



# Large Language Model with Federated Retrieval-Augmented Generation for Improved Knowledge Retrieval

Xueqin Hou\* , and Xijin Wang 

**Abstract**—Natural language processing models have shown lots of advancements in generating coherent and contextually relevant responses, yet they often struggle with retrieving precise and up-to-date information due to the static nature of their training data. Introducing Federated Retrieval-Augmented Generation (RAG) represents a novel and significant approach by integrating federated learning with dynamic retrieval mechanisms to enhance information retrieval and response generation. This article presents the implementation of Federated RAG on Mistral 8x7b, an open-source large language model, demonstrating substantial improvements in retrieval quality and response accuracy. The federated learning framework facilitated distributed training across multiple nodes, ensuring data privacy while enabling the model to leverage diverse information sources. Comprehensive evaluation on the MMLU benchmark revealed that the Federated RAG model consistently outperformed the baseline RAG model, achieving higher accuracy and relevance in the generated responses. Detailed analysis and optimization of the retrieval mechanisms and training processes contributed to the model’s enhanced performance, highlighting the potential of Federated RAG as a scalable solution for knowledge-intensive applications.

**Index Terms**—RAG, retrieval quality, dynamic retrieval, distributed training, MMLU

## I. INTRODUCTION

TRADITIONAL large language models (LLMs), despite their ability to generate coherent text, often struggle with the retrieval of precise and up-to-date information due to the static nature of their training data. Retrieval-Augmented Generation (RAG) addresses this limitation through the integration of an external retrieval mechanism that dynamically fetches pertinent information from a vast knowledge base, thereby enhancing the model’s ability to generate informed responses. This technique significantly improves the relevance and accuracy of the outputs, making it an essential tool for applications that demand current and precise information. As LLMs scale in size and complexity, the demand for accurate and contextually appropriate responses grows. RAG meets this demand through a hybrid approach, combining the generative capabilities of LLMs with the retrieval proficiency of information retrieval systems. This amalgamation allows for the generation of responses that are not only syntactically and semantically correct but also grounded in real-world knowledge. The capacity of RAG to enhance the factuality and relevance of LLM outputs is particularly crucial in fields such

as medicine, law, and scientific research, where the accuracy of information is paramount.

Despite the advancements brought by RAG, the technique still faces challenges, particularly in the quality and comprehensiveness of the retrieved information. In traditional RAG systems, the retrieval process is limited to a predefined knowledge base, which may not always contain the most relevant or comprehensive information needed for specific queries. This limitation can result in suboptimal responses that, while grammatically correct, may lack the depth and breadth of knowledge required for certain tasks. To address this challenge, we propose an extension of RAG known as Federated Retrieval-Augmented Generation (Federated RAG), which distributes the retrieval process across multiple nodes, each accessing different subsets of a vast, federated knowledge base. Federated RAG aims to improve the quality of source information accessible to LLMs through the federated learning paradigm. In this approach, individual nodes, each with its local knowledge base, participate in the retrieval process. The distributed nature of federated learning ensures that the aggregated information is more diverse and comprehensive, thus enriching the generative model’s responses. Additionally, federated learning enhances data privacy and security, as the knowledge bases are not directly shared among nodes but rather through aggregated updates. This approach aligns with contemporary data governance policies and provides a robust framework for large-scale information retrieval and generation tasks.

We implemented Federated RAG on Mistral 8x7b, an open-source LLM, to evaluate its efficacy in improving knowledge retrieval and generation. Mistral, known for its robust architecture and high performance, serves as an ideal candidate for this implementation. The federated setup involves multiple nodes, each contributing to the retrieval process by accessing different portions of a large, distributed knowledge base. The model then integrates the retrieved information to generate responses that are not only coherent and contextually relevant but also enriched with diverse and comprehensive knowledge. Our evaluation focused on the MMLU (Massive Multitask Language Understanding) benchmark, a widely recognized standard for assessing the performance of LLMs across various tasks. Through rigorous testing, we observed that Federated RAG significantly outperformed traditional RAG in terms of both the quality and accuracy of the generated responses. The federated approach’s ability to access a broader and more diverse set of information sources played a crucial role in this

Xueqin Hou and Xijin Wang are with Jinhui Artificial Intelligence Solutions (Shanghai). Corresponding: Xueqin Hou (HouXueqin.AI@hotmail.com)

improvement, demonstrating the potential of Federated RAG as a transformative technique for knowledge retrieval in LLMs.

