2025 Annual International Conference on Data Science, Machine Learning and Blockchain Technology (AICDMB) | 979-8-3315-4216-0/25/$31.00 ©2025 IEEE | DOI: 10.1109/AICDMB64359.2025.11277692

# Retrieval Augmented Generation (RAG) using LLMs

Madiha Vahaj
*Computer Science and Engineering*
*Amity School of Engineering and Technology*
*Noida, India*
*madiha.vahaj@gmail.com*

Syed Mehran Raza
*Computer Science and Engineering*
*Amity School of Engineering and Technology*
*Noida, India*
*razamehran451@gmail.com*

Vibha Nehra
*Computer Science and Engineering*
*Amity School of Engineering and Technology*
*Noida, India*
*vnehra@amity.edu*

***Abstract*— Large Language Models (LLMs) have transformed AI with their ability to generate human-like text, but challenges such as hallucinations, outdated knowledge, and contextual inaccuracies persist. Retrieval-Augmented Generation (RAG) systems address these limitations by integrating real-time information retrieval with LLMs, enhancing the accuracy, relevance, and trustworthiness of generated content. This research aims to develop and evaluate RAG systems to improve the reliability of AI-generated outputs. To assess the impact of RAG, a comparative analysis is conducted on LLMs, including Llama, Mistral, Falcon, and T5. Performance has been evaluated by comparing results on similar tasks with and without RAG, focusing on metrics such as accuracy, contextual understanding, and domain-specific relevance. The findings demonstrate that RAG significantly enhances content generation, resolving critical challenges associated with LLMs. This study underlines the potential of retrieval-augmented techniques to improve the reliability and applicability of AI-generated content across various domains.**

***Keywords—Retrieval Augmented Generation, LLM, Llama, Falcon, Mistral, T5***

## I. Introduction

### A. Background

Large language models (LLMs) have transformed natural language processing (NLP), allowing artificial intelligence systems to produce human-like text, summarize data, and respond to inquiries spanning multiple domains [8]. These models—Llama, Mistral, Falcon, and T5—rely on pre-trained information gleaned from enormous databases. Though they have great capacity, LLMs also face major difficulties like hallucinations—where the models generate false or erroneous information—and an incapacity to access real-time or domain-specific knowledge [9]. Dealing with these constraints becomes essential as artificial intelligence applications call for precise, current, context-aware outputs.

Three basic stages define the RAG: retrieval, in response to user inquiries, where pertinent information is extracted from outside sources; augmentation, in which the retrieved data is fluidly combined with the input prompt; and generation, in which an LLM processes augmented input to generate the final response. Emerging as a potential solution to these problems is Retrieval-Augmented Generation (RAG) [6]. RAG systems can provide more accurate, relevant, anchored in factual data replies by combining external knowledge retrieval techniques with the generating capacity of LLMs. Particularly useful for applications needing real-time or domain-specific knowledge, this hybrid method combines the benefits of information retrieval systems and LLMs [6][7].

### B. Problem Statement

This work intends to solve these problems by means of a Retrieval-Augmented Generation (RAG) system that improves LLM results by external knowledge retrieval. This work also intends to evaluate, with and without RAG integration, the performance of several LLMs: Llama, Mistral, Falcon, and T5. Evaluation of their respective strengths and shortcomings in producing accurate, context-aware, and domain-relevant replies, therefore illustrating the additional benefit of retrieval-based augmentation.

### C. Project Objectives

The primary objective of this project is to design, develop, and evaluate a Retrieval-Augmented Generation (RAG) system that improves the accuracy, relevance, and contextual understanding of AI-generated content. Specifically, the project aims to:

- Implement an efficient retrieval mechanism for extracting relevant information from external resources.
- Develop methods to seamlessly integrate retrieved data with LLM inputs.
- Compare the performance of various LLMs (Llama, Mistral, Falcon, and T5) with and without RAG enhancement.
- Evaluate the effectiveness of the RAG system across different domains and task types using metrics such as accuracy, contextual relevance, and factual reliability.

### D. Scope

This research sets the foundation for developing Retrieval-Augmented Generation (RAG) systems that enhance the capabilities of Large Language Models (LLMs). Future improvements to this model can focus on several key areas:

- **Advanced Retrieval Mechanisms**: Developing more sophisticated algorithms to improve the efficiency and

accuracy of retrieving relevant information, including handling ambiguous or incomplete queries.

