# Understanding the gap in evaluation frameworks and metrics for RAG-based question-answering system

Nakul Mehta[1,*], Adegboyega Ojo[2] and Edward Curry[3]

[1]CRT in AI, Data Science Institute, University of Galway, Ireland
[2]School of Public Policy & administration, Carleton University, Canada
[3]Data Science Institute, University of Galway, Ireland

**Abstract**

Large Language Models (LLMs) have been extensively employed for answer generation. Particularly, Retrieval-Augmented Generation (RAG) systems have emerged as a dominant paradigm for enhancing LLM outputs with external knowledge, yet comprehensive evaluation of these systems remains fragmented and inadequate. This paper presents a systematic analysis of existing evaluation frameworks and metrics for RAG-based Question Answering (QA) systems, addressing three core research questions: the landscape of current evaluation approaches, their reliability across eight critical linguistic dimensions, and fundamental gaps requiring novel solutions. We synthesize metrics spanning classical NLP, and LLM-based evaluation paradigms, then empirically assess their robustness using a synthetically generated dataset targeting Negation, Morphology, Reordering, Paraphrasing, Active-Passive Voice, Direct-Indirect Speech, Synonyms, and Numerics. Our results demonstrate that embedding-based metrics like BERTScore exhibit superior overall performance but fail catastrophically on numerical perturbations and negation flips. We find that no single metric class suffices for comprehensive evaluation, with precision-recall asymmetries in syntactic transformations and near-chance performance of context-based metrics exposing critical vulnerabilities. These findings reveal urgent needs for hybrid evaluation frameworks, reasoning path traceability, combined pipeline assessment, and harmlessness evaluation, laying groundwork for the next generation of evaluation methodologies.

**Keywords**

Evaluation, LLM, RAG, Metrics, Frameworks, Question Answering (QA), LLM as Judge

## 1. Introduction

Large Language Models (LLMs) have become a constant part of the research and have grown exponentially popular [1]. However, even the largest models struggle when it comes to memorize knowledge that is substantial but sparse in the training corpus [2]. Moreover, the concepts such as hallucination and knowledge cut-off limits their abilities to provide precise output. To address this gap, Retrieval Augmented Generation (RAG) [3] emerged as a solution that helps to extend the knowledge of the LLMs by incorporating external knowledge sources (could be domain specific) to tailor their response. Since 2022, many RAG systems have shown significant progress in extending and tailoring the answering generation capabilities of LLMs [4].

Standard RAG system contains three stages: 1) Indexing 2) Retrieval 3) Generation [3]. However, these can be further categorized into Indexing, pre-retrieval, retrieval, post-retrieval, and generation. [5]. This modularity introduces flexibility and numerous customization options [4] but also raises concerns about the way they could be evaluated or compared to determine the best possible system. During the RAG evaluation, every component/stage of the system gets evaluated under a workflow but this presents several challenges including the responsibility of choosing the best possible evaluation technique to assess the quality of the system. An output generated by a RAG based QA systems could be mix of both correct and incorrect answers, while also some of them being part of LLM hallucination. In such scenario, evaluating the efficacy of these system become paramount [6].

*ELMKE: Evaluation of Language Models in Knowledge Engineering (2026)*
*Corresponding author.
n.mehta3@universityofgalway.ie (N. Mehta); adegboyegaojo@cunet.carleton.ca (A. Ojo); edward.curry@universityofgalway.ie (E. Curry)
0009-0008-5374-6320 (N. Mehta); 0000-0003-0565-2115 (A. Ojo); 0000-0001-8236-6433 (E. Curry)

Traditional evaluation criteria includes the usage of metrics such as precision, recall, F1-score, ROUGE-L, and BLEU. However, they focus mainly on the lexical overlap failing to grasp the semantic meaning or overall complexity of the overall task especially in context to concepts such as grounding, hallucination and reasoning [7]. Therefore, automated approaches such as evaluation of these RAG based QA systems using LLM have started to emerge [8]. These automated methods tests the ability of the system in terms of metrics such as answer relevance, faithfulness, context precision, context recall, context relevance, claim precision, claim recall, aiming to capture the semantic depth, hallucination, grounding and overall correctness of the RAG system [8, 2, 5, 9, 10]. However, for systems with large complexity that may have long or complex responses, new methods are required for comprehensive evaluation [4]. Furthermore, several works adopt self-defined evaluation metrics for RAG assessment, underscoring the absence of widely accepted benchmarking standards for evaluating RAG systems [11]. In most cases the evaluation of RAG's application in various downstream tasks and with different retrievers may yield divergent results [1]. This need introduces a gap in the evaluation techniques due to the absence of benchmark and presence of entropy and diversity in the evaluation criteria.

Understanding this gap, our work aims to provide the overall view of the existing evaluation techniques in RAG based QA systems with major focus on the generation aspect (i.e., the answer correctness) of the RAG system. With our work, we aim to answer the following research questions:

- **RQ1:** What are the existing techniques and metrics used to evaluate the LLM based answer generation (in-context to a RAG based QA systems)?
- **RQ2:** What is the reliability of these techniques in evaluating the output based on the 8 aspects of traditional NLP evaluation named Negation, Morphology, Reordering, Paraphrasing, Active-Passive voice, Direct-Indirect speech, Synonyms and Numerics?
- **RQ3:** What are the gaps and challenges in the existing techniques? Is there a need for new metric or technique?

## 2. Related Work

This section entails the related work in assessing the evaluation of RAG based QA techniques in terms of both the answer generation and traditional NLP aspects such as Negation, Morphology, Synonyms, Direct-Indirect speech, Active-passive voice, Reordering, Paraphrasing, and Numerics. The overall performance of the RAG system is affected by multiple factors, such as the retrieval model, construction of the external knowledge base, and language model [11]. The exact RAG evaluation criteria also vary from study to study. To the best of our knowledge, there has been a significant amount of work done in devising new metrics or frameworks for the evaluation but few works discuss about assessing all these new measures especially from the 8 traditional aspects of NLP as mentioned above.

Authors such as Goodrich et al. [9], Celikyilmaz et al. [12], Ito et al. [13] focused on evaluating the accuracy or the correctness aspect of the generated text. These works are similar to the presented work but they do not focus on either the overall RAG evaluation techniques or the 8 aspects. Goodrich et al. [9] presented the idea of factual accuracy proposing a model based metric for its estimation. Introducing the Named Entity Recognition (NER) in combination with transformer encoder layer, the work aims to compare the model based metrics to the ROGUE-L, the BLEU score and the OpenIE based precision metrics. On the contrary, Celikyilmaz et al. [12] and Ito et al. [13] provided a broader overview of the evaluation of text generation. The survey presented by Celikyilmaz et al. [12] presents human-centric, untrained automatic (n-grams) and machine-learning based metrics evaluation comparing them in terms of the tasks named machine translation, image captioning, speech recognition, summarization , document generation, question generation, dialogue response, and if they are distance based or similarity based. Furthermore, it discusses one of the 8 aspects - paraphrasing by calculating Paraphrase Identification (PI) based on text similarity and presents ideas on lexical cohesion. Similarly Ito et al. [13] presented a survey on the reference-free metrics with similar focus on the tasks as mentioned above such as machine translation, summarization, etc.