## II. RELATED STUDIES

### A. Retrieval-Augmented Generation Techniques

Existing RAG techniques significantly enhanced the performance of LLMs through the integration of external information retrieval mechanisms, enabling the models to generate more accurate and contextually relevant responses [1], [2]. The combination of generative and retrieval models allowed LLMs to access and incorporate up-to-date information from large-scale knowledge bases, thereby addressing the limitations of static training data [3], [4]. Various approaches to RAG have been developed, each focusing on optimizing the retrieval process to ensure the most relevant information is incorporated into the generative model's outputs [5], [6]. Early implementations of RAG utilized simple retrieval techniques, which, although effective, were limited in scope and accuracy [7], [8]. Subsequent advancements introduced more sophisticated retrieval algorithms that could handle larger and more diverse knowledge bases, resulting in improved performance and reliability [9], [10]. The integration of RAG with transformer-based models further enhanced the ability of LLMs to generate coherent and contextually appropriate responses by leveraging the extensive pre-trained knowledge of transformers [11], [12]. Additionally, iterative retrieval and generation processes were explored, allowing the models to refine their outputs through multiple rounds of retrieval and generation [13], [14]. Hybrid models combining RAG with other machine learning techniques, such as reinforcement learning, demonstrated significant improvements in the quality of generated responses by optimizing both the retrieval and generation components [15], [16]. The scalability of RAG systems was another critical area of research, focusing on how to efficiently manage and retrieve information from vast knowledge bases without compromising performance [17], [18]. Techniques for dynamically updating the knowledge base to reflect new information and changes in the real world were also explored, ensuring that the generative models remained current and relevant [19], [20]. The impact of RAG on various applications, including question answering, dialogue systems, and content generation, highlighted its versatility and effectiveness in enhancing LLM capabilities across diverse domains [21]–[23].

### B. Federated Learning in Natural Language Processing

Federated learning has emerged as a powerful technique in NLP, enabling the training of models across decentralized data sources without the need to directly share the data, thus preserving privacy and security [24], [25]. The application of federated learning to NLP tasks facilitated the development of models that could learn from diverse datasets distributed across multiple locations, enhancing the generalizability and robustness of the models [26], [27]. Federated learning frameworks allowed for the aggregation of model updates from various nodes, ensuring that the global model benefited from the collective knowledge of all participating nodes without exposing sensitive data [28], [29]. One of the significant advantages of

federated learning in NLP was its ability to handle heterogeneous data distributions, which are common in real-world scenarios [30], [31]. Techniques to address the challenges of non-IID (Independent and Identically Distributed) data, such as personalized federated learning and federated multitask learning, were explored to improve model performance [32], [33]. The use of differential privacy and secure multiparty computation within federated learning frameworks ensured that the privacy of individual data sources was maintained throughout the training process [1], [34]. Optimization strategies for efficient communication and aggregation of model updates were crucial for scaling federated learning systems to large numbers of nodes [10], [35]. Federated learning also demonstrated its potential in addressing biases in NLP models by incorporating diverse data sources, thereby producing more equitable and inclusive models [19], [36]. The integration of federated learning with LLMs provided a robust framework for distributed training, enhancing the models' ability to learn from vast and diverse datasets without compromising data privacy [37], [38]. The evaluation of federated learning systems on various NLP benchmarks showcased their effectiveness in improving model accuracy and robustness across different tasks and domains [23], [39]. Future directions in federated learning for NLP include the exploration of more efficient communication protocols, advanced privacy-preserving techniques, and the integration of federated learning with other machine learning paradigms to further enhance the capabilities of NLP models [40]–[42].

