Enhancing Machine Learning Comprehension: A System for Reading and Understanding Machine Learning Articles

Nikhil Mylarusetty
Departement of Applied Data Science
San Jose State University
San Jose, USA
nikhil.mylarusetty@sjsu.edu

Ming-Hwa Wang
Departement of Applied Data Science
San Jose State University
San Jose, USA
ming-hwa.wang@sjsu.edu

Manidedeepya Chennapragada
Departement of Applied Data Science
San Jose State University
San Jose, USA
mandedeepya.chennapragada@sjsu.edu

***Abstract*— The surge in academic publications along with complexity of papers within the domain of Machine Learning (ML) poses a challenge in efficiently comprehending the abundance information. This study aims to address this challenge by developing a user-friendly interface focused on improving the efficiency and effectiveness of analyzing ML research papers. Leveraging the papers from the arXiv database repository and a QASPER questionnaire dataset, this research integrates research papers seamlessly into the Pinecone vector database for streamlined data storage and retrieval. Additionally, a Large Language Model (LLM) is created, incorporating llama2, Mistral 7B, and retrieval augmented generation techniques to enhance natural language processing capabilities. Evaluation of the LLM's efficiency is conducted through measures such as context relevancy, faithfulness, answer relevancy, context recall, and harmfulness (EVALS), with the best resulting scores indicating that the RAG with Custom Fine Tuned LLaMa model achieved Context Relevancy: 0.4918, Faithfulness: 0.9074, Answer Relevancy: 0.8743, Context Recall: 0.9167, and Harmfulness: 0. Users are provided with the opportunity to engage with the material through a Question and Answer system which helps in facilitating a deeper understanding of the technical content by navigating the complexities of ML papers. This research contributes to improving comprehension of ML literature, thereby fostering innovation and advancement in the field.**

Keywords*— *Machine Learning, Pinecone vector database, Large Language Model (LLM), llama 2, Mistral 7B, Retrieval Augmented Generation, Natural Language Processing, Question and Answer system, EVALS

I. Introduction (*Heading 1*)

The realm of Machine Learning (ML) research is witnessing an exponential surge in academic publications, with an average of approximately 100 papers published globally each day, totaling around 36,500 papers annually. This inundation of research underscores the pressing challenge of efficiently comprehending the extensive ML literature. In response, a reading and understanding system is built for Machine Learning Research Papers where this research emerges with the aim of enhancing and comprehension of ML research papers

Motivated by the escalating demand among researchers, students, and professionals for tools facilitating rapid access to pertinent information within the ML literature, this research seeks innovative solutions to address this need. Traditional methods of literature review often struggle to keep pace with the mounting volume and intricacy of ML research papers, necessitating the utilization of advanced technologies to streamline the process. Central to the research's approach is the deployment of Large Language Models (LLMs) to augment comprehension of ML research papers. These models leverage state-of-the-art natural language processing (NLP) techniques to facilitate precise text comprehension and generation, thereby enhancing user assistance.

In support of its objectives, the research aggregates a comprehensive dataset from the [12] arXiv database repository, encompassing a diverse array of ML research papers. These papers undergo meticulous processing into smaller text chunks to enable efficient handling and analysis. Additionally, integration of a [13] QASPER (questionnaire) dataset aids in fine-tuning the LLMs. A notable innovation of the research involves the conversion of text chunks into embeddings, creating a numerical representation imbued with semantic meaning. These embeddings are stored in a vector database, enabling semantic search functionality, which retrieves relevant text chunks and research papers based on user queries. Rigorous evaluation of the research's efficacy encompasses measures such as context relevancy, faithfulness, answer relevancy, context recall, and harmfulness [11]. By furnishing researchers and practitioners with a user-friendly tool to access, comprehend, and engage with ML literature, this paper endeavours to foster innovation and advancement in the field.

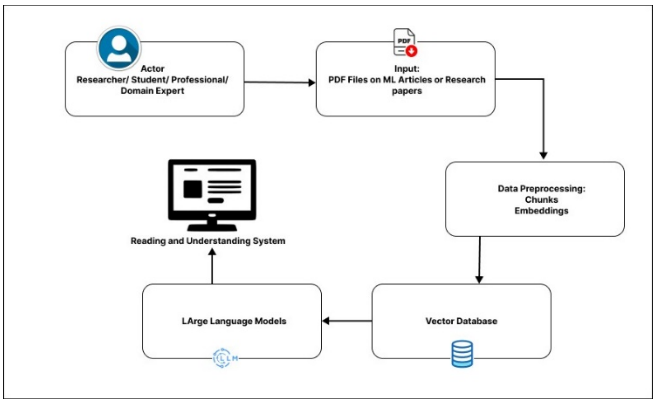
Actor
Researcher/ Student/ Professional/ Domain Expert
PDF
Input:
PDF Files on ML Articles or Research papers
Data Preprocessing:
Chunks
Embeddings
Reading and Understanding System
LArge Language Models
LLM
Vector Database

Fig. 1. Overview of Proposed System Architecture

In essence, the "A Reading Understanding System for Machine Learning Articles" paper addresses a critical exigency in the ML research domain by furnishing a practical solution to efficiently navigate and comprehend the extensive ML literature, thereby propelling progress and innovation in the field.