- **Multi-Modal Integration**: Extending the model to handle multi-modal data, enabling it to process and integrate information from text, images, audio, and video sources.
- **Robust Evaluation Metrics**: Introducing more comprehensive evaluation frameworks that account for domain-specific nuances, robustness to noisy data, and long-context comprehension.
- **Human-in-the-Loop Systems**: Incorporating mechanisms for user feedback to refine and validate model outputs, ensuring greater reliability and trustworthiness.

These advancements will further enhance the model's ability to deliver accurate, context-aware, and domain-specific responses, solidifying its applicability across diverse industries and use cases.

## II. Literature Review

Table 1: Literature Review

| S.No | Summary | Gaps |
|---|---|---|
| [1] | LLMs face challenges in arithmetic, common sense, and ethics, but skilled prompting, quality control, and integration with live data and domain-specific models offer potential for their future development. | Lacks analysis on retrieval algorithm limitations, LLM selection and fine-tuning, augmentation strategies, handling ambiguous prompts, and comprehensive evaluation metrics. |
| [2] | Retrieval-based LMs offer superior performance, adaptability, and potential across diverse domains, with a focus on practical applications and future directions. | Lacks evaluation on domain-specific performance, scalability, LLM selection, and handling ambiguous queries, which are crucial for real-world applications. |
| [3] | RAG-Ex, framework for LLMs, shows high accuracy and user agreement, though further optimizations are needed for efficiency and robustness. | Lacks discussion on retrieval mechanism efficiency, data integration with LLMs, fine-tuning, comparison with traditional LLMs, and addressing common LLM limitations, all of which are key aspects of your project. |
| [4] | Retrieval-augmented LLMs enhance query-answering systems by addressing multi-modal contexts, leveraging contrastive learning, and offering flexibility and scalability through advanced indexing. | Lacks evaluation on system performance (accuracy, fluency, relevance), LLM selection, handling ambiguous queries, and comparison with state-of-the-art systems. |
| [5] | Introduces the Retrieval-Augmented Generation Benchmark (RGB) to evaluate LLMs, finding that while RAG improves accuracy, challenges remain in rejecting irrelevant information and integrating data. | Over looks varied retrieval methods, real-world scenarios, smaller models, temporal knowledge, human-in-the-loop feedback, and scalability concerns in RAG systems. |
| [6] | RAG enhances LLMs by integrating external data for more accurate answers, evolving through three paradigms, and requiring future improvements in robustness, fine-tuning, multi-modal data handling, and evaluation metrics. | Research on RAG lacks comprehensive synthesis, effective evaluation methods, robustness to noise, integration with long contexts, hybrid approaches, and exploration of scaling laws, along with practical engineering challenges. |

As highlighted in Table I, the field of Retrieval-Augmented Generation (RAG) represents a significant advancement in overcoming the critical limitations of large language models (LLMs). LLMs face challenges in arithmetic precision, common sense reasoning, and ethical considerations, often struggling with simple computational tasks and maintaining consistent logic across domains [1].

Retrieval-based language models offer a promising solution, outperforming parametric LLMs with fewer parameters. These models enhance performance and flexibility in updating knowledge, adapting effectively to tasks like dialogue systems, semantic parsing, and machine translation, with potential in multilingual, multimodal, and code retrieval domains [2]. Frameworks like RAG-Ex [3] enable approximate explanations correlated with downstream performance, improving response accuracy and transparency.

However, significant gaps persist in retrieval algorithm limitations, handling out-of-domain queries, and integrating noisy data [4]. Integrating retrieved information with LLM inputs remains complex, with limited studies on optimal strategies [3]. RAG systems also face challenges in rejecting irrelevant information, integrating diverse data sources, managing factual inconsistencies, and showing sensitivity to query complexity [5].

The research community calls for better evaluation metrics, robust retrieval mechanisms, and improved information integration techniques [6]. Future directions include developing efficient retrieval algorithms, seamless integration methods, and comprehensive evaluation frameworks to create contextually aware and reliable language systems that address current LLM limitations [1][6].

## III. Methodology

This section outlines the methodology for developing and evaluating the Retrieval-Augmented Generation (RAG) system, detailing the preparatory phase, model implementation, and comparative analysis framework.

