Knowledge Ply Chat

Retrieval-Augmented Generation for Enhancing LLMs with Non-Parametric Knowledge

[1]M Krishna Satya Varma
Assistant Professor
Department of Information Technology
Sagi Rama Krishnam Raju Engineering College
Bhimavaram, India

[2]Koteswara Rao; [3]Sai Ganesh; [4]Venkat Sai Koushik;
[5]Rama Krishnam Raju
Department of Information Technology
Sagi Rama Krishnam Raju Engineering College
Bhimavaram, India

Abstract:- Despite their ability to store information and excel at many NLP tasks with fine-tuning, large language models tend to have issues about accurately accessing and altering knowledge, which leads to performance gaps in knowledge-intensive tasks compared to domain-specific architectures. Additionally, these models face problems when it comes to having transparent decision-making processes or updating their world knowledge. To mitigate these limitations, we propose a Retrieval Augmented Generation (RAG) system by improving the Mistral7B model specifically for RAG tasks. The novel training technique includes Parameter-Efficient Fine-Tuning (PEFT) which enables efficient adaptation of large pre-trained models on-the-fly according to task-specific requirements while reducing computational costs. In addition, this system combines pre-trained embedding models that use pre-trained cross-encoders for effective retrieval and reranking of information. This RAG system will thus leverage these state-of-the-art methodologies towards achieving top performances in a range of NLP tasks such as question answering and summarization.

Keywords:- *Component: RAG, PEFT, Cross Encoders.*

I. INTRODUCTION

Large language models, despite their ability to retain factual knowledge and perform remarkably well in various NLP tasks after fine-tuning, can still have difficulty accessing and manipulating knowledge accurately. This issue causes performance gaps in knowledge-intensive tasks when compared to specialized architectures. Furthermore, these models struggle to maintain transparent decision-making processes and keep up with current global knowledge. To address these limitations, we proposed creating a Retrieval Augmented Generation (RAG) system specifically designed to improve the Mistral7B model. Using Parameter-Efficient Fine-Tuning (PEFT) with Low-Rank Adaptation (LoRA) [4], the system optimizes the adaptation of large pre-trained models to task-specific needs while minimizing computational costs. Furthermore, the system incorporates pre-trained embedding models for effective knowledge retrieval and reranking via pre-trained cross-encoders. Furthermore, the system incorporates pre-trained embedding models for effective knowledge retrieval and reranking via pre-trained cross-encoders. Advanced techniques will be used to achieve cuttingedge performance across a wide range of NLP tasks, including question answering and summarization. Finally, this study contributes to the development of scalable and efficient RAG models, as well as future AI-driven applications, by encouraging more reliable and contextually accurate knowledge utilization. We want to build scalable and efficient RAG models by focusing on an innovative training method called Parameter-Efficient Fine-Tuning (PEFT) and Low-Rank Adaptation (LoRA). We will demonstrate our system's capabilities through extensive experimentation, with a focus on improvements to knowledge-intensive tasks and enhanced contextual knowledge utilization in language generation tasks. We believe that our development will foster transparency and trust in responsible AI, addressing pressing concerns related to decision-making processes and knowledge maintenance in large language models.

II. LITERATURE SURVEY

Large Language Models (LLMs) have revolutionized various tasks in Natural Language Processing (NLP). However, a key limitation remains: their inability to effectively access and utilize external knowledge, especially in knowledge-intensive domains [1] This can lead to inaccurate or incomplete outputs when dealing with complex questions or tasks requiring information.[1] highlights the challenges associated with LLMs, emphasizing their limitations in knowledge access and manipulation. They introduce RAG models as a solution, combining a pre-trained LLM (parametric memory) with a non-parametric memory represented by an external knowledge base accessed through a neural retriever. This allows RAG models to surpass traditional parametric seq2seq models and task-specific architectures in knowledge-intensive NLP tasks, especially open domain question answering [1]. Additionally, RAG models generate more specific, diverse, and factual language compared to traditional LLMs [1]. The effectiveness of RAG models has been demonstrated across various NLP applications.[2] explored RAG models for math question answering. Their system incorporates content from a math textbook, achieving improved answer quality by leveraging retrieved information during the response generation process. The study highlights user preference for RAG-generated responses if they are not overly reliant on textbook content. In healthcare, [3] presented a pioneering approach for disease phenotyping in Electronic Health Records (EHRs) using