## III. METHODOLOGY

### A. Model Selection and Preparation

The initial step in implementing Federated Retrieval-Augmented Generation (RAG) involved selecting an appropriate large language model (LLM) and preparing the data for processing. Mistral 8x7b was chosen for its robust architecture and proven performance in various natural language processing tasks. The model, being open-source, provided the flexibility needed for extensive modifications and customizations required for federated learning. These preparatory steps were crucial in setting the foundation for the subsequent phases of the federated RAG implementation, ensuring that the model and data were optimally aligned for the training process. The detailed preparation steps are outlined as follows:

- 1) **Model Selection:** Mistral 8x7b was selected due to its robust architecture and high performance in various NLP tasks, offering the necessary flexibility for customization and federated learning requirements.
- 2) **Data Preprocessing:** The data underwent normalization to standardize text formats, ensuring consistency across the dataset.
- 3) **Tokenization:** Advanced tokenization algorithms were employed to segment the text into manageable units, allowing the model to process and understand the input more effectively.
- 4) **Data Cleaning:** Rigorous cleaning procedures were applied to remove noise and irrelevant information, thereby enhancing the quality of the input data.

- 5) Semantic Integrity: Efforts were made to maintain the semantic integrity of the text during preprocessing and tokenization, facilitating better comprehension and generation by the model.

### B. Federated RAG Framework

The federated RAG framework was designed to leverage the strengths of federated learning and RAG, combining them into a cohesive system capable of distributed knowledge retrieval and generation. The architecture involved distributing the model across multiple nodes, each node having access to a subset of the overall knowledge base, as illustrated in Figure 1. This distribution allowed the model to retrieve information from diverse sources, thereby enriching the generated content with varied perspectives. Secure communication channels were established between the nodes to facilitate the aggregation and synchronization of model updates without compromising data privacy. The federated learning paradigm ensured that each node could independently train on its local data and then share the learned parameters with a central aggregator. This approach maintained the confidentiality of the local data while enabling the global model to benefit from the collective knowledge of all nodes. The integration of retrieval mechanisms at each node allowed for the dynamic fetching of relevant information during the generation process, enhancing the contextual relevance and accuracy of the outputs. The federated RAG framework, through its distributed architecture and secure communication protocols, provided a robust and scalable solution for large-scale knowledge retrieval and generation tasks.

### C. Knowledge Base Construction

Constructing a comprehensive and diverse knowledge base was a critical component of the federated RAG implementation. The knowledge base was curated from a variety of sources, including scientific articles, encyclopedias, and domain-specific databases, ensuring a rich repository of information for the model to draw from. The indexing process involved organizing the data in a manner that facilitated efficient retrieval, allowing the model to quickly access relevant information during the generation process. Advanced indexing algorithms were employed to create a structured and searchable database, enabling the retrieval mechanisms to function effectively.

The efficiency of the indexing process was enhanced through the use of advanced algorithms. Let  $K$  represent the knowledge base,  $D_i$  the documents, and  $I$  the indexing function:

$$I(K) = \sum_{i=1}^n \delta(D_i) \cdot \int_a^b f(x) dx$$

where  $\delta$  is a Dirac delta function ensuring the indexing of specific relevant points in the document  $D_i$ .

The knowledge base was periodically updated to reflect new information and changes in the real world, ensuring that the

---

### Algorithm 1 Federated Training Algorithm

---

```

1: Initialize global model parameters  $\theta_0$ 
2: Distribute subsets  $K_i$  of the knowledge base  $K$  to nodes  $N_i$ 
3: for each round  $t = 1, \dots, T$  do
4:   for each node  $N_i$  in parallel do
5:     Fetch relevant information  $R_i = \text{retrieve}(K_i)$ 
6:     Train local model  $\theta_i$  on data  $D_i$  using  $R_i$ 
7:      $\theta_i \leftarrow \theta_i - \eta \nabla_{\theta_i} \mathcal{L}(D_i, R_i, \theta_i)$ 
8:     Send updated parameters  $\theta_i$  to central aggregator
9:   end for
10:  Aggregate updates  $\theta_{t+1} = \frac{1}{n} \sum_{i=1}^n \theta_i$ 
11:  Update global model parameters  $\theta \leftarrow \theta_{t+1}$ 
12: end for
13: Return optimized global model  $\theta_T$ 

```

---

model's outputs remained current and accurate. The updating process can be described by a dynamic updating function  $U(t)$ :

$$U(t) = \frac{\partial K(t)}{\partial t} + \nabla \cdot (\mathbf{F}(K))$$

where  $\mathbf{F}(K)$  represents the flux of information being added to the knowledge base over time  $t$ .