II. RELATED WORK

A. Literature Survey

In recent years, the field of natural language processing (NLP) has experienced a transformation, marked by remarkable progress and the emergence of advanced systems to diverse domains. Among these advancements, the integration of large language models (LLMs) with knowledge graphs stands out as a significant research, given by the work of the paper [1] in the domain of healthcare. Guo and colleagues' contribution have focused on enhancing medical question answering systems, leveraging the robust language understanding capabilities of LLMs and the structured domain-specific knowledge provided by knowledge graphs. By amalgamating these cutting-edge technologies, their research has paved the way for the development of more accurate and contextually rich systems capable of addressing various medical queries with precision and depth.

Similarly, in the dynamic landscape of news dissemination, in the research [2] have spearheaded efforts to develop sophisticated question-answering systems tailored specifically to the complexities of the news domain. In this approach they integrated named entity recognition, topic modeling, and document retrieval algorithms, represents a comprehensive solution for extracting comprehensive and timely information from diverse news sources. In the domain of text summarization, this paper [3] have made significant progress by refining pretrained encoder models such as BERT and Transformer-based architectures. Their method focuses on generating high-quality abstractive summaries has made a major change in the efficiency of information extraction from textual data, enabling users to distill key insights from extensive documents with ease. Moreover, the introduction of pointer-generator networks by [4] has bridged the gap between extractive and abstractive summarization techniques, offering a novel approach to generating informative summaries that capture essential content from source texts. By leveraging hybrid loss functions and coverage mechanisms, their research has facilitated the production of summaries that strike a balance between fidelity to the source material and conciseness. These seminal research endeavors underscore the pivotal role of NLP and LLM applications across diverse domains, ranging from healthcare and news dissemination to text summarization. By using advanced techniques such as representation learning, dynamic adaptation, and containerization, researchers are reshaping the landscape of NLP, ushering in a new era of more efficient and effective information processing.

In summary, the contributions of these researchers not only advance the in NLP but also lay the groundwork for future innovations in information processing and knowledge sharing. Their work helps as a evidence to the transformative potential of NLP and LLM applications, offering invaluable insights and solutions to complex challenges in various domains.

B. Technical Survey

Recent research in the fields of machine learning (ML) and natural language processing (NLP) has witnessed significant strides, particularly in the exploration of large language models (LLMs) and their specialized applications [5]. Notably, studies have delved into various methodologies for LLM training, with a particular focus on domain adaptation using domain-specific corpora. This approach has showcased superior performance across domains such as biomedical, finance, and law, highlighting its potential for real-world applications[6]. Moreover, the integration of LLMs with Visual Language Models (VLMs) has opened up new avenues for semantic scene understanding, particularly in the context of scenes captured by Unmanned Aerial Vehicles (UAVs) [6]. This integration has enabled zero-shot semantic scene understanding, revolutionizing industries with its broad applications.

In the domain of Open-Domain Question Answering (ODQA), advancements in the Retrieval Augmented Generation (RAG) model have demonstrated promising results, particularly in domain adaptation scenarios[7]. The RAG model has significantly enhanced performance across diverse domains, showcasing its adaptability and effectiveness in handling various types of queries. Furthermore, the utilization of LLMs in healthcare decision-making processes has garnered significant attention. Studies have explored the potential for personalized diagnostic and clinical guidance through the integration of clinical data and medical sources, paving the way for more efficient and accurate healthcare solutions [7]. In parallel, efforts have been made to improve paraphrase generation techniques, with a specific focus on parameter-efficient methods and the application of reinforcement learning (RL) techniques. These advancements aim to enhance text generation capabilities and foster more natural and coherent paraphrases[3]. Additionally, the development of specialized LLMs tailored to specific domains, such as renewable energy, underscores the importance of domain-specific models in advancing research and applications in specialized areas. These specialized models play a crucial role in addressing domain-specific challenges and driving innovation in niche fields.

Collectively, these findings highlight the diverse approaches and contributions in recent advancements in ML and NLP. They underscore the ongoing efforts to push the boundaries of AI capabilities through innovative methodologies and specialized model development, paving the way for transformative applications across various domains.