RAG models. Their method outperforms traditional techniques in identifying disease phenotypes, particularly for rare diseases. This application showcases the potential of RAG models to revolutionize disease diagnosis and improve clinical research by leveraging retrieved snippets from EHRs. These studies showcase the versatility and promise of RAG models in augmenting LLM capabilities across diverse domains.

III. SYSTEM OVERVIEW

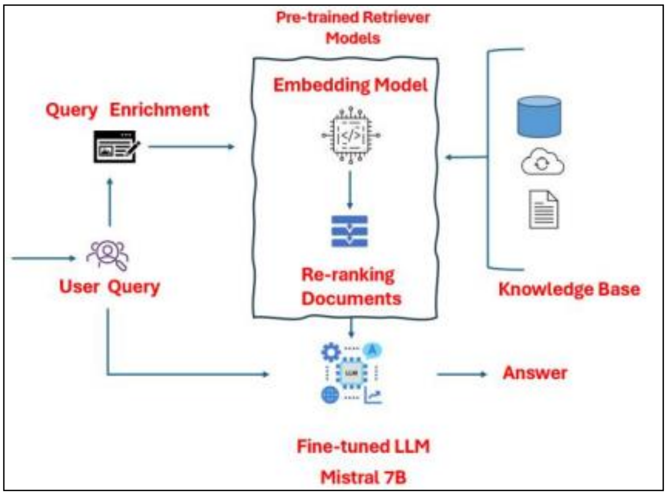

Fig 1: Overview of our System Design.

Our approach involved using pre-trained retriever models (Embeddings + Cross-Encoders) and fine-tuning a transformer decoder-only model, Mistral 7B (Generator), for the Retrieval- Augmented Generation (RAG) task. To enhance query performance and improve the retrieval of related documents from external sources, we generated multiple similar queries. We utilized cosine similarity to select the top-K documents. For the final prediction, we treated retrieved documents as latent variables and marginalized over sequence-to-sequence predictions given different documents.

❖ *Our Proposed System Delivers the Following Contributions:*

A. *Mistral 7B Model Fine-Tuning:* Mistral 7B model will be fine-tuned for RAG tasks to improve knowledge access and manipulation and promote reliable contextual knowledge utilisation in knowledge-intensive tasks.

B. *RAG Pipeline:* We created an advanced RAG pipeline using cutting-edge research, including pre-trained embedding models and cross-encoders, to handle NLP tasks efficiently and effectively like question answering and summarization.

C. *RAG Pipeline Evaluation:* We will conduct rigorous experiments to assess our developed pipeline. This thorough evaluation will demonstrate our system's superior performance when compared to baseline solutions that prioritise accountability, accessibility, and responsible AI.

The following sections, we will elaborate the responsibility and design of each of the components.

A. *Mistral 7B Model Fine-Tuning*

We utilize two datasets provided by Neural Bridge AI: the Retrieval-Augmented Generation (RAG) Dataset 12000 and the Hallucinations Dataset, forming a comprehensive collection of 13,000 English entries containing context, questions, and answers. This dataset is optimized for RAG-focused models and is divided into a 10,400-instance training set and a 2,600-instance test set. The context is derived from Falcon RefinedWeb, while questions and answers are generated using the advanced capabilities of GPT-4. Our data preprocessing stage tailors an instruction prompts template for the Mistral 7B instruct model to effectively leverage its pre-trained knowledge and fine-tune it on RAG tasks. This strategic preprocessing aims to improve the model's capacity for context manipulation and reliable, contextually accurate generation of answers in knowledge-intensive tasks.