### *A. Preparatory Phase*

*1) Comprehensive Literature Review:* A thorough literature review on Retrieval-Augmented Generation (RAG) and Large Language Models (LLMs) was conducted. This review encompassed recent advances, challenges, and existing approaches in the field, informing the design and implementation of the RAG system. Key themes included the limitations of current LLMs, the role of retrieval mechanisms, and successful integration strategies.

*2) Selection of LLM Architectures:* Four state-of-the-art LLM architectures, specifically Mistral [10][11], Falcon [12], Llama[11], and T5 [13], were selected for comparative

analysis. The selection criteria focused on availability, performance benchmarks, and adaptability to RAG integration.

### B. Model Implementation

*1) Development of Non-RAG Baseline Models:* Non-RAG baseline models were developed to establish a performance benchmark for each selected LLM.

*2) Integration of RAG Techniques:* RAG techniques were integrated with each selected language model. This involved augmenting the model's input with the retrieved information before generating the output, ensuring that the LLM could leverage real-time data or domain-specific knowledge to improve the quality of responses.

### C. Comparative Analysis Framework

*1) Definition of Performance Metrics:* Comprehensive performance metrics were established to evaluate the outputs of both RAG-augmented and non-RAG models. Metrics included:

*a) Accuracy:* Assessment of factual correctness using precision and recall.

*b) Contextual Relevance:* Measured through metrics such as BLEU, ROUGE, or human evaluations.

*2) Creation of Diverse Questions:* A diverse set of questions was created across multiple domains, including general knowledge, technical subjects, and industry-related queries. This diverse question set benchmarked the models' performance in generating accurate and contextually relevant responses under varying conditions.

*3) Systematic Comparative Analysis:* A systematic approach was adopted to compare the performance of RAG-enhanced models with their non-RAG counterparts. Each LLM's responses to the same questions were analyzed to evaluate improvements achieved through RAG integration.

## IV. Result and Discussions

This study compared the performance of four large language models (LLMs)—Mistral, Llama, T5, and Falcon—were evaluated in this work on a set of domain-specific questions. Each model was assigned three questions and given the identical reference text. Standard metrics for assessing language generation tasks, rouge-l (roule) and bleu scores were used to evaluate the models' performances. The aim was to evaluate their capacity to provide three separate questions covering many knowledge domains accurate and fluid replies. Three domain-specific questions were chosen to test the models' performance:

- Q1: "What are the main features of Alzheimer's disease, and what structures are involved?" (Medical Domain)
- Q2: "What are the time complexities of the Quick Sort algorithm?" (Computer Science Domain)
- Q3: "What are the qualifications required to become the President of India?" (Civic Knowledge Domain)

Two widely used metrics were employed to evaluate the quality of the generated responses:

- **RouLE ( ROUGE-L Recall-Oriented Understudy of Gisting Evaluation)**: It measures the longest common word sequence, computed by the Longest Common Subsequence (LCS) algorithm [15]. This metric assesses the relative length of the model-generated output, with higher values indicating better performance in estimating the expected output length relative to the input. RouLE is measured for multiple queries (Q1, Q2, and Q3).
- **BLEU (Bilingual Evaluation Understudy)**: BLEU is a widely used metric in machine translation and text generation, which evaluates the n-gram precision between the generated output and reference texts. Higher BLEU scores indicate better alignment with the reference outputs [14]. BLEU scores are also calculated for different questions (Q1, Q2, and Q3).

Table 2: Model Performance Comparison (LLM With RAG)

| Model | Q1 | | Q2 | | Q3 | |
|---|---|---|---|---|---|---|
| | RouLe | BLEU | RouLe | BLEU | RouLe | BLEU |
| Mistral | 0.6486 | 0.2852 | 0.8888 | 0.1431 | 0.9692 | 0.7256 |
| Falcon | 0.6486 | 0.2852 | 0.3950 | 0.0140 | 0.8474 | 0.7121 |
| Llama | 0.2790 | 0.0731 | 0.2077 | 0.0108 | 0.9218 | 0.8292 |
| T5 | 0.3157 | 0.0073 | 0.8888 | 0.6391 | 0.9137 | 0.7997 |