III. DATA ENGINEERING

A. Data Collection

The data collection process for this research involves two primary sources: the arXiv database and the QASPER dataset. From the arXiv database, which hosts over 2 million research articles, we extract papers specifically focused on machine learning within computer science. This systematic extraction aims to compile a specialized dataset containing the latest advancements in the field. Concurrently, we utilize the Hugging Face/QASPER dataset, designed for question answering in Natural Language Processing (NLP) and ML literature. This dataset comprises 5,049 meticulously crafted questions formulated by NLP experts, tailored to extract information exclusively available in the full text of research papers. The data collected from these sources is stored using two distinct mechanisms: a vector database for research papers from arXiv, facilitating efficient retrieval based on user queries, and the Hugging Face for the QASPER dataset, ensuring scalability and redundancy across multiple zones. Accessibility and security of the research dataset are paramount considerations, with continuous updates and efforts to bolster data security, including secure SSH connections.

B. Data Pre-processing

1) ArXiv Dataset: Our data pre-processing commenced with extracting the text from the unstructured PDFs into organized data, aiming to facilitate comprehensive research and analysis. The primary focus was on extracting metadata and content from PDFs, also eliminating duplicate PDFs from datasets. Duplicate papers were identified and removed to maintain data accuracy. The detection and removal of duplicate content involved using MD5 hash values to identify repetitive PDFs, ensuring each paper contributed unique information. Further the data pre-processing for PDF files involved a series of steps where we aimed to extract text while removing unnecessary information like images and tables. In this process, the directory containing the PDF files to be processed to train the model are selcted where each PDF file is then loaded and its content are extracted using a library such as PyMuPDF (fitz) to access the text on each page. Once the text is extracted which is followed by the next step to filter out unwanted inclusions, such as images and tables, which are often included in PDF documents along with textual content. This is achieved through a combination of techniques. Firstly, the length of each extracted text block is checked to identify and remove short blocks, which are likely to be captions or labels associated with images rather than meaningful text. Additionally, a check for non-textual characters is performed on each text block. Blocks containing a high density of non-textual characters, such as symbols or punctuation marks, are removed, as they are indicative of image or table content. This filtering process helps to ensure that only meaningful text is retained for further processing. To further refine the extraction process, image processing techniques can be employed using libraries like OpenCV. These techniques can be used to analyze the content of images embedded within the PDF and identify regions that are likely to contain non-textual content. Once these regions are identified, optical character recognition (OCR) can be applied using a library such as Tesseract to extract any textual content embedded within the images. Finally, the remaining text blocks are consolidated into coherent paragraphs or sections, resulting in a cleaned and processed text ready for further analysis or use. Overall, this data pre-processing process combines text extraction, filtering, image processing, and OCR techniques to effectively extract text from PDF files while removing unnecessary information, thereby facilitating the efficient processing and analysis of PDF documents

2) QASPER Dataset: This research utilizes the QASPER dataset, a comprehensive collection of scholarly papers structured in a CSV format with columns for title, abstract, full_text, qas, and figures_and_tables, where each cell is represented in JSON format. The 'qas' column, pivotal to this study, contains questions posed about the papers and their corresponding answers, essential for training a Large Language Model (LLM). The preprocessing of the 'qas' column involves several intricate steps to prepare it for fine-tuning the LLM. Initially, the 'qas' column is extracted from the dataset, isolating it for further transformation. Prior to the extraction of the 'qas' column the original dataset has been checked using pre-processing techniques to remove irrelevant or erroneous entries if necessary, ensuring the quality and accuracy of the data.