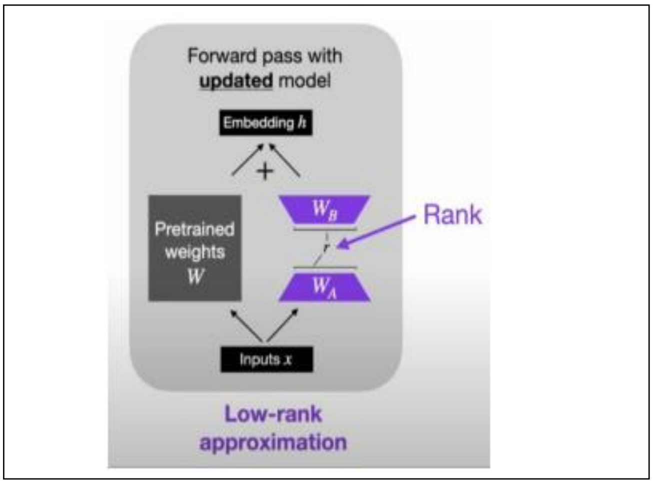

Fig 2: LoRA: low-Rank Adaptation of Large Language Models.

To make fine-tuning more efficient, LoRA's approach is to represent the weight updates with two smaller matrices (called update matrices) through low-rank decomposition. These new matrices can be trained to adapt to the new data while keeping the overall number of changes low. The original weight matrix remains frozen and doesn't receive any further adjustments. To produce the results, both the original and the adapted weights are combined [4]. We employed the Parameter-Efficient Fine-Tuning (PEFT) approach with Low-Rank Adaptation (LoRA) to fine-tune the Mistral 7B instruct model for Retrieval Augmented Generation tasks. We added trainable rank decomposition matrices to the query, key, and value matrices of the model's self- attention layers. The rank of the decomposition matrices was set to 32, and the alpha parameter was set to 64. These values were chosen based on a hyperparameter search to find the optimal configuration for the RAG tasks. The fine-tuning process entailed adjusting the model's parameters to optimize its performance on the RAG tasks. We used a learning rate of 2e-5, a batch size of 2, and a gradient accumulation step of 1 for fine-tuning over 5200 steps, with a warmup ratio of 0.03.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24APR285

B. *RAG Pipeline:*

- *Query Enrichment:*

Enhance the initial user query by paraphrasing it, adding synonymous terms, and expanding it in a meaningful way using Mistral 7B. This will increase the chance of retrieving more relevant and diverse documents for a given query.

- *Pre-Trained Embedding Model*

Utilized a sentence-transformers model ("multi-qa-mpnet-base-dot-v1") designed for semantic search, mapping sentences and paragraphs to a 768-dimensional dense vector space. Trained on a corpus of 215M (question, answer) pairs sourced from diverse datasets.

- *Reranking Model*

In our project, we enhance query result relevance through a custom reranker named MixedBreadAI/mxbai-embed-2d-large-v1, which employs a Cross Encoder architecture. This model intricately assesses pairs of queries and documents, assigning relevance scores that reflect each document's alignment with the query's intent. By leveraging advanced natural language processing techniques, the reranked reorders the initial broad set of documents, prioritizing the top-N most pertinent ones to the user's query. Our integration of this reranking process into the retrieval pipeline ensures delivery results that closely match the query's semantic nuances.

- *Finetuned-Mistral7B(Text-Generation):*

Our core generation model, Mistral 7B, stands as a pinnacle in natural language processing, boasting 7.3 billion parameters and outperforming previous benchmarks. Engineered by Mistral AI, Mistral 7B [5] integrates Grouped-query Attention (GQA) and Sliding Window Attention (SWA) mechanisms, ensuring efficient inference and adept handling of longer sequences. Released under the Apache 2.0 license, Mistral 7B offers unrestricted deployment and fine-tuning capabilities with publicly available datasets, showcasing remarkable generalization without proprietary data reliance. Optimized for scalability and efficiency, it exhibits superior performance across a spectrum of tasks including Commonsense Reasoning, World Knowledge, Reading Comprehension, Math, and Code, solidifying its position as a cutting-edge solution in the field of natural language understanding.

C. *RAG Pipeline Evaluation*

To evaluate the generation quality of the RAG pipeline, we computed the ROUGE, BERT scores for the generated responses of the model with and without fine-tuning on the test dataset. The ROUGE score measures the overlap of n-grams between the generated Answer and the reference Answer, while the BERTScore [6] computes the similarity of two text sequences using BERT embeddings. We generated responses for a set of input queries in the test dataset using both the finetuned and non-fine-tuned models. We then computed the ROUGE, BERT scores for each generated response using the rouge-score, bert-score Python packages, respectively. The evaluation phase of our research project involved a comprehensive analysis of the Retrieval Augmented Generation (RAG) pipeline's performance, with a specific focus on the impact of finetuning the Mistral 7B model.

The results obtained from this analysis are succinctly presented in Table 1 and Table 2, which showcase the performance metrics using Rouge and Bert scores, respectively. The results of the evaluation are presented in the following tables.

Table 1: Evaluation Results [Rogue Score] for the RAG Pipeline with and without Fine-Tuning on the Test Dataset.

Model	Mistral 7B Pretrained	Mistral 7B Finetuned
Rouge 1	0.20	**0.41**
Rouge 2	0.13	**0.35**
Rouge L	0.18	**0.39**

Table 1 details the Rouge scores, a standard metric for evaluating the quality of text generation. The Rouge scores assess the overlap between the generated answer and reference answer, with Rouge1 measuring unigram overlap, Rouge2 for bigrams, and RougeL for the longest common subsequence. The fine-tuned Mistral 7B model exhibits a substantial improvement in all Rouge metrics when compared to the pretrained model. Specifically, the Rouge1 score nearly doubles from 0.20 to 0.41, indicating a significant enhancement in capturing the most critical unigram overlaps. Similarly, the Rouge2 score, capturing higher-order n-gram overlaps, sees a more than twofold increase from 0.13 to 0.35, while the RougeL score, reflecting the model's ability to capture sentence-level structures, improves from 0.18 to 0.39.

Table 2: Evaluation Results [Bert Score] for the RAG Pipeline with and without Fine-Tuning on the Test Dataset.

Model	Mistral 7B Pretrained	Mistral 7B Finetuned
Precision	0.12	**0.40**
Recall	0.121	**0.38**
F1 Score	0.099	**0.32**

Table 2 presents the Bert scores, which are instrumental in evaluating semantic similarity and are based on contextual embeddings. The enhancement in both Rouge and Bert scores post-fine-tuning suggests that the RAG pipeline, when optimized with task-specific data, can be optimized to generate text that is not only lexically closer to the reference summaries but also semantically more coherent and precise. The improved recall and precision further imply that the fine-tuned model is more effective at retrieving relevant information and generating responses that maintain a high degree of factual accuracy. The fine-tuning process yields dramatic improvements in precision (from 0.12 to 0.40), recall (from 0.121 to 0.38), and F1 Score (from 0.099 to 0.32). These scores collectively indicate that the finetuned model is more adept at producing semantically relevant and accurate content, aligning closely with the reference material.

IV. CONCLUSION

In conclusion, the evaluation results underscore the efficacy of fine-tuning the Mistral 7B model within the RAG framework for improving its performance on knowledge-intensive NLP tasks. The stark contrast in performance metrics before and after fine-tuning emphasizes the value of task-specific optimization in large language models. These findings advocate for the integration of fine-tuning as a necessary step towards enhancing the practical utility of LLMs in real-world applications where accuracy and up-to-date knowledge are paramount. This research contributes to the ongoing discourse on the importance of model adaptation and continual learning to maintain the relevance and effectiveness of pre-trained language models in an ever-evolving data landscape.