Works such as [14], [15], [16], [17], [18], [4], [10], [19], [20] focused on evaluating the LLMs and RAG based systems in domain of QA and ontologies. Authors such as Brehme et al. [4], Imtiyaz et

al. [10], Roychowdhury et al. [19], and Oro et al. [20] gave specific focus on the RAG based systems. Some focused on specific metrics such as answer relevance, answer correctness, and BERT matching score in comparison to the human judgement [20], while others showed the idea of answer relevance, faithfulness, context relevance, and answer correctness [19] but focuses majorly on metrics numbers in case of different LLMs than the actual metrics. On the other hand, Brehme et al. [4] and Imtiyaz et al. [10] presented specific metrics based on the different stages of RAG (indexing, retrieval, generation) and overall RAG evaluation framework respectively. Brehme et al. [4] specifically focuses on stating the existing metrics in the literature with less focus on their evaluation or definitions whereas Imtiyaz et al. [10] described the existing RAG evaluation frameworks which assess the overall qualtiy of the systems taking LLM-as-Judge. Other works presents the evaluation of LLM based on metrics such as accuracy (both human and LLM based), F-score (F1-score and F2-score), precision, recall, error score, BLEU, and BERTScore while also presenting the application domains and future research for LLMs. Other works on RAG such as by Gao et al. [1], Zhao et al. [21], and Hu et al. [22] works on assessing the RAG systems. However, their main focus is on the RAG techniques or methods rather than specifically addressing the metrics and metrics evaluation aspect.

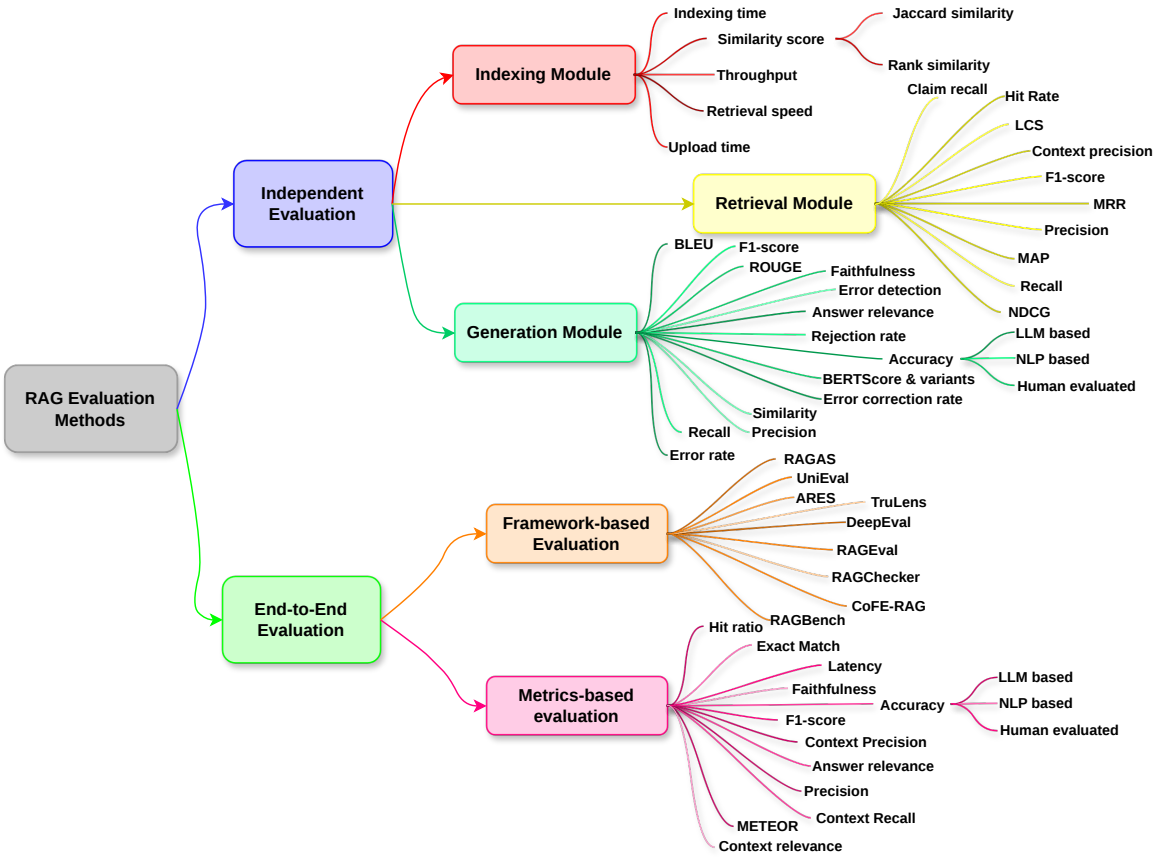


**Figure 1:** Taxonomy for evaluating the RAG based QA systems

Specific works such as [23], and [24] focuses on developing an understanding of LLMs for evaluation. Hannah et al. [24] showcased and stressed on using LLMs as comparative agents rather than just the evaluators. Their work presents the efficacy of LLM based comparison in terms of zero-shot and few-shot prompting with experimental results on LLM generated dataset in collaboration with the domain expert. Contrastively, Malin et al. [23] focused on the hallucination aspect of the LLMs, showcasing the faithfulness of the LLMs by presenting details of multiple datasets and their faithfulness metric.

## 3. Understanding RAG evaluation metrics and frameworks

Evaluation of RAG systems presents a challenge in itself [25]. Based on the analysis of multiple evaluation frameworks and metrics, a RAG system can be evaluated on its individual stages such indexing, retrieval, and generation or in end-to-end fashion. This end-to-end evaluation could be further bifurcated into either overall metrics based evaluation or framework-based evaluation. Most of the studies have started to opt LLM-as-Judge for the RAG evaluation, computing metrics such as faithfulness, answer relevance, context relevance, context recall, and many more. Frameworks such as ARES [8], RAGAS [2], RAGChecker [6], and CoFE-RAG [26] are becoming prominent baselines for evaluation of the RAG approaches. Despite these LLM-as-Judge approaches, traditional metrics such as precision, recall, F1-score, BLEU, METEOR, ROUGE, etc., are also some of the metrics that are used for the evaluation purposes. This section describes the RAG frameworks and the NLP metrics that are used to evaluate the RAG systems from multiple aspects. Fig. 1 describes the taxonomy of RAG evaluation techniques.

