Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 4, April : 2024

DOCUMENT BASED LLM POWERED CHATBOT ASSISTANT

Ms.P.Sravani[1], Chitti Hemanth Kumar[2], Gudla Manikanta[3], Chenna Mastan Reddy[4,] Gallela Nirmala[5]

[1] Assistant Professor,Department of Computer Science & Engineering, Raghu Engineering College, Vishakhapatnam, Andhra Pradesh

[2,3,4,5] Students of B-TECH, Raghu Engineering College , Vishakhapatnam, Andhra Pradesh

Email:- hemanthkumarchitti2002@gmail.com,gmanikanta135@gmail.com, 20981a0532@raghuenggcollege.in, 20981a0555@raghuenggcollege.in

ABSTRACT

The project's goal is to create a dual-functional question-answering assistant that can respond to inquiries in-depth by utilizing both document-based and image-based techniques. This novel approach tackles a variety of queries, including those requiring visual context or involving intricate issues represented by pictures, by fusing the strength of Large Language Models (LLMs) with cutting-edge image processing techniques. The assistant's document-based component analyzes and extracts data from large document collections using LLMs, such as those from the Lang- Chain framework. By depending on the data in the target documents, this method aims to produce correct and context-aware answers, improving the system's capacity to respond accurately to a range of queries. The image-based component presents a brand-new technique for handling and analyzing visual information. The assistant can recognize and comprehend issues or situations shown in photos by integrating image recognition and analysis capabilities. This feature offers a novel approach to interact with and resolve non-text-based problems by enabling the assistant to respond to inquiries that are specifically linked to the visual material. When these two elements are combined, a flexible system that can respond to a wide range of questions is produced. The assistant is prepared to provide pertinent and precise replies, regardless of whether the query is dependent on textual data or calls for understanding a visual issue. This dual capability improves the assistant's usefulness in a variety of fields and applications in addition to increasing its capabilities. The project offers a strong and creative solution for customers looking for in-depth answers to their questions, marking a substantial advancement in the field of question-answering systems. This assistant is a helpful tool for a wide range of applications since it uses powerful image processing techniques and LLMs to deliver a highly customizable and tailored solution that can meet specific needs and domains.

Keywords: educational support system, intelligent chatbot, document query engine, fine-tuning, sentence Transformer, parameter-efficient fine-tuning, QLoRA, FAISS, content transformation, contextual embeddings, adaptive retrieval, content processing, PDF extraction, table analysis.

1. INTRODUCTION

The project's goal is to create a dual-functionality question-answering assistant that can deliver thorough replies to inquiries by combining document-based and image-based methods. Through the use of sophisticated image processing techniques for visual content and Large Language Models (LLMs) for textual information, this novel approach seeks to improve the capabilities of current question-answering systems. The intention is to develop a flexible assistant that can answer a broad variety of queries, regardless of whether they need to comprehend a visual problem or are based on textual information. This dual functionality makes the assistant a valuable tool for a variety of areas, including customer service, information search, and personal support. It also increases the assistant's usefulness across multiple domains and applications.

1. Document-based Question Answering: Utilizing LLMs and RAG to extract and generate answers from extensive document collections, ensuring precise and context-aware responses to a wide range of queries.
2. Image-based Question Answering: Incorporating advanced image processing and recognition capabilities to understand and respond to questions based on visual content. This aspect is crucial for addressing queries that require understanding or interpreting images.

By focusing on these areas, the project aims to create a comprehensive and adaptable question-answering assistant that can cater to a broad spectrum of user needs, making it a valuable asset in the field of AI-driven information retrieval and personal assistance systems.