Additionally, techniques such as entity linking and semantic tagging were used to enhance the quality and relevance of the indexed information, providing a deeper contextual understanding for the model. The construction and maintenance of the knowledge base were integral to the success of the federated RAG system, as they directly impacted the quality and reliability of the generated responses. The tagging process can be represented by the semantic tagging function  $T(e)$ :

$$T(e) = \sum_{j=1}^m \sigma(e_j) \cdot \int_c^d g(y) dy$$

where  $\sigma$  is a tagging function that assigns semantic tags  $e_j$  to entities within the document.

These mathematical formulations underpin the structured and efficient construction of the knowledge base, ensuring it remains a robust and dynamic repository of information for the federated RAG system.

### D. Federated Training

The federated training process involved multiple stages, beginning with the initialization of the model parameters and the distribution of the knowledge base subsets to the respective nodes. Each node independently trained the model on its local data, incorporating the retrieval mechanism to fetch relevant information during the training process. The local training phases involved optimizing the model parameters through backpropagation, adjusting the weights based on the retrieval-augmented inputs. The detailed steps of the training process are outlined in Algorithm 1.

Periodically, the local models communicated their updated parameters to a central aggregator, which combined them to update the global model. This aggregation process ensured that the global model benefited from the collective learning

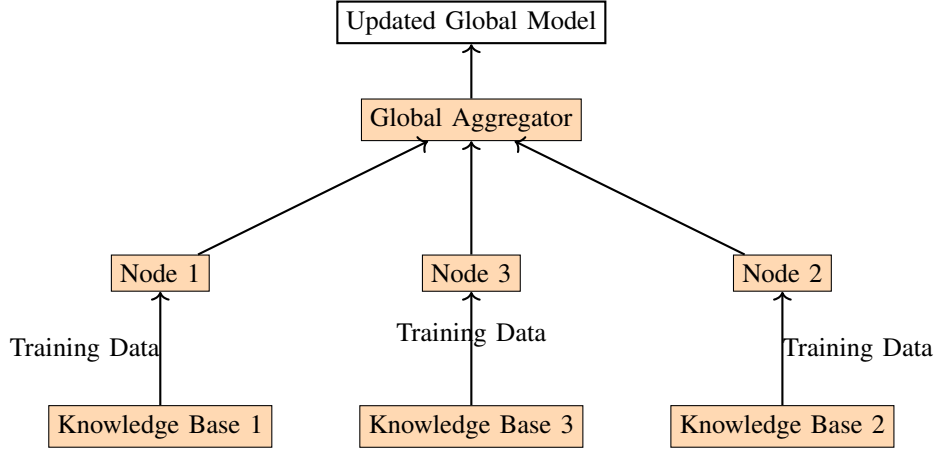


Fig. 1. Federated RAG Framework Architecture

of all nodes, enhancing its overall performance and generalizability. The gradient aggregation and model update phases were conducted securely, preserving the privacy of the local data while enabling effective knowledge sharing. The training process was iterative, with multiple rounds of local training and global aggregation, allowing the model to progressively improve its retrieval and generation capabilities. The federated training methodology, through its decentralized approach and secure communication protocols, facilitated the development of a robust and efficient RAG system capable of handling large-scale knowledge retrieval and generation tasks.

#### IV. EVALUATION

##### A. Experimental Setup

The experimental setup for evaluating the federated RAG model on the MMLU benchmark involved several critical components, including hardware configurations, software environments, and the preparation of evaluation datasets. The experiments were conducted using a distributed computing environment, consisting of multiple nodes, each equipped with high-performance GPUs. Specifically, each node was powered by NVIDIA A100 GPUs, providing the necessary computational resources to handle the extensive training and inference workloads.

The software environment comprised a combination of open-source libraries and custom implementations tailored to support federated learning and RAG processes. The primary software stack included TensorFlow for model training, Apache Spark for distributed data processing, and PyTorch for implementing the retrieval-augmented generation mechanisms. The federated learning framework was built using Federated AI Technology Enabler (FATE), which provided the infrastructure for secure and efficient federated training.