### 3.1. Framework level

Considering the framework based evaluation techniques such as RAGAS [2], ARES [8], DeepEval [27], RAGEval [7], RAGChecker [6], CoFE-RAG [26], and RAGBench [28], one common aspect is the usage of the LLM. Metrics such as faithfulness (which checks if the answer is grounded in the given context or not), answer relevance (which checks that the generated answer should address the actual question), context relevance (refers to the idea that the retrieved context should be focused, containing as little irrelevant information as possible), or hallucination (which checks if the generated answer is a figment of LLM imagination or from the provided context) utilizes LLMs ability to analyze the effectiveness the RAG system. This makes these metrics more LLM dependent i.e., the choice of LLM would affect the output of the metrics. Additionally, the usage of LLM adds a layer of black-box nature to the whole calculation, questioning the efficacy of the metrics itself. Most of the frameworks, such as RAGAS, DeepEval do not require any specific fine tuning, attaching less computational overload to the overall calculation when compared to frameworks such as ARES. Another common overlap between metrics based evaluation is the lack of testing. Most of the frameworks such as RAGAS, RAGEval, CoFE-RAG, and RAGBench usually show tests on only one dataset or their own synthetically generated dataset; thus limiting their strength. A more detailed summary of the frameworks is provided in Table 1.

Table 1 aims to develop an understanding of RAG evaluation frameworks in terms of concepts such semantic depth, transparency, grounding, and robustness to noise along with technical aspects such as if the framework is dependent on LLM, or requires fine tuning an LLM, does it also measures or evaluates the RAG stages (indexing, retrieval, and generation) and the dataset on which the frameworks was tested to establish its generalizability, scalability and efficacy. From all, RAGChecker stands out for exhaustively assessing retriever and generator components covering claim recall, context utilization, noise sensitivity, hallucination, and self-knowledge thereby strongly addressing grounding and robustness, though some claim level faithfulness computations overlap conceptually with RAGAS. CoFE-RAG introduces a multi-granularity keyword strategy and blends traditional NLP metrics (BLEU, ROUGE, recall) with faithfulness and correctness, but validation on synthetic rather than benchmark datasets limits external validity. RAGBench similarly proposes new utilization and adherence metrics under the TRACe framework and introduces a synthetic dataset derived from open-book QA sources, yet lacks benchmarking against established datasets. RAGAS operationalizes faithfulness, answer relevance, and context relevance in a fully automated pipeline, offering partial semantic coverage but with high LLM dependency and limited transparency due to reliance on proprietary APIs; it also does not rigorously verify answer correctness. ARES extends this by incorporating fine-tuning on human-annotated datasets and reporting confidence intervals, improving statistical reliability but increasing human effort and domain adaptation costs. Similarly, DeepEval provides customizable metrics and API support, enabling structured evaluation of contextual precision and recall, though it remains constrained by closed-source LLM usage and requires access to retrieved context and reference outputs. In contrast, RAGEval introduces scenario specific dataset generation and granular sentence-level metrics such as Effective Information Rate (EIR) and

| Framework | Metrics | Notes | Semantic Depth | LLM Dependency | Transparency | Grounding | Fine tuning | Robustness to noise | Retriever | Indexing | Generation | Dataset | Limitations |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| RAGAS [2] | Faithfulness, AR, context relevance | Fully automated pipeline | P | H | N | Y | N | Y | Y | N | Y | WikiEval | Does not properly checks for answer correctness Restricted to OpenAPI |
| ARES [8] | Faithfulness, AR, context relevance | Requires human annotated dataset for fine tuning the LLMs & Provides confidence intervals | P | H | N | Y | Y | Y | Y | N | Y | Natural Questions, HotpotQA, FEVER, and Wizards of Wikipedia, MultiRC and ReCoRD | Does not properly checks for answer correctness Require human annotators for domain specific evaluation |
| DeepEval [27] | AR, Faithfulness, Contextual Relevancy, Contextual Precision, Contextual Recall | Allows to create customised metrics & Provides APIs | P | Y | N | Y | N | Y | Y | N | Y | Not mentioned | Limited to the use of closed source LLMs |
| RAGEval [7] | Recall, Effective Information Rate (EIR), Completeness, Hallucination, Irrelevance | Can generate scenario specific RAG evaluation datasets & Has metrics that checks at both sentence level and granular level | P | P | P | Y | N | Y | Y | N | Y | DragonBall Dataset (synthetically created) | All of its generator assessing metrics depends on the key points the system creates using the LLM which is bottle neck for testing the generator capabilities. |
| RAG Checker [6] | Precision, Recall, Context precision, Claim recall, Context utilization, Noise sensitivity, Hallucination, Self-knowledge, Faithfulness | Evaluates 8 RAG systems & Exhaustive generator evaluation metrics | Y | Y | P | Y | N | Y | Y | N | Y | ClapNQ, NovelQA, RObustQA(Writing, BioASQ, Finance, Lifestyle, Recreation, Science, Technology), KIWI | The claim retailment process is similar to the way Faithfulness is calculated in RAGAS - Redundancy |
| CoFE-RAG [26] | Accuracy, Recall, BLEU, ROGUE, Faithfulness, Correctness | Introduced multi-granularity keywords approach | Y | P | P | Y | N | Y | Y | N | Y | Synthetic dataset containing factual, analytical, comparative and tutorial based queries | Not tested on any benchmark dataset. Involves testing on synthetic dataset |
| RAGBench [28] | Utilization, Relevance, Adherence, and Completeness | Introduces a new dataset & TRACe evaluation framework | Y | Y | P | Y | N | N | Y | N | Y | Synthetic dataset created from open-book question answer datasets | Not tested on any benchmark dataset. Involves testing on synthetic dataset |

Table 1: **Comparison of the RAG evaluation frameworks**
AR: Answer Relevance, Y: Yes, N: No, P: Partial, L: Low, M: Medium, H: High, - : Not relevant

hallucination detection, offering broader semantic inspection; however, its generator-side assessment depends heavily on LLM-derived key points, creating a methodological bottleneck.

Across all frameworks, a common pattern emerges: higher semantic depth and grounding capabilities correlate with increased LLM dependence, reduced transparency, and greater computational or annotation overhead. Moreover, many systems rely on synthetic or internally generated datasets, raising concerns about generalizability and real-world robustness despite methodological sophistication. Additionally, a common gap in the existing frameworks is the lack of testing for the indexing stage. Indexing serves as one of the important factors for the RAG systems, as the efficient retrieval depends on how the context was indexed, whats the speed of the indexing database, and what is the throughput of the database.

### 3.2. Metrics level

To further step a bit deeper, Table 2 provides a metrics level analysis of all the metrics mentioned in Fig. 1 from the ground truth need to the answer grounding. It provides details if the metric - requires some existing gold standard dataset to calculate or a label to compare with (needs reference), is able to handle the semantic changes in the text output such as negation or reordering (covers semantic depth), is dependent on any platform or does it changes if the amount of test data changes (scales cheaply), changes when the type of test data changes (dataset agnostic), depends on the retrieved context (context agnostic), needs LLM for calculation (LLM Dependent), is a white box or a black box technique (Transparency), is able to check if the generated answer is derived from the ground-truth (Grounding) and what amount of human effort is required to calculate or code (Low, Medium or High). It reveals a division between system-level performance metrics and semantic or generation-quality metrics, each addressing fundamentally different evaluation goals. Infrastructure-oriented measures such as Indexing Time, Throughput, Retrieval Speed, Upload Time, and Latency primarily assess operational efficiency rather than semantic correctness. They generally do not require reference answers, are not LLM dependent, and demand low human effort because they rely on measurable system statistics. However, they are rarely dataset or context agnostic, as performance often depends on storage mechanisms, indexing strategies, or infrastructure constraints. Their transparency is typically high since calculations are straightforward and reproducible, yet they provide no grounding guarantees. In contrast, retrieval focused ranking metrics such as Hit Rate, MRR, MAP, and NDCG require references and evaluate how effectively relevant information is surfaced. These scale relatively well due to fixed ranking formulations, but they only weakly capture semantic depth, as they measure position and overlap rather than reasoning quality. Their dependence on LLMs is conditional-stronger when embeddings or generative judgments are involved and human effort increases when retrieval mechanisms become complex or require annotation.