The advent of LLMs like GPT-4 and BERT has marked a new era in information retrieval, moving beyond the limitations of traditional keyword-based searches to introduce a more intuitive, human-like interaction with data. These models are not just about retrieving data based on input queries; they are about understanding the context, the subtleties of language, and the underlying intent behind each query. This transformative approach allows for generative question answering (GQA), where complex queries are met with comprehensive, relevant, and accurate answers.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 4, April : 2024

2. LITERATURE SURVEY

Retrieval-Augmented Generation for Large Language Models: The survey begins with a comprehensive review of RAG for LLMs, discussing the challenges faced by LLMs, such as hallucination, outdated knowledge, and non-transparent reasoning processes. It introduces RAG as a promising solution that incorporates external database knowledge to improve the accuracy and credibility of LLM-generated responses. The review covers the progression of RAG paradigms, including Naive RAG, Advanced RAG, and Modular RAG, and scrutinizes the tripartite foundation of RAG frameworks: retrieval, generation, and augmentation techniques. It also highlights state-of-the-art technologies in these components and presents an up-to-date evaluation framework and benchmark.

Recent Developments in RAG: The survey further delves into recent studies that have contributed to the advancement of RAG systems. It mentions several notable works, such as Recomp, Demonstrate-Search-Predict, Replug, Augmented Large Language Models with Parametric Knowledge Guiding, Self-RAG, Benchmarking large language models in retrieval-augmented generation, Ragas, Ares, and Murag. These studies have focused on improving RAG systems through various mechanisms, including compression, selective augmentation, self-reflection, and the integration of parametric knowledge guiding. They have also contributed to the development of benchmarking frameworks and the evaluation of RAG systems.

MultiModal RAG: The survey highlights the importance of multi-modal RAG in enhancing visual question answering and text-to-image generation. It mentions papers like MuRAG, REVEAL, and Re-Imagen, which have focused on integrating visual and textual data to improve the capabilities of RAG systems. These studies underscore the potential of multi-modal RAG in addressing the limitations of text-only RAG systems and in providing more comprehensive and accurate responses to queries that involve both textual and visual content.

In summary, the literature survey provides a detailed overview of the current state of research in RAG for LLMs, emphasizing the advancements in RAG systems, the integration of external knowledge, and the development of multi-modal RAG. It also outlines the challenges and future research directions in this field, highlighting the potential of RAG to significantly enhance the capabilities of LLMs in answering complex, knowledge-intensive questions.

2.1.Interactive document Summarizer using LLM technology [1].

The development of Large Language Models (LLMs) and generative AI has seen significant advancements, with notable milestones like ChatGPT in 2022, making these technologies more accessible to the public. Generative AI is a hot topic in the technology sector, offering a wide range of subjects for design science research. The thesis focuses on the properties and applications of LLMs and generative AI, as well as the capabilities and potential of Retrieval Augmented Generation (RAG). It involves the creation of a software application capable of interactive discussions within the context of given documents, utilizing modern development tools and environments. The application is tested with suitable test material and compared to existing applications with similar functionality, demonstrating its effectiveness. The implementation method resulted in an application that is comparable to commercial solutions, capable of summarizing lengthy documents into clear, concise text and answering proposed questions with accurate and relevant information.

2.2.PEARL: Personalizing Large Language Model Writing Assistants [2].

The paper introduces PEARL, a retrieval-augmented Large Language Model (LLM) writing assistant designed to personalize LLM outputs to match the author's communication style and specialized knowledge, addressing a significant barrier in the development of writing assistants. PEARL is personalized with a generation-calibrated retriever, trained to select historic user-authored documents for prompt augmentation, aiming to best personalize LLM generations for a user request. The training of the retriever includes a novel method for selecting training data that identifies user requests likely to benefit from personalization and documents that provide that benefit, as well as a scale-calibrating KL-divergence objective to ensure the retriever closely tracks the benefit of a document for personalized generation. The paper demonstrates the effectiveness of PEARL in generating personalized workplace social media posts and Reddit comments, showcasing its potential in practical applications.

2.3.Emory Infobot: A LLM-powered Virtual Assistant for University Students [3].