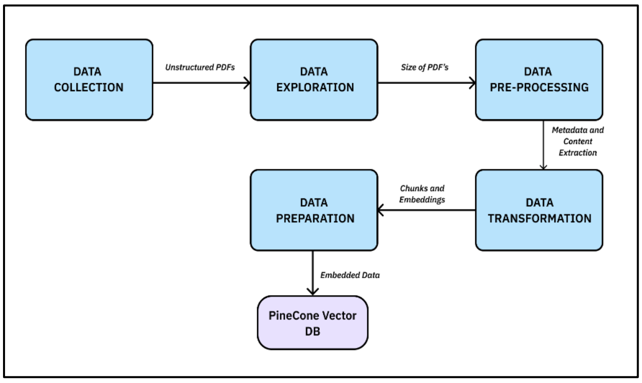

Fig. 2. Data Process for arXiv Dataset

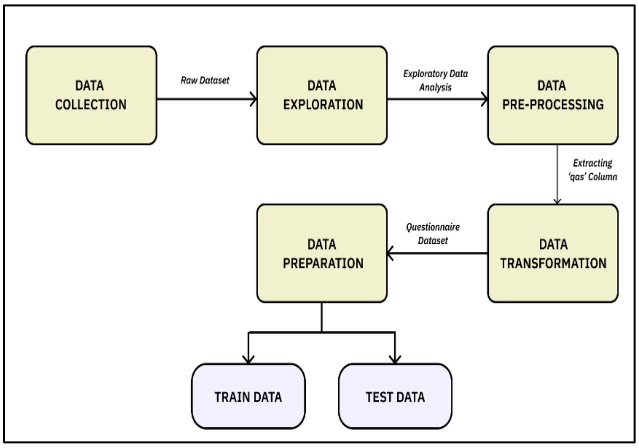

Fig. 3. Data Process for QASPER Dataset

C. Data Transformation

1) ArXiv Dataset: Metadata extraction involved retrieving basic information like title, author(s), publication date, etc., crucial for document sorting and retrieval. The content extraction step was pivotal, as it contained substantive information necessary for analysis. Leveraging the PyPDF2 library, we efficiently extracted metadata and content, storing metadata in JSON format and content as plain text files. To enhance the handling and comprehension of lengthy technical documents, we implemented a method to divide PDFs into manageable text chunks. Using the 'PyPDFDirectoryLoader' tool from the 'langchain' framework, over 1,000 PDFs were given as an input and split into chunks of 200 characters each with a 20-character overlap, resulting in 195,000 text chunks. This division maintains crucial information while making the content more accessible for further processing. The next step involved converting these text chunks into embeddings, numerical vectors that capture the context and meaning of the text. We utilized the 'HuggingFaceEmbeddings' package with the 'sentence-transformers/all-MiniLM-L6-v2' pre-trained language model to create these embeddings. These embeddings were then managed and stored in Pinecone's vector storeage, which helps in fast retrieval and matching of user queries with relevant text chunks. Embeddings are used in enabling computers to understand and interact with textual data effectively. They act as a bridge between language and computational processes, ensuring that our system can process and retrieve information accurately. This approach not only makes text more understandable to computers but also enhances the overall functionality and efficiency of our system in handling technical documents related to machine learning.

2) QASPER Dataset: The data transformation process for the QASPER dataset involves several steps to extract structured information from the 'qas' column, which is originally formatted as JSON. This column contains a variety of information related to each question-answer pair, including the question itself, whether the answer is 'Yes' or 'No,' the actual answer, evidence supporting the answer, and highlighted evidence. Additionally, the 'qas' column contains the Paper ID, which uniquely identifies each paper in the dataset, and the Paper Title. To transform this data into a more usable format, a customized Python script is employed. The script parses the JSON structure of the 'qas' column to extract the aforementioned fields, such as the Paper ID, Paper Title, Question, Yes/No indicator, Answer, Evidence, Highlighted Evidence, and a Merged Title and Question column. This process involves careful handling of the JSON format to ensure all relevant information is correctly extracted and formatted into a new dataset. The extracted dataset from 'qas' column used in this study consists of several columns as mentioned. To streamline the data for analysis and improve its usability, a detailed transformation process was implemented. Initially, the focus was on extracting specific information from the dataset. This involved parsing the columns related to Question, Answer, and Paper Title to isolate the relevant content. For the Question column, the text was extracted to provide a clear and concise representation of the questions posed in each paper. Similarly, the Answer column was populated with the corresponding answers, ensuring that the information was relevant to the specific question and paper. Additionally, the transformation process involved incorporating the Paper Title into the dataset, linking each question and answer pair to its respective paper. This linkage enables to easily identify the context in which each question was asked and answered, providing valuable insights into the research papers' content. Overall, the data transformation process resulted in a more structured and informative dataset, facilitating easier analysis and interpretation of the research papers' content. The extracted fields can be used to train and evaluate models, improve question-answering capabilities, and gain insights into the content and structure of the original papers. This process enhanced the usability and utility of the QASPER dataset for research and development purpose in this research.

IV. METHODOLOGY

A. LLAMA-2-7B Variant

According to their documentation [9] LLAMA model's algorithm is an adaptation of the Transformer model proposed in the paper 'Attention is All you Need' [8], which is designed to effectively process input text by converting individual words or tokens into high-dimensional vectors called embeddings. These embeddings capture semantic and syntactic characteristics, forming the basis for subsequent layers in the model. Notably, LLAMA [9] incorporates rotary positional embeddings (RoPE) [16], which encode absolute position using rotation matrices while integrating relative position dependencies in self-attention. This enhances the model's understanding of sequential and positional context. Additionally, LLAMA [9] adopts pre-normalization using RMSNorm [17] for input sub-layers, improving training stability. The model employs both scaled dot-product attention and grouped-query attention mechanisms to handle extended context lengths and generate coherent language. Furthermore, LLAMA [9] replaces the traditional ReLU activation function with SwiGLU [18] for improved processing of complex language structures. Llama 2 [20], an upgraded version of Llama 1 [9], which was trained using a new set of publically available data where the updated version has a pretraining corpus by 40%, doubled the model's context length, and implemented grouped-query attention[19].