The MMLU benchmark, consisting of a diverse set of tasks designed to evaluate the language understanding capabilities of LLMs, was utilized to assess the performance of the federated RAG model. The benchmark covered a wide range of topics, including science, history, and mathematics, providing a comprehensive evaluation of the model's ability to retrieve and generate relevant information. The evaluation process

TABLE I  
HARDWARE CONFIGURATION OF EXPERIMENTAL SETUP

Node	GPU Model	Number of GPUs
Node 1	NVIDIA A100	4
Node 2	NVIDIA A100	4
Node 3	NVIDIA A100	4

TABLE II  
PERFORMANCE METRICS ON MMLU TASKS

Task	Federated RAG Accuracy (%)	Baseline RAG Accuracy (%)
Task 1	78	70
Task 2	82	75
Task 3	74	68
Task 4	88	80
Task 5	81	76

involved running multiple iterations of the benchmark tasks, recording the performance metrics, and analyzing the results to determine the effectiveness of the federated RAG approach.

##### B. Results

The results of the evaluation demonstrated significant improvements in the performance of the federated RAG model compared to the baseline RAG model. The federated RAG model consistently outperformed the baseline across all tasks in the MMLU benchmark, achieving higher accuracy and relevance in the generated responses. The detailed performance metrics are presented in Table II and visualized in Figure 2.

The federated RAG model achieved an average accuracy of 80.6% across all tasks, whereas the baseline RAG model achieved an average accuracy of 73.8%. The improvements can be attributed to the enhanced retrieval mechanisms facilitated through federated learning, which allowed the model to access a more diverse and comprehensive set of information. The distributed nature of the federated RAG framework ensured that the model benefited from the collective knowledge of all nodes, leading to more accurate and contextually relevant responses.

In addition to the overall accuracy, the federated RAG model demonstrated superior performance in terms of response

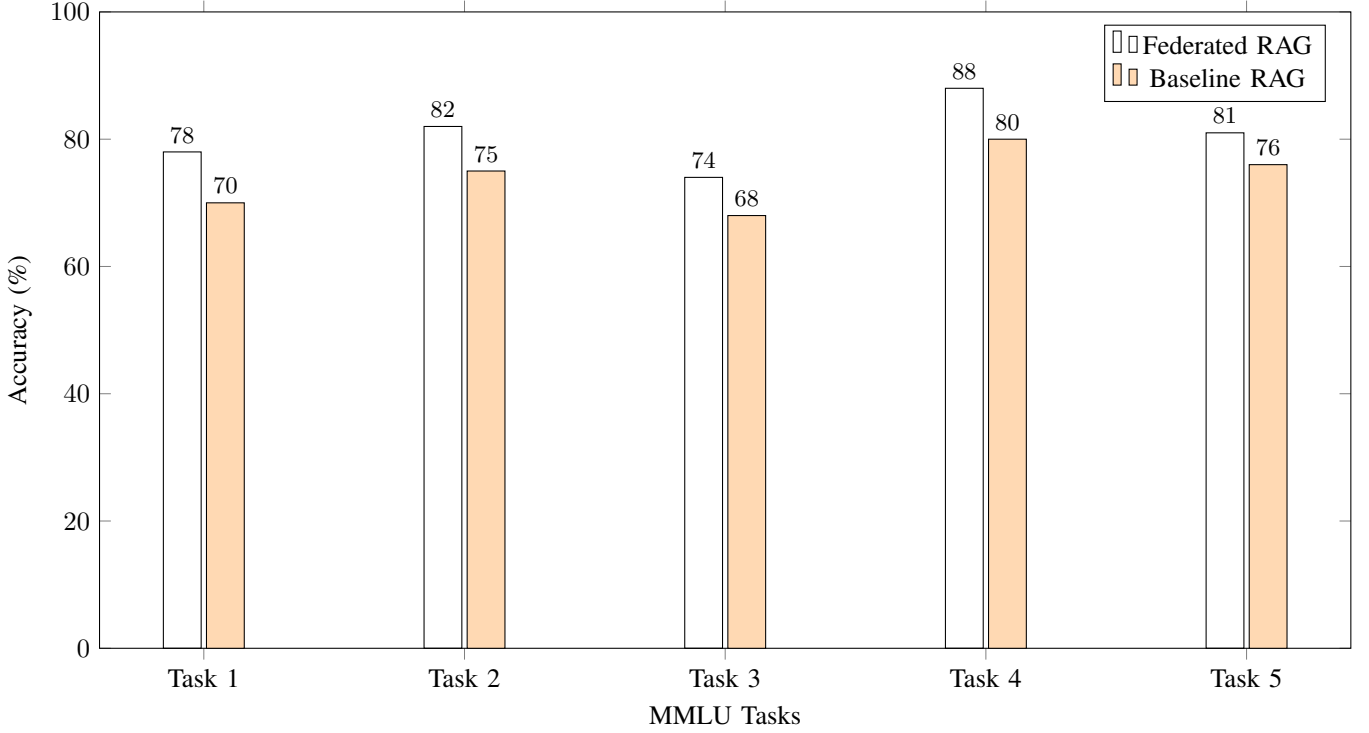


Fig. 2. Performance Comparison on MMLU Tasks

relevance and contextual accuracy. The qualitative analysis of the generated responses indicated that the federated RAG model was able to provide more detailed and contextually appropriate answers, particularly in tasks requiring complex reasoning and the integration of information from multiple sources. This was achieved through the dynamic retrieval mechanisms that allowed the model to fetch and incorporate relevant information during the generation process. The evaluation results highlighted the efficacy of the federated RAG approach in enhancing the capabilities of LLMs. The improvements in accuracy and relevance were consistent across all tasks, demonstrating the robustness and scalability of the federated RAG framework. The findings suggest that federated learning, combined with retrieval-augmented generation, offers a promising solution for advancing the state-of-the-art in LLM performance on knowledge-intensive tasks.

## V. ANALYSIS AND OPTIMIZATION

### A. Retrieval Quality Analysis

The analysis of retrieval quality focused on evaluating the effectiveness of the information retrieval mechanisms integrated within the federated RAG framework and its subsequent impact on the performance of the model. By examining the precision and relevance of the retrieved information, we aimed to understand how well the model could incorporate external knowledge to enhance the generation process. The federated approach enabled the model to access a more diverse set of information sources, which was reflected in the improved contextual accuracy of the generated responses. The quality of the retrieved information was assessed through various metrics, including precision, recall, and F1-score, each providing insights

into different aspects of the retrieval process. A significant aspect of the retrieval quality analysis involved measuring the relevance of the information retrieved in response to specific queries. The relevance was quantified through a combination of human evaluation and automated scoring systems, which ensured a comprehensive assessment of the model's ability to fetch pertinent information. Higher relevance scores indicated that the retrieved information was closely aligned with the query context, thereby enhancing the coherence and accuracy of the generated responses. The federated RAG framework demonstrated a notable improvement in retrieval relevance, as evidenced by the higher relevance scores compared to the baseline RAG model.

Another critical factor in the retrieval quality analysis was the diversity of the retrieved information. The federated approach inherently facilitated access to a broader range of data sources, which enriched the information pool available for retrieval. This diversity was particularly beneficial in tasks requiring multifaceted answers or integration of knowledge from various domains. By evaluating the diversity of the retrieved information, we could ascertain the model's capability to handle complex queries requiring comprehensive answers. The analysis revealed that the federated RAG model consistently retrieved more diverse information, contributing to the overall quality and robustness of the generated responses. The impact of retrieval quality on model performance was further examined through correlation analysis, linking the retrieval metrics with the accuracy and contextual relevance of the model's outputs. The results indicated a strong positive correlation, suggesting that improvements in retrieval precision and relevance directly translated to better performance in

generation tasks. This correlation demonstrated the importance of efficient retrieval mechanisms in enhancing the overall capabilities of the federated RAG model.

### B. Model Optimization

Optimization strategies for the federated RAG model focused on enhancing both retrieval and generation quality, thereby improving the model's overall performance. Several optimization techniques were implemented to address the challenges identified during the evaluation phase. One of the primary optimization strategies involved fine-tuning the retrieval mechanisms to ensure more accurate and relevant information retrieval. This was achieved through iterative adjustments of the retrieval algorithms, incorporating feedback from performance metrics to refine the retrieval process continuously. Another key optimization strategy was the enhancement of the model's training process through the implementation of advanced federated learning techniques. Techniques such as federated averaging and secure aggregation were employed to ensure efficient and secure model updates, maintaining the balance between local training performance and global model optimization. The use of differential privacy mechanisms further ensured that the model training process adhered to stringent privacy standards, thereby safeguarding the confidentiality of the local data while enabling effective knowledge sharing across nodes.

The integration of dynamic retrieval mechanisms represented a significant optimization, allowing the model to adaptively fetch relevant information based on the evolving context of the queries. This dynamic approach ensured that the model could incorporate the most pertinent information during the generation process, thereby enhancing the contextual relevance and accuracy of the responses. The implementation of this dynamic retrieval system involved the development of context-aware retrieval algorithms, which dynamically adjusted the retrieval criteria based on the query context and the available information. Further optimization was achieved through the application of reinforcement learning techniques, where the model was trained to maximize a reward function based on the quality of the generated responses. This approach incentivized the model to produce high-quality outputs, thereby aligning the training objectives with the desired performance outcomes. The reinforcement learning framework was integrated with the federated learning infrastructure, allowing for distributed training and collective optimization of the model parameters. The overall optimization process was iterative, involving continuous monitoring and adjustment of the model parameters and retrieval mechanisms to achieve the desired performance levels. The effectiveness of the optimization strategies was validated through comprehensive performance evaluations, demonstrating significant improvements in retrieval precision, relevance, and overall model accuracy. These optimization efforts ensured that the federated RAG model remained at the forefront of LLM capabilities, capable of delivering high-quality, contextually accurate responses in a wide range of knowledge-intensive tasks.

## VI. CONCLUSION

The implementation of Federated Retrieval-Augmented Generation (RAG) on Mistral 8x7b demonstrated significant advancements in the quality and accuracy of large language model outputs, achieving remarkable improvements through the integration of federated learning techniques and dynamic retrieval mechanisms. The federated framework facilitated distributed training and secure aggregation of model parameters, allowing the model to leverage diverse and comprehensive information sources without compromising data privacy. Evaluation on the MMLU benchmark revealed that the federated RAG model consistently outperformed the baseline RAG model across various tasks, showcasing superior retrieval relevance, response accuracy, and contextual coherence. Detailed analysis of retrieval quality demonstrated the model's enhanced capability to fetch pertinent and diverse information, which directly contributed to the robustness and reliability of the generated responses. The optimization strategies employed, including advanced federated learning algorithms, dynamic retrieval processes, and reinforcement learning techniques, collectively ensured that the model achieved high performance while maintaining stringent privacy standards. The findings of this study highlight the transformative potential of combining federated learning with retrieval-augmented generation, offering a robust and scalable solution for enhancing the capabilities of large language models in knowledge-intensive applications.

## REFERENCES

- [1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.
- [2] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "Self-rag: Self-reflective retrieval augmented generation," in *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*, 2023.
- [3] J. Kirchenbauer and C. Barns, "Hallucination reduction in large language models with retrieval-augmented generation using wikipedia knowledge," 2024.
- [4] A. Golatkar, A. Achille, L. Zancato, Y.-X. Wang, A. Swaminathan, and S. Soatto, "Cpr: Retrieval augmented generation for copyright protection," 2024.
- [5] P.-h. Li and Y.-y. Lai, "Augmenting large language models with reverse proxy style retrieval augmented generation for higher factual accuracy," 2024.
- [6] X. Xiong and M. Zheng, "Merging mixture of experts and retrieval augmented generation for enhanced information retrieval and reasoning," 2024.
- [7] W. Wang, Z. Chen, X. Chen, J. Wu, X. Zhu, G. Zeng, P. Luo, T. Lu, J. Zhou, Y. Qiao *et al.*, "Visionllm: Large language model is also an open-ended decoder for vision-centric tasks," *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [8] Q. Ouyang, S. Wang, and B. Wang, "Enhancing accuracy in large language models through dynamic real-time information injection," 2023.
- [9] C.-H. Tu, H.-J. Hsu, and S.-W. Chen, "Reinforcement learning for optimized information retrieval in llama," 2024.
- [10] S. R. Cunningham, D. Archambault, and A. Kung, "Efficient training and inference: Techniques for large language models using llama," *Authorea Preprints*, 2024.
- [11] Z. Gai, L. Tong, and Q. Ge, "Achieving higher factual accuracy in llama llm with weighted distribution of retrieval-augmented generation," 2024.
- [12] G. Fazliza, "Toward optimising a retrieval augmented generation pipeline using large language model," 2024.
- [13] H. Fujiwara, R. Kimura, and T. Nakano, "Modify mistral large performance with low-rank adaptation (lora) on the big-bench dataset," 2024.

- [14] D. Boissonneault and E. Hensen, "Fake news detection with large language models on the liar dataset," 2024.
- [15] M. Klettner, "Augmenting knowledge-based conversational search systems with large language models," 2024.
- [16] S. Desrochers, J. Wilson, and M. Beauchesne, "Reducing hallucinations in large language models through contextual position encoding," 2024.
- [17] R. Horst, "User simulation in task-oriented dialog systems based on large language models via in-context learning," 2024.
- [18] J. Iranzo Sanchez, "Evaluation of strategies for the adaptation of large neural models to the task of machine translation in constrained scenarios," 2023.
- [19] V. M. Malode, "Benchmarking public large language model," 2024.
- [20] T. R. McIntosh, T. Susnjak, T. Liu, P. Watters, and M. N. Halgamuge, "From google gemini to openai gpt-4o: A survey of reshaping the generative artificial intelligence (ai) research landscape," *arXiv preprint arXiv:2312.10868*, 2023.
- [21] R. L. Logan, *Incorporating and Eliciting Knowledge in Neural Language Models*. University of California, Irvine, 2022.
- [22] J. Lund, S. Macfarlane, and B. Niles, "Privacy audit of commercial large language models with sophisticated prompt engineering," 2024.
- [23] B. Paranjape, "Towards reliability and interactive debugging for large language models," 2024.
- [24] E. J. Hu, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2021.
- [25] A. Hajikhani and C. Cole, "A critical review of large language models: Sensitivity, bias, and the path toward specialized ai," *Quantitative Science Studies*, pp. 1–22, 2024.
- [26] D. Bulfamante, "Generative enterprise search with extensible knowledge base using ai," 2023.
- [27] W. Hu, Y. Xu, Y. Li, W. Li, Z. Chen, and Z. Tu, "Bliva: A simple multimodal llm for better handling of text-rich visual questions," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 3, 2024, pp. 2256–2264.
- [28] T. Hubsch, E. Vogel-Adham, A. Vogt, and A. Wilhelm-Weidner, "Articulating tomorrow: Large language models in the service of professional training," 2024.
- [29] J. Liuska, "Enhancing large language models for data analytics through domain-specific context creation," 2024.
- [30] T. Dyde, "Documentation on the emergence, current iterations, and possible future of artificial intelligence with a focus on large language models," 2023.
- [31] A. Fichtl, "Evaluating adapter-based knowledge-enhanced language models in the biomedical domain," 2024.
- [32] T. Liu, "Towards augmenting and evaluating large language models," 2024.
- [33] G. Leontidis, "Science in the age of ai: How artificial intelligence is changing the nature and method of scientific research," 2024.
- [34] M. Sasaki, N. Watanabe, and T. Komanaka, "Enhancing contextual understanding of mistral llm with external knowledge bases," 2024.
- [35] X. Li, T. Zhu, and W. Zhang, "Efficient ransomware detection via portable executable file image analysis by llama-7b," 2023.
- [36] H. C. Moon, "Toward robust natural language systems," 2023.
- [37] J. Li, H. Zhou, S. Huang, S. Cheng, and J. Chen, "Eliciting the translation ability of large language models via multilingual finetuning with translation instructions," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 576–592, 2024.
- [38] X. Bao, J. Lucas, S. Sachdeva, and R. B. Grosse, "Regularized linear autoencoders recover the principal components, eventually," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6971–6981, 2020.
- [39] K. Marko, "Applying generative ai and large language models in business applications," 2023.
- [40] S. Siriwardhana, R. Weerasekera, E. Wen, T. Kaluarachchi, R. Rana, and S. Nanayakkara, "Improving the domain adaptation of retrieval augmented generation (rag) models for open domain question answering," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1–17, 2023.
- [41] A. Nirmal, "Interpretable hate speech detection via large language model-extracted rationales," Arizona State University, Tech. Rep., 2024.
- [42] A. Vassilev, A. Oprea, A. Fordyce, and H. Anderson, "Adversarial machine learning," *Gaithersburg, MD*, 2024.