Emory Infobot is a question-answering system developed for Emory University students, utilizing a Retrieval Augmented Generation (RAG) system to provide accurate information on various college-related topics. This system combines text generation from large language models with document retrieval from a dataset of question-answer pairs related to Emory University. To evaluate its performance, the system underwent testing with real Emory students, covering a broad spectrum of queries including academic schedules, campus facilities, student life, and extracurricular activities. The testing process involved iterative use of different sets of data, with each set being used as a training set in turn, followed by classification of the remaining sets. The system's development and testing are supported by a range of references on RAG, including overviews of large language models, studies on improving math question-answering with RAG, and tools for evaluating RAG systems, enhancing the understanding of its underlying technology and applications.

3. IMPLEMENTATION STUDY

The existing systems for the project, as outlined in the provided sources, primarily focus on either document-based or image-based question-answering approaches, with some attempts to integrate both functionalities.

Document-based Systems:

The document-based component of the project leverages Large Language Models (LLMs) for question-answering. A notable example is the use of the LangChain framework in conjunction with LLMs like LLaMA 2, which is pretrained and fine-tuned with extensive data, making it highly capable of analyzing and extracting information from vast document collections. This approach ensures precise and context-aware answers by drawing from the information within the target documents. The integration of LLMs with the LangChain framework and tools like FAISS (Facebook AI Similarity Search) for efficient similarity search and clustering of dense vectors enhances the system's ability to provide accurate responses to a variety of questions.

Image-based Systems:

The image-based component introduces a novel method for processing and interpreting visual data. A notable example is the use of Retrieval Augmented Generation (RAG) and LLMs to build an image question-answering system. This system is capable of answering queries using data from image-based sources like infographics, tapping into the rich information embedded in images. The process involves text extraction from images, text embedding, context retrieval, and response generation. The system's effectiveness is influenced by the number of infographics used as contexts, which must be carefully balanced to avoid overwhelming the LLM with noise or lacking necessary information.

Integration Challenges:

While there have been advancements in both document-based and image-based question-answering systems, the integration of these two components into a single system presents a significant challenge. Existing systems either focus on textual information or visual content, but there is a lack of comprehensive solutions that can effectively handle both types of queries. Achieving this integration requires overcoming technical and conceptual challenges related to the integration of LLMs and image processing techniques, as well as ensuring the system's ability to accurately interpret and respond to a wide range of queries.

3.1 PROPOSED METHODOLOGY

The proposed system for the dual-functionality question-answering assistant, leveraging Retrieval Augmented Generation (RAG) and Large Language Models (LLMs), is designed to enhance the capabilities of LLMs in answering questions related to data, files, or documents by providing the necessary context. This system aims to develop a comprehensive solution capable of answering queries using data from both textual and image-based sources, thereby tapping into the rich information embedded in images and documents. Here's an overview of the proposed system.

3.1.1. Input Processing: The system begins by receiving a query, which can be either textual or visual in nature. For textual queries, the input is processed using LLMs, which are capable of understanding and generating responses based on the context and semantics of the text. For visual queries, the system employs advanced image processing techniques to extract relevant features and information from the visual content.

3.1 2. Retrieval: The system indexes a series of related documents by chunking them, generating embeddings of the chunks, and indexing them into a vector store. At inference, the query is also embedded in a similar way. The relevant documents are obtained by comparing the query against the indexed vectors, also denoted as "Relevant Documents".

3.1.3. Generation: The relevant documents are combined with the original prompt as additional context. The combined text and prompt are then passed to the LLM for response generation, which is then prepared as the final output of the system to the user. This process ensures that the system can pull the relevant information needed for the model to answer the question appropriately, even when the LLM lacks direct knowledge of the current events or specific context.

3.1.4. Augmentation: The system incorporates the "Augmentation" aspect of RAG, which involves enhancing the capabilities of LLMs by providing them with additional context or information that is not directly present in the input query. This can include external knowledge bases, additional documents, or other relevant information that can help the LLM generate more accurate and comprehensive responses.

