Evaluating Retrieval-Augmented Generation for Question Answering with Large Language Models

Ermelinda Oro[1,2,*], Francesco Maria Granata[2], Antonio Lanza[2], Amir Bachir[2], Luca De Grandis[2] and Massimo Ruffolo[1,2]

[1]National Research Council, Institute for High Performance Computing and Networking, via P. Bucci 8/9C, Rende (CS), 87036, Italy
[2]Altilia srl, TechNest Start-up Incubator of University of Calabria, Piazza Vermicelli, Rende (CS), 87036, Italy

Abstract

We present a comprehensive framework for evaluating retrieval-augmented generation (RAG) systems designed for question-answering tasks using large language models (LLMs). The proposed framework integrates document ingestion, information retrieval, answer generation, and evaluation phases. Both ground truth-based and reference-free evaluation metrics are implemented to provide a multi-faceted assessment approach. Through experiments across diverse datasets like NarrativeQA and a proprietary financial dataset (FinAM-it), the reliability of existing metrics is investigated by comparing them against rigorous human evaluations. The results demonstrate that ground truth-based metrics such as BEM and RAGAS Answer Correctness exhibit a moderately strong correlation with human judgments. However, reference-free metrics still struggle to capture nuances in answer quality without predefined correct responses accurately. An in-depth analysis of Spearman correlation coefficients sheds light on the interrelationships and relative effectiveness of various evaluation approaches across multiple domains. While highlighting the current limitations of reference-free methodologies, the study underscores the need for more sophisticated techniques to better approximate human perception of answer relevance and correctness. Overall, this research contributes to ongoing efforts in developing reliable evaluation frameworks for RAG systems, paving the way for advancements in natural language processing and the realization of highly accurate and human-like AI systems.

Keywords

Retrieval Augmented Generation (RAG), Question Answering (QA), Retrieval, Large Language Model (LLM), Evaluation

1. Introduction

Retrieval-Augmented Generation (RAG) systems, which integrate information retrieval with natural language generation, have shown promise in enhancing language models' capabilities. However, evaluating their performance remains challenging, particularly when ground truth data is unavailable, impeding accurate assessments of system utility. To address this challenge, we present a comprehensive framework designed to facilitate the rigorous evaluation of RAG systems for question-answering tasks. Our framework integrates document ingestion, retrieval, generation, and evaluation phases, leveraging state-of-the-art technologies to optimize accuracy and relevance. We implement both ground truth-based and reference-free evaluation metrics, providing a multi-faceted approach to assessing system outputs. Through an extensive series of experiments spanning diverse domains and datasets we investigate the reliability and validity of existing evaluation methodologies. Specifically, we examine the correlation between various metrics and rigorous human evaluations, shedding light on their strengths, limitations, and potential for improvement. Our findings reveal that while ground truth-based metrics like BEM and RAG Answer Correctness exhibit moderate alignment with human judgments, reference-free metrics still struggle to accurately capture answer quality nuances without predefined correct responses. By analyzing Spearman correlation coefficients, we elucidate the interrelationships and relative effectiveness of different evaluation approaches across multiple domains.

This research makes the following key contributions: (i) presenting a comprehensive framework for evaluating RAG systems with state-of-the-art components, (ii) implementing and comparing diverse ground truth-based and reference-free evaluation metrics, (iii) conducting rigorous experiments across multiple datasets to assess metric reliability against human judgments, and (iv) analyzing the strengths and limitations of existing metrics, highlighting the need for advanced reference-free evaluation techniques that better approximate human perception.

The rest of the paper is organized as follows: Section 2 discusses related work. Section 3 presents the method. Section 4 shows the experimental evaluation and Section 5 concludes the work.