B. Fine-Tuned LLAMA Model

The The fine-tuned LLAMA-2-7b variant is a state-of-the-art model that incorporates several innovative techniques to enhance its conversational abilities as shown in Fig.5. for the base model LLAMA-2-7B [20]. One of the key innovations is the use of Quantized Low-Rank Attention (QLoRA)[14], which is designed to optimize the attention mechanism. As mentioned in Fig.4 given by [14] the QLoRA achieves this by reducing the computational complexity and memory usage of the attention mechanism. It does so by setting the LoRA [15] attention dimension to 64, which allows for a more efficient computation of attention weights [15]. Additionally, QLoRA [14] introduces an alpha parameter to regulate the influence of the low-rank approximation, enabling finer control over the attention mechanism [14]

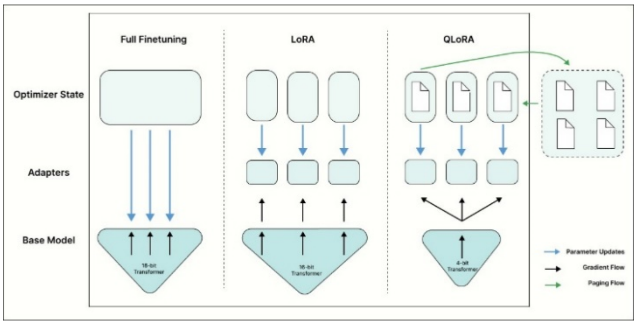

Fig. 4. Full Finetuning , LoRA and QLoRA Working Techniques

In addition to QLoRA, the fine-tuned LLAMA-2-7b model employs several memory optimization strategies to further enhance its performance. We implemented 4-Bit Precision Model Loading to decrease the model's memory footprint by loading model parameters with 4-bit precision. This allows the model to consume less memory, increasing its efficiency and ability to handle larger datasets. Another memory optimization method employed by the model is Mixed-Precision Training with FP16, which uses the benefits of reduced precision (FP16) arithmetic to accelerate training and reduce memory usage.

Furthermore, the training configuration of the fine-tuned LLAMA-2-7b model is carefully designed to balance learning optimization and computational efficiency. This includes optimizations applied to the optimizer and weight decay parameters, which are crucial for achieving high performance in deep learning models. By optimizing these parameters, the model can achieve better convergence and higher accuracy while using fewer computational resources.

C. *Mistral 7B*

The Mistral 7B model, a large language model (LLM), represents a significant advancement in natural language processing, balancing performance and computational efficiency. As referred by [10] Mistral 7B is built upon the transformer architecture, Mistral 7B exhibits several innovative features, including a substantial scale with 32 layers and a width of 4096 units per layer[10]. Multi-head attention mechanisms [22] and grouped-query attention [19] enhance the model's ability to process complex language patterns effectively. Mistral 7B employs the GELU activation function [21] and a carefully designed output layer to generate coherent and contextually appropriate language.

D. *Fine-Tuned Mistral Model*

The fine-tuned Mistral-7b model undergoes comprehensive optimization to enhance its conversational abilities as mentioned in the Fig.6. The optimization process for the fine-tuned Mistral-7b model involves several key techniques aimed at improving its conversational abilities. One crucial optimization is the integration of Quantized Low-Rank Attention (QLoRA), a method that reduces the computational complexity of attention mechanisms by quantizing attention weights and decomposing the attention matrix into low-rank factors [14] as mentioned in Fig4. This approach helps in reducing the computational cost of the model while maintaining its performance[14]. Another important optimization technique is the use of memory-efficient loading techniques such as "bitsandbytes." This technique minimizes the memory footprint of the model by storing and loading parameters in a compact format, reducing the amount of memory required during training and inference.

Furthermore, configuration adjustments are made to optimize learning and memory usage. This includes tuning batch sizes to balance computational efficiency and model performance, adjusting gradient accumulation steps to manage memory constraints, and applying gradient clipping to prevent exploding gradients during training.

In addition to these techniques, the "paged_adamw_32bit" optimizer is employed with specific learning rate and weight decay parameters. This optimizer is designed to handle large models efficiently by using a paged parameter update strategy, which reduces memory consumption and improves training stability.

Collectively, these optimizations enhance the adaptability and performance of the Mistral-7b model in specialized conversational tasks. By reducing computational complexity, minimizing memory usage, and optimizing learning parameters, the model is better equipped to handle complex conversational scenarios while maintaining efficiency and stability during training.

E. *Retrieval Augmented Generation*

Retrieval-Augmented Generation (RAG) is an approach that enhances the capabilities of large language models (LLMs) by integrating external knowledge retrieval as given in paper [24]. While LLMs excel in generating text, translating languages, and answering questions, they can sometimes falter in factual accuracy and staying current with evolving information. [24] RAG tackles these challenges through a structured process. RAG," or Retrieval-Augmented Generation, is a technique used in artificial intelligence, particularly in natural language processing. It combines two main components: retrieval of information and generation of responses.

When RAG receives a question or a prompt, the first step is to fetch relevant information [24]. Leveraging techniques like semantic similarity, this component sifts through the retrieved data to pinpoint the most relevant passages, ensuring that only highly informative content is selected.

The retrieved documents or snippets are then seamlessly integrated with the original prompt, forming an augmented input. This augmentation significantly enriches the context available to the LLM, enabling it to better grasp the user's intent and understand the factual background pertinent to the query [24].

Subsequently, this augmented input is fed into the LLM for generation. [24] By leveraging both its inherent knowledge and the externally retrieved information, the LLM produces a response that is not only more accurate and factually grounded but also tailored precisely to the user's specific query. The synergy between the LLM's capabilities and the retrieved knowledge enhances the overall quality of generated responses [24].

In our research using the RAG approach the information that is sent to four types of language models, referred to here as Base Lamma 2, Base Mistral LLM, Custom Fine-tuned LLAMA 2, Custom Fine-tuned Mistral 7B. These custom fine-tuned models have been previously fine-tuned using a dataset specifically designed for questions and answers of the Machine Learning articles.

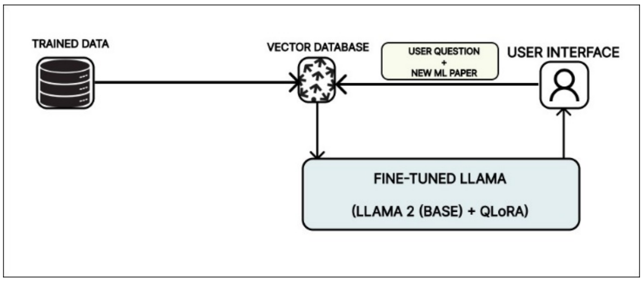

Fig. 5. RAG + Fine-tuned Llama 2

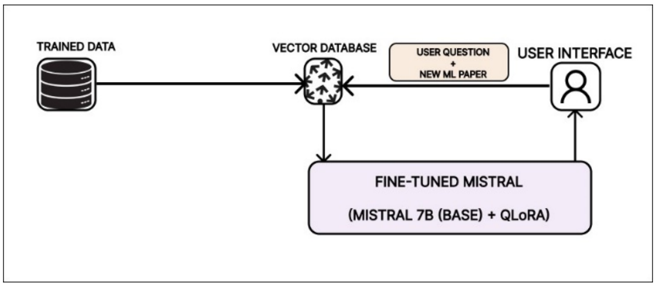

Fig. 6. RAG + Fine-tuned Mistral 7B

V. EVALUATION AND RESULTS

A. *Evaluation*

Evaluations (evals) are crucial tasks which are used to assess the quality of the output produced by Large Language

Model (LLM) systems. This involves in generating output from input prompts and comparing it against ideal answers to evaluate the system's performance. Evals are instrumental in measuring accuracy, which is essential for LLM systems deployed in various applications. These metrics will help identify blind spots or areas of weakness in real-time by evaluating model outputs based on actual user interactions, enabling ongoing improvements and optimizations. Evals is used as the best metric for comparisons between fine-tuned models and their base foundational model, highlighting improvements in accuracy over time.

In our study, to comprehensively assess the performance of our proposed system, we employed the RAGAS EVALS [23] framework. This involved creating a dedicated dataset comprising questions, retrieved contexts, answers, and ground truth information. The dataset facilitated a thorough evaluation of the system's capability to retrieve relevant information from academic papers and generate accurate and contextually relevant responses

The evaluation process utilized a comprehensive set of metrics, categorized under Retriever Evaluation and Generation Evaluation, to measure the system's effectiveness in various aspects

i) Retriever Evaluation

Context Precision [23]: This metric assesses how accurately the retriever system selects relevant information from the retrieved data in response to a query. It measures the ratio of pertinent information to the total retrieved information, indicating the system's ability to filter out irrelevant data and forward only relevant content for further processing.

Context Recall [23]: Context recall evaluates the system's capability to retrieve all relevant information available in the data source. It measures the completeness of information retrieval, highlighting the system's effectiveness in sourcing necessary data to generate accurate responses, even if it involves retrieving a broader set of documents.

ii) Generation Evaluation

Faithfulnessb[23]: Faithfulness gauges the factual accuracy of the answers generated by the system. It involves comparing generated answers with ground truth data to ensure that the system maintains a high standard of truthfulness and reliability in its responses.

Answer Relevancy [23]: Answer relevancy measures how well the generated answers address the posed questions. It ensures that the system not only provides factually accurate responses but also tailors them appropriately to the specific context of the question, enhancing utility for the user.

iii) Overall Answer Quality

AspectCritique [23]: AspectCritique evaluates various dimensions of answer quality beyond traditional metrics like precision and recall. It includes aspects such as harmfulness, coherence, conciseness, and absence of malicious content, providing a holistic assessment of the system's outputs to ensure accuracy, relevance, safety, and user-friendliness.

B. *RAG with Custom Fine-tuned LLAMA 2 Model*

The integration of the Retrieval-Augmented Generation (RAG) framework with our custom fine-tuned LLAMA 2 model yielded a substantial improvement in faithfulness, reaching a score of 0.9074. This high score suggests that the fine-tuned model excels at providing information that is true to the given context, ensuring factual accuracy in its responses.