3.1. 5. Evaluation and Improvement: The system includes mechanisms for evaluating the accuracy and effectiveness of the generated answers. This involves comparing the assistant's responses against correct answers or benchmarks to assess its performance. The evaluation process also includes feedback loops that allow the system to learn and improve over time, enhancing its capabilities and accuracy.

3.1.6. Applications: The proposed system is designed for applications across various domains, including medical diagnostics, e-commerce, and virtual personal assistants. It can analyze medical images and answer specific questions about diagnoses, treatment

options, or patient conditions. In e-commerce, it can improve product descriptions by analyzing images and generating descriptive captions, enhancing the shopping experience. Virtual personal assistants can benefit from the broad understanding of the system to provide comprehensive assistance to visually impaired individuals by describing images and answering related questions.

.4 METHODOLOGY and Alogritham

To develop a dual-functionality question-answering assistant that leverages both document-based and image-based approaches, the project can be broken down into several key modules, each addressing a specific aspect of the system's functionality. Here's a step-by-step description of the modules based on the provided abstract and the information from the sources:

4. 1. **Data Collection and Preprocessing**

- Objective: Gather and prepare a comprehensive dataset of texts or articles relevant to potential queries, as well as a collection of images and their associated questions.

-Methods: Collect data from various sources, including documents, articles, and image-based datasets. Preprocess the data by cleaning, tokenizing, and organizing the text for detailed analysis. For images, ensure they are labeled with relevant questions or captions.

4.2. **Text Indexing**

- Objective: Create an index or database for quick document retrieval based on user questions.

- Methods: Use techniques like TF-IDF or word embeddings for efficient indexing of the text data. This step is crucial for enabling the system to quickly identify relevant documents or sections in response to user queries.

4.3. **Image Indexing and Processing**

- Objective: Develop a method for indexing and processing images to facilitate their retrieval and analysis in response to visual queries.

- Methods: Implement image indexing techniques similar to those used for text, ensuring that images can be quickly retrieved based on visual content. For processing, employ advanced image recognition and analysis capabilities to understand the content and context of images [2].

4.4. **Question Processing Module**

- Objective: Analyze user questions to understand their intent and context.

- Methods: Use tokenization, semantic analysis, and other NLP techniques to parse and understand the user's question. This module should be capable of handling both textual and visual queries, leveraging M-LLMs for multimodal understanding [1][2].

4.5**. Retrieval Module**

- Objective: Search through indexed documents and images to identify the most relevant sections or documents for the queries.

- Methods: Implement retrieval algorithms that can efficiently search through both text and image indexes, applying techniques like named entity recognition or information extraction to pinpoint answers within these documents or images [1].

4.6. **Answer Generation Module**

- Objective: Generate accurate and contextually relevant answers to user queries.

- Methods: Utilize M-LLMs for generating responses to both textual and visual queries. For textual queries, the system should be able to generate coherent answers based on the retrieved documents. For visual queries, the system should analyze the image and question simultaneously, extracting relevant features from both modalities, and synthesizing them into a cohesive understanding [2].

4.7. **User Interface**

- Objective: Develop an interface or app for easy user interaction.

- Methods: Create a user-friendly interface that allows users to pose questions and receive precise answers. This interface should support both textual and visual input, ensuring that users can interact with the system in a natural and intuitive manner.

4.8. **Continuous Evaluation and Feedback**

- Objective: Assess and improve the system’s performance through continuous evaluation and feedback gathering.

- Methods: Implement mechanisms for evaluating the system's accuracy and relevance of answers. Collect user feedback to identify areas for improvement and refine the system's capabilities.

4.2 Algorithm

The proposed algorithm for the dual-functionality question-answering assistant, combining document-based and image-based approaches, leverages the capabilities of MultiModal Large Language Models (MM-LLMs). This algorithm is designed to enhance the performance of question-answering systems by integrating both textual and visual data processing

1. Input Processing: The algorithm begins by receiving a query that can be either textual or visual in nature. For textual queries, the input is processed using LLMs, which are capable of understanding and generating responses based on the context and semantics of the text. For visual queries, the input is processed using advanced image processing techniques to extract relevant features and information from the visual content .

2. Feature Extraction and Fusion: For textual queries, the LLMs extract features from the input text, which may include semantic understanding, contextual relevance, and the ability to generate responses. For visual queries, the algorithm employs image processing techniques to extract features such as object recognition, scene understanding, and visual content description. The extracted features from both textual and visual inputs are then fused to create a unified representation that captures the essence of the query .

3. Query Understanding and Answer Generation: The fused features are used to understand the query's intent and context. This understanding is crucial for generating accurate and relevant answers. The algorithm leverages the capabilities of MM-LLMs to generate responses that are not only contextually accurate but also comprehensive, addressing the query's intent and providing detailed answers.

4. Evaluation and Improvement: The algorithm includes mechanisms for evaluating the accuracy and effectiveness of the generated answers. This involves comparing the assistant's responses against correct answers or benchmarks to assess its performance. The evaluation process also includes feedback loops that allow the system to learn and improve over time, enhancing its capabilities and accuracy.

5. Adaptation and Scaling: The proposed algorithm is designed to be adaptable and scalable, capable of handling a wide range of queries across various domains. It leverages the scalability of LLMs and advanced image processing techniques to ensure that the system can efficiently process and respond to queries as the volume and complexity of data increase

5 RESULTS AND SCREEN SHOTS

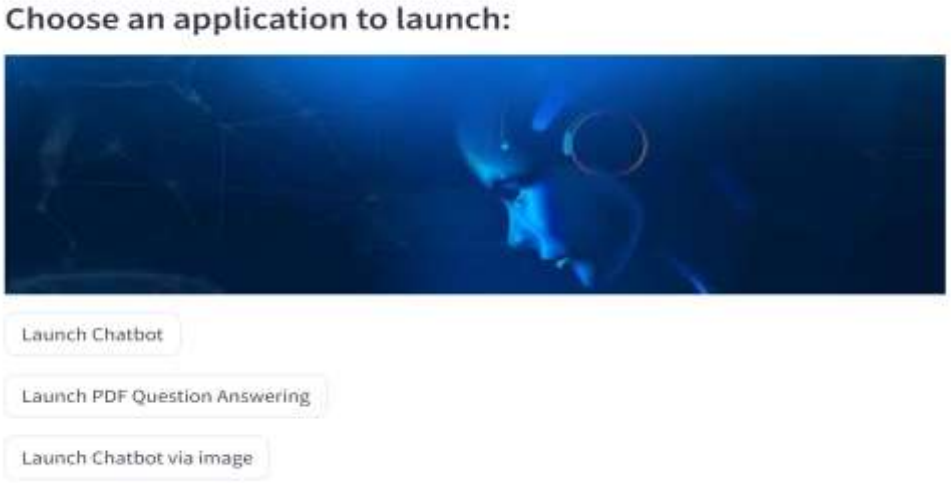

Fig:1 Home page

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 4, April : 2024

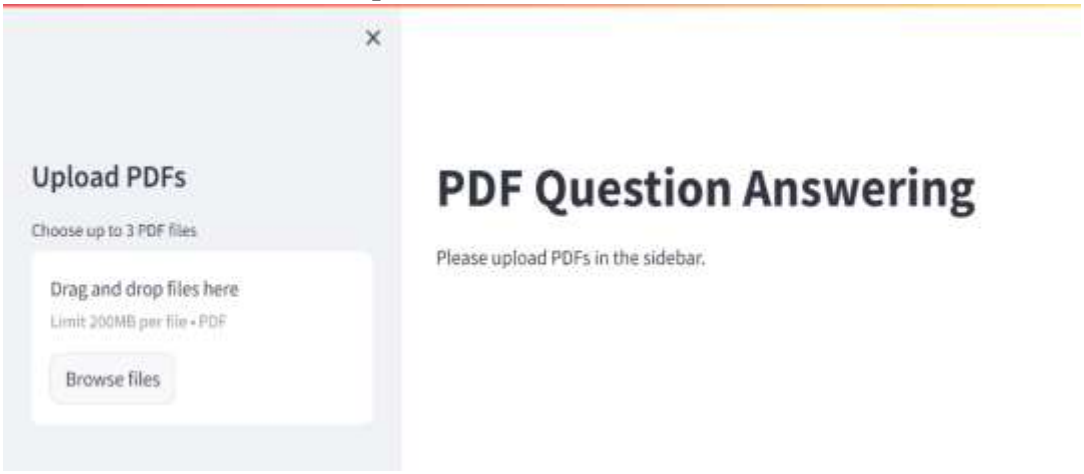

Fig 2:pdf Chabot

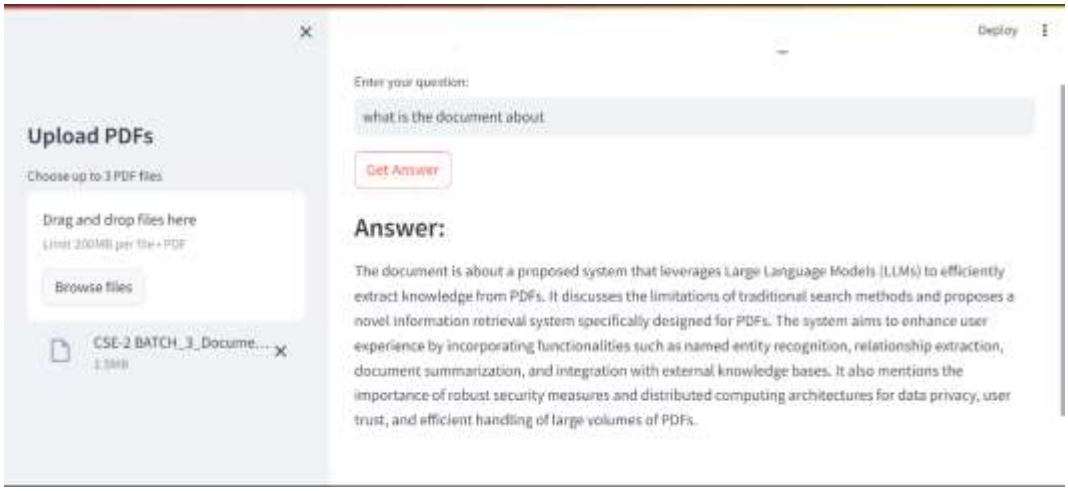

Fig 3:- *working of the assistant on pdf documents*

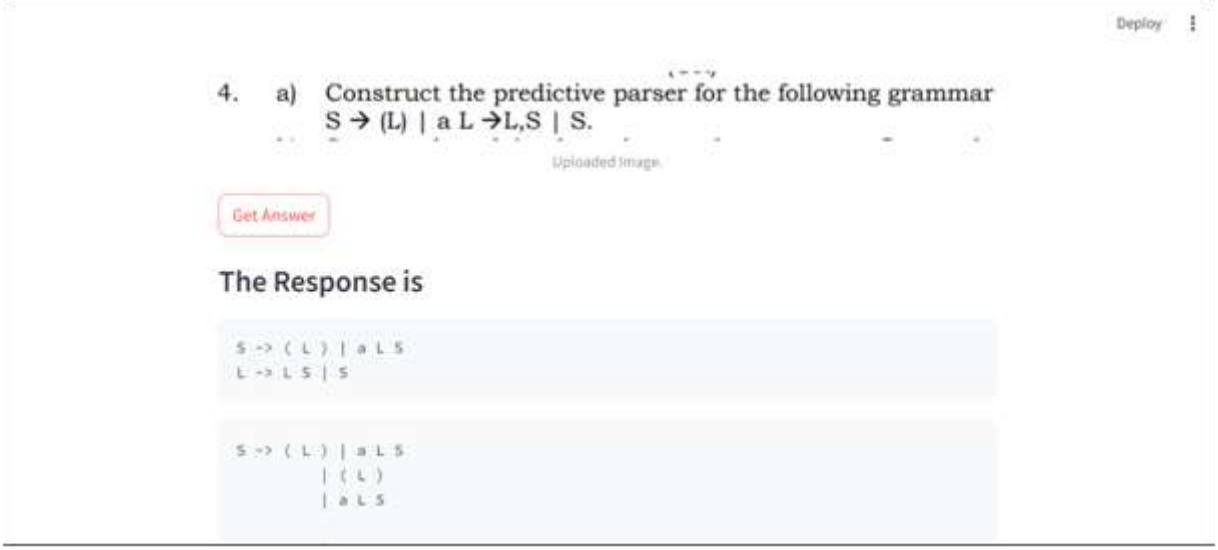

Fig 4:- *working of image based assistant*

6. CONCLUSION AND FUTURE SCOPE

The conclusion of the project involves developing a sophisticated chatbot that leverages both image and document processing capabilities, utilizing Google Vision API for image analysis and OpenAI's Large Language Model (LLM) along with Faiss for document processing. This integration allows the chatbot to understand and respond to a wide array of queries, making it a versatile tool for various applications. Multimodal Interaction: Expanding the chatbot's capabilities to handle more types of media, such as audio and video, by integrating additional APIs or models.
Advanced Natural Language Processing (NLP): Enhancing the chatbot's understanding of user queries through more advanced NLP techniques, including context-aware models.
Customization and Personalization: Allowing users to customize the chatbot's behavior and responses based on their preferences, potentially training the LLM on user-specific data for more personalized interactions.
Integration with Other Services: Extending the chatbot's functionality by integrating it with other services and platforms, such as social media or e-commerce sites, to provide more context-specific information and recommendations.
By focusing on these future extensions, the project can evolve into a more powerful and versatile chatbot that can cater to a wide range of applications and user needs, making it a valuable tool for both individuals and organizations.

7. REFERENCES

[1]S.Mangrulkar,Sayak Paul,"Parameter efficient Fine-Tuning of Billion-Scale Models on Low Resource Hardware", Github, February 2023

[2]Edward Hu, Yelong Shen, Philip Wallis, Lu Wang,"LoRA: Low-Rank Adaptation Of Large Language Models", Microsoft Corporation, 2023

[3]Shalini Dhote, "Parameter-Efficient Fine-Tuning of Large Language Models with LoRA and QLoRA",August, 2023

[4]Keivalya Pandya, Prof.Dr.Mehfuza Holia, "Automating Customer Service using Langchain", Cornell University, October, 2023.

[5]Adith Sreeram, P. Jithendra Sai, "An Effective Query System Using LLMs and Langchain",June, 2023

[6]Zhi-Hua Zhou, "A Theoretical Perspective of Machine with Computational Resource Concerns", Cornell University, May, 2023

[7]Shih-Yang Liu,Chien-Yi Wang,Hongxu Yin,Min-Hung Chen,"DORA:Weights-Decomposed Low-Rank Adaption",February,2024

[8]Niklas Donges, "What is Transfer Learning? Exploring the popular Deep learning Approach",September,2022

[9]Derrick Mwiti, "Transfer Learning Guide: A Practical Tutorial With Examples for Text in Keras", netune.ai, August, 2023

[10]Jeonghoon Kim,Jung Hyun Lee,Se Jung Kwon, "Memory-Efficient Fine-Tuning of Compressed Large Language Models via sub-4-bit Integer Quantization", March, 2023