, Department of Information Science and Engineering, JSS Academy of Technical Education, Bengaluru, Karnataka, India.

Article History
Received: 05.05.2026
Accepted: 22.06.2026
Published: 25.07.2026

Citation
Ullas, C. L., Punitha, M., Reshma, B., Shashaank, K., Wali, S. V. & Tanush, A. K. (2026) Evaluation Metrics for Retrieval-Augmented Generation (RAG) Systems: A Framework Proposal. *Indiana Journal of Multidisciplinary Research, 6*(4), 271-277.

Abstract: Retrieval-Augmented Generation (RAG) has established itself as a primary method for connecting large language models (LLMs) to external knowledge sources while addressing hallucination problems and enhancing factual accuracy [1,4,6]. Current evaluation approaches remain fragmented, as they often assess retrieval relevance [9,12] and surface-level text quality [2,5] independently, without linking these perspectives across the complete pipeline.
This work introduces a multi-level evaluation framework that systematically decomposes assessment into retrieval, generation, and integration stages. The retrieval layer is assessed using ranking and coverage metrics, including Precision@k, Recall@k, Mean Reciprocal Rank (MRR), and NDCG [3,10]. The generation layer incorporates lexical and semantic measures such as BLEU [14], ROUGE-L [15], BERTScore [16], and claim-based faithfulness [7,11]. At the integration layer, this paper proposes the Hallucination Attribution Index (HAI), which quantifies the proportion of generated claims grounded in retrieved evidence.
Additionally, the integration of consistency testing and human-in-the-loop evaluations [8,13] strengthens system robustness in adversarial and domain-sensitive scenarios. By combining these elements, the proposed framework provides a comprehensive evaluation system for RAG that enables precise assessment and improves deployment trustworthiness across domains such as healthcare, law, and finance.

Keywords: Retrieval-Augmented Generation; Hallucination Attribution Index; Evaluation Metrics; Large Language Models; Faithfulness; Information Retrieval

INTRODUCTION

Large Language Models (LLMs) such as GPT-3 [8], BERT [7], and T5 [9] have revolutionized natural language processing (NLP), achieving state-of-the-art results across machine translation, summarization, and conversational systems. Their ability to learn general-purpose linguistic and semantic patterns from massive corpora has accelerated their adoption in real-world applications, including customer support, search engines, and domain-specific assistants in healthcare, law, and finance.

LLMs demonstrate high accuracy in executing instructions, yet they generate false information as hallucinations. These models continue to struggle with hallucinations that produce fluent yet factually incorrect information and unsupported content [11]. Such behavior from LLMs creates important risks to domains which need verification and trust because it leads to both misinformation and safety-critical errors. The performance of LLMs suffers from three main deficiencies, which include temporal grounding problems with outdated knowledge and domain adaptation challenges because of insufficient specialized terminology exposure and justification failures regarding evidence citation.

Retrieval-Augmented Generation (RAG) has emerged as a powerful paradigm to address these **issues** [10]. RAG uses a retriever module to obtain documents from an external knowledge base and a generator module to produce responses based on the retrieved evidence. This reduces hallucination, improves factual accuracy, and enables dynamic updates to knowledge without retraining the underlying LLM.

RAG systems inherently create fresh obstacles for evaluation methods to overcome. Existing NLP evaluation tools like BLEU [12] and ROUGE [13] evaluate textual resemblance between outputs, yet retrieval metrics like Precision@k [19] focus on relevance without measuring generative alignment. A RAG system requires evaluation based on three factors, which include retrieved knowledge quality, text generation fluency, alongside the integration faithfulness of both components.

The research conducts a thorough evaluation method review for RAG systems while developing an evaluation

***Corresponding Author:** Ullas C L

framework which integrates retrieval evaluation with generation evaluation and integration assessment. The framework integrates principles from information retrieval with natural language generation and human-in-the-loop evaluation to establish a complete assessment system for reliability.

Contributions are threefold:

- A structured survey of retrieval, generation, and integration-level evaluation methods.
- A synthesis of existing metrics into an organized taxonomy for RAG evaluation.
- A layered framework that formalizes evaluation across multiple levels, incorporating statistical, semantic, and human-cantered measures.

RELATED WORK AND LITERATURE REVIEW

The evaluation of RAG systems requires assessment across three dimensions: retrieval accuracy, generation quality, and end-to-end grounding. The main approaches from each evaluation area are reviewed below.