While the context relevancy and answer relevancy scores showed a slight increase at 0.4918 and 0.9743, respectively, the context recall score experienced a small decrease to 0.9167. Importantly, the fine-tuned model maintained a harmfulness score of zero, indicating no generation of harmful content, thus ensuring the safety and reliability of its outputs.

C. *RAG with Custom Fine-tuned Mistral Model*

The integration of RAG with our custom fine-tuned Mistral model resulted in a faithfulness score of 0.2441, indicating an improvement over the base Mistral model. However, this score was lower than the fine-tuned LLAMA 2 model, suggesting room for further enhancement in terms of factual accuracy.

The context relevancy score decreased to 0.4172, while the answer relevancy increased to 0.9592. The model showcased a high context recall score of 0.9583 and maintained a zero-harmfulness score, ensuring the safety and reliability of its outputs.

D. *Comaprative Analysis*

The evaluation results across the different models are summarized in Table I and graphically represented in Fig. 7, facilitating direct comparisons between the base LLaMa, base Mistral, fine-tuned LLaMa, and fine-tuned Mistral models. These comparisons highlight the relative strengths and weaknesses of each model configuration.

TABLE I. RESULTS TABLE

Model	**Model Evaluation**				
	Context Relevancy	***Faithfulness***	***Answer Relevancy***	***Context Recall***	***Harmfulness***
RAG + Base LLaMa	0.4667	0.0833	0.9461	1	0
RAG + Base Mistral	0.5238	0.2	0.918	1	0
RAG + Custom FT LLaMa	0.4918	0.9074	0.9743	0.9167	0
RAG + Custom FT Mistral	0.4172	0.2441	0.9592	0.9583	0

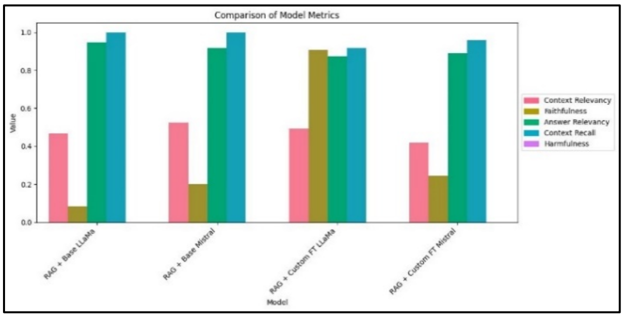

Fig. 7. Graphical Model Evaluation Comparision

E. Discussion and Implications

The evaluation results demonstrate the effectiveness of our proposed system in understanding academic papers and generating accurate and relevant responses. The fine-tuned LLAMA 2 model exhibited superior performance in terms of faithfulness, suggesting its ability to provide factual information that aligns with the given context more accurately than the other models evaluated.

While both fine-tuned models showed improvements over their base counterparts, the fine-tuned LLAMA 2 model outperformed the fine-tuned Mistral model in several key metrics, including faithfulness and context recall. However, the fine-tuned Mistral model performed better in terms of answer relevancy, indicating its strength in tailoring responses to specific questions more effectively.

It is noteworthy that both fine-tuned models maintained excellent performance in avoiding harmful content generation, underscoring their reliability and safety for practical applications.

The hybrid approach of integrating the Retrieval-Augmented Generation (RAG) framework with fine-tuned language models (LLMs) demonstrated its effectiveness in leveraging the strengths of both components. The fine-tuned LLMs provided enhanced factual accuracy and context understanding, while the RAG framework facilitated efficient retrieval and utilization of relevant information from the academic papers.

Moving forward, potential areas for further improvement could focus on enhancing the context relevancy and context recall aspects of the system, as well as exploring techniques to strike an optimal balance between faithfulness and answer relevancy. Additionally, expanding the evaluation dataset and incorporating more diverse academic domains could provide valuable insights into the system's generalization capabilities across different fields of study.

VI. CONCLUSION AND FUTUREWORK

A. Conclusion:

In conclusion, our study showcases the successful integration of the Retrieval-Augmented Generation (RAG) framework with custom fine-tuned language models (LLAMA 2 and Mistral) to enhance conversational abilities, particularly in understanding academic papers. Through rigorous fine-tuning and evaluation, we have demonstrated significant enhancements in the faithfulness and context relevancy of generated content, indicating the efficacy of our approach in aligning responses with given contexts. While the fine-tuned LLAMA 2 model showcased superior faithfulness scores compared to Mistral, both variants exhibited strong capabilities in maintaining context recall and answer relevancy. Notably, our models achieved a zero score in harmfulness, affirming their reliability in generating safe and appropriate content. These results underscore the promise of fine-tuned language models integrated with retrieval-augmented generation frameworks to advance natural language understanding and generation systems. This advancement is crucial for deeper comprehension of complex academic papers, ultimately fostering innovation in machine learning research.