Ital-IA 2024: 4th National Conference on Artificial Intelligence, organized by CINI, May 29-30, 2024, Naples, Italy
*Corresponding author.
ermelinda.oro@icar.cnr.it (E. Oro);
francesco.granata@altiliagroup.com (F. M. Granata);
antonio.lanza@altiliagroup.com (A. Lanza);
amir.bachir@altiliagroup.com (A. Bachir);
luca.degrandis@altiliagroup.com (L. D. Grandis);
massimo.ruffolo@altiliagroup.com (M. Ruffolo)
0000-0002-5529-1007 (E. Oro); 0000-0003-4425-753X (F. M. Granata); 0000-0002-2875-4133 (L. D. Grandis); 0000-0002-4094-4810 (M. Ruffolo)

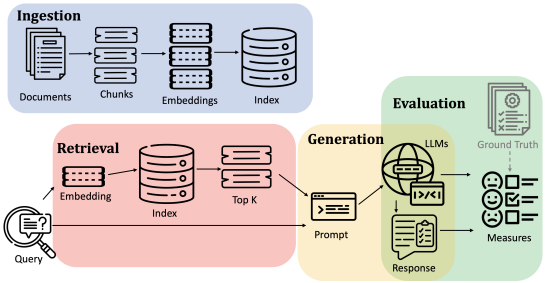

Figure 1: The simplified figure of the implemented RAG System.

2. Related Work

RAG systems have been implemented in various forms [1, 2, 3, 4, 5], incorporating advanced strategies like document splitting, chunking, retrieval, and diverse models for embedding and language generation, including proprietary and open-source models from platforms like HuggingFace[1]. We have also explored different variants of RAG systems, however, this paper's primary focus is not to introduce a novel RAG system or methodology but to comprehensively evaluate the effectiveness of Large Language Model (LLM)-derived metrics, emphasizing reference-free approaches.

Several prior works have proposed frameworks and novel metrics that leverage the capabilities of LLMs [6, 7, 8, 9, 10, 11]. Unlike these existing solutions, which aim to score different RAG systems or propose new evaluation methods, metrics, or datasets, our research is specifically targeted at evaluating the potential satisfaction of end-user customers who receive the evaluation scores generated by such systems.

By concentrating on the practical utility and interpretability of evaluation metrics from the perspective of end-users, our study diverges from the conventional approach of optimizing technical performance alone. Instead, we strive to bridge the gap between state-of-the-art evaluation techniques and the real-world expectations of customers who rely on these systems for decision-making and information retrieval.

3. Method

3.1. Framework for RAG and evaluation

This paper introduces a framework for running and evaluating a RAG system for efficiently processing and responding to natural language queries. The system integrates state-of-the-art technologies to enhance answer accuracy and relevance. The process is segmented into four main phases: **Ingestion**: Input documents are processed into manageable chunks, leveraging techniques like document layout analysis for PDFs. The chunks are embedded into high-dimensional vectors capturing their semantic essence and ingested into a vector store for efficient similarity search. **Retrieval**: Upon receiving a query, its vector form undergoes similarity search in the vector store to identify the k most relevant chunks. This narrows down the information to the most pertinent chunks for answer generation. **Generation**: A Large Language Model (LLM) synthesizes information from the retrieved chunks to construct a coherent and natural-sounding answer to the query. **Evaluation**: A two-sided approach employs both ground-truth dependent and independent metrics. Ground-truth dependent metrics assess correctness against predefined answers, while ground-truth independent metrics evaluate answer relevance without a predefined set. This dual approach enables a comprehensive assessment of performance, correctness, and overall text quality. The system can receive human evaluations of question-answer pairs to evaluate metric reliability and alignment with expectations.

3.2. Evaluation Strategies

In our RAG system, we implemented and tested a wide range of evaluation metrics. Specifically, our system incorporates metrics for assessing individual RAG components like Information Retrieval (IR) and Answer Generation, as well as the overall pipeline. For IR, we used classical metrics such as Recall@K, Precision@K, mAP, MRR, and nDCG. For answer generation, the implemented metrics were divided into two categories: Syntactic metrics evaluate formal response aspects, including BLEU [12], ROUGE [13], Precision, Recall, F1, and Exact Match [14]. These focus on text properties rather than semantic meaning. Semantic metrics evaluate response meaning, including BERT score [15] and BEM score [16]. BEM is preferred over BERT due to reported correlation with human evaluations and our empirical findings. **LMM-derived Metrics**: We implemented in our framework the RAG triad of metrics for the three main steps of an RAG's execution [6]: (i) Context relevance that assesses if the passage returned is relevant for answering the given query. (ii) Groundedness that assesses if the generated answer is faithful to the retrieved passage or if it contains hallucinated or extrapolated statements beyond the passage. (iii) Answer relevance that assesses if the generated answer is relevant given the query and retrieved passage. In addition, we implemented the Answer correctness that exploits LLMs and gold answers to measure the factual correctness of an answer. In this paper, only a subset of metrics are considered and compared for assessing the quality of the answers (see Section 4.2).

[1] https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

Manual evaluation. To verify the reliability of automated evaluation metrics, we implemented a rigorous manual evaluation process to assess the relevance, accuracy, and coherence of the answers generated by our RAG system. This manual evaluation was conducted by three independent human annotators, each with expertise in the domain of the questions posed to the system. For each evaluation session, the annotators were presented with the question, the corresponding answer generated by the RAG system, and the ground truth provided by the original dataset or the customer answers. The primary task for each annotator was to assess the quality of the generated answer in relation to the posed question, employing a discrete scoring 5-point likert scale. The criteria for scoring were as follows: 1. **Very Poor**: The generated answer is totally incorrect or irrelevant to the question. This case indicates a failure of the system to comprehend the query or retrieve pertinent information. 2. **Poor**: The generated answer is predominantly incorrect but with glimpses of relevance suggesting some level of understanding or appropriate retrieval. 3. **Neither**: The generated answer mixes relevant and irrelevant information almost equally, showcasing the system's partial success in addressing the query. 4. **Good**: The generated answer is largely correct but includes minor inaccuracies or irrelevant details, demonstrating a strong understanding and response to the question. 5. **Very Good**: Reserved for answers that are completely correct and fully relevant, reflecting an ideal outcome where the system accurately understood and responded to the query. The annotators conducted their assessments independently to ensure unbiased evaluations. Upon completion, the scores for each question-answer pair were collected and compared. In cases of discrepancy, a consensus discussion was initiated among the annotators to agree on the most accurate score. This consensus process allowed for mitigating individual bias and considering different perspectives in evaluating the quality of the generated answers. This manual evaluation process helps particularly in assessing the reliability and validity of our system's automated evaluation metrics. By comparing the human-generated scores against the results produced by these automated measures, we can determine the extent to which the automatic metrics accurately reflect human judgment and perception of answer quality.

4. Experiments

Considering different domains (Section 4.1), we investigate the reliability of a subset of existing metrics (Section 4.2) for evaluating a RAG system (Section 3.1). We explore the feasibility of adopting reference-free metrics and the correlation among them and the human evaluation (Section 3.2).

4.1. Datasets

NarrativeQA - English. A subsample of the NarrativeQA dataset [17] was used, with 50 book-related and 50 movie script-related questions (1% of the test set), spanning 41 unique books and 42 unique movie scripts. This allowed evaluating the RAG system's performance across two distinct narrative content types.

Financial Asset Management - Italian. The FinAM-it dataset, created by Altilia, consists of 50 question-answer pairs from Italian asset management documents on topics like investment strategies, risk management, and regulatory compliance. The questions are complex and diverse, often requiring information from multiple paragraphs, with detailed, conversational-style answers.

4.2. Metrics

Table 1
Naming and classification of metrics shown in the experimental evaluation

Acronym	Name - Framework	Type
BEM	BEM score - TensorFlow	GT-based
AR TruLens	Answer Relevance - TruLens	GT-free
AR RAGAS	Answer Relevance - RAGAS	GT-free
AC RAGAS	Answer Correctness - RAGAS	GT-based

In this paper we focus on evaluating the generated answer's quality of the entire pipeline.

In our analysis, we considered the **BEM score** (BERT matching score) [15], which we experimented is the most satisfying among the classic metrics. It is a metric that uses a BERT model [18] trained to solve an answer equivalence task, this task is solved by training a classifier that tells if two given answers are equivalent and returns the equivalence score. We use the variation of the BERT score *Answers and questions* that exploits the two answers and the question as model input. This variation results in performing better [16].

In addition, we considered novel LLM-derived metrics developed in the RAGAS [6] and Truelens[2] systems. These metrics offer evaluations both ground truth-based and reference-free. In particular, from RAGAS we used the two main metrics that focus on answers: Answer Correctness and Answer Relevance. More in detail: (i) **Answer Correctness**[3]: This metric measures the factual correctness of an answer and needs the presence of a ground truth. It employs an LLM to extract factual statements from both the predicted answer and the ground truth labeling them as True Positives if are present in both the answers, False Negatives if are present only in

[2]https://www.trulens.org/
[3]https://docs.ragas.io/en/latest/concepts/metrics/answer_correctness.html

the ground truth, and False Positives if they are present only in the prediction. Then a final F1 score is calculated, this score in the range $(0, 1)$ is the Answer Correctness. (ii) **Answer Relevance** [4]: This metric measures how pertinent the generated answer is to the prompt given to the LLM in the generation step. It computes a score in the range $(0, 1)$ as the mean of the cosine similarities between the original question and a set of artificial questions generated by an LLM on the basis of the predicted answer and the given context. The formula of the score is the following: $AnswerRelevance = \frac{1}{N}\sum_{i=1}^{N} cosine(E_o, E_{g_i})$ where E_o is the embedding of the original generated answer and E_{g_i} is the embedding of the i-th generated question. From TruLens we used the implemented Answer Relevance metric that prompts an LLM to evaluate the relevance of the answer with respect to the input prompt that includes context and question. The score that the LLM assigns to each answer is in the range $(0, 1)$.

To study the interrelationships and relative effectiveness among various evaluation metrics, we exploit the Spearman correlation coefficient. The **Spearman Rank Correlation** [19] is a non-parametric measure that assesses the statistical dependence between the rankings of two variables. It tells how well the relationship between these variables can be described using a monotonic function. This measure is computed on ranked data, allowing for the analysis of both ordinal variables and continuous variables that have been converted into ranks. The Spearman Rank Correlation coefficient is denoted by ρ, and its value ranges from -1 to 1 inclusive, where 1 indicates perfect positive correlation, 0 indicates no correlation, and -1 indicates perfect negative correlation.

4.3. Settings

For this implementation, we employed OpenAI models for the embedding, retrieval, and generation stages of the RAG and to implement evaluations with RAGAS and TruLens. The Ingestion step produced chunks of 1024 characters, balancing semantic integrity with avoiding irrelevant or redundant information. Larger chunks may capture more context but increase noise, while smaller sizes may sacrifice contextual information. These chunks were embedded using OpenAI's *text-embedding-ada-002*[5], a state-of-the-art transformer model for generating high-quality text embeddings. For retrieval within the vector store, the system identified the 10 most similar embeddings to previously indexed chunks. During generation, we employed the GPT-4-Turbo model[6] with the following prompt structure:

```
You are a chatbot having a
    conversation with a human.
Given the following extracted parts
    of a long document and a question,
     create a final answer.
If you don't know the answer, just
    say that you don't know, don't try
     to make up an answer.
Context: {CONTEXT}
Chat history: {CHAT_HISTORY}
Human: {HUMAN_INPUT}
Chatbot:
```

This prompt provided the model with instructions, context, and encouraged concise, truthful answers without fabrication.

4.4. Results

For both books and movies subsamples from the NarrativeQA dataset, as can be seen in table 2 and table 3, human judgment shows a moderately strong Spearman correlation with BEM (0.735 and 0.704) and AC RAGAS scores across both GPT-3.5-turbo (0.718, 0.792), and GPT-4-turbo models (0.67 and 0.781). This indicates that these ground truth-based metrics are more aligned with human perception of answer quality. Reference-free metrics show poor correlation with human judgment, especially AR RAGAS (0.234 and 0.483), highlighting the fact that evaluating an answer without ground truth is still a challenging problem for Large Language Models. The analysis of the FinAM-it dataset as it can be seen in table 4 shows generally lower correlations across all metrics, with the highest correlation being observed between human judgment and AC RAGAS gpt-4-turbo (0.531). This could be related to the fact that the FinAM-it dataset presents more challenging and diverse content that is more difficult to evaluate. Extending the analysis on all the datasets at once, it can be seen that all the metrics have still difficulties to approximate the human evaluation in a robust and reliable way.

5. Conclusion

Our exploration into evaluating Retrieval Augmented Generation (RAG) systems via ground truth-based and reference-free metrics was driven by the need for reliable evaluation frameworks, particularly for scenarios lacking ground truth data. Our evaluation framework's implementation has demonstrated its potential for facilitating a more comprehensive understanding of these systems' capabilities in such situations. Through rigorous experimentation across different domains and datasets, including NarrativeQA and a specialized industrial dataset, we

[4] https://docs.ragas.io/en/latest/concepts/metrics/answer_relevance.html
[5] https://openai.com/blog/new-and-improved-embedding-model
[6] https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo

Table 2
Spearman correlations on NarrativeQA books subsample

Metrics	Human Judgement	BEM	AR TruLens gpt-3.5-turbo	AR RAGAS gpt-3.5-turbo	AC RAGAS gpt-3.5-turbo	AR TruLens gpt-4-turbo	AR RAGAS gpt-4-turbo	AC RAGAS gpt-4-turbo
Human Judgement	1.000	**0.735**	0.436	0.234	**0.718**	0.420	0.150	**0.670**
BEM	**0.735**	1.000	0.185	0.224	**0.740**	0.405	-0.026	**0.713**
AR TruLens gpt-3.5-turbo	0.436	0.185	1.000	0.197	0.274	0.477	0.178	0.224
AR RAGAS gpt-3.5-turbo	0.234	0.224	0.197	1.000	0.129	0.156	**0.633**	0.121
AC RAGAS gpt-3.5-turbo	**0.718**	**0.740**	0.274	0.129	1.000	0.238	0.093	**0.854**
AR TruLens gpt-4-turbo	0.420	0.405	0.477	0.156	0.238	1.000	0.122	0.108
AR RAGAS gpt-4-turbo	0.150	-0.026	0.178	**0.633**	0.093	0.122	1.000	0.097
AC RAGAS gpt-4-turbo	**0.670**	**0.713**	0.224	0.121	**0.854**	0.108	0.097	1.000

Table 3
Spearman correlations on NarrativeQA movies subsample

Metrics	Human Judgement	BEM	AR TruLens gpt-3.5-turbo	AR RAGAS gpt-3.5-turbo	AC RAGAS gpt-3.5-turbo	AR TruLens gpt-4-turbo	AR RAGAS gpt-4-turbo	AC RAGAS gpt-4-turbo
Human Judgement	1.000	**0.704**	0.565	0.483	**0.792**	0.213	0.411	**0.781**
BEM	**0.704**	1.000	0.522	0.428	**0.752**	0.235	0.358	**0.746**
AR TruLens gpt-3.5-turbo	0.565	0.522	1.000	0.390	0.476	0.270	0.422	0.473
AR RAGAS gpt-3.5-turbo	0.483	0.428	0.390	1.000	0.403	0.406	**0.738**	0.421
AC RAGAS gpt-3.5-turbo	**0.792**	**0.752**	0.476	0.403	1.000	0.228	0.358	**0.977**
AR TruLens gpt-4-turbo	0.213	0.235	0.270	0.406	0.228	1.000	0.456	0.200
AR RAGAS gpt-4-turbo	0.411	0.358	0.422	**0.738**	0.358	0.456	1.000	0.379
AC RAGAS gpt-4-turbo	**0.781**	**0.746**	0.473	0.421	**0.977**	0.200	0.379	1.000

compared various evaluation methodologies against human judgment. While ground truth-based metrics like BEM and AC RAGAS showed moderate to strong correlation with human judgments across different domains and models, reference-free metrics still face significant challenges in achieving similar correlation levels. This highlights the current limitations of automated metrics in capturing nuanced aspects of human judgment, suggesting an urgent need for further refinement of reference-free evaluation methods. The Spearman correlation analysis reveals that while some metrics align more closely with human assessments, there is still significant room for improvement, especially for more challenging and diverse content like the FinAM-it dataset. These findings under-

Table 4
Spearman correlations on FinAM-it dataset

Metrics	Human Judgement	BEM	AR TruLens gpt-3.5-turbo	AR RAGAS gpt-3.5-turbo	AC RAGAS gpt-3.5-turbo	AR TruLens gpt-4-turbo	AR RAGAS gpt-4-turbo	AC RAGAS gpt-4-turbo
Human Judgement	1.000	0.208	0.178	0.153	0.053	0.280	0.230	0.531
BEM	0.208	1.000	0.214	0.209	0.276	0.001	0.203	0.278
AR TruLens gpt-3.5-turbo	0.178	0.214	1.000	0.412	0.433	0.181	0.446	0.300
AR RAGAS gpt-3.5-turbo	0.153	0.209	0.412	1.000	0.463	-0.191	0.608	0.130
AC RAGAS gpt-3.5-turbo	0.053	0.276	0.433	0.463	1.000	-0.099	0.243	0.255
AR TruLens gpt-4-turbo	0.280	0.001	0.181	-0.191	-0.099	1.000	-0.009	0.245
AR RAGAS gpt-4-turbo	0.230	0.203	0.446	0.608	0.243	-0.009	1.000	0.157
AC RAGAS gpt-4-turbo	0.531	0.278	0.300	0.130	0.255	0.245	0.157	1.000

Table 5
Spearman correlations on all datasets

Metrics	Human Judgement	BEM	AR TruLens gpt-3.5-turbo	AR RAGAS gpt-3.5-turbo	AC RAGAS gpt-3.5-turbo	AR TruLens gpt-4-turbo	AR RAGAS gpt-4-turbo	AC RAGAS gpt-4-turbo
Human Judgement	1.000	**0.627**	0.423	0.323	0.536	0.314	0.287	**0.653**
BEM	**0.627**	1.000	0.310	0.266	**0.654**	0.249	0.155	**0.711**
AR TruLens gpt-3.5-turbo	0.423	0.310	1.000	0.346	0.303	0.302	0.375	0.302
AR RAGAS gpt-3.5-turbo	0.323	0.266	0.346	1.000	0.213	0.201	**0.682**	0.198
AC RAGAS gpt-3.5-turbo	0.536	**0.654**	0.303	0.213	1.000	0.208	0.139	**0.813**
AR TruLens gpt-4-turbo	0.314	0.249	0.302	0.201	0.208	1.000	0.250	0.187
AR RAGAS gpt-4-turbo	0.287	0.155	0.375	**0.682**	0.139	0.250	1.000	0.169
AC RAGAS gpt-4-turbo	**0.653**	**0.711**	0.302	0.198	**0.813**	0.187	0.169	1.000

score the complexity of accurately evaluating RAG systems and the importance of considering domain-specific factors in metric development and selection. The observed limitations can have practical consequences, such as inaccurate system performance assessments, leading to suboptimal deployment decisions and reduced user satisfaction. Looking forward, our study emphasizes developing more nuanced and sophisticated evaluation frameworks that can better approximate human judgment. This entails improving existing metrics' accuracy and reliability and exploring new methodologies to effectively capture qualitative aspects of generated answers.

While our evaluation framework provides valuable insights, we acknowledge several limitations: (i) Current reference-free metrics still struggle to match human judgment, necessitating further refinement. (ii) Metric performance suffers for challenging, domain-specific datasets, highlighting the need for domain-aware or adaptive approaches. (iii) Our analysis covered a subset of available metrics; exploring a wider range, including leveraging advanced LLMs and additional context, is needed. (iv) Results should be validated across different RAG configurations and domains for broader applicability. (v) Despite rigorous human evaluation, inherent subjectivity and potential biases may have impacted findings. We view these limitations as opportunities to contribute to developing more reliable, accurate, and human-like evaluation frameworks that can drive advancements in natural language processing capabilities and the realization of highly effective RAG systems across diverse domains.

References

[1] K. Guu, K. Lee, Z. Tung, P. Pasupat, M. Chang, Retrieval augmented language model pre-training, in: ICML 2020, 13-18 July 2020, Virtual Event, volume 119 of *Proceedings of Machine Learning Research*, PMLR, 2020, pp. 3929–3938. URL: http://proceedings.mlr.press/v119/guu20a.html.

[2] O. Khattab, C. Potts, M. Zaharia, Relevance-guided supervision for openqa with colbert, 2021. arXiv:2007.00814.

[3] K. Shuster, S. Poff, M. Chen, D. Kiela, J. Weston, Retrieval augmentation reduces hallucination in conversation, 2021. arXiv:2104.07567.

[4] S. Huo, N. Arabzadeh, C. Clarke, Retrieving supporting evidence for generative question answering, in: SIGIR-AP, ACM, 2023, pp. 11–20. URL: http://dx.doi.org/10.1145/3624918.3625336. doi:10.1145/3624918.3625336.

[5] T. Zhang, S. G. Patil, N. Jain, S. Shen, M. Zaharia, I. Stoica, J. E. Gonzalez, Raft: Adapting language model to domain specific rag, 2024. arXiv:2403.10131.

[6] S. Es, J. James, L. Espinosa-Anke, S. Schockaert, Ragas: Automated evaluation of retrieval augmented generation, 2023. arXiv:2309.15217.

[7] Y. Tang, Y. Yang, Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries, 2024. arXiv:2401.15391.

[8] M. Gao, X. Hu, J. Ruan, X. Pu, X. Wan, Llm-based nlg evaluation: Current status and challenges, 2024. arXiv:2402.01383.

[9] Z. Zhang, M. Fang, L. Chen, Retrievalqa: Assessing adaptive retrieval-augmented generation for short-form open-domain question answering, 2024. arXiv:2402.16457.

[10] V. Katranidis, G. Barany, Faaf: Facts as a function for the evaluation of rag systems, 2024. arXiv:2403.03888.

[11] J. Saad-Falcon, O. Khattab, C. Potts, M. Zaharia, Ares: An automated evaluation framework for retrieval-augmented generation systems, 2024. arXiv:2311.09476.

[12] C.-Y. Lin, E. Hovy, Automatic evaluation of summaries using n-gram co-occurrence statistics, in: Human Language Technology Conference of the North American Chapter of the ACL, 2003, pp. 150–157. URL: https://aclanthology.org/N03-1020.

[13] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, ACL, Barcelona, Spain, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013.

[14] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, SQuAD: 100,000+ questions for machine comprehension of text, in: J. Su, K. Duh, X. Carreras (Eds.), EMNLP, ACL, Austin, Texas, 2016, pp. 2383–2392. URL: https://aclanthology.org/D16-1264. doi:10.18653/v1/D16-1264.

[15] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, 2020. arXiv:1904.09675.

[16] J. Bulian, C. Buck, W. Gajewski, B. Boerschinger, T. Schuster, Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation, 2022. arXiv:2202.07654.

[17] T. Kočiský, J. Schwarz, P. Blunsom, C. Dyer, K. M. Hermann, G. Melis, E. Grefenstette, The narrativeqa reading comprehension challenge, 2017. arXiv:1712.07040.

[18] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.

[19] K. F. Weaver, V. Morales, S. L. Dunn, K. Godde, P. F. Weaver, Pearson's and Spearman's Correlation, John Wiley and Sons, Ltd, 2017, pp. 435–471. doi:https://doi.org/10.1002/9781119454205.ch10.