V. FUTURE ENHANCEMENTS AND RESEARCH DIRECTIONS

- Real-time Information Integration: Investigate methods for integrating real-time information sources (e.g., news feeds, social media) into the RAG pipeline. This would allow LLMs to access and leverage the latest information for even more dynamic responses.
- Multimodal Retrieval: Explore incorporating non-textual information like images or audio into the retrieval process. This would enable RAG systems to access and utilize a broader range of knowledge sources.
- Address potential biases present in the underlying datasets used to train both the LLM and the retrieval models. Develop methods to ensure fairness and mitigate the risk of biased information influencing the RAG system's outputs.
- Explore techniques for explaining the reasoning process behind the information retrieved and the answer generated by the RAG system. This transparency would be crucial for building user trust and ensuring responsible AI development.

By exploring these future research directions, we can further unlock the potential of RAG systems and pave the way for the development of truly intelligent and knowledgeable LLMs.

REFERENCES

[1]. P. Lewis et al., “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” May 2020, [Online]. Available: http://arxiv.org/abs/2005.11401 .

[2]. Z. Levonian et al., “Retrieval-augmented Generation to Improve Math Question-Answering: Trade-offs Between Groundedness and Human Preference,” Oct. 2023, [Online]. Available: http://arxiv.org/abs/2310.03184 .

[3]. W. E. Thompson et al., “Large Language Models with Retrieval-Augmented Generation for Zero-Shot Disease Phenotyping,” Dec. 2023, [Online]. Available: http://arxiv.org/abs/2312.06457 .

[4]. E. J. Hu et al., “LoRA: Low-Rank Adaptation of Large Language Models,” Jun. 2021, [Online]. Available: http://arxiv.org/abs/2106.09685 .

[5]. A. Q. Jiang et al., “Mistral 7B,” Oct. 2023, [Online]. Available: http://arxiv.org/abs/2310.06825 .

[6]. T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “BERTScore: Evaluating Text Generation with BERT,” Apr. 2019, [Online]. Available: http://arxiv.org/abs/1904.09675

[7]. T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “BERTScore: Evaluating Text Generation with BERT,” Apr. 2019, [Online]. Available: http://arxiv.org/abs/1904.09675 .

[8]. D. S. Sachan, S. Reddy, W. Hamilton, C. Dyer, and D. Yogatama, “End-to-End Training of Multi-Document Reader and Retriever for Open-Domain Question Answering,” Jun. 2021, [Online]. Available: http://arxiv.org/abs/2106.05346 .

[9]. A. Ziletti and L. D’Ambrosi, “Retrieval augmented text-to-SQL generation for epidemiological question answering using electronic health records,” Mar. 2024, [Online]. Available: http://arxiv.org/abs/2403.09226 .

[10]. W. E. Thompson et al., “Large Language Models with Retrieval-Augmented Generation for Zero-Shot Disease Phenotyping,” Dec. 2023, [Online]. Available: http://arxiv.org/abs/2312.06457 .

[11]. S. S. Manathunga and Y. A. Illangasekara, “Retrieval Augmented Generation and Representative Vector Summarization for large unstructured textual data in Medical Education,” Aug. 2023, [Online]. Available: http://arxiv.org/abs/2308.00479 .

[12]. J. Li, Y. Yuan, and Z. Zhang, “Enhancing LLM Factual Accuracy with RAG to Counter Hallucinations: A Case Study on Domain-Specific Queries in Private Knowledge-Bases,” Mar. 2024, [Online]. Available: http://arxiv.org/abs/2403.10446 .

[13]. E. Melz, “Enhancing LLM Intelligence with ARM-RAG: Auxiliary Rationale Memory for Retrieval Augmented Generation,” Nov. 2023, [Online]. Available: http://arxiv.org/abs/2311.04177 .

[14]. J. Lála, O. O’Donoghue, A. Shtedritski, S. Cox, S. G. Rodriques, and A. D. White, “PaperQA: Retrieval-Augmented Generative Agent for Scientific Research,” Dec. 2023, [Online]. Available: http://arxiv.org/abs/2312.07559 .

[15]. A. J. Yepes, Y. You, J. Milczek, S. Laverde, and R. Li, “Financial Report Chunking for Effective Retrieval Augmented Generation,” Feb. 2024, [Online]. Available: http://arxiv.org/abs/2402.05131 .