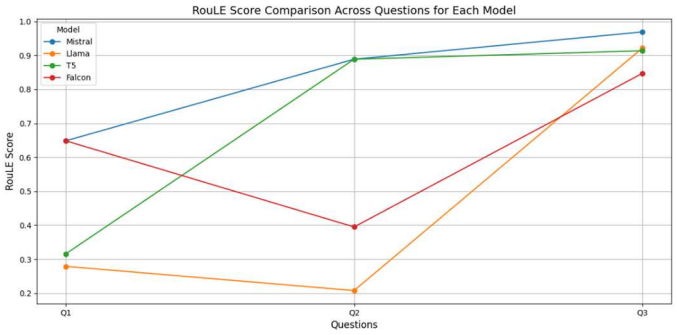


**Fig. 1.** RouLe score comparison across models for multiple questions

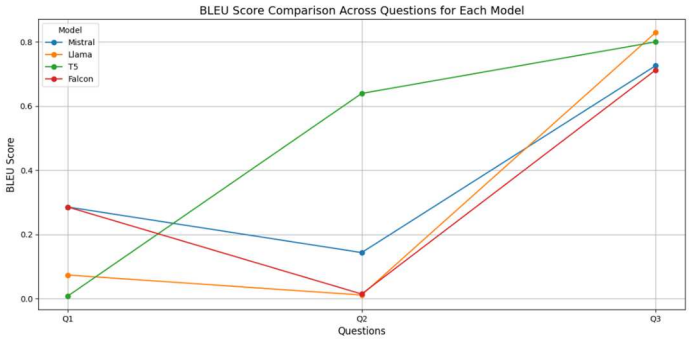


**Fig. 2.** BLEU score comparison across models for multiple questions

Table II shows the LLM models' performance when RAG is implemented. With high results on all three questions (Q1: 0.648, Q2: 0.889, Q3: 0.969), **Mistral** shows an excellent

RouLE performance. These numbers show Mistral's consistency in producing outputs with rather aligned lengths with regard for expectations. The BLEU ratings for Mistral vary, nevertheless (Q1: 0.285, Q2: 0.143, Q3: 0.726). The Q3 BLEU score is especially higher, indicating in that case greater alignment with reference outputs. While the BLEU scores in Fig. 2 indicate a more varying trajectory, culminating at Q3, Mistral's RouLE values in Fig. 1 show a consistent rise from Q1 to Q3. These findings show its capacity to properly address factual as well as domain-specific queries.

**Llama** shows a significant drop in performance across the RouLE metrics, with lower values for Q1 (0.279) and Q2 (0.208), though it reaches a relatively higher value for Q3 (0.922). BLEU scores for Llama are also low for Q1 (0.073) and Q2 (0.011), but the BLEU score for Q3 is 0.829, demonstrating an improvement and suggesting better handling of complex, non-technical questions. Fig.1 and Fig.2 illustrates this drop in performance, with Llama's RouLE and BLEU scores both starting low and showing a noticeable rise at Q3.

**T5** exhibits a balanced performance with relatively strong RouLE scores (Q1: 0.316, Q2: 0.889, Q3: 0.914), indicating that it performs well in length estimation. However, its BLEU scores are inconsistent, with very low values for Q1 (0.007) and Q3 (0.0006), though it shows a somewhat higher BLEU score for Q2 (0.639), suggesting it generates better text for that query. In Fig.2, T5's BLEU scores display a steep contrast between Q1 and Q2, with a dramatic drop at Q3, while its RouLE scores remain relatively consistent. T5 showed the lowest performance overall, particularly in BLEU scores. These results suggest difficulties in generating precise and fluent responses to domain-specific queries, indicating a need for further optimization.

**Falcon** performs similarly to Mistral, showing high RouLE scores for Q1 (0.649) and Q3 (0.847), but a noticeable dip for Q2 (0.395). Its BLEU scores mirror the RouLE trend to some extent, with higher values for Q3 (0.712) compared to Q1 (0.285) and Q2 (0.014). Fig.1 and Fig.2 illustrates the dip in Q2 for both RouLE and BLEU, with Q3 showing the highest values across both metrics. Falcon's performance declined in Q2, indicating potential limitations in handling algorithmic or technical queries.

Table 3: RouLe and BLEU Similarity Index of LLM Models with RAG and without RAG

| S.No. | Model | Question | RouLE_ Similarity | BLEU_ Similarity |
|---|---|---|---|---|
| 1 | Mistral | Q1 | 0.226804 | 0.027918 |
| | | Q2 | 0.277108 | 0.019496 |
| | | Q3 | 0.318182 | 0.117017 |
| 2 | Falcon | Q1 | 0.736842 | 0.258984 |
| | | Q2 | 0.337349 | 0.016208 |
| | | Q3 | 0.400000 | 0.178185 |
| 3 | T5 | Q1 | 0.600000 | 0.086334 |
| | | Q2 | 0.740741 | 0.022678 |
| | | Q3 | 0.187500 | 0.000597 |

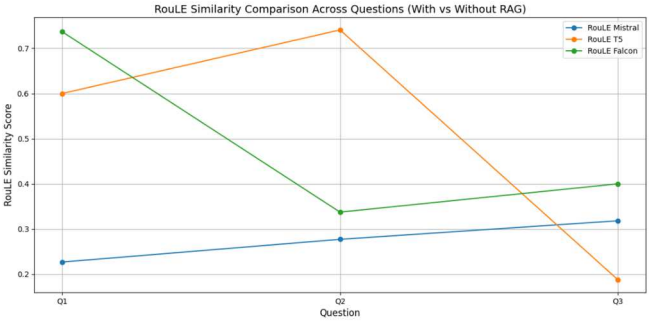


**Fig. 3.** BLEU similarity comparison across models (with vs without RAG)

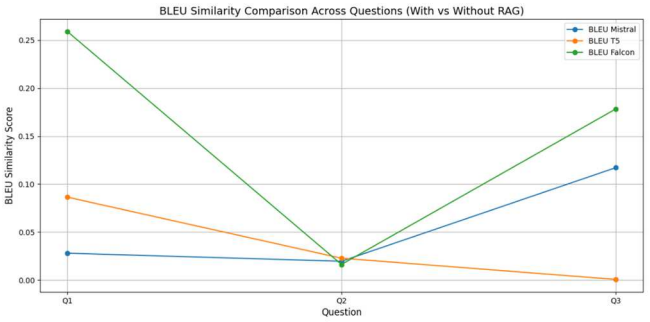


**Fig. 4.** RouLe similarity comparison across models (with vs without RAG)

Analyzed for three questions (Q1, Q2, Q3) across three models—Mistral, T5, and Falcon—were the similarity scores for produced outputs. Two aspects were the alignment of produced content with reference outputs depending on RouLE scores (measuring length estimate) and BLEU scores (evaluating n-gram accuracy). The findings are presented in Table III and graphically shown in Figures 3 and 4, therefore offering a thorough comparison of the performance of the models with and without Retrieval-Augmented Generation (RAG) application.

**Mistral** shows moderate similarity to the reference outputs in terms of RouLE and BLEU across all questions. For Q1, the RouLE score is 0.227 and the BLEU score is 0.028, indicating a lower alignment with both length expectations and reference text. For Q2, the values are 0.277 (RouLE) and 0.019 (BLEU), which remain low. For Q3, Mistral shows an increase in RouLE (0.318) and BLEU (0.117), but these scores are still below those of some other models, suggesting that while Mistral's generated text length aligns reasonably with the reference, the precision of its generated content (as measured by BLEU) could be improved.

**T5** has higher similarity scores in both metrics compared to Mistral. For Q1, it has a strong RouLE score of 0.600 and a BLEU score of 0.086, suggesting that T5's generated text has a better length estimation and n-gram alignment with reference outputs. In Q2, T5 further improves, achieving a RouLE score of 0.741 and a BLEU score of 0.023. However, in Q3, the model sees a slight drop in both metrics (RouLE: 0.188, BLEU: 0.0006), indicating lower alignment for that particular question. Fig.3 and Fig.4 shows the steady improvement in T5's RouLE

scores, while its BLEU scores peak at Q1, before declining at Q3.

**Falcon** stands out in terms of similarity scores, particularly for Q1, where it achieves a high RouLE score of 0.737 and a relatively higher BLEU score of 0.259. Falcon also performs well for Q2 with RouLE at 0.337 and BLEU at 0.016. For Q3, Falcon demonstrates a solid alignment with reference content, with RouLE at 0.400 and BLEU at 0.178, suggesting its strength in producing more accurate and relevant outputs. Fig.3 and Fig.4 highlights Falcon's consistent performance, with high similarity scores at Q1 and a moderate drop for Q2 before stabilizing for Q3.

In summary, Mistral, Falcon, and T5 exhibit distinct strengths across the experiments. Mistral performs well in terms of output length estimation (RouLE) but has varying BLEU scores that suggest it might struggle with precise content generation. T5 demonstrates solid performance with good length estimation and stronger BLEU scores in certain cases, while Falcon maintains consistent performance with higher alignment to reference outputs, particularly for Q1 and Q3. Llama, on the other hand, generally lags behind in both RouLE and BLEU across most questions, highlighting areas where its performance could be improved.

## V. CONCLUSION AND FUTURE SCOPE

This study demonstrates the significant potential of Retrieval-Augmented Generation (RAG) systems in overcoming Large Language Models' (LLMs') constraints like hallucinations, out-of-date knowledge, and contextually shallow outputs. Real-time retrieval systems help RAG systems improve the dependability, contextual relevance, and correctness of AI-generated replies. Using measures like BLEU and ROUGE, a comparison of LLMs—including Mistral, Falcon, Llama, and T5—showcases how consistently RAG integration improves performance across several domains. Among the models, Mistral and Falcon showed very strong accuracy in facts and contextual alignment. These results show the need of coupling the generating capacity of LLMs with external knowledge retrieval, so providing a viable route for creating AI systems that provide dependable, context-aware, and domain-specific solutions for a wide spectrum of uses. The promising outcomes of this study open several avenues for future exploration in Retrieval-Augmented Generation (RAG) systems. Further research can focus on integrating multi-modal retrieval to enhance LLMs with data from visual, audio, or structured sources. Improving real-time retrieval efficiency and optimizing fusion strategies for retrieved knowledge with prompts can yield more contextually aware outputs. Additionally, deploying human-in-the-loop frameworks for iterative feedback and refinement will boost reliability and user trust. Finally, expanding the evaluation across diverse domains and languages will validate the scalability and generalizability of RAG-enhanced models in real-world applications.

## REFERENCES


[1] Teubner, Timm, Christoph M. Flath, Christof Weinhardt, Wil van der Aalst, and Oliver Hinz. "Welcome to the era of chatgpt et al. the prospects of large language models." *Business & Information Systems Engineering* 65, no. 2 (2023): 95-101.

[2] Asai, Akari, Sewon Min, Zexuan Zhong, and Danqi Chen. "Retrieval-based language models and applications." In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts)*, pp. 41-46. 2023.

[3] Sudhi, Viju, Sinchana Ramakanth Bhat, Max Rudat, and Roman Teucher. "Rag-ex: A generic framework for explaining retrieval augmented generation." In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2776-2780. 2024.

[4] Wang, Mengzhao, Haotian Wu, Xiangyu Ke, Yunjun Gao, Xiaoliang Xu, and Lu Chen. "An Interactive Multi-modal Query Answering System with Retrieval-Augmented Large Language Models." *arXiv preprint arXiv:2407.04217* (2024).

[5] Chen, Jiawei, Hongyu Lin, Xianpei Han, and Le Sun. "Benchmarking large language models in retrieval-augmented generation." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 16, pp. 17754-17762. 2024.

[6] Gao, Yunfan, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. "Retrieval-augmented generation for large language models: A survey." *arXiv preprint arXiv:2312.10997* (2023).

[7] Lewis, Patrick, et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks." *Advances in neural information processing systems* 33 (2020): 9459-9474.

[8] Naveed, Humza, et al. "A comprehensive overview of large language models." *arXiv preprint arXiv:2307.06435* (2023).

[9] Hadi, Muhammad Usman, et al. "A survey on large language models: Applications, challenges, limitations, and practical usage." *Authorea Preprints* (2023).

[10] Hamzah, Farizal, and Nuraini Sulaiman. "Multimodal integration in large language models: A case study with mistral llm." (2024).

[11] Hou, Guangyu, and Qin Lian. "Benchmarking of commercial large language models: Chatgpt, mistral, and llama." (2024).

[12] Almazrouei, Ebtesam, et al. "The falcon series of open language models." *arXiv preprint arXiv:2311.16867* (2023).

[13] Mastropaolo, Antonio, et al. "Studying the usage of text-to-text transfer transformer to support code-related tasks." *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*. IEEE, 2021.

[14] Reiter, Ehud. "A structured review of the validity of BLEU." *Computational Linguistics* 44.3 (2018): 393-401.

[15] Barbella, Marcello, and Genoveffa Tortora. "Rouge metric evaluation for text summarization techniques." *Available at SSRN 4120317* (2022).