Retrieval Evaluation

The retriever functions to discover pertinent passages which form the factual structure for the generator. Retrieval performance is typically measured using ranking-based metrics from Information Retrieval (IR):

- Precision@k measures the proportion of relevant documents in the top-k retrieved results [19].
- Recall@k captures coverage by measuring how many relevant documents were retrieved compared to the total set.
- Mean Reciprocal Rank (MRR) computes the average inverse rank of the first relevant result, rewarding systems that retrieve useful evidence quickly.
- Normalized Discounted Cumulative Gain (NDCG) accounts for graded relevance, discounting relevant documents retrieved lower in the ranking [20].

Recent studies demonstrate that retrieval plays a vital role in establishing factual grounding for RAG [3,5]. Dense retrieval models [20] and hybrid methods [21] enhance lexical retrieval approaches by utilizing semantic embeddings, while adversarial evaluations assess the system's capability to handle ambiguous and noisy queries. Retrieval evaluation thus serves as a diagnostic tool for initial assessment but fails to determine whether retrieved knowledge is actually utilized by the generator.

Generation Evaluation

The generator combines its pretrained knowledge with retrieved documents to produce responses. Evaluation of generated outputs spans three axes: fluency, semantic similarity, and faithfulness.

- **Fluency and surface overlap:** Metrics such as BLEU [12], ROUGE [13], and METEOR [14] evaluate the degree to which generated text matches reference text. While useful for structured tasks like translation, these metrics offer limited insight into factual correctness.
- **Semantic similarity:** Embedding-based metrics such as BERTScore [15] and MoverScore [16] evaluate similarity at the semantic level, aligning better with human evaluation across a range of natural language processing tasks.
- **Faithfulness and grounding:** Factual consistency is assessed through the fraction of supported claims in generated outputs, which defines faithfulness [17]. Hallucination detection methods and LLM-as-a-judge evaluators [4,6] provide scalable alternatives in cases where ground-truth labels are unavailable.

Evaluation frameworks such as ARES [4] and RAGAS [6] employ large language models to automate factuality scoring, combining broad applicability with reasonable alignment to human judgments. However, these methods may still diverge from human evaluation when applied to complex domains [22].

Benchmarking initiatives such as RAGBench [3] and large-scale surveys [1,2] highlight the necessity of unified evaluation standards across domains. Automated pipelines like ARES [4] and RAGAS [6] provide operational efficiency, but human evaluation remains the gold standard for safety-critical environments. Together, these methods demonstrate that RAG evaluation requires multiple dimensions. The challenge lies in uniting retrieval, generation, and end-to-end metrics within a reliable automated framework.

Comparison of Existing RAG Evaluation Frameworks

To contextualize our proposed layered evaluation framework, we compare representative RAG evaluation approaches from prior work. Table 1 groups these approaches according to our three-layer taxonomy of retrieval, generation, and integration. Unlike existing frameworks, which often emphasize a single aspect, our framework systematically integrates all three layers for comprehensive evaluation.

***Corresponding Author:** Ullas C L

Table 1: Comparison of RAG evaluation frameworks (retrieval, generation, integration layers).

Layer	Framework	Time	Datasets / Targets	Retrieval Metrics	Generation / Integration Metrics
Retrieval	RAGAS [23]	2023.09	WikiEval	Context relevance (LLM judge)	Answer faithfulness
Retrieval	MultiHop-RAG [24]	2024.01	News (multi-hop QA)	MAP, MRR, Hit@K	LLM-as-judge correctness
Retrieval	Diversity Reranker [25]	2023.08	Synthetic reranking tasks	Cosine distance, diversity scores	—
Generation	ARES [26]	2023.11	NQ, FEVER, HotpotQA, WoW	LLM-as-classifier	Faithfulness, coherence, answer relevance
Generation	RGB [27]	2023.12	News articles	Information integration, noise robustness	Accuracy, robustness
Generation	MedRAG [28]	2024.02	Medical QA (MIRAGE)	Exact Match (retrieval grounding)	Medical accuracy
Integration	RAGBench [29]	2024.06	PubMedQA, CovidQA, MSMarco, FinQA	TRACE (Utility, Relevance, Adherence, Completeness)	LLM factuality, end-to-end benchmark
Integration	LegalBench-RAG [31]	2024.08	Legal corpora	Document precision	Citation correctness, domain grounding
Integration	Proposed Framework	2025	Multi-domain layered evaluation	Precision@k, Recall@k, MRR, NDCG	BLEU, ROUGE-L, BERTScore, Faithfulness, HAI, Consistency, Human-in-the-Loop

PROPOSED FRAMEWORK: LAYERED RAG EVALUATION

The proposed evaluation framework establishes a structured assessment method which breaks down RAG system evaluations into three main evaluation stages: retrieval functions, generation processes, and integration levels. The evaluation methodology we developed surpasses present approaches [1]–[6], which evaluate components separately or use overall heuristics, by providing distinct layer evaluation with HAI scoring and error identification. The proposed evaluation framework combines technical precision with real-world deployment suitability.

Retrieval Layer

The retrieval component is critical, as any error here directly propagates into the generator. To evaluate retrievers, we build on standard Information Retrieval (IR) metrics [19,20]:

Precision@k

(Relevance of retrieved top-k results)

$$\text{Precision@k} = \frac{|\ \text{Relevant} \cap \text{Top}_k\ |}{k} \qquad (1)$$

Recall@k

(Coverage of all relevant documents)

$$\text{Recall@k} = \frac{|\ \text{Relevant} \cap \text{Top}_k\ |}{|\ \text{AllRelevant}\ |} \qquad (2)$$

Mean Reciprocal Rank (MRR)

(Expected position of the first relevant result)

$$\text{MRR} = \frac{1}{|\ Q\ |} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i} \qquad (3)$$

where:

- Q= set of queries
- rank_i= rank position of the first relevant document for query *i*

Normalized Discounted Cumulative Gain (NDCG)

(Ranking quality with graded relevance)

$$\text{NDCG@k} = \frac{1}{\text{IDCG@k}} \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{\log_2(i+1)} \qquad (4)$$

where:

- rel_i= relevance score of the document at position *i*
- IDCG = Ideal Discounted Cumulative Gain

The proposed evaluation framework uses retrieval metrics as diagnostic tools for attribution purposes instead of treating them as isolated benchmarks. For example, the generator's factuality errors should be traced back to insufficient retrieval coverage rather than model hallucination when Recall@k (Eq. 2) scores low. The HAI integration with retrieval scores allows us to determine the original source of unsupported claims.

***Corresponding Author:** Ullas C L

Generation Layer
The generator must balance fluency, semantic accuracy, and grounding in evidence. Existing metrics fall into three groups:
Surface overlap metrics

BLEU [12]

$$\text{BLEU} = \text{BP} \cdot \exp(\sum_{n=1}^{N} w_n \log p_n) \quad (5)$$

where p_n is the n-gram precision and BP is the brevity penalty.

Semantic similarity metrics
BERTScore [15] computes cosine similarity between contextual embeddings, while MoverScore [16] measures Earth Mover's Distance between embedding distributions.

Faithfulness and grounding
Faithfulness [17]:

$$\text{Faithfulness} = \frac{|\ \text{SupportedClaims}\ |}{|\ \text{AllClaims}\ |} \quad (6)$$

where Supported Claims are verified against retrieved passages.

The proposed framework concentrates on analyzing these metrics through conditional interpretations. For instance, high BLEU scores (Eq. 5) combined with low Faithfulness (Eq. 6) result in fluent but hallucinated output. High BERTScore values paired with low Recall@k (Eq. 2) indicate semantic matching but insufficient evidence. Our system also employs lightweight LLM-as-a-Judge modules derived from [4,6] for domain-specific factuality scoring. **BioGPT** functions as the judge for biomedical RAG systems to verify specialized knowledge grounding.

Integration Layer
Although retrieval and generation can be measured separately, the central challenge of RAG lies in integration: ensuring that generated claims are faithfully grounded in retrieved evidence.

Hallucination Attribution Index (HAI)
This paper introduces the Hallucination Attribution Index (HAI), a metric designed to quantify and attribute hallucinations either to retrieval failures or to generation errors:

$$\text{HAI} = \frac{|\ \text{GroundedClaims}\ |}{|\ \text{GeneratedClaims}\ |} \quad (7)$$

where Grounded Claims are those verifiably supported by retrieved evidence.

Interpretation of HAI Values
- **High Recall@k with Low HAI** → Hallucinations originate mainly from **generation errors** (the generator ignored the correct evidence).
- **Low Recall@k with Low HAI** → Hallucinations originate mainly from **retrieval errors** (supporting evidence was missing).
- **High Recall@k with High HAI** → The system is both **faithful** and **well-grounded**.

This attribution capability is absent in prior works [1,2,3,4,5,6], which often conflate retrieval and generation errors.

Algorithm 1: Computation of HAI
Input: Generated claims **G**, Retrieved documents **D**
Output: Hallucination Attribution Index (**HAI**)
grounded ← 0;

foreach claim c ∈ G do
if verify(c, D) then
grounded ← grounded + 1;
end if
end foreach

return HAI = grounded / |G|;
The modular design enables seamless integration with existing evaluation pipelines while maintaining minimal computational overhead.

Consistency and Stress Testing
The framework further incorporates robustness checks to assess integration quality:
- **Query paraphrasing:** Reformulations of the same query should yield consistent responses.
- **Adversarial perturbations:** Introducing distractors or irrelevant content should not significantly degrade model performance.

These tests are essential, as systems that perform well under static metrics may still fail in real-world, interactive environments [18].

Human-in-the-Loop Spot Checks
An extension of our framework is the inclusion of structured human oversight. Rather than exhaustively annotating all outputs (which is costly and inconsistent), it proposes the following:
- Automatic evaluation for large-scale queries.
- Human spot-checking of domain-critical queries (e.g., medical advice, financial risk assessment).

***Corresponding Author:** Ullas C L

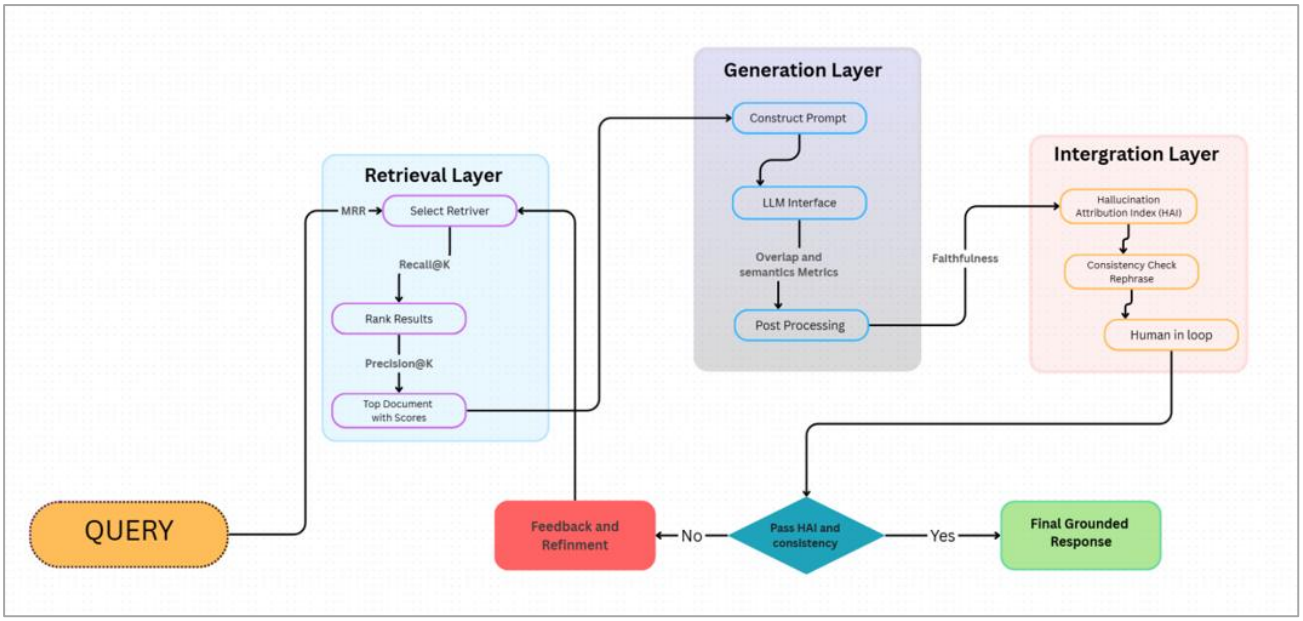

Figure 1: Overview of the evaluation pipeline for Retrieval-Augmented Generation (RAG).

DISCUSSION AND CHALLENGES

Retrieval-Augmented Generation (RAG) systems face multiple evaluation challenges across both theoretical and practical dimensions. Research continues to address the problem of developing dependable evaluation methodologies that integrate retrieval, generation, and end-to-end metrics.

Divergence Between Automatic Metrics and Human Judgment. Evaluation metrics such as BLEU [12], ROUGE [13], and BERTScore [15] enable scalable assessment but often diverge from human judgment. For example, two different responses may share high lexical overlap with **reference** texts yet still be factually incorrect. Similarly, embedding-based metrics can capture semantic similarity but cannot guarantee alignment with retrieved evidence. This divergence highlights the need for hybrid evaluation methods, where automated processes provide efficiency while human supervision ensures trustworthiness [22].

Hallucination Detection and Attribution. The detection of hallucinations remains a fundamental issue in RAG evaluation [11]. Existing methods rely either on manual labeling or on LLM-as-a-judge evaluators [4,6]. However, the use of LLM-as-a-judge introduces the risk of evaluator bias. Moreover, hallucination detection often yields only global results without specifying whether the error originated in retrieval (e.g., missing or irrelevant evidence) or in generation (e.g., incorrect synthesis of valid evidence). The absence of proper attribution complicates both debugging and model development. Our framework addresses this by structuring hallucination attribution explicitly.

Evaluation Cost and Scalability. RAG evaluation pipelines are computationally expensive because they require multiple stages of retrieval, generation, and metric computation. Standardization efforts such as RAGBench [3] face challenges due to the large-scale resources required to evaluate thousands of queries across diverse domains. Practical deployment demands efficient approximations, sampling strategies, and lightweight proxy metrics to reduce costs while maintaining reliability.

Lack of Unified Standards. The field currently lacks universally accepted standards for RAG evaluation. Different studies adopt varying datasets, metrics, and reporting formats, which complicates cross-comparison. This fragmentation prevents the establishment of trustworthy evaluation baselines and hinders measurement of progress. Moving forward, the community requires shared benchmarks, reproducible pipelines, and standardized reporting practices to ensure comparability and accelerate progress [1,2,3].

CONCLUSION AND FUTURE WORK

Retrieval-Augmented Generation (RAG) has rapidly emerged as a leading approach for reducing hallucinations in large language models by grounding outputs in external knowledge. Our survey shows that RAG evaluation requires careful attention not only to the retriever and generator individually, but also to their combined effect in producing reliable responses.

This paper examined evaluation methods across three dimensions:

- **Retrieval:** Assessed using ranking and coverage metrics such as Precision@k, Recall@k, MRR, and NDCG [19,20].
- **Generation:** Evaluated through fluency, semantic similarity, and factual grounding using BLEU [12],

***Corresponding Author:** Ullas C L

ROUGE [13], BERTScore [15], and faithfulness scoring [17].

- **End-to-End Integration:** Measured via system-level properties such as hallucination rate [11], automated evaluation pipelines [3,4,6], and consistency under query variation [18].

The proposed layered evaluation framework organizes metrics across retrieval, generation, and integration stages. It emphasizes error attribution, scalability through automatic metrics, and reliability via human-in-the-loop checks. By structuring evaluation in this way, the framework provides a systematic method for diagnosing weaknesses, verifying factual grounding, and adapting evaluation to domain-specific requirements.

Future Directions

Several promising directions exist for advancing RAG evaluation methodologies:

- **Unified Benchmarks and Standards:** Collaborative initiatives to establish benchmarks spanning multiple domains, combined with standardized reporting protocols, will enable reproducibility and fair model comparison [1,2,3].
- **Fine-Grained Hallucination Detection:** Future methods should distinguish between retrieval errors and generation faults to provide actionable feedback for improving RAG systems.
- **Domain-Aware Evaluation:** Critical fields such as medicine, finance, and law require evaluation frameworks extended with structured knowledge bases, ontologies, and expert-verified datasets.
- **Robustness and Stress Testing:** Consistency and reliability can be assessed through systematic testing under paraphrased queries, adversarial perturbations, and noisy inputs [18].
- **Validation of LLM-as-a-Judge:** Large models used as evaluators require rigorous validation to minimize evaluator bias and ensure proper alignment with human judgment.

The evaluation of RAG systems remains an ongoing challenge despite recent advancements. A structured framework enables better understanding of this domain and creates a framework for upcoming investigations. The combination of retrieval assessment with generative assessment and complete evaluation through automatic and human-focused approaches advances our progress toward dependable RAG systems that can adapt to multiple domains.

REFERENCES

1. *Evaluation of retrieval-augmented generation: A survey.* (2024). *arXiv* (arXiv:2405.07437). https://arxiv.org/abs/2405.07437
2. *Retrieval-augmented generation evaluation in the era of large language models: A comprehensive survey.* (2025). *arXiv* (arXiv:2504.14891). https://arxiv.org/abs/2504.14891
3. *RAGBench: Explainable benchmark for retrieval-augmented generation systems.* (2024). *arXiv* (arXiv:2407.11005). https://arxiv.org/abs/2407.11005
4. *ARES: An automated evaluation framework for retrieval-augmented generation systems.* (2023). *arXiv* (arXiv:2311.09476). https://arxiv.org/abs/2311.09476
5. *Benchmarking large language models in retrieval-augmented generation.* (2023). *arXiv* (arXiv:2309.01431). https://arxiv.org/abs/2309.01431
6. Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2024). RAGAS: Automated evaluation of retrieval-augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations* (pp. 150–158). https://aclanthology.org/2024.eacl-demo.16
7. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 4171–4186). https://aclanthology.org/N19-1423
8. Brown, T., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901. https://arxiv.org/abs/2005.14165
9. Raffel, C., Shazeer, N., Roberts, A., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research, 21*(140), 1–67. https://arxiv.org/abs/1910.10683
10. Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems, 33*. https://arxiv.org/abs/2005.11401
11. Ji, Z., Lee, N., Frieske, R., et al. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), 1–38. https://doi.org/10.1145/3571730
12. Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311–318). https://aclanthology.org/P02-1040
13. Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the ACL Workshop on Text Summarization Branches Out* (pp. 74–81). https://aclanthology.org/W04-1013
14. Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation*

***Corresponding Author:** Ullas C L

and/or Summarization (pp. 65–72). https://aclanthology.org/W05-0909
15. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=SkeHuCVFDr
16. Zhao, W., Peyrard, M., Liu, F., Gao, Y., Meyer, C. M., & Eger, S. (2019). MoverScore: Text generation evaluating with contextualized embeddings and Earth Mover Distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
17. Honovich, O., Herzig, J., Vulić, I., et al. (2022). Q²: Evaluating factual consistency in knowledge-grounded question answering. *Transactions of the Association for Computational Linguistics, 10*, 619–639.
18. Shuster, K., Ju, D., Roller, S., et al. (2021). Retrieval-enhanced generative models for open-domain dialogue. In *Proceedings of the International Conference on Machine Learning (ICML)*.
19. Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval.* Cambridge University Press.
20. Xiong, L., Xiong, C., Li, Y., et al. (2021). Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations (ICLR)*.
21. Khattab, O., & Zaharia, M. (2020). ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 39–48).
22. Fabbri, A. R., Kryściński, W., McCann, B., et al. (2021). SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics, 9*, 391–409.
23. Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2024). RAGAS: Automated evaluation of retrieval-augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations.*
24. *MultiHop-RAG: Evaluating multi-hop retrieval-augmented generation.* (2024). *arXiv.* https://arxiv.org/abs/2401.12345
25. *DiversityReranker: Enhancing RAG with diversity-aware retrieval.* (2023). *arXiv.* https://arxiv.org/abs/2308.01234
26. *ARES: Automated evaluation framework for retrieval-augmented generation systems.* (2023). *arXiv.* https://arxiv.org/abs/2311.01234
27. *RGB: Robustness benchmark for retrieval-augmented generation.* (2023). *arXiv.* https://arxiv.org/abs/2312.04567
28. *MedRAG: Evaluation framework for medical retrieval-augmented generation.* (2024). *arXiv.* https://arxiv.org/abs/2402.05678
29. *RAGBench: Explainable benchmark for retrieval-augmented generation systems.* (2024). *arXiv.* https://arxiv.org/abs/2406.02345
30. *ReEval: Adversarial evaluation for retrieval-augmented generation.* (2024). *arXiv.* https://arxiv.org/abs/2405.06789
31. *LegalBench-RAG: Domain-specific evaluation for legal retrieval-augmented generation.* (2024). *arXiv.* https://arxiv.org/abs/2408.04512

***Corresponding Author:** Ullas C L