B. Future Work

Moving forward, several avenues for improvement and expansion emerge from our study. First and foremost, there is a need for continued refinement of the Mistral model to enhance its performance and applicability, given its broader potential impact. This could involve further fine-tuning, exploring diverse training datasets, and refining the model architecture. Moreover, there is a significant opportunity to enhance the user interface (UI) of our system. Improvements in UI design can greatly enhance accessibility and usability, leading to broader adoption and increased effectiveness in real-world scenarios. This could include intuitive features for inputting queries, visual aids for understanding model outputs, and interactive elements for refining generated responses.

Additionally, future research could focus on exploring the scalability of our approach to handle larger and more diverse datasets, as well as investigating the integration of additional modalities such as images or graphs to further enrich the understanding and generation capabilities of the models. By addressing these areas of improvement and expansion, we can continue to push the boundaries of natural language processing and contribute to the advancement of conversational AI systems tailored for complex domains like academic paper understanding.

References

[1] Guo, Q., Cao, S., & Yi, Z. (2022). A medical question answering system using large language models and knowledge graphs. *International Journal of Intelligent Systems*, *37*(11), 8548–8564. https://doi.org/10.1002/int.22955

[2] Darapaneni, N., Chetan, P., Paduri, A. R., Gaddala, A., Tiwari, G., Basu, S., & Parvathaneni, S. (2021). Building a question and answer system for news domain. *2021 2nd International Conference on Secure Cyber Computing and Communications (ICSCCC)*. https://doi.org/10.1109/icsccc51823.2021.9478180

[3] Liu, Y., & Lapata, M. (2019). Text Summarization with Pretrained Encoders. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1908.08345

[4] See, A., Liu, P. J., & Manning, C. D. (2017). Get To The Point: Summarization with Pointer-Generator Networks. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1704.04368

[5] De Curtò, J., De Zarzà, I., & Calafate, C. T. (2023). Semantic Scene Understanding with Large Language Models on Unmanned Aerial Vehicles. Drones, 7(2), 114. https://doi.org/10.3390/drones7020114

[6] Siriwardhana, S., Weerasekera, R., Wen, E., Kaluarachchi, T., Rana, R., & Nanayakkara, S. (2023). Improving the domain adaptation of Retrieval Augmented Generation (RAG) models for open domain question answering. Transactions of the Association for Computational Linguistics, 11, 1–17. https://doi.org/10.1162/tacl_a_00530

[7] Wang, C., Ong, J., Wang, C. Y. J., Ong, H., Cheng, R. R., & Ong, D. (2023). Potential for GPT Technology to optimize future clinical Decision-Making using Retrieval-Augmented Generation. Annals of Biomedical Engineering

[8] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. arXiv (Cornell University), 30, 5998–6008. https://doi.org/10.48550/arXiv.1706.03762

[9] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, É., & Lample, G. (2023). LLAMA: Open and Efficient Foundation Language Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2302.13971

[10] Jiang, A. Q., Sablayrolles, A., Arthur, M., Bamford, C., Chaplot, D. S., De Las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M., Stock, P., Scao, T. L., Lavril, T., Wang, T. J., Lacroix, T., & Sayed, W. E. (2023). Mistral 7B. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2310.06825

[11] Y. Liu, D. Iter, X. Yi - chong, S. Wang, R. Xu, and Z. C, "G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment," arXiv (Cornell University), Mar. 2023, doi: 10.48550/arxiv.2303.16634.

[12] “arXiv.org e-Print archive.” https://arxiv.org/

[13] “Qasper Dataset — Allen Institute for AI.” https://allenai.org/data/qasper

[14] Dettmers, Tim, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. "Qlora: Efficient finetuning of quantized llms." arXiv preprint arXiv:2305.14314 (2023).

[15] Hu, Edward J., Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. "Lora: Lowrank adaptation of large language models." arXiv preprint arXiv:2106.09685 (2021).

[16] Su, J., Lu, Y., Pan, S., Wen, B., & Liu, Y. (2021). RoFormer: Enhanced Transformer with Rotary Position Embedding. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2104.09864

[17] Zhang, B., & Sennrich, R. (2019). Root mean square layer normalization. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1910.07467

[18] Shazeer, N. (2020). GLU variants improve transformer. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2002.05202

[19] Ainslie, J., Lee-Thorp, J., De Jong, M., Zemlyanskiy, Y., Lebrón, F., & Sanghai, S. (2023). GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2305.13245

[20] Touvron, H., Martin, L., Stone, K. H., Albert, P. J., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., . . . Scialom, T. (2023). Llama 2: Open foundation and Fine-Tuned chat models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2307.09288

[21] Hendrycks, D., & Gimpel, K. (2016). Gaussian Error Linear Units (GELUS). *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1606.08415

[22] Cordonnier, J., Loukas, A., & Jäggi, M. (2020). Multi-Head attention: collaborate instead of concatenate. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2006.16362

[23] Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2023). RAGAS: Automated Evaluation of Retrieval Augmented Generation. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2309.15217

[24] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP tasks. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2005.11401