

A project report on

LLM BASED NEWS CHAT Q&A TOOL

Submitted in partial fulfillment for the award of the degree of

Bachelor of Technology in Computer Science and Engineering with Specialization in Artificial Intelligence and Machine Learning

by

SHUVADIPTA BISWAS (21BAI1294)

PRIYAM CHOWDHURY (21BAI1267)



VIT[®]
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)
CHENNAI

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

November, 2024

A project report on

LLM BASED NEWS CHAT Q&A TOOL

Submitted in partial fulfillment for the award of the degree of

Bachelor of Technology in Computer Science and Engineering with Specialization in Artificial Intelligence and Machine Learning

by

SHUVADIPTA BISWAS (21BAI1294)

PRIYAM CHOWDHURY (21BAI1267)



SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

November, 2024



VIT[®]

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

CHENNAI

DECLARATION

I hereby declare that the thesis entitled "LLM BASED NEWS RESEARCH TOOL" submitted by PRIYAM CHOWDHURY (21BAI1267), for the award of the degree of Bachelor of Technology in Computer Science and Engineering with specialization in Artificial Intelligence and Machine Learning, Vellore Institute of Technology, Chennai is a record of bonafidework carried out by me under the supervision of Dr Sweetlin Lemalatha

I further declare that the work reported in this thesis has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university.

Place: Chennai

Date: 20/11/24

Priyam Chowdhury
Signature of the Candidate



VIT

Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)
CHENNAI

CERTIFICATE

This is to certify that the report entitled "LLM Based News Research Tool" is prepared and submitted by Priyam Chowdhury (21BAI1267) to Vellore Institute of Technology, Chennai, in partial fulfillment of the requirement for the award of the degree of Bachelor of Technology in Computer Science and Engineering with Specialization in Artificial Intelligence and Machine Learning is a bonafide record carried out under my guidance. The project fulfills the requirements as per the regulations of this University and in my opinion meets the necessary standards for submission. The contents of this report have not been submitted and will not be submitted either in part or in full, for the award of any other degree or diploma and the same is certified.

Signature of the Guide:

Name: Dr./Prof. Date:

Signature of the Examiner

Name: Dr. V. Nool

Date: 20/11/2024.

Signature of the Examiner

Name: Dr. K. DEVI

Date: 20.11.24

Approved by the Head of Department,

B.Tech. CSE with Specialization in Artificial

Intelligence and Machine Learning

Name: Dr. Sweetlin Hemalatha C

Date:



ABSTRACT

This project will be on the development of an application that performs news research using LangChain and OpenAI. The application is to make reading of news articles easy by allowing the user to input multiple news article URLs and get summaries based on those articles. The application further enables the user to ask specific questions about the news, and the LLM retrieves the answer based on the news articles. This saves the user time and effort by directly receiving specific relevant answers from the news articles instead of having to read through the entire article. LangChain is a software framework helping in integrating large language models (LLMs) into applications. Owing to its characteristic role as an integration framework for language models, the use-cases for LangChain overlap with those of the language model per se, including document analysis and summarization, chatbots, and code analysis. Using LangChain and OpenAI's API has a strong power that derives capabilities upon those already well-capable of handling NLP aspects. In LangChain, building applications that analyze, summarize, or even generate content based on language doesn't necessarily take much effort at all. LangChain enables the connection of language models like GPT through OpenAI to some external data sources, and one can even perform a summary generation, question-answering, and even extracting insights from huge text datasets. This integration utilizes the advanced language models in OpenAI and gives them the workflow and data management capabilities of LangChain, thereby adding more functionality and scaling applications effectively, especially with tasks involving multiple sources or complex pipelines of language processing. It will also employ Vector Databases in word embeddings so the model can understand the context and meanings asked by the user and how it will retrieve an answer.

ACKNOWLEDGEMENT

It is my pleasure to express with deep sense of gratitude to Dr. C. Sweetlin Hemalatha (Professor, SCOPE HoD- B.Tech.CSE - AI & ML) School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, for her constant guidance, continual encouragement, understanding; more than all, she taught me patience in my endeavor. My association with her is not confined to academics only, but it is a great opportunity on my part of work with an intellectual and expert in the field of Artificial Intelligence and Machine Learning

It is with gratitude that I would like to extend my thanks to the visionary leader Dr. G. Viswanathan our Honorable Chancellor, Mr. Sankar Viswanathan, Dr. Sekar Viswanathan, Dr. G V Selvam Vice Presidents, Dr. Sandhya Pentareddy, Executive Director, Ms. Kadhambari S. Viswanathan, Assistant Vice-President, Dr. V. S. Kanchana Bhaaskaran Vice Chancellor, Dr. T. Thyagarajan Pro-Vice Chancellor, VIT Chennai and Dr. P. K. Manoharan, Additional Registrar for providing an exceptional working environment and inspiring all of us during the tenure of the course.

Special mention to Dr. Ganesan R, Dean, Dr. Parvathi R, Associate Dean Academics, Dr. Geetha S, Associate Dean Research, School of Computer Science and Engineering, Vellore Institute of Technology, Chennai for spending their valuable time and efforts in sharing their knowledge and for helping us in every aspect.

In jubilant state, I express ingeniously my whole-hearted thanks to Dr. Sweetlin Hemalatha C, Head of the Department, B.Tech. CSE with Specialization in Artificial Intelligence and Machine Learning and the Project Coordinators for their valuable support and encouragement to take up and complete the thesis.

My sincere thanks to all the faculties and staffs at Vellore Institute of Technology, Chennai who helped me acquire the requisite knowledge. I would like to thank my parents for their support. It is indeed a pleasure to thank my friends who encouraged me to take up and complete this task.

Place: Chennai

Date:

Name of the student

CONTENTS

CONTENTS	7
LIST OF FIGURES	8
LIST OF ACRONYMS	9
CHAPTER 1 INTRODUCTION	
1.1INTRODUCTION	10
1.2OVERVIEW	11
1.3CHALLENGES PRESENT	13
1.4PROJECT STATEMENT	14
1.5OBJECTIVES	15
1.6 SCOPE OF THE PROJECT	16
CHAPTER 2 BACK GROUND	
2.1INTRODUCTION	17
2.2 SURVEY	19
CHAPTER 3 PROPOSED WORK	
3.1OVERVIEW	24
3.2ARCHITECTURE	24
3.3DATASET	27
3.4PROCEDURE	27
3.5 RESULTS	29
CHAPTER 4 CONCLUSION AND FUTURE WORK	
4.1 CONCLUSION	30
4.2 FUTURE WORK	30
REFERENCES	36

LIST OF FIGURES

1.1 PROCESS DIAGRAM	24
1.2 ARCHITECTURAL WORKFLOW DIAGRAM	25
1.3 SAMPLE QUERY OUTPUT	28
1.4 WORD COUNT PER DOCUMENT	29

LIST OF ACRONYMS

LLM - Large Language Model

API – Application Programming InterfaceAI – Artificial Intelligence

GPT – Generative Pre-training TransformerURL – Uniform Resource Locator

XAI – Explainable Artificial Intelligence

Chapter 1

1.1 INTRODUCTION

1.1.1 BACKGROUND OF NEWS ANALYSIS IN THE DIGITAL ERA

Digital media has exponentially grown with news being the preferred information source for millions worldwide. More changes in news production and consumption due to the advancement in digital publishing in terms of access and speed have culminated in information saturation. Most of the time, it is daunting to sift through thousands of articles, opinions, and analyses of opinion pieces and go through mountains of content in order to find something of real value. Traditional search engines only enable locating articles based on keywords, and it does not contain the sophistication to derive deeper understanding or context-driven analysis. In this case, an AI-led tool is capable of scanning news articles and bringing well-informed information and impressions to the users in the most efficient manner.

1.1.2 ROLE OF LARGE LANGUAGE MODELS IN MODERN NEWS ANALYSIS

Large Language Models like OpenAI's GPT models introduce a superior aspect of language comprehension, which allows these systems to be better at understanding, summarizing, and generating human-like text. These LLMs allow processing of news analysis by enabling the language structures to be complex enough to find nuanced sentiments. In fact, nuanced meanings from a host of sources can easily be derived. Therefore, they are well-situated for tasks such as summarizing articles in respect of key topics or answers that may be needed with very high accuracy for specific questions. Processing large volumes of text, these models will help users in gaining insights without manually reading each article, thereby speeding up the process of research.

1.1.3 IMPORTANCE OF CONTEXTUAL AND REAL-TIME ANALYSIS

News today in the world is alive and always unfolding events within real time, as well as new information popping up fast. Traditional static analysis tools can't keep abreast with this tempo. The value to users who must track an event or analyze a multi-sided story is an answer that doesn't just read but also understands and remembers context between articles. Contextual understanding allows the tool to respond much more accurately, so users will be able to move around in narratives across multiple sources of information, check facts, and see how complex events unfold.

1.1.4 TECHNOLOGICAL FOUNDATIONS: OPENAI, LANGCHAIN, AND VECTOR DATABASES

As an overview, the project incorporates a few of the new and most high-technology means to achieve the above-stated objectives. LangChain orients these interactions between models that way in a series of language models, which will mainly act as the key analysis and generation components. Vector databases allow for a built-in approach to enable the decision to allow for fast storage and retrieval of contextual information. Although each of these processes can operate independently, the interaction and interconnection of these elements are the key, intricate features of the system that allow it to carry out meaningful operation of search and analyse user input, provide the user with considered and contextually relevant output.

1.1.5 PURPOSE AND RELEVANCE OF THE PROJECT

The purpose of this tool for news research is to develop a streamlined, intelligent solution for interaction with news. This will offer a more modern way for journalists, researchers, as well as general users, to maneuver information so that they can look beyond what is perceived on the surface and really discover deeper meanings and connections. In an era where false information is rampant and people are time-bound, something that gives quick, accurate, and relevant answers is of immense value, so the project is not just timely but also very meaningful.

1.2 OVERVIEW

1.2.1 SYSTEM ARCHITECTURE AND KEY COMPONENTS

The news research tool architecture is built mainly around three components: Current types of AI models include LLMs from OpenAI, LangChain for organizing the work, and vector databases for storing context.

Translation is handled by the language model for text conscience and creation. Whereas LangChain is a system that addresses more complicated questions so the device can keep multiple-step conversations active. Vector databases enable the context to be transferred from an article to another so that users can ask question which can be answered conditionally to two or more articles. In total, these are semantic systems that can perform complex question answering and abstracting.

1.2.2 USER INTERACTION FLOW

The tool aims at providing a seamless, intuitive user experience. The user journey starts when one or more URLs of news articles are inputted. After that, the tool uses the LLM to summarize each article. Users can then input specific questions that the model will process in the context of the article(s) to generate accurate answers. The interaction flow is highly flexible, thereby allowing users to pursue further questions resulting from an original one, refine the inquiry itself, or compare insights coming from several sources.

1.2.3 FUNCTIONALITY OF CORE TECHNOLOGIES

All these technologies have a specific purpose within the system. OpenAI's LLM executes processes on language, interprets meaning, and then it presents summaries and answers. LangChain structures these answers in such a way that multi-step interactions can be performed where the model could process the information step by step, always refining answers based on the changing inputs from users. The vector database makes sure that there's always efficient storage and retrieval of contextual information so that the responses remain coherent while discussing multiple articles or cross-referencing analysis.

1.2.4 CROSS-ARTICLE COMPARISON AND ANALYSIS

Another advantage of this tool is the capability to cross-article compare information. This facility allows users to trace how stories are changing, or which aspects contradict each other, at least to have a better view about what people had written on one topic. The vector databases retain each article's context, allowing users to draw associative lines between the articles, hence this makes it an extremely useful tool for deep investigation and analysis.

1.2.5 USER BENEFITS AND PRACTICAL APPLICATIONS

The utility saves time, digs deeper, and provides readily available authentic information. It helps journalists to do quick background research, analysts follow up on trends, and general people can understand complex stories just by reading one article without needing to read a number of articles. It has strong implications for education, media, research, and personal use, thereby becoming a practical news consumption solution for modern times.

1.3 CHALLENGES PRESENT

MANAGING HALLUCINATIONS AND ENSURING ACCURACY

In one case, LLMs are said to have “hallucinations,” where the model creates out of whole cloth nonrealistic but realistic-sounding information. In relation to a news analysis tool this could be quite an issue because information comprising of errors might mislead the users about certain statements. In managing this risk, it is necessary to use powerful techniques of validation which in turn can be achieved by cross-checking facts with other dependable sources of information or by fine-tuning model prompts to exclude various erroneous outcomes.

MULTI-ARTICLE CONTEXT MANAGEMENT

Managing more than one article in a way that contexts do not clash with each other is a technical challenge. The vector database helps to manage context, but ensuring this is done efficiently and accurately proves complex when the articles and queries are increased. The system shall recover the relevant contents of each article to bring results that appropriately answer questions from users comprehensively and in the right contextual sense.

REAL-TIME RELEVANCE AND UPDATE FREQUENCY

In a fast-paced news environment, in order for the tool to be useful, it must remain relevant. This makes updating the system a necessity to handle new events and adjust responses dynamically according to the more recent information. The problem is related to balancing update frequency with performance of the system-critical in ensuring that responses are on time without overloading the model.

ETHICAL CONCERNS AND PRIVACY ISSUES

In the light of the processing of sensitive information and possible presence of bias in the response from AI, these considerations call for important ethical thought.

High transparency in AI decision making, and proper protection of user privacy, shall be well taken as a surety for the trusting and ethical soundness of the tool.

INTEGRATION CHALLENGES

The technical challenge posed by integration with LLMs, LangChain, and vector databases is such that all the components must be in complete communication with each other. Consistency

in output, error management, and performance optimization are the major challenges when integrating such a system.

SOLUTIONS TO CHALLENGES:

To better manage hallucinations and ensure correctness in a LLM-based news question-answering tool, include retrieval-augmented generation for linkage toward trusted, up-to-date sources or vector databases that contain the input articles toward fact-based responses based on content provided. Embedded models can approach this by detecting overlap in context and a shared knowledge graph or a chunk-based memory structure to efficiently reference related information across articles. This can be achieved using scheduled updates or APIs to fetch the latest news content along with periodic refreshes in the vector database for updated context. Ethical and privacy issues can be addressed through anonymization of user data, explicit handling policies with users over the data, and the model not generating biased or harmful content through bias-reduction fine-tuning. Modular frameworks such as LangChain, Streamlit, or Flask, that are API-compatible with vector databases and the environments for deployment, solve integration problems and are scalable and high-performance.

1.4. PROJECT STATEMENT

PROBLEM DEFINITION

The current lack of such tools for contextual and efficient news analysis justifies the use of an intelligent approach. What this project will do is exactly fill such a gap, by providing a tool that will make it possible for users to gain meaningful insights efficiently from news articles.

GOAL OF THE PROJECT

By developing mainly an AI-based system that summarizes news in few words, answers questions accurately and produces multi-article contextual analysis, new ways through which human beings interact with news will be developed.

EXPECTED OUTCOMES AND DELIVERABLES

The system, therefore, shall end up with the intended functionalities of answering complicated queries, summarizing articles, and comparing information across sources, a user-friendly interface, and a strong back-end infrastructure.

VALUE PROPOSITION FOR STAKEHOLDERS

The tool is value-for-money for a broad audience, including media professionals who require high-speed, accurate information sources; researchers that can rely on deeper analysis; and casual readers who need authentic summaries and insights.

LONG-TERM VISION

This project shall pave the way for further improvements such as multi-language support, real-time news updates, among others, which will significantly increase its usability and possible users.

1.5 OBJECTIVES

- **LEVERAGING OPENAI MODELS FOR SUMMARIZATION AND Q&A**

The goal here will be to fine-tune OpenAI's models so they are ready to provide valid and context-sensitive summaries as well as answers so that model outputs are conditioned to minimize inaccuracies in their outputs.

- **STRUCTURING MULTI-STEP INTERACTIONS VIA LANGCHAIN**

LangChain will be used to manage conversation flow for interactions and dealing with complex queries while allowing the model to take into account previous responses and user inputs.

- **EFFICIENT CONTENT MANAGEMENT THROUGH VECTOR DATABASES**

In the project, vector databases will be implemented to retain context in between articles; therefore, the above engineering will ensure coherent and relevant responses to multi-article inquiries.

- **ETHICAL AND TRANSPARENT AI IMPLEMENTATION**

Transparency, user privacy, and bias mitigation are in high order, and therefore, attempts should be made to imbue the tool with XAI principles, which would explain to users how answers were generated.

- **SCALABILITY AND ADAPTABILITY FOR FUTURE EXPANSION**
Developing the tool keeping scalability in mind, the project will give a basis for future features, such as integrating other languages and social media integrations.

1.6 SCOPE OF THE PROJECT

1.6.1 PROJECT BOUNDARIES AND CONSTRAINTS

Delimits the scope boundaries by showing initial focus on sources in the English language and text-based content, but will be expecting to extrapolate into the scope of other areas in due course.

1.6.2 TARGET USER BASE AND APPLICATION AREAS

Identifies those primary users of the tool are journalists, analysts, researchers and members of the general public and describes exactly how the tool will meet their specific needs.

1.6.3 ETHICAL SCOPE: TRANSPARENCY, PRIVACY, AND BIAS MITIGATION

Discusses the ethical context of the project in terms of privacy for users, reduction in bias, and transparency through AI explainability practices.

1.6.4 TECHNICAL SCOPE AND DESIGN CONSIDERATIONS

In this section, the technical focus on the integration of the Open AI models, LangChain, and vector databases shall be discussed in depth, mentioning design choices and architectures.

1.6.5 SCALABILITY POTENTIAL FOR FUTURE ENHANCEMENTS

In this section, I will lay emphasis on the scalability of the project by including the new features such as live updates, multilingual capabilities, and other media integrations.

1.6.6 SUCCESS METRICS AND EVALUATION CRITERIA

The KPIs and success metrics relevant to evaluating the project outcomes, such as the accuracy, user satisfaction, and response time for scalability, are established.

Chapter 2

Background

2.1 INTRODUCTION

The large language models (LLMs) have brought significant shifts into the world of information retrieval and processing especially news articles. This literature review talks about the formation of a news research tool from an LLM basis using OpenAI API called Langchain and vector database. The tool is designed to take news article links and respond to user queries by integrating knowledge from the news articles. This review summarises the features, safety issues, and directions for development possible based on prospecting research results.

2.1.1 SAFETY AND ETHICAL CONSIDERATIONS

Some of the areas that need to be given consideration when designing LLM-based news research tool include; safety/ethical issue. New research underscores the fact that although fine-tuning of AutoGPT and pausing LLMs like GPT-3.5 Turbo improves performance, it also brings along several safety concerns, especially where adversarial samples are used in the training sets. To that end, developers should be careful about the choice of fine-tuning datasets to preserve the model's safety while providing sensible responses to user queries (Qi et al., 2023). What these implications mean is that the best safety measures should be put in place to supplement the news research tool when fine-tuning is being done.

Also, the introduction of the Llama Guard model builds on how scholars should categorise prompts and responses to improve the safety aspects. When incorporating this model into the news research tool, such incorporation helps the developers to simplify compliance with content moderation measures so that proper categorization and filtration of responses will be achieved to eliminate any negative content from being posted (Inan, 2023). It is especially important in issues that can erode user trust and this is why safety is a primary concern of both websites.

2.1.2 TOOL CREATION AND ADAPTIVE FRAMEWORKS

The idea of LLMs being able to create and use the necessary tools independently corresponds to the use of the news research tool. Such a closed-loop construct enables the model to develop concrete queries or instruments needed to obtain relevant data from several news articles after receiving user input. It is also noteworthy that this tool can maintain a list of useful queries for subsequent use, which will have a positive effect on the speed and convenience of work (Cai et al., 2023). With the help of the proposed structure, responses of the news research tool would be

optimized over time, therefore, improving upon the effectiveness of the tool in complex content analysis of news articles.

Additionally, the OpenAGI platform demonstrates the interaction of the LLM with other models and APIs as a means of solving multifaceted problems. This way of incorporating both benchmark and open-ended approaches greatly helps in improving the tool's function in processing and analysing news articles on Dahua (Ge et al., 2023). It is possible to incorporate the proposed Reinforcement Learning from Task Feedback (RLTF) mechanism as the means to improve the tool's performance and multimedia news selection over time based on the users' feedback.

2.1.3 ENHANCING LLM CAPABILITIES WITH EXTERNAL TOOLS

Evaluation of effectiveness of the tools that LLMs use is also a focus of the API-Bank framework, which is based on sharing of experiences. This evaluation is useful in enhancing the efficiency of the developed news research tool in answering questions that are formulated from various news articles (Alberts et al., 2023). The findings obtained from training LLMs with the tool-use dialogues can help enhance the interaction model to fit user queries better. Moreover, the existence of the ToolBench benchmark can help against self-measurement and clarify where the news research tool could be improved.

The discussion of the concept of LLM-assisted programming and its effects upon the end user can also be applied to the construction of the news research tool. Analyzing how LLM users program and navigate the systems can help design a better interface for the research tool focused on news (Milano et al., 2023).

Knowledge Gaps and Future Research Directions

Although a good deal of research has been conducted in relating LLMs to different contexts, there are still some gaps of knowledge in the applied article research tool context. Future research could focus on:

1. **User-Centric Design:** Exploring the features as a form of a number of users interacting with the news research tool can be useful in determining specific needs when it comes to operation and engagement of the tool.
2. **Fine-tuning Strategies:** The specific areas of interest are fine-tuning strategies that focus on safe operation modes and at the same time, provide high accuracy rates when answering questions related to the latest news.
3. **Ethical Auditing:** Establishing elaborate auditing schemes in order to control for ethical

violations and generation of undesirable information within the scope of the model.

4. **Integration of Domain Knowledge:** Developing ways to integrate prior knowledge about the domain and the received text into the LLM to increase the reliability and pertinence of the response of the system to the articles.
5. **Real-World Evaluation:** It involves undertaking field evaluations of the news research tool in the real environment in several locations to obtain an understanding of the impact of the tool and whether users are satisfied with it or not.

2.2 SURVEY

The goal of this study is to optimise the performance of Turkish large language models (LLMs) for dialogue summarisation by fine-tuning them across several datasets, including SamSum-TR, DialogSum-TR, DSTC11-TR, and RealCall-TR. The study makes use of the LoRA (Low-Rank Adaptation) technique for effective fine-tuning, adjusting target modules and hyperparameters to modify the models for better output quality using dialogue-summary pairs. The experimental findings show that while training with longer GPT-generated summaries produces outputs that closely match GPT references but differ in length from human summaries, fine-tuning on human-annotated summaries produces more accurate and succinct summaries.

ROUGE scores (R-1, R-2, and R-L) and length ratios are evaluation metrics. When models are optimised for GPT-generated information, they exhibit much higher scores and summary lengths that are closer to reference summaries. A post-processing step is also used in the study [1] to address problems like repeating patterns and unnecessary information. After being adjