Evaluation and Comparison of Open-Source LLMs Using Natural Language Generation Quality Metrics

Dzenan Hamzic, Markus Wurzenberger, Florian Skopik, Max Landauer
AIT Austrian Institute of Technology
Vienna, Austria
firstname.lastname@ait.ac.at

Andreas Rauber
TU Wien
Vienna, Austria
rauber@ifs.tuwien.ac.at

***Abstract*—The rapid advancement of Large Language Models (LLMs) has transformed natural language processing, yet comprehensive evaluation methods are necessary to ensure their reliability, particularly in Retrieval-Augmented Generation (RAG) tasks. This study aims to evaluate and compare the performance of open-source LLMs by introducing a rigorous evaluation framework. We benchmark 20 LLMs using a combination of established metrics such as BLEU, ROUGE, BERTScore, along with and a novel metric, RAGAS. The models were tested across two distinct datasets to assess their text generation quality. Our findings reveal that models like nous-hermes-2-solar-10.7b and mistral-7b-instruct-v0.1 consistently excel in tasks requiring strict instruction adherence and effective use of large contexts, while other models show areas for improvement. This research contributes to the field by offering a comprehensive evaluation framework that aids in selecting the most suitable LLMs for complex RAG applications, with implications for future developments in natural language processing and big data analysis.**

***Index Terms*—Large Language Model, Retrieval-Augmented Generation, Natural Language Generation Evaluation, LLM Benchmarking, LLM Evaluation**

I. Introduction

The rapid development of Large Language Models (LLMs) has led to significant advancements in natural language processing (NLP) and generation capabilities [1]–[3]. These models, trained on vast amounts of text data, have demonstrated remarkable proficiency in tasks ranging from text completion and summarization to question-answering and creative writing [4]. Despite these advancements, there remains a critical need to evaluate and compare the quality of generation output from diverse LLMs to ensure their reliability and accuracy in practical applications [5]–[7].

The importance of robust evaluation mechanisms for LLMs cannot be overstated. As these models become increasingly integrated into various domains such as healthcare, finance, and education, the potential impact of their outputs grows exponentially [8]. Inaccurate or biased generations could lead to misinformation, unfair decision-making, or even safety risks in critical applications [9]. Moreover, the complexity and opacity of LLMs make it challenging to predict or interpret their outputs, further emphasizing the need for comprehensive evaluation techniques [10].

Recent research has highlighted several key areas where LLM evaluation is crucial. These include assessing factual accuracy [11], and measuring the model's ability to understand and generate context-appropriate responses [12]. Additionally, as LLMs are often fine-tuned or adapted for specific tasks, there is a need to evaluate their performance in both general and domain-specific contexts [13].

This paper aims to address these evaluation needs by systematically assessing the generation quality of various LLMs using the RAG approach [14] and selected metrics. RAG is a technique that combines language model-based generation with information retrieval to enhance the quality and accuracy of text generation. By incorporating relevant information retrieved from external knowledge sources, RAG models are able to generate more coherent, factual, and contextually appropriate outputs compared to standalone language models. RAG combines the strengths of retrieval-based and generation-based methods, potentially offering a more robust framework for evaluating LLM outputs [15]. By incorporating external knowledge retrieval into the generation process, RAG allows for a more nuanced evaluation of how LLMs leverage and integrate information, which is crucial for assessing their real-world applicability [16].

Our study will employ a diverse set of evaluation metrics, including BLEU [17], ROUGE [18], and more recent metrics like BERTScore [19] and Faithfulness [20]. These metrics will be used to assess various aspects of generation quality.

By conducting this comprehensive evaluation, we aim to provide valuable insights into the strengths and limitations of different LLMs, contribute to the ongoing efforts in improving LLM evaluation methodologies. This research is timely and crucial as the field of AI continues to evolve rapidly, with new and more powerful language models being developed at an unprecedented pace [21].

This work aims to address two fundamental research objectives in the context of leveraging LLMs for RAG applications:

1) Selection of an appropriate open-source LLM for RAG Applications: One of the critical challenges in the development of RAG systems is the selection of an appropriate LLM. This research seeks to explore the criteria and methodologies that should guide the choice of an LLM for a specific RAG application. By examining various factors such as model size, context length, and performance on specific tasks, this study aims to provide

979-8-3503-6248-0/24/$31.00 ©2024 IEEE

a comprehensive framework for selecting an LLM for a RAG application.
2) Comparative analysis and ranking of LLM performance in RAG generation: The second objective of this research is to conduct a detailed comparative analysis of various LLMs within the context of RAG applications. This analysis explores how different LLMs perform in generating high-quality and contextually relevant responses when augmented with retrieval mechanisms. Furthermore, it involves a systematic ranking of these models that bases on a set of carefully selected metrics.

This paper makes three key contributions to the evaluation of LLMs[1]:

1) Comprehensive LLM Generation Evaluation: Utilizing the RAGAS framework [20] in combination with conventional NLP evaluation metrics, we systematically assess LLMs based on Faithfulness [20], Answer Relevance [20], Answer Similarity [20], Answer Correctness [20], BLEU [17], ROUGE [18], and BERTScore [19], providing a thorough analysis of their output quality.
2) Benchmarking open-source LLMs: We benchmark a number of popular open-source LLMs, selected from Replicate [22], including models like Llama 3-8B [23], Mixtral 8x7b instruct [24] and Llama 2-7B-Chat [25]. This process identifies reliable and high-performing models for various applications.
3) Analysis of Metric Correlations: This research analyzes the correlations between evaluation metrics across datasets, identifying metrics like BERTScore and ROUGE as consistent, while others, like Faithfulness and Answer Relevancy, show low correlations. These findings highlight the importance of selecting metrics based on the specific task and dataset.

Consequently, this research will contribute to the broader understanding of LLM performance in terms of text generation, providing valuable insights for researchers, developers, and practitioners in the field of NLP and artificial intelligence (AI). The findings will guide the selection and deployment of LLMs in various applications, helping to improve the likelihood that these models produce faithful, relevant, and accurate responses.

The remainder of this paper is structured as follows: Sect. 2 reviews current evaluation methods for language models, focusing on the limitations of traditional metrics and the advantages of using LLMs as automated evaluators. Sect. 3 describes the selection of models, datasets, and evaluation metrics. Sect. 4 presents performance results, model rankings, and visual analyses, highlighting key strengths and weaknesses across different metrics. Sect. 5 summarizes findings and suggests future research directions.

[1]https://github.com/dzenanh/Evaluation-and-Comparison-of-Open-Source-LLMs-Using-Natural-Language-Generation-Quality-Metrics

II. Background and related work

Evaluating the performance of large language models (LLMs) is a complex task, particularly in the field of Natural Language Processing (NLP), due to the intricate nature of human language. Traditional metrics like BLEU and ROUGE, which are the most frequently reported NLP performance metrics [26], originally designed for translation and summarization tasks, have shown limited applicability and low correlation with human judgment across broader NLP tasks [26]. To address these limitations, recent approaches have explored using LLMs as evaluators, a method known as "LLM-as-a-judge."

The "LLM-as-a-judge" approach offers significant advantages in scalability and cost-effectiveness. Automated evaluations using strong LLMs like GPT-4 reduce the need for human involvement, allowing for the efficient handling of large datasets. This method also achieves high agreement with human judgments, matching human-level concordance in many cases, thus serving as a reliable and objective substitute for human judges [27]. Additionally, LLM judges provide detailed feedback and can mitigate human biases, enhancing the transparency and objectivity of evaluations. These evaluations, conducted with LLM judges, are reproducible, assuming deterministic LLM behavior, thus providing consistent and reliable results.

This method is particularly valuable in evaluating open-ended tasks, where traditional benchmarks often fall short. The "LLM-as-a-judge" approach complements existing benchmarks by focusing on human-centric evaluation metrics, ensuring a comprehensive assessment of LLM capabilities [27]. For instance, the RAGAs framework [20] integrates this approach within RAG systems, automating the evaluation of generated content based on predefined criteria.

Recent studies further emphasize the challenges in evaluating RAG systems. Chen et al. [28] developed the RAG Benchmark (RGB) to evaluate four key RAG capabilities: noise resilience, rejection of irrelevant information, integration of retrieved knowledge, and robustness against counterfactuals. Their findings highlight significant limitations in current LLMs, underscoring the need for continued research.

Yu et al. [29] propose a Unified Evaluation Process for RAG, emphasizing the need for benchmarks that balance retrieval accuracy and generative quality. Their survey identifies gaps in current methods, particularly the inadequacy of existing metrics for evaluating faithfulness and accuracy in generation, and the need for more diverse and comprehensive datasets that reflect real-world scenarios—both of which are addressed in this research. The paper suggests future research directions to enhance the effectiveness of RAG system evaluations.

Similarly, Friel et al. [30] introduced RAGBench, a large-scale benchmark designed to address evaluation challenges in RAG systems across various industry domains. Their findings reveal that evaluations of relevance, faithfulness, and correctness conducted using LLMs often yield less accurate

assessments compared to those conducted using fine-tuned models like DeBERTa, highlighting the need for more reliable evaluation methods.

Incorporating these insights, the integration of "LLM-as-a-judge" within RAG frameworks not only enhances the scalability and reliability of evaluations but also ensures that LLM-generated content aligns with human preferences, making these models more practical and effective in real-world scenarios.

III. Methodology

This section explains how we selected and evaluated the language models used in our study. We start by describing the criteria for choosing models from the Replicate [22] platform, ensuring a diverse and cost-effective selection. Next, we outline the data and settings used to test these models, making sure the evaluation is fair and consistent.

A. LLM Choosing Criteria

Open-source LLMs were chosen from the cloud-based platform Replicate, a startup that leverages cloud technology to run machine learning models. Replicate stands out for its simplicity and developer-friendly approach, offering a wide range of pre-trained models with easy customization and deployment, all within a flexible pay-per-use pricing model. Its strong support for open-source models, built-in version control, and straightforward API integration make it an attractive choice for developers and researchers who want to experiment with or deploy AI models without dealing with complex infrastructure management. We selected and utilized all 40 recommended open-source LLMs available on Replicate[2], ranging from 130 million to 70 billion parameters, ensuring a robust and comprehensive set for evaluation.

Alternative approaches to using Replicate include:

1) Running models from Hugging Face[3] locally: This option involves downloading and running the models directly on your own hardware. While it offers full control over the models and their configurations, it requires a powerful computer with substantial processing power and memory, especially for handling large models with up to 70 billion parameters. This setup can be challenging for those without access to high-end hardware.
2) Running models from Hugging Face on major cloud providers (AWS, Azure, GCP): This approach leverages the scalability and flexibility of cloud computing to run models. However, it requires specialized expertise to set up and manage the cloud infrastructure effectively. Tasks such as configuring virtual machines, managing storage, and optimizing costs can be complex and time-consuming, making this option more suitable for teams with cloud experience.

After selecting the LLMs from Replicate, the models were further refined by checking if they were evaluated on six key benchmarks (IFEval [32], BBH [34], MATH [35], GPQA [36], MuSR [37], and MMLU-PRO [38]) using the Eleuther AI Language Model Evaluation Harness [31]. This unified framework tests generative language models on a wide range of evaluation tasks, ensuring that the selected models have undergone rigorous testing across diverse scenarios. Only the models that were tested with all six benchmarks were selected, which was crucial to ensure that the final set of models not only met the initial selection criteria but also demonstrated strong performance across important evaluation benchmarks. This sub-selection process led to a final set of 20 Open-Source LLM (Table I) models for evaluation.

The selected LLMs vary widely in their characteristics, offering a diverse set of options in terms of model size, context handling capabilities, and performance across key benchmarks. The models range from 2 billion to 70 billion parameters, with context sizes spanning from 8,000 to 32,000 tokens. Performance on benchmarks such as IFEval, BBH, and MMLU-PRO highlights the models' varying strengths, with some excelling in specific tasks like mathematical reasoning (MATH Level 5) or factual correctness (GPQA). This diversity enables a comprehensive evaluation across different NLP tasks, ensuring that the chosen models provide robust and reliable insights for a broad range of applications.

We operate under the assumption that if an LLM is trained or fine-tuned on a specific type of dataset, it will deliver comparably good results when tested on that same type of dataset. For example, if a model is trained with a mathematics dataset, it is expected to perform better on a mathematics test compared to models that were neither trained nor fine-tuned on such datasets. With this in mind, we considered the six key benchmarks from the Eleuther AI Language Model Evaluation Harness [31] as the defining characteristics or properties of an LLM.

1) IFEval [32] is a dataset specifically created to assess a model's capability to adhere to clear instructions, such as "incorporate keyword x" or "utilize format y." The primary emphasis is on the model's compliance with formatting directives rather than the substance of the output, enabling the application of precise and stringent evaluation metrics [33].
2) BBH (Big Bench Hard) [34] consists of 23 tasks from the BigBench dataset, designed to evaluate language models using objective metrics. These tasks include multistep arithmetic, algorithmic reasoning, language comprehension, and general knowledge. Performance on BBH aligns closely with human preferences, providing valuable insights into model capabilities [33].
3) MATH [35] is a collection of high-school level competition problems compiled from various sources and consistently formatted using LaTeX for equations and Asymptote for figures. The generated content must adhere to a precise output format. Only the most challenging level 5 questions from MATH were retained, referred to as MATH Lvl 5 [33].
4) GPQA [36] (Graduate-Level Google-Proof Q&A Bench-

[2]https://replicate.com/collections/language-models
[3]https://huggingface.co/LLMs

mark) is a knowledge dataset with questions created by PhD-level experts in biology, physics, and chemistry. Designed to be difficult for non-experts, the dataset has undergone rigorous validation to ensure complexity and accuracy. Access to GPQA is tightly controlled to prevent data contamination, and plain text examples are not shared in adherence to the authors' guidelines [33].

5) MuSR [37] (Multistep Soft Reasoning) is a novel dataset composed of algorithmically generated problems, each approximately 1,000 words long. The problems encompass murder mysteries, object placement challenges, and team allocation optimizations. To solve these problems, models must combine reasoning with the ability to parse long-range context. Most models perform only slightly better than random chance on this dataset [33].
6) MMLU-PRO [38] (Massive Multitask Language Understanding - Professional) is an enhanced version of the MMLU dataset, traditionally used for multiple-choice knowledge assessments. Recent studies highlighted problems in the original MMLU, such as noisy data with some unanswerable questions and a reduction in difficulty due to advancements in model capabilities and increased data contamination. MMLU-Pro addresses these concerns by providing 10 answer choices instead of 4, incorporating more reasoning-based questions, and undergoing expert review to minimize noise. Consequently, MMLU-Pro is of superior quality and presents a greater challenge compared to the original dataset [33].

In all of these evaluations, a higher score indicates better performance, with scores ranging from 0 to 100.

B. Data and Settings

In this work, two Question-and-Answer datasets are utilized, each consisting of 50 pairs of questions and answers.

The first dataset (DSA), as described in [39], originally comprised 107 question-answer pairs generated using ChatGPT-4. The authors ensured that the generation process followed specific criteria to maintain technical precision, provide a sufficient challenge, and align with potential user inquiries directed towards a RAG system. To ensure that the evaluation data accurately reflect the performance of RAG techniques in real-world scenarios, each Q&A pair underwent human inspection and review to validate its relevance and accuracy. These pairs were derived from a subset of 13 papers included in a dataset containing 423 selected research papers focused on the topics of AI and LLMs, drawn from arXiv [40]. From this dataset, 50 Question-Answer pairs were randomly selected.

The PDF documents were processed using Langchain's file-loader[4], which, by default, chunks the documents by page. These chunks were then converted into embedding vectors using SBERT's "multi-qa-MiniLM-L6-cos-v1" model [41] during the process of loading the documents into the Chroma vector database.

[4]https://sj-langchain.readthedocs.io/en/latest/document_loaders/langchain.document_loaders.pdf.PyPDFDirectoryLoader.html

```
template = """Answer the question
based only on the following context:
{context}

Question: {question} """
```

Fig. 1. RAG Prompt Template.

For the second dataset (DSB), a literature search was conducted using Langchain's document-loader[5], an ArXiv[6] API wrapper. The query "security penetration testing" was used, resulting in 200 documents being loaded. This topic was intentionally chosen to differ from the AI-related first dataset, ensuring a broader evaluation scope. To ensure that none of the LLMs used in this work were trained on the selected documents, papers released after April 18, 2024 (the release date of the Llama 2 models), were chosen. This filtering resulted in a final selection of 11 papers. From these, a Q&A dataset with 60 items was generated using state-of-the-art RAGAs Q&A generation capabilities, with GPT-4o as the LLM for generation. The final 50 Q&A pairs were randomly selected.

The textual data from these documents was divided into chunks of size 2000 characters and an overlap of 200. The applied separators have been "\n\n", "\n", "(?<=\.)", and " ". . The resulting document chunks were then converted into embeddings and stored in the Chroma vector database. The prompt template (Fig. 1) is formatted as follows:

The `context` variable is constructed by concatenating the content of the 2 documents most similar to the question. The contents of these documents are joined together with "\n\n" between each page.

GPT-4o was used as the "LLM-as-a-Judge" in this experiment. The LLM was initialized with temperature 0, which sets the randomness of the predictions to a minimum, and top_p 1, which ensures that the model considers the entire probability distribution for generating the next token. The models under test were initialized with the following parameters: temperature 0.1, max_length 500, and top_p 1.

C. Evaluation Metrics

The evaluation framework used in this research consists of RAGAs, BLEU, ROUGE, and BERTScore. We selected these metrics because they collectively provide a balanced assessment of both surface-level accuracy (BLEU, ROUGE) and deeper semantic alignment (BERTScore), while RAGAs specifically evaluates the faithfulness, relevance, and correctness of generated answers in RAG tasks. These metrics were chosen for their proven effectiveness in capturing both traditional and advanced aspects of text quality, avoiding others due to their narrower focus or lower relevance to the comprehensive evaluation needed for this study.

[5]https://python.langchain.com/v0.2/docs/integrations/document_loaders/arxiv/

[6]arxiv.org

TABLE I
COMPARISON OF DIFFERENT MODELS ACROSS VARIOUS EVALUATION METRICS.

Model	# Parameter(billions)	Context Size (thousands)	IFEval	BBH	MATH Lvl 5	GPQA	MUSR	MMLU-PRO
gemma-2b	2.0	8.0	30.88	8.25	2.72	0.67	7.56	4.06
gemma-2b-it	2.0	8.0	26.90	5.21	0.45	3.80	3.03	3.92
gemma-7b	7.0	8.0	26.59	21.12	6.42	4.92	10.98	21.64
gemma-7b-it	7.0	8.0	38.68	11.94	1.59	4.59	12.53	7.72
llama-2-13b-chat	13.0	4.0	45.62	4.91	0.98	1.10	3.56	10.26
llama-2-70b-chat	70.0	4.0	49.58	4.61	0.91	0.56	4.17	15.92
llama-2-7b-chat	7.0	4.0	39.65	4.49	0.64	0.56	3.48	7.52
meta-llama-3-70b	70.0	8.0	48.71	16.54	19.69	16.01	21.27	41.21
meta-llama-3-70b-instruct	70.0	8.0	80.99	50.19	23.34	4.92	10.92	46.74
meta-llama-3-8b	8.0	8.0	14.55	24.50	3.25	7.38	6.24	24.55
meta-llama-3-8b-instruct	8.0	8.0	28.24	28.24	8.69	7.61	6.60	29.60
mistral-7b-instruct-v0.1	7.0	8.0	45.02	13.79	1.51	0.00	5.77	15.34
mistral-7b-instruct-v0.2	7.0	8.0	52.94	22.91	6.04	3.47	7.61	19.08
mistral-7b-v0.1	7.0	8.0	38.26	22.02	2.49	5.59	10.68	22.36
mixtral-8x7b-instruct-v0.1	56.0	32.0	53.45	9.71	12.11	3.60	11.14	19.86
nous-hermes-2-solar-10.7b	10.7	8.0	52.79	19.54	5.21	5.82	13.82	27.31
qwen1.5-14b	14.0	32.0	29.05	30.06	16.47	5.93	10.46	30.06
qwen1.5-7b	7.0	32.0	26.84	23.08	4.46	6.49	9.16	21.29
yi-34b-chat	34.0	4.0	46.99	37.62	4.31	11.74	8.36	34.37
yi-6b	6.0	4.0	28.93	19.41	1.51	2.57	7.04	22.12
yi-6b-chat	6.0	8.0	33.95	17.00	0.68	5.93	3.57	22.92

For the generation component in RAG, RAGAs includes a Faithfulness metric (1) to measure hallucinations and an Answer Relevancy metric (2) to assess the relevance of the answers to the questions. BLEU and ROUGE are also utilized to provide additional conventional NLP performance metrics.

Faithfulness (F) evaluates how factually consistent the generated answer is with the provided context. It is derived from both the answer and the retrieved context, and the results are scaled to a range of $(0, 1)$, where higher values indicate better performance [42].

The faithfulness of the answer $a_s(q)$ to the context $c(q)$, where q denotes the question, as the condition where the statements made in the answer can be logically derived from the context. To measure faithfulness, an LLM is first used to extract a set of statements, $S(a_s(q))$, from the answer. This process simplifies longer sentences into shorter, more precise assertions [20].

In other words, an answer is considered faithful if every claim made within it can be substantiated by the context. To assess this, a list of claims from the generated answer is compiled. Each claim is then verified against the provided context to determine whether it can be logically derived from it. This involves evaluating whether the claim is a direct or reasonably inferred conclusion from the information in the context, taking into account both explicit details and implicit connections that are clearly supported by the context [42].

Faithfulness score [42], F, is defined in equation (1):

$$F = \frac{|V|}{|S|}, \tag{1}$$

where, $|V|$ is the number of statements that were supported according to the LLM and $|S|$ is the total number of statements.

The metric **Answer Relevance (AR)** (2) measures the appropriateness of the generated answer to the presented prompt. Answers that are partial or include extraneous information receive lower scores. This metric is calculated based on the question, the context, and the answer, and is defined as the mean cosine similarity between the original question and a number of artificial questions, which are reverse-engineered based on the answer [42].

$$\text{AR} = \frac{1}{N} \sum_{i=1}^{N} cos(E_{G_i}, E_O), \tag{2}$$

where, E_{G_i} represents the embedding of the generated question i, E_O denotes the embedding of the original question, and N refers to the number of generated questions, which is set to 3 by default [42].

An answer is considered relevant when it directly and appropriately responds to the original question. The assessment of relevance does not focus on factual accuracy but rather penalizes answers that are incomplete or contain unnecessary details. To compute the relevance score, the LLM is prompted multiple times to generate suitable questions for the given answer. The mean cosine similarity between these generated questions and the original question is then measured. The rationale is that if the generated answer accurately addresses the initial question, the LLM should produce questions from the answer that closely match the original question [42].

Answer Similarity (AS) metric is determined through the following three-step process:

1) Convert the ground truth answer into a vector using the designated embedding model.
2) Transform the generated answer into a vector using the same embedding model.
3) Calculate the cosine similarity between these two vectors.

Mathematically, the metric can be represented by the equation (3):

$$\text{AS} = cos(A_{gt}, A_G), \tag{3}$$

where, A_{gt} is the ground truth answer converted to embedding, and A_G is the embedding of the generated answer [42].

The notion of AS involves evaluating how closely the generated answer aligns semantically with the ground truth. This assessment results in a score ranging from 0 to 1, where a higher score indicates a closer match between the generated answer and the ground truth. Assessing semantic similarity provides insights into the quality of the generated response. This evaluation, along with the AR, employs an embedding-based approach to determine the semantic similarity score [42].

Answer Correctness (AC) involves evaluating the accuracy of the generated answer in relation to the ground truth. This assessment produces scores ranging from 0 to 1, where a higher score signifies a closer match between the generated answer and the ground truth, indicating greater correctness.

AC includes two key components: semantic similarity and factual similarity between the generated answer and the ground truth. These components are integrated using a weighted scheme to determine the answer correctness score.

AC is determined by combining factual accuracy with the semantic similarity between the provided answer and the ground truth. Factual accuracy measures the factual agreement between the generated answer and the ground truth. This measurement is based on the following concepts:

- TP (True Positive): Facts or statements that are found in both the ground truth and the generated answer.
- FP (False Positive): Facts or statements that appear in the generated answer but are absent in the ground truth.
- FN (False Negative): Facts or statements that are present in the ground truth but missing from the generated answer.

In equation (4), we now employ the F1 score formula to measure correctness.

$$\text{F1 Score} = \frac{|TP|}{(|TP| + 0.5 \cdot (|FP| + |FN|))} \tag{4}$$

The **RAGAS Score** in this work is calculated using the equally weighted sum of its individual components (F, AR, AS and AC), which assess the generation components of the RAG system.

$$\text{RAGAS} = \frac{1}{4}\left(F + AR + AS + AC\right), \tag{5}$$

where, F is the Faithfulness score, AR is the Answer Relevance score, AS is the Answer Similarity score, and AC is the Answer Correctness score.

The **BLEU** (Bilingual Evaluation Understudy) is the most frequently used NLP-specific metric in research [26]. It evaluates the similarity between a generated sentence and a reference sentence. The BLEU score ranges from 0, indicating a complete mismatch, to 1, indicating a perfect match. This metric was developed specifically for the evaluation of automated machine translations. Moreover, it correlates highly with human evaluation [17]. The calculation of the BLEU score (7) also requires the inclusion of a brevity penalty (BP) (6).

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}, \tag{6}$$

where, r is the reference corpus length, and c is the translation length.

Then,

$$\text{BLEU} = \text{BP} \cdot e^{\sum_{n=1}^{N} w_n \log p_n}, \tag{7}$$

where, w_n are the weights assigned to each n-gram precision, summing to one, and p_n is the n-gram precision.

Achieving a perfect score of 1 in translation evaluations is rare, unless the translation exactly matches the reference. As a result, even human translators are unlikely to achieve a score of 1 [17]. In this research, the BLEU score uses 4-grams with equally assigned weights, i.e., 25% for each n-gram.

The **ROUGE** (Recall-Oriented Understudy for Gisting Evaluation) metric (8) is a commonly used tool for evaluating the quality of machine-generated text, such as summaries, translations, or other forms of text generation. ROUGE works by comparing the words or sequences of words in the generated text to those in a reference or "gold standard" text. It measures how much of the reference text's content is captured by the generated text, focusing on recall, which means it looks at how much of the important information from the reference is included in the output [18]. ROUGE-L is often chosen for evaluating machine-generated text because it focuses on the Longest Common Subsequence (LCS), which is what the "L" stands for, between the generated text and the reference text. This is particularly useful in tasks like summarization and translation, where the structure and order of information are important. ROUGE-L considers both precision and recall, but it places emphasis on recall, which means it captures how much of the important information from the reference text is included in the output [43].

$$\text{ROUGE-L} = F_\beta = \frac{(1+\beta^2) \cdot \text{R}_{LCS} \cdot \text{P}_{LCS}}{\text{R}_{LCS} + \beta^2 \cdot \text{P}_{LCS}}, \tag{8}$$

where, $\text{R}_{LCS} = \frac{\text{LCS}(X,Y)}{|Y|}$ and $\text{P}_{LCS} = \frac{\text{LCS}(X,Y)}{|X|}$.

Here, $\text{LCS}(X,Y)$ is the length of the longest common subsequence between the reference text Y and the generated text X, R_{LCS} is the recall of the LCS, and P_{LCS} is the precision of the LCS. β is a parameter that determines the relative importance of recall and precision (typically, $\beta = 1$, meaning equal importance is given to both).

BERTScore, introduced by Zhang et al. (2020) [19] is a text generation evaluation metric that leverages pre-trained BERT embeddings to measure the similarity between generated text and reference text. Unlike conventional metrics that rely on exact n-gram matching, BERTScore uses contextualized word embeddings to capture semantic meaning, allowing for a more nuanced evaluation. It computes precision (9), recall (10), and

F_1-score (11) by comparing the embeddings of the candidate and reference sentences, making it particularly effective for tasks like summarization and translation [19].

$$\text{Precision} = \frac{1}{|C|} \sum_{x \in C} \max_{y \in R} \cos(x, y), \tag{9}$$

where, C is the set of candidate tokens, R is the set of reference tokens, and $\cos(x, y)$ is the cosine similarity between the BERT embeddings of token x from the candidate and token y from the reference.

$$\text{Recall} = \frac{1}{|R|} \sum_{y \in R} \max_{x \in C} \cos(y, x) \tag{10}$$

where R is the set of reference tokens, C is the set of candidate tokens, and $\cos(y, x)$ is the cosine similarity between the BERT embeddings of token y from the reference and token x from the candidate.

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{11}$$

where, Precision (9), and Recall (10) are calculated as shown above.

IV. Evaluation

In this section, we evaluate the performance of selected LLMs across multiple datasets. The evaluation includes analyzing metric correlations, ranking the LLMs, and comparing the top and bottom performers. Each step is detailed in the following subsections.

A. Correlation Analysis of Evaluation Metrics

In the first step, we analyzed the correlations between various evaluation metrics across two datasets (Fig. 2), DSA and DSB introduced in Sect. III-B, to assess the consistency and reliability of these metrics in evaluating machine-generated text. The correlation analysis revealed several strong relationships, indicating that the metrics are highly consistent both across datasets and within individual datasets.

Faithfulness exhibited a very strong positive correlation between DSA and DSB ($r = 0.91$), suggesting that the faithfulness scores are remarkably stable across different datasets. This consistency indicates that the metric is reliable for evaluating the faithfulness of generated text, regardless of the dataset used.

Answer Correctness also showed a strong correlation between DSA and DSB ($r = 0.87$), which implies that the correctness of answers is consistently assessed across different datasets. Additionally, there is a high correlation between answer correctness and answer similarity within both DSA ($r = 0.75$) and DSB ($r = 0.86$). This suggests that when an answer is deemed correct, it is also likely to be similar to the reference answers, highlighting a close relationship between these two evaluation dimensions.

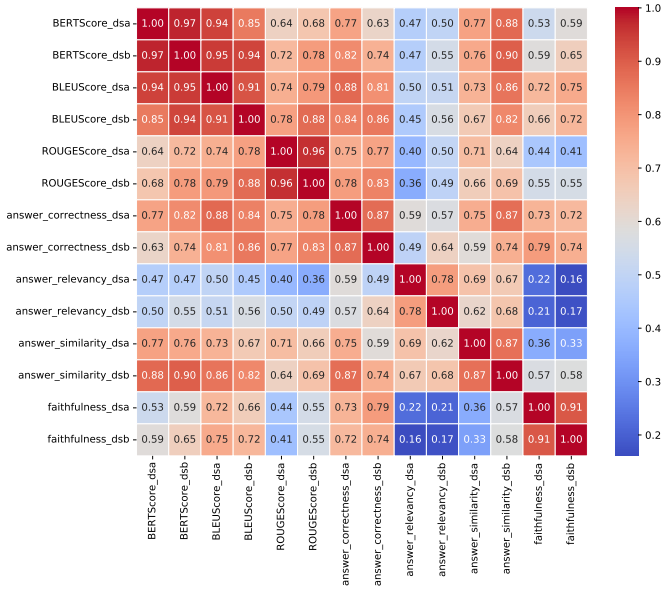

Fig. 2. Correlation across all metrics for the two datasets DSA and DSB.

Answer Similarity maintained a strong positive correlation between DSA and DSB ($r = 0.87$), further confirming the consistency of this metric across datasets. This consistency is crucial for ensuring that the similarity between generated and reference answers is evaluated uniformly across different contexts.

The **BLEU, ROUGE, and BERTScore** metrics, which are commonly used for evaluating the quality of text generation, also exhibited strong correlations across datasets. BLEU scores between DSA and DSB showed a correlation of $r = 0.85$, while ROUGE scores were even more consistent with a correlation of $r = 0.96$. BERTScore, which leverages contextual embeddings, demonstrated the highest consistency across datasets with a correlation of $r = 0.97$. Moreover, within DSB, ROUGE and BERTScore were highly correlated ($r = 0.94$), indicating that these two metrics often agree on the quality of the generated text.

These strong correlations suggest that the evaluated metrics are robust and reliable indicators of text generation quality, providing consistent results across different datasets.

However, not all metrics showed strong correlations across the datasets, highlighting potential sensitivities to specific dataset characteristics. For instance, the correlation between **Faithfulness** and **Answer Relevancy** in DSA was notably low ($r = 0.22$), indicating that a model's ability to produce factually accurate content does not necessarily align with its ability to generate relevant responses. This trend is consistent in DSB, where the correlation between Faithfulness and Answer Relevancy remains low ($r = 0.17$). These low correlations suggest that models may excel in ensuring factual accuracy while struggling to maintain relevance in their responses, or vice versa.

Additionally, **Answer Relevancy** itself showed gener-

TABLE II
LLM PERFORMANCE RANKINGS ACROSS METRICS - DSA.

Model	BLEU	ROUGE	BERT	RAGAS	Average
nous-hermes-2-solar-10.7b	0.038 (1)	0.231 (2)	0.872 (5)	0.677 (1)	**2.25**
qwen1.5-14b	0.036 (2)	0.219 (4)	0.875 (1)	0.646 (3)	**2.50**
mistral-7b-instruct-v0.1	0.032 (4)	0.246 (1)	0.873 (2)	0.639 (5)	**3.00**
qwen1.5-7b	0.032 (4)	0.216 (6)	0.873 (2)	0.649 (2)	**3.50**
gemma-2b-it	0.034 (3)	0.218 (5)	0.873 (2)	0.601 (8)	**4.50**
mixtral-8x7b-instruct-v0.1	0.031 (6)	0.221 (3)	0.866 (6)	0.640 (4)	**4.75**
meta-llama-3-70b-instruct	0.030 (7)	0.207 (8)	0.865 (7)	0.629 (6)	**7.00**
mistral-7b-instruct-v0.2	0.023 (9)	0.216 (6)	0.862 (8)	0.560 (10)	**8.25**
llama-2-70b-chat	0.026 (8)	0.203 (10)	0.859 (9)	0.610 (7)	**8.50**
llama-2-13b-chat	0.016 (10)	0.173 (16)	0.847 (11)	0.590 (9)	**11.50**
llama-2-7b-chat	0.016 (10)	0.178 (14)	0.849 (10)	0.525 (13)	**11.75**
gemma-7b	0.013 (12)	0.205 (9)	0.842 (12)	0.481 (17)	**12.50**
gemma-2b	0.011 (13)	0.197 (11)	0.834 (15)	0.493 (16)	**13.75**
yi-34b-chat	0.010 (14)	0.128 (18)	0.842 (12)	0.558 (11)	**13.75**
meta-llama-3-8b	0.007 (17)	0.189 (12)	0.817 (17)	0.556 (12)	**14.50**
yi-6b	0.006 (18)	0.180 (13)	0.810 (20)	0.514 (14)	**16.25**
meta-llama-3-70b	0.008 (16)	0.177 (15)	0.814 (19)	0.508 (15)	**16.25**
mistral-7b-v0.1	0.010 (14)	0.151 (17)	0.841 (14)	0.392 (20)	**16.25**
yi-6b-chat	0.004 (19)	0.088 (19)	0.833 (16)	0.478 (18)	**18.00**
gemma-7b-it	0.002 (20)	0.025 (20)	0.815 (18)	0.459 (19)	**19.25**

TABLE III
LLM PERFORMANCE RANKINGS ACROSS METRICS - DSB.

Model	BLEU	ROUGE	BERT	RAGAS	Average
mistral-7b-instruct-v0.1	0.201 (1)	0.405 (1)	0.896 (1)	0.716 (2)	**1.25**
meta-llama-3-70b-instruct	0.187 (2)	0.397 (2)	0.890 (3)	0.710 (3)	**2.50**
nous-hermes-2-solar-10.7b	0.179 (3)	0.379 (4)	0.890 (3)	0.726 (1)	**2.75**
gemma-2b-it	0.150 (4)	0.385 (3)	0.892 (2)	0.662 (7)	**4.00**
qwen1.5-7b	0.126 (6)	0.332 (6)	0.883 (5)	0.672 (5)	**5.50**
mixtral-8x7b-instruct-v0.1	0.136 (5)	0.345 (5)	0.879 (7)	0.671 (6)	**5.75**
qwen1.5-14b	0.112 (7)	0.325 (7)	0.882 (6)	0.693 (4)	**6.00**
mistral-7b-instruct-v0.2	0.097 (9)	0.314 (9)	0.875 (8)	0.629 (8)	**8.50**
llama-2-70b-chat	0.101 (8)	0.301 (11)	0.871 (9)	0.590 (10)	**9.50**
gemma-7b	0.081 (10)	0.316 (8)	0.861 (10)	0.557 (14)	**10.50**
llama-2-13b-chat	0.072 (12)	0.270 (13)	0.861 (10)	0.593 (9)	**11.00**
gemma-2b	0.081 (10)	0.313 (10)	0.852 (13)	0.554 (15)	**12.00**
llama-2-7b-chat	0.064 (13)	0.259 (15)	0.856 (12)	0.540 (16)	**14.00**
yi-34b-chat	0.042 (14)	0.185 (18)	0.852 (13)	0.580 (11)	**14.00**
meta-llama-3-70b	0.042 (14)	0.256 (16)	0.833 (17)	0.566 (13)	**15.00**
yi-6b	0.034 (18)	0.284 (12)	0.825 (19)	0.572 (12)	**15.25**
mistral-7b-v0.1	0.042 (14)	0.216 (17)	0.847 (15)	0.450 (20)	**16.50**
meta-llama-3-8b	0.038 (17)	0.267 (14)	0.831 (18)	0.527 (18)	**16.75**
yi-6b-chat	0.014 (19)	0.117 (19)	0.840 (16)	0.499 (19)	**18.25**
gemma-7b-it	0.003 (20)	0.046 (20)	0.825 (19)	0.540 (16)	**18.75**

ally lower correlations with other metrics, particularly with **BERTScore** ($r = 0.47$ in DSA and $r = 0.55$ in DSB) and **ROUGE** ($r = 0.36$ in DSA and $r = 0.49$ in DSB), suggesting that relevance may be influenced by factors not fully captured by these traditional metrics.

B. Ranking of Language Models

Next, we determine which LLMs performed best across both datasets. To achieve this, we ranked the LLMs according to their BLEU, ROUGE, BERT, and RAGAS scores, and then calculated the average rank of these ranked scores to determine the overall performance. Table II displays the ranking for the DSA dataset and Tab. III displays the results for the DSB dataset.

The Spearman's rank correlations are 0.93 for BLEU, 0.94 for ROUGE, 0.95 for BERT, and 0.91 for RAGAS, all with p-values of 0.0000. The strong Spearman's rank correlations across BLEU, ROUGE, BERT, and RAGAS metrics confirm the consistency of performance rankings between DSA and DSB. High-performing models in DSA tend to excel in DSB, while lower-ranked models consistently underperform, highlighting the robustness of the evaluation and the need for targeted improvements.

The analysis reveals that certain LLMs consistently perform well across both datasets, securing top positions in the rankings. Notably, the models **nous-hermes-2-solar-10.7b**, **mistral-7b-instruct-v0.1**, **qwen1.5-7b**, and **gemma-2b-it** consistently appear in the top five, demonstrating their robustness and reliability across multiple evaluation metrics.

Conversely, some models consistently rank lower across the datasets. Specifically, **yi-6b**, **mistral-7b-v0.1**, **yi-6b-chat**, and **gemma-7b-it** frequently appear in the bottom five, indicating a need for further optimization to improve their performance relative to the other models evaluated.

C. Comparative Analysis of Top and Bottom Performers

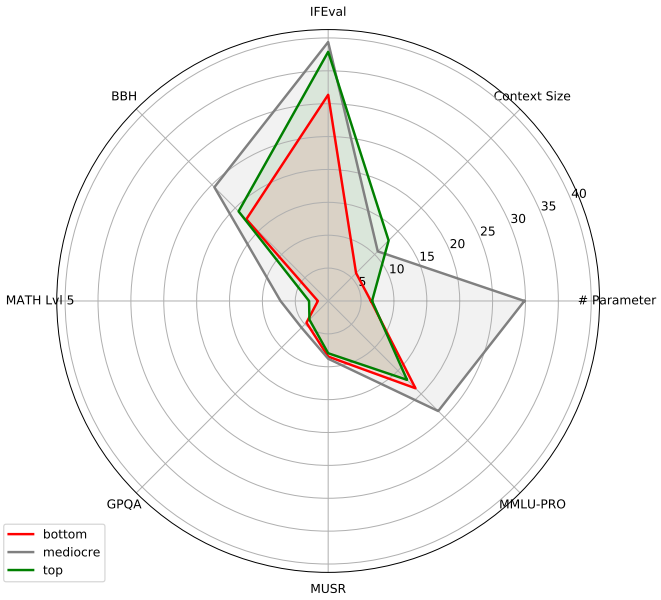

Fig. 3. Top vs. bottom performers with LLM Characteristics.

The radar plot in Fig. 3 illustrates the performance across multiple evaluation benchmarks, including IFEval, BBH, MATH Level 5, GPQA, MUSR, and MMLU-PRO, as well as LLM characteristics, such as Context Size and the number of parameters, for top, mediocre, and bottom-performing models. The top models, marked by the green area, demonstrate balanced performance, with noticeable strengths in IFEval and Context Size. These strengths suggest their proficiency in adhering to strict instructions and managing large contexts, making them suitable for complex RAG tasks.

The mediocre performing models, depicted in grey, show strengths in specific external benchmark tasks, such as IFEval and BBH, where larger parameter sizes are beneficial for complex reasoning or handling large contexts. However, when evaluated on our datasets (DSA and DSB) using key metrics like BLEU, ROUGE, BERT, and RAGAS, these models underperform compared to the top models. This suggests that while these models excel in specific external benchmarks, they do not consistently deliver high-quality text generation across the broader range of tasks reflected in our datasets. Consequently, these models can be seen as expert performers

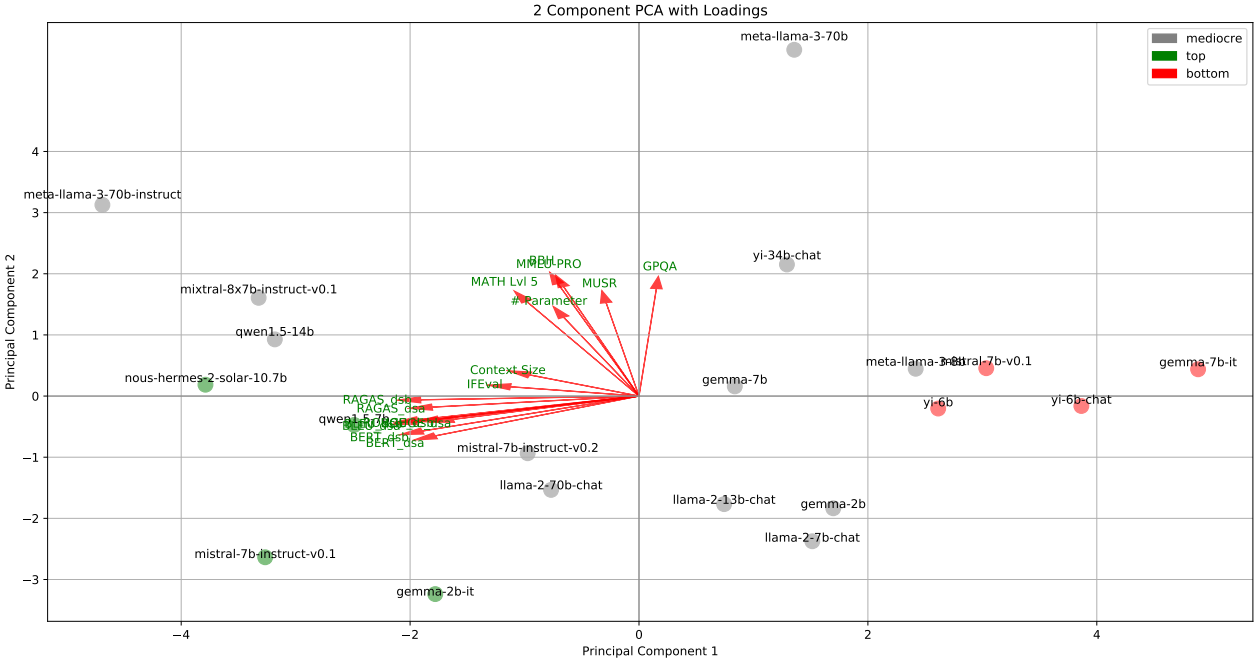

Fig. 4. Principal Component Analysis.

in their specialized domains but less general in their applicability, which limits their classification as top performers in the broader evaluations conducted in this study.

The bottom performers, shown in red, lag behind, especially in MATH Level 5 and BBH, highlighting their struggles with mathematical reasoning and broader comprehension tasks. The clear gap in performance across metrics underscores the challenges these models face in complex scenarios.

In a 2-component PCA with loadings (Fig. 4), each axis (Principal Component 1 and Principal Component 2) represents a combination of original variables that maximize the variance in the data. The loadings for each variable (represented by arrows) indicate how much each original variable contributes to the principal components. Variables that align closely with the axes of the principal components have a strong influence on those components. In this biplot, the loadings for metrics such as IFEval, Context Size, BLEU, ROUGE, BERT, and RAGAS point in similar directions, indicating that these metrics are positively correlated and contribute similarly to the principal components. Conversely, variables like MATH Lvl 5, number parameter, and BBH align more closely with Component 2, suggesting that they capture different aspects of model performance. Notably, the PCA explained variance is 0.72, indicating that these two principal components account for 72% of the total variance in the data, which provides a robust summary of the relationships between the evaluation metrics and model characteristics.

Top-performing models spread more across both Principal Component 1 and Principal Component 2, indicating that they excel across a wider range of performance dimensions. In contrast, bottom performers are more clustered along the second component, with most variation occurring only along the first component, reflecting their more limited and inconsistent performance. Interestingly, IFEval, though linked to instruction adherence, aligns more closely with the first group of variables, indicating that LLMs excelling in IFEval and handling large contexts are well-suited for producing high-quality RAG outputs.

V. Conclusion and Future Work

This study introduces a framework for evaluating open-source LLMs in RAG tasks, using metrics like BLEU, ROUGE, BERTScore, and the new RAGAS metric. By analyzing 20 LLMs across two datasets, we found that models like nous-hermes-2-solar-10.7b and mistral-7b-instruct-v0.1 consistently performed well, making them ideal for tasks requiring strong overall performance. In contrast, models like yi-6b and gemma-7b-it struggled, especially in more complex tasks, suggesting they are better suited for simpler applications.

Although this study didn't directly test instruction-following or context management, the strong performance of the top models suggests they handle these tasks well. The RAGAS metric, which evaluates multiple aspects of text quality, offers a more comprehensive assessment. However, it needs further testing to become a widely accepted standard.

While ethical issues like bias in LLM outputs are important, this study didn't address them directly. Future research should include evaluations of these concerns, along with other factors like energy efficiency related to model size. Expanding the evaluation framework to include more diverse datasets will also help tailor LLMs to specific tasks more effectively.

ACKNOWLEDGMENTS

Funded by the European Union under GA no. 101121403 (NEWSROOM) and GA no. 101121418 (EUCINF). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them. Co-funded by the Austrian FFG Kiras project ASOC (905301).

REFERENCES

[1] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Barnes, and A. S. Mian, "A comprehensive overview of large language models," *ArXiv*, abs/2307.06435, 2023.

[2] T. B. Brown et al., "Language models are few-shot learners," *ArXiv*, abs/2005.14165, 2020.

[3] C. Raffel, N. M. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, pp. 140:1–140:67, 2019.

[4] A. Chowdhery et al., "PaLM: Scaling language modeling with pathways," *J. Mach. Learn. Res.*, vol. 24, pp. 240:1–240:113, 2022.

[5] H. Wang, S. Zhao, Z. Qiang, B. Qin, and T. Liu, "Beyond the answers: Reviewing the rationality of multiple choice question answering for the evaluation of large language models," *ArXiv*, abs/2402.01349, 2024.

[6] M. T. Ribeiro, T. S. Wu, C. Guestrin, and S. Singh, "Beyond accuracy: Behavioral testing of NLP models with CheckList," in *Proc. Annu. Meet. Assoc. Comput. Linguist.*, 2020.

[7] S. Gehrmann, E. Clark, and T. Sellam, "Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text," *J. Artif. Intell. Res.*, vol. 77, pp. 103-166, 2023.

[8] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" in *Proc. 2021 ACM Conf. Fairness, Accountability, and Transparency*, 610–623, 2021.

[9] L. Weidinger et al., "Ethical and social risks of harm from language models," *ArXiv*, abs/2112.04359, 2021.

[10] M. Danilevsky et al., "A survey of the state of explainable AI for natural language processing," in *Proc. 1st Conf. Asia-Pacific Chapter Assoc. Comput. Linguist. and 10th Int. Joint Conf. Natural Language Process.*, K.-F. Wong, K. Knight, and H. Wu, Eds., 447–459, Suzhou, China: Assoc. Comput. Linguist., 2020. [Online]. Available: https://aclanthology.org/2020.aacl-main.46.

[11] W. Kryscinski et al., "Evaluating the factual consistency of abstractive text summarization," in *Proc. 2020 Conf. Empirical Methods in Natural Language Process. (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds., 9332–9346. Online: Assoc. Comput. Linguist., 2020. [Online]. Available: https://aclanthology.org/2020.emnlp-main.750. [Online]. Available: https://doi.org/10.18653/v1/2020.emnlp-main.750.

[12] D. Khashabi et al., "UnifiedQA: Crossing format boundaries with a single QA system," *ArXiv*, abs/2005.00700, 2020.

[13] S. Gururangan et al., "Don't stop pretraining: Adapt language models to domains and tasks," *ArXiv*, abs/2004.10964, 2020.

[14] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," *ArXiv*, abs/2005.11401, 2020.

[15] S. Borgeaud et al., "Improving language models by retrieving from trillions of tokens," in *Proc. Int. Conf. Mach. Learn.*, 2021.

[16] F. Petroni et al., "KILT: A benchmark for knowledge intensive language tasks," in *Proc. North Amer. Chapter Assoc. Comput. Linguist.*, 2020.

[17] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meet. Assoc. Comput. Linguist.*, Philadelphia, Pennsylvania, USA, 2002, pp. 311–318.

[18] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. ACL-04 Workshop*, Barcelona, Spain, 2004, pp. 74-81.

[19] T. Zhang et al., "BERTScore: Evaluating text generation with BERT," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2020.

[20] S. Es et al., "RAGAS: Automated evaluation of retrieval augmented generation," *ArXiv*, abs/2309.15217v1, 2023.

[21] R. Bommasani et al., "On the opportunities and risks of foundation models," *ArXiv*, abs/2108.07258, 2021.

[22] Replicate. Accessed June 15, 2024. https://www.replicate.com.

[23] Meta AI, "Introducing LLaMA 3." Accessed June 15, 2024. https://ai.meta.com/blog/meta-llama-3/.

[24] A. Q. Jiang et al., "Mixtral of experts," *ArXiv*, abs/2401.04088, 2024.

[25] H. Touvron et al., "Llama 2: Open foundation and fine-tuned chat models," *ArXiv*, abs/2307.09288, 2023.

[26] K. Blagec et al., "A global analysis of metrics used for measuring performance in natural language processing," *ArXiv*, abs/2204.11574, 2022.

[27] L. Zheng et al., "Judging LLM-as-a-judge with MT-Bench and Chatbot Arena," *ArXiv*, abs/2306.05685, 2023.

[28] J. Chen et al., "Benchmarking large language models in retrieval-augmented generation," in *Proc. AAAI Conf. Artif. Intell.*, 2023.

[29] H. Yu et al., "Evaluation of retrieval-augmented generation: A survey," *ArXiv*, abs/2405.07437, 2024.

[30] R. Friel, M. Belyi, and A. Sanyal, "RAGBench: Explainable Benchmark for Retrieval-Augmented Generation Systems, " *ArXiv*, abs/2407.11005, 2024.

[31] EleutherAI, "Language model evaluation harness." GitHub. Accessed August 6, 2024. [Online]. Available: https://github.com/EleutherAI/lm-evaluation-harness.

[32] J. Zhou et al., "Instruction-following evaluation for large language models," *ArXiv*, abs/2311.07911, 2023.

[33] Hugging Face, "Open LLM leaderboard," Hugging Face, n.d. Accessed August 6, 2024. [Online].'"latex Available: https://huggingface.co/docs/leaderboards/open_llm_leaderboard/about.

[34] M. Suzgun et al., "Challenging BIG-Bench tasks and whether chain-of-thought can solve them," in *Proc. Annu. Meet. Assoc. Comput. Linguist.*, 2022.

[35] D. Hendrycks et al., "Measuring mathematical problem solving with the MATH dataset," *ArXiv*, abs/2103.03874, 2021.

[36] D. Rein et al., "GPQA: A graduate-level google-proof Q&A benchmark," *ArXiv*, abs/2311.12022, 2023.

[37] Z. Sprague et al., "MuSR: Testing the limits of chain-of-thought with multistep soft reasoning," *ArXiv*, abs/2310.16049, 2023.

[38] Y. Wang et al., "MMLU-Pro: A more robust and challenging multi-task language understanding benchmark," *ArXiv*, abs/2406.01574, 2024.

[39] M. Eibich, S. Nagpal, and A. Fred-Ojala, "ARAGOG: Advanced RAG output grading," *ArXiv*, abs/2404.01037, 2024.

[40] J. Calam, "AI arXiv dataset." Available: https://huggingface.co/datasets/jamescalam/ai-arxiv. Accessed June 24, 2024.

[41] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-networks," in *Proc. Conf. Empirical Methods Natural Language Process.*, 2019.

[42] S. Es et al., "RAGAS: Metrics." Available: https://docs.ragas.io/en/stable/concepts/metrics. Accessed June 3, 2024.

[43] C.-Y. Lin, "Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics," in *Proc. 43rd Annu. Meet. Assoc. Comput. Linguist. (ACL '05)*, Ann Arbor, Michigan, USA, 2005.