On the generation and semantic evaluation side, metrics such as F1-score, precision, recall, BLEU, ROUGE, METEOR, exact match, and BERTScore variants depend heavily on reference answers and emphasize lexical or embedding-based similarity. While they scale cheaply and remain largely dataset agnostic, they capture limited semantic depth and often fail to assess reasoning, grounding, or factual faithfulness. Embedding-based metrics introduce partial opacity due to black-box representations, whereas lexical metrics remain transparent but shallow. More recent evaluation paradigms such as context relevance, answer relevance, faithfulness, context precision/recall, error detection, error correction rate, rejection rate, and accuracy - shift focus toward grounding and reasoning quality. These are typically LLM dependent, require moderate to high human effort (especially for annotation or model calibration), and scale less efficiently due to model inference costs. However, they offer stronger semantic coverage and, in some cases, explicit grounding verification. Notably, faithfulness and error-based metrics emphasize alignment between generated answers and retrieved context, directly addressing hallucination risks, an area where traditional overlap metrics underperform. Overall, the table highlights a trade-off landscape: inexpensive, transparent metrics offer scalability but shallow semantic insight, whereas grounding-aware and reasoning-focused metrics provide deeper evaluation at higher computational and human cost.

| Method | Needs Reference | Semantic Depth | Scales Cheaply | Dataset Agnostic | Context Agnostic | LLM Dependent | Human Effort | Transparency | Grounding |
|---|---|---|---|---|---|---|---|---|---|
| Indexing time [29] | N | - | P (depends on storage) | N (depends on dataset dimensions) | - | N | L (only to define the code) | P (storage dependent) | - |
| Similarity [30, 31, 32] | Y | N (doesn't cover all aspects) | Y (fixed dimensional vectors) | Y | Y | Y | L-M | P (Black box embeddings) | N |
| Throughput [29] | - | - | N (fixed as per system) | N | - | - | L | Y | - |
| Retrieval speed [29] | Y | - | Y (with proper indexing method) | Y | - | N | L | Y | - |
| Upload time [29] | - | - | N | P (depends on data structure ) | - | N | L | Y | - |
| Hit Rate [33] | Y (needs starting point) | - | P (depends on the data) | N | N | N (retriever dependent) | L-H (depends on retrieval) | Y | P |
| LCS [34, 35] | Y | P (very limited) | N (expensive for longer texts) | Y | - | N | L | Y | N |
| Latency [36] | N | - | N (infrastructure dependent) | Y | N | Y | L | Y | N |
| F1-score [25, 9] | Y | N | Y | P (depends on dataset) | P (indirectly depends on context) | P (depends if LLM used) | L-M | P (depends on LLM or direct math) | P (if LLM used yes else no) |
| Precision [25, 9] | Y | N | Y | P (depends on dataset) | P (indirectly depends on context) | P (depends if LLM used) | L-M | P (depends on LLM or direct math) | P (if LLM used yes else no) |
| Recall [25, 9] | Y | N | Y | P (depends on dataset) | P (indirectly depends on context) | P (depends if LLM used) | L-M | P (depends on LLM or direct math) | P (if LLM used grounding else no) |
| MRR [37] | Y | N | Y | Y | N | P (depends if LLM used) | L-M | P (depends if LLM used) | - |
| MAP [38] | Y | N | Y | Y | N | P (depends if LLM used) | M-H | P (depends if LLM used) | - |
| NDCG [39] | Y | N | Y | Y | N | N | M-H | Y | - |
| BLEU [40] | Y | P (limited) | Y | Y | Y | N | L-M | Y | N |
| Context Relevance [2] | N | N | N | Y | N (depends on quality of context) | Y | L-H | N | N |
| Answer relevance [2] | N | N | N | Y | P (depends on question) | Y | L-H | N | N |
| Faithfulness [2] | P (requires context) | Y | N | Y | N | Y | L-H | N | Y |
| Error detection [41] | Y | Y | P (depends on model size) | N | N | Y | M-H | N | Y |
| Error correction rate [41] | Y | Y | P (depends on model size) | N | N | Y | M-H | N | Y |
| Rejection rate [41] | Y | Y | P (depends on model size) | N | N | Y | M-H | N | Y |
| Accuracy [33, 19] | Y | P (if eval by NLP/LLM/ human) | P (depends if LLM is used) | Y | Y | P (depends if LLM used) | L-H | P (depends on eval method) | P (depends on eval method) |
| BERTScore & variants [42, 43] | Y | P (limited) | N | Y | Y | Y | L-M | N | N |
| ROUGE [44, 29, 33] | Y | P (limited) | Y | Y | Y | N | L | Y | N |
| Exact Match [25] | Y | N | Y | Y | Y | N | L | Y | N |
| METEOR [45] | Y | P (limited) | Y | Y | Y | N | L-M | Y | N |
| Context Precision [2] | Y | N | Y | Y | N | N | L | Y | - |
| Context Recall [2] | Y | N | Y | Y | N | N | L | Y | P(only for gold context) |

**Table 2**
**Comparison of various metrics for RAG system**
Y: Yes, N: No, P: Partial, L: Low, M: Medium, H: High, - : Not relevant

### 3.3. Summary

This section aims to answer RQ1 using Table 1 and Table 2 that highlights a comprehensive comparative analysis of the evaluation frameworks & metrics for RAG revealing heterogeneity in how various metrics and frameworks address the fundamental challenges of measuring the RAG performance. Overall they underscores that no single metric adequately balances all evaluation criteria - semantic depth, dataset agnostic, scales cheaply, context agnostic, LLM dependent, human effort, transparency, grounding, robust to noise, fine tuning, indexing, and generation. Additionally, a common gap is lack of testing for the indexing stage in popular frameworks. This analysis suggests that robust RAG assessment necessitates composite evaluation frameworks that strategically combine cheap, scalable lexical metrics for filtering with expensive, LLM-based semantic verification for final quality assurance, while carefully documenting transparency trade-offs inherent in each layer of the evaluation stack.

## 4. Evaluating the metrics

This section evaluates the performance of the metrics across eight traditional NLP aspects: Negation, Reordering, Paraphrasing, Synonyms, Direct-Indirect Speech, Active-Passive Voice, Morphology, and Numerics. Our objective is to analyze the effectiveness of the evaluation metrics listed in Table 2, with a primary focus on assessing the quality of the generated answers. To support this analysis, we consider the following metrics: EM, Token_F1 (or F1-score), Token_Precision(or Precision), Token_Recall(or Recall), METEOR, BLEU, BERTScore_P, BERTScore_R, BERTScore_F1, Faithfulness, Context_Precision, Context_Recall, Token_Accuracy, Meaning_Accuracy_LLM (Accuracy using LLM), Meaning_Accuracy_Emb (Accuracy using similiarity), ROUGE_1, ROUGE_2, ROUGE_L and new metric called Fuzzy_Accuracy[1]. These metrics are examined to understand how effectively they capture variations in linguistic structure and meaning across the selected NLP aspects while evaluating the correctness and quality of generated responses.

### 4.1. Dataset construction and description

The dataset was created through a systematic synthetic data generation pipeline designed to evaluate model robustness across multiple linguistic dimensions. The process began with four established benchmark datasets: HotpotQA [2], Natural Question [3], 2WikiMultiHop[4], and Long Form QA[5]. From these sources, we initially collected 250 samples each, specifically extracting columns containing context, supporting context, and reference answers. This curated subset served as the foundation for generating synthetic variations. The expansion strategy employed an LLM (Llama-3.2-3B-Instruct[6]) to create controlled perturbations of the original answers. For each of the 250 base samples, the pipeline generated variations across eight distinct linguistic aspects plus one control condition. These aspects included semantic modifications such as enhancing semantic depth, substituting synonyms or antonyms, altering morphology, introducing negation, paraphrasing content, reordering sentence structures, converting to passive voice, transforming to indirect speech, and manipulating numerical values. This multiplicative approach-250 samples × 9 conditions (8 aspects + 1 control)-yielded a theoretical maximum of 9000 variations. Following generation, the dataset underwent manual curation to ensure quality and representation. Finally, 500 samples were selected that optimally represented all eight aspects while eliminating redundancy. Each curated sample was then manually evaluated to verify alignment between the original answers and their synthetic counterparts, with these assessments recorded in a dedicated “Analysis” column.

[1] For more details refer to supplemental material: https://github.com/darkpheonix2/Metrics/blob/main/Supplemental_material.pdf

[2] https://hotpotqa.github.io/

[3] https://ai.google.com/research/NaturalQuestions/dataset

[4] https://huggingface.co/datasets/xanhho/2WikiMultihopQA

[5] https://huggingface.co/datasets/akoksal/LongForm

[6] https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct
The choice of LLM was based on popularity on HuggingFace and the model size

The resulting dataset structure, contains multiple columns tracking the experimental design. Core fields include the original context and supporting context from the source benchmarks, the reference answer, and the generated response. Crucially, each row is tagged with an 'Aspect' column indicating which linguistic transformation was applied-ranging from specific modifications like "Negation" or "Paraphrasing" to the unaltered "Control" condition. This structured format facilitates systematic analysis of how different linguistic perturbations affect model performance across question-answering tasks. [7]

## 4.2. Implementation results

The evaluation results reveal significant performance disparities across different evaluation metrics and linguistic transformation aspects, highlighting the nuanced challenges in assessing semantic equivalence under controlled perturbations. We employed ROC-AUC as the primary evaluation criterion as it provides a threshold-independent measure of metric discriminative ability, quantifying how effectively each metric separates semantically preserved transformations (positive class) from meaning-altering perturbations (negative class) across all possible operating points. Moreover, AUC aggregates performance across the full spectrum of decision boundaries, making it particularly suitable for comparing evaluation metrics with heterogeneous score distributions.

Looking at the overall AUC rankings (Fig. 2), BERTScore variants dominate the top positions, suggesting that embedding-based approaches, particularly those leveraging contextualized representations from pre-trained language models, are substantially more effective at distinguishing between semantically preserved transformations and meaning-altering modifications compared to traditional n-gram overlap metrics. The strong performance of BERTScore_P specifically indicates that precision oriented semantic similarity measurement, may be particularly robust across diverse linguistic variations. The particularly poor performance of Token_Recall (0.4308), and Token_Accuracy (0.4344) at the bottom of the rankings demonstrates that metrics emphasizing exhaustive coverage or exact string matching are fundamentally inadequate for this evaluation paradigm, as they penalize valid linguistic variations that preserve core meaning while potentially failing to detect subtle semantic shifts.

The aspect-specific breakdown [8] exposes critical vulnerabilities in current evaluation methodologies. The Numbers aspect presents the most dramatic metric failures, with Token_F1 collapsing to 0.111, Token_Precision at 0.222, and Meaning_Accuracy_Emb plummeting to 0.111 (which has highest overall AUC). This catastrophic performance reflects a fundamental challenge: numerical perturbations (changing 100 to 101) create minimal surface-level disruption while substantially altering factual content, and most metrics fail to capture this sensitivity. Conversely, Faithfulness achieves its highest score (0.854) on this aspect, suggesting that specialized fact-checking or entailment-based metrics may be necessary for numerical accuracy assessment. The near-random performance of Context_Precision and Context_Recall (0.500) across most aspects indicates these metrics provide no discriminative power for semantic equivalence evaluation, essentially performing at chance level.

For Voice, Token_Precision achieves 0.811 while Token_Recall falls to 0.184, creating an huge F1 disparity. This pattern emerges because passive constructions typically expand sentence length and introduce auxiliary verbs, causing precision-oriented metrics to reward the retained content words while recall metrics penalize the structural expansion. Indirect Speech shows an even more extreme version of this phenomenon, with Token_Precision at 0.893 but Token_Recall at 0.094 (the lowest recall score across all aspects). The substantial word count increases and syntactic restructuring required for indirect speech conversion (adding "that" clauses, shifting pronouns, backshifting tenses) devastate recall-based measurements while preserving precision. BERTScore variants maintain relatively balanced performance on these aspects (0.740-0.846 range), demonstrating their superior handling of syntactic variation through semantic embedding space alignment rather than token alignment.

[7] Note: A more detailed explanation for the dataset can be found in the supplementary material: https://github.com/darkpheonix2/Metrics/blob/main/Supplemental_material.pdf

[8] For more details refer to the supplementary material

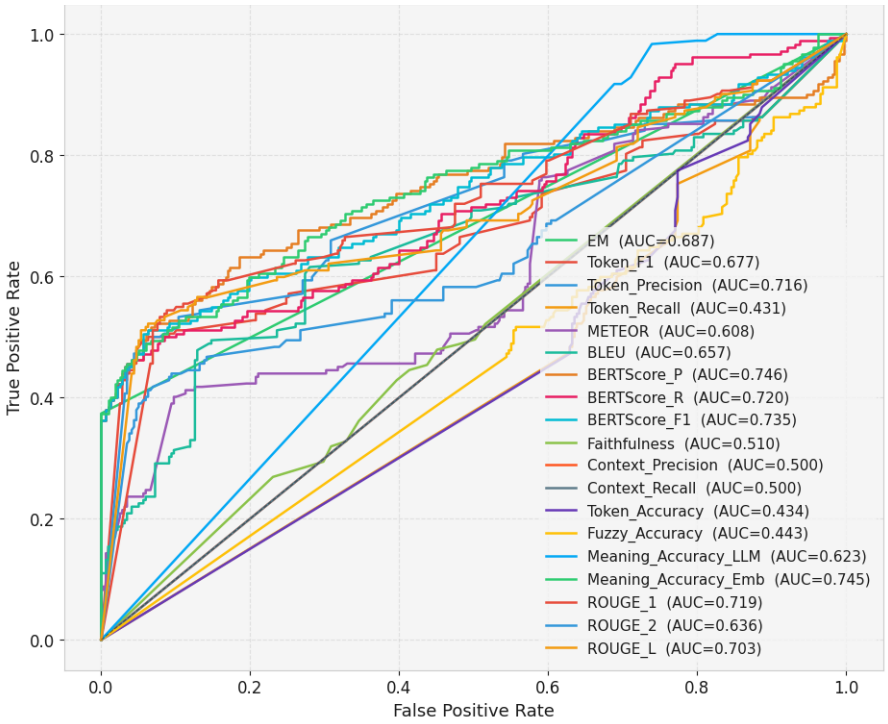


**Figure 2:** Overall ROC-AUC for multiple metrics

The Negation exposes fundamental limitations in embedding-based semantic similarity approaches. BERTScore_R achieves only 0.297 and BERTScore_P reaches merely 0.381, among the lowest scores for these metrics across any aspect. This dramatic failure occurs because negation flips the truth value of propositions while maintaining high lexical overlap and similar contextual embeddings. Interestingly, Meaning_Accuracy_LLM achieves its highest score (0.695) on this aspect, suggesting that large language model-based evaluators may possess better logical reasoning capabilities to detect polarity reversals compared to static embedding comparisons. Faithfulness also performs relatively well (0.542), indicating that entailment-based or contradiction-detection approaches may be necessary for robust negation handling.

The Synonyms and Paraphrasing demonstrate the most balanced metric performance, reflecting the semantic preservation goals of these transformations. Meaning_Accuracy_LLM achieves its peak performance on Synonyms (0.828), indicating that LLM-based evaluators excel when lexical substitution maintains propositional content. BERTScore_R reaches 0.771 on this aspect, substantially higher than its overall average, suggesting that recall-oriented semantic similarity is particularly effective when core concepts remain intact despite vocabulary changes. However, Fuzzy_Accuracy performs poorly on both aspects (0.331 and 0.324), demonstrating that approximate string matching fails to recognize valid paraphrases as semantically equivalent.

Morphology shows relatively consistent performance across BERTScore variants (0.648-0.675) and traditional metrics, with METEOR achieving 0.631 through its synonym-aware matching. Reordering demonstrates the critical importance of handling non-sequential similarity, with Token_Recall at a mere 0.406 while BERTScore maintains more robust performance (0.416-0.444). The Meaning_Accuracy_Emb score of 0.705 on Reordering suggests that sentence-level embedding averaging may partially compensate for word order changes, though this remains below its overall average.

These results collectively argue for a hybrid evaluation framework that combines complementary metric classes. The consistent superiority of BERTScore variants suggests that contextualized embedding similarity should form the foundation of semantic equivalence assessment. However, the specific failures (numerical sensitivity, negation detection, and extreme syntactic transformations) indicate that specialized modules are necessary. Fact-checking components for numerical accuracy, natural lan-

guage inference models for negation and contradiction detection, and potentially LLM-based judges for complex semantic reasoning would address the blind spots of embedding-based approaches. The complete inadequacy of the non-discriminative performance of context-based metrics (Context_Precision, Context_Recall) suggests these should be excluded from robust evaluation protocols. The dramatic precision-recall trade-offs observed in syntactic transformations also imply that F1 aggregation may obscure important metric behaviors, and that aspect-specific metric selection may ultimately prove more effective than universal metric application.

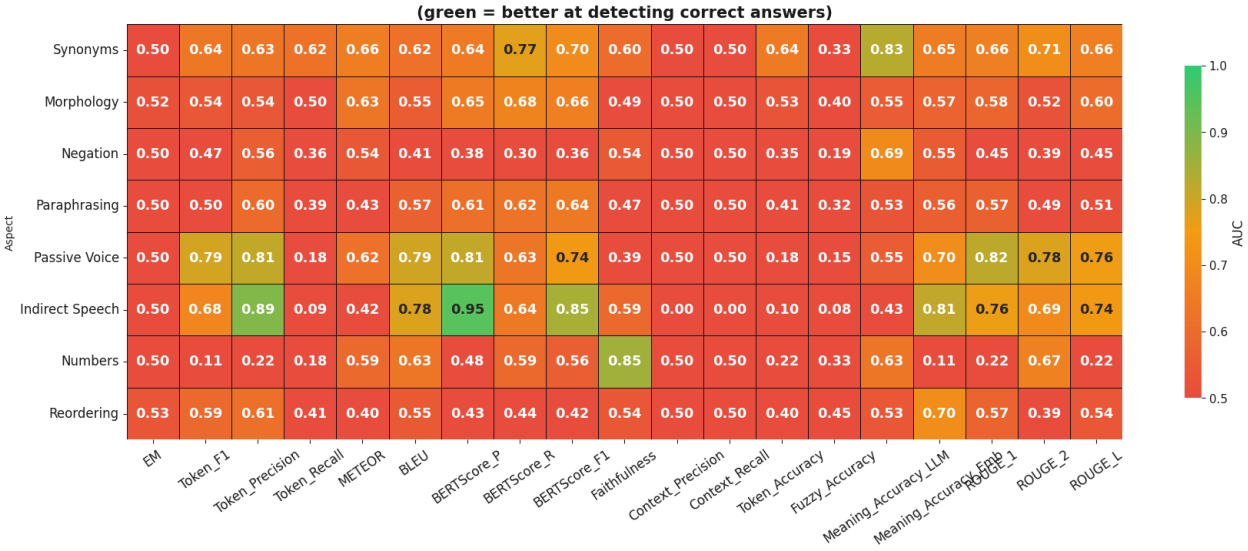


**Figure 3:** AUC per metric per aspect

### 4.3. Summary

This section aimed to answer RQ2. Based on the overall results especially as observed in Fig. 3, it can be stated that no existing evaluation criteria can fully assess the quality of the generated answer. However, combinations of Token_Precision, BLEU, Meaning_Accuracy_LLM, and Meaning_Accuracy_Emb can be used to capture the correctness of the generated response, as each of these metrics shows an AUC of less than or equal to 0.5 in only one aspect. The combination should be designed so that the weakness of one metric is compensated by the strength of another. For example, employing both Meaning_Accuracy_LLM and Meaning_Accuracy_Emb, and computing their harmonic mean, could help capture all aspects of correctness. This is because each metric compensates for the other's limitations (Meaning_Accuracy_LLM underperforms on Direct-Indirect Speech, while Meaning_Accuracy_Emb) struggles with Numerics. Additionally, Faithfulness also serves as a good metrics when measuring the grounding as this allows to check the numeric facts from the context properly.

## 5. Discussion and directions

This section aims to highlight the overall gaps in the existing evaluation metrics and frameworks, targeting RQ3, along with providing directions for future evaluation.

**Factual Correctness and Metric Limitations** Our empirical analysis reveals a fundamental tension in automated semantic equivalence assessment: while embedding-based metrics like BERTScore demonstrate superior overall robustness across diverse linguistic transformations, they exhibit critical blind spots precisely where surface-level similarity diverges from semantic fidelity. Most notably, these metrics fail catastrophically on numerical perturbations and negation flips, where Meaning_Accuracy_LLM demonstrates complementary strengths.

**The Inadequacy of Single-Metric Approaches** The near-chance performance of Context_Precision and Context_Recall across all aspects indicates that document-level relevance metrics fail to capture sentence-level semantic preservation, while the bottom-tier performance of exact matching metrics (Token_Accuracy 0.434) confirms that string similarity is fundamentally inadequate for evaluating linguistically diverse but semantically equivalent outputs. Most critically, the numeric aspect,as also supported by Oberst et al. [46],exposes a dangerous failure mode where minimal surface disruption causes most metrics to maintain moderate scores despite substantial factual alteration, with only Faithfulness (0.854) successfully detecting this semantic shift.

**Semantic Depth versus Scalability** Traditional NLP metrics demonstrate strong scalability and dataset agnosticism but suffer from limited semantic coverage, as they primarily rely on n-gram overlap or embedding similarity rather than genuine understanding of meaning. Conversely, LLM-dependent metrics offer deeper semantic evaluation, yet they sacrifice computational efficiency and transparency, operating as "black box" evaluators that require substantial human effort for prompt engineering and validation. This dichotomy creates a problematic trade-off for practitioners: systems requiring rigorous semantic validation must accept higher latency, infrastructure costs, and reduced interpretability, while scalable solutions remain confined to surface-level lexical comparisons that may miss critical semantic errors in generated content.

**Reference Dependency and Contextual Adaptability** While classical information retrieval metrics (MRR, MAP, NDCG) and traditional NLP scores require ground-truth references that are often expensive to curate, reference-free alternatives such as latency, indexing time, and certain RAGAS metrics eliminate this bottleneck but provide no absolute quality assurance. Furthermore, the context agnosticism observed in our analysis exposes architectural vulnerabilities, methods like LCS and throughput metrics fail to adapt across varying contextual requirements, whereas semantic similarity approaches maintain flexibility but embed hidden assumptions about vector space geometry that may not generalize across domains.

**Structural Gaps in Current Evaluation Frameworks** A significant structural blind spot exists in how current frameworks conceptualize the RAG pipeline, treating retrieval and generation as discrete black boxes rather than interconnected systems where indexing quality fundamentally constrains downstream performance. This methodological homogeneity implies that existing frameworks may be optimizing locally optimal solutions while missing global system interactions that determine real-world deployment efficacy. The absence of rigorous indexing evaluation represents a critical oversight, as retrieval quality serves as the foundational bottleneck for all subsequent generation tasks.

## 5.1. Problems and Future Directions

**Combined Pipeline Evaluation:** Our analysis indicates an urgent need for combined evaluation pipelines that consider RAG at both the fundamental level and as an integrated framework to assess overall efficacy. This necessitates the use of multiple complementary metrics for verifying factual correctness beyond semantic and surface-level checks. Such multi-metric approaches could be further explored in task-specific domains, where the appropriateness of individual metrics may vary based on application requirements.

**Reasoning Path Evaluation:** Few existing metrics adequately address the evaluation of reasoning processes in LLMs, highlighting the need for concepts involving traceability of retrieval graphs, coherence verification, step-by-step faithfulness assessment, utilization metrics [28], and explicit reasoning scores.

**Confidence Calibration and Comparative Evaluation:** As suggested by Saad et al. [8], incorporating confidence intervals alongside point estimates could strengthen evaluation reliability. Furthermore, given that LLMs demonstrate superior performance in comparative tasks versus absolute scoring [12, 24], Likert-scale-based evaluation frameworks may offer more robust assessment than direct score generation.

**Harmlessness and Guardrails:** Beyond correctness evaluation, minimal attention has been devoted to guardrails that filter content based on user demographics such as age or qualification. Comprehensive

RAG evaluation should include testing for societal impact, bias propagation, and harm potential across diverse user populations [5].

## 6. Conclusion

This study presents a comprehensive analysis of evaluation metrics for Retrieval-Augmented Generation systems, demonstrating that no single metric class suffices for robust semantic equivalence assessment. Our empirical findings reveal critical blind spots in embedding-based approaches for numerical and negation perturbations, the inadequacy of exact matching metrics for linguistically diverse outputs, and fundamental trade-offs between scalability and semantic depth across traditional and LLM-based evaluators. These results underscore the necessity of hybrid evaluation frameworks that combine complementary metric classes, embedding-based similarity for general robustness, specialized fact-verification for numerical accuracy, and LLM-based judges for complex reasoning, while highlighting urgent open problems including combined pipeline evaluation, reasoning path traceability, and harmlessness assessment that demand attention in future RAG evaluation research.

For reproducibility of the research, all the data, the supplementary material and the analysis can be found in the GitHub link[9]: https://github.com/darkpheonix2/Metrics

## 7. Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT, Grammarly in order to: Grammar and spelling check, Paraphrase and reword. After using these tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

## 8. Acknowledgments

This publication has emanated from research conducted with the financial support of Taighde Éireann – Research Ireland under Grant No. 18/CRT/6223. This work was also supported by Canada Research Chair Program.

## References

[1] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, H. Wang, H. Wang, et al., Retrieval-augmented generation for large language models: A survey, arXiv preprint arXiv:2312.10997 2 (2023) 32.

[2] S. Es, J. James, L. E. Anke, S. Schockaert, 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, 2024, pp. 150–158.

[3] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, Advances in neural information processing systems 33 (2020) 9459–9474.

[4] L. Brehme, T. Ströhle, R. Breu, Can llms be trusted for evaluating rag systems? a survey of methods and datasets, in: 2025 IEEE Swiss Conference on Data Science (SDS), IEEE, 2025, pp. 16–23.

[5] Honest, harmless, helpful evals, https://www.trulens.org/getting_started/core_concepts/honest_harmless_helpful_evals, 2024. Accessed: 2026-02-28.

[6] D. Ru, L. Qiu, X. Hu, T. Zhang, P. Shi, S. Chang, C. Jiayang, C. Wang, S. Sun, H. Li, et al., Ragchecker: A fine-grained framework for diagnosing retrieval-augmented generation, Advances in Neural Information Processing Systems 37 (2024) 21999–22027.

[9] https://github.com/darkpheonix2/Metrics

[7] K. Zhu, Y. Luo, D. Xu, Y. Yan, Z. Liu, S. Yu, R. Wang, S. Wang, Y. Li, N. Zhang, et al., Rageval: Scenario specific rag evaluation dataset generation framework, in: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2025, pp. 8520–8544.

[8] J. Saad-Falcon, O. Khattab, C. Potts, M. Zaharia, Ares: An automated evaluation framework for retrieval-augmented generation systems, in: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), 2024, pp. 338–354.

[9] B. Goodrich, V. Rao, P. J. Liu, M. Saleh, Assessing the factual accuracy of generated text, in: proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, 2019, pp. 166–175.

[10] I. Imtiyaz, M. Parvaz, S. Mandha, F. Farooq, A. K. Kolankeh, Assessing rag: A comprehensive review of evaluation frameworks, in: 2025 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE), IEEE, 2025, pp. 490–495.

[11] Y. Lyu, Z. Li, S. Niu, F. Xiong, B. Tang, W. Wang, H. Wu, H. Liu, T. Xu, E. Chen, Crud-rag: A comprehensive chinese benchmark for retrieval-augmented generation of large language models, ACM Transactions on Information Systems 43 (2025) 1–32.

[12] A. Celikyilmaz, E. Clark, J. Gao, Evaluation of text generation: A survey, arXiv preprint arXiv:2006.14799 (2020).

[13] T. Ito, K. van Deemter, J. Suzuki, Reference-free evaluation metrics for text generation: A survey, arXiv preprint arXiv:2501.12011 (2025).

[14] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, et al., A survey on evaluation of large language models, ACM transactions on intelligent systems and technology 15 (2024) 1–45.

[15] D. Hamzic, M. Wurzenberger, F. Skopik, M. Landauer, A. Rauber, Evaluation and comparison of open-source llms using natural language generation quality metrics, in: 2024 IEEE International Conference on Big Data (BigData), IEEE, 2024, pp. 5342–5351.

[16] S. Joshi, Evaluation of large language models: Review of metrics, applications, and methodologies (2025).

[17] M. Llugiqi, F. J. Ekaputra, M. Sabou, From experts to llms: Evaluating the quality of automatically generated ontologies, in: 2nd Workshop on Evaluation of Language Models in Knowledge Engineering (ELMKE), co-located with ESWC-25, to appear. URL: https://ceur-ws. org, volume 3977, 2025.

[18] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, C. Zhu, G-eval: Nlg evaluation using gpt-4 with better human alignment, in: Proceedings of the 2023 conference on empirical methods in natural language processing, 2023, pp. 2511–2522.

[19] S. Roychowdhury, S. Soman, H. Ranjani, N. Gunda, V. Chhabra, S. K. Bala, Evaluation of rag metrics for question answering in the telecom domain, arXiv preprint arXiv:2407.12873 (2024).

[20] E. Oro, F. M. Granata, A. Lanza, A. Bachir, L. De Grandis, M. Ruffolo, Evaluating retrieval-augmented generation for question answering with large language models., in: Ital-IA, 2024, pp. 12–17.

[21] S. Zhao, Y. Yang, Z. Wang, Z. He, L. K. Qiu, L. Qiu, Retrieval augmented generation (rag) and beyond: A comprehensive survey on how to make your llms use external data more wisely, arXiv preprint arXiv:2409.14924 (2024).

[22] Y. Hu, Y. Lu, Rag and rau: A survey on retrieval-augmented language model in natural language processing, arXiv preprint arXiv:2404.19543 (2024).

[23] B. Malin, T. Kalganova, N. Boulgouris, A review of faithfulness metrics for hallucination assessment in large language models, IEEE Journal of Selected Topics in Signal Processing (2025).

[24] G. Hannah, J. de Berardinis, T. R. Payne, V. Tamma, A. Mitchell, E. Piercy, E. Johnson, A. Ng, H. Rostron, B. Konev, Large language models as knowledge evaluation agents, in: 2nd Workshop on Evaluation of Language Models in Knowledge Engineering (ELMKE), co-located with ESWC-25, to appear. URL: https://ceur-ws. org, volume 3977, 2025.

[25] Y. Li, Y. Zhang, Question answering on squad 2.0 dataset, in: Department of Energy Resources Engineering and Department of Electrical Engineering, Stanford University, 2018.
[26] J. Liu, R. Ding, L. Zhang, P. Xie, F. Huang, Cofe-rag: A comprehensive full-chain evaluation framework for retrieval-augmented generation with enhanced data diversity, arXiv preprint arXiv:2410.12248 (2024).
[27] C. AI, Rag evaluation quickstart, 2026. URL: https://deepeval.com/docs/getting-started-rag, accessed: 2026-02-28.
[28] R. Friel, M. Belyi, A. Sanyal, Ragbench: Explainable benchmark for retrieval-augmented generation systems, arXiv preprint arXiv:2407.11005 (2024).
[29] S. Kukreja, T. Kumar, V. Bharate, A. Purohit, A. Dasgupta, D. Guha, Performance evaluation of vector embeddings with retrieval-augmented generation, in: 2024 9th international conference on computer and communication systems (ICCCS), IEEE, 2024, pp. 333–340.
[30] R. Singh, S. Singh, Text similarity measures in news articles by vector space model using nlp, Journal of The Institution of Engineers (India): Series B 102 (2021) 329–338.
[31] L. Caspari, K. G. Dastidar, S. Zerhoudi, J. Mitrovic, M. Granitzer, Beyond benchmarks: Evaluating embedding model similarity for retrieval augmented generation systems, arXiv preprint arXiv:2407.08275 (2024).
[32] J. Risch, T. Möller, J. Gutsch, M. Pietsch, Semantic answer similarity for evaluating question answering models, in: Proceedings of the 3rd Workshop on Machine Reading for Question Answering, 2021, pp. 149–157.
[33] R. S. Wahidur, S. Kim, H. Choi, D. S. Bhatti, H.-N. Lee, Legal query rag, IEEE Access (2025).
[34] Y. Hui, Y. Lu, H. Zhang, Uda: A benchmark suite for retrieval augmented generation in real-world document analysis, Advances in Neural Information Processing Systems 37 (2024) 67200–67217.
[35] M. Paterson, V. Dančík, Longest common subsequences, in: International symposium on mathematical foundations of computer science, Springer, 1994, pp. 127–142.
[36] A. Reynolds, F. Corrigan, Improving real-time knowledge retrieval in large language models with a dns-style hierarchical query rag, Authorea Preprints (2024).
[37] X. V. Li, F. Sanna Passino, Findkg: Dynamic knowledge graphs with large language models for detecting global trends in financial markets, in: Proceedings of the 5th ACM International Conference on AI in Finance, 2024, pp. 573–581.
[38] Y. Tang, Y. Yang, Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries, arXiv preprint arXiv:2401.15391 (2024).
[39] Y. Shi, H. Shang, X. Zi, W. Xu, Y. Feng, M. Xu, Answering narrative-driven recommendation queries via a retrieve–rank paradigm and the ocg-agent, in: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 2025, pp. 13192–13213.
[40] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.
[41] J. Chen, H. Lin, X. Han, L. Sun, Benchmarking large language models in retrieval-augmented generation, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, 2024, pp. 17754–17762.
[42] M. Hanna, O. Bojar, A fine-grained analysis of bertscore, in: Proceedings of the Sixth Conference on Machine Translation, 2021, pp. 507–517.
[43] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, arXiv preprint arXiv:1904.09675 (2019).
[44] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text summarization branches out, 2004, pp. 74–81.
[45] S. Banerjee, A. Lavie, 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 and/or summarization, 2005, pp. 65–72.
[46] D. Oberst, How to evaluate llms for rag?, Medium (2023). URL: https://medium.com/@darrenoberst/how-accurate-is-rag-8f0706281fd9, accessed: 2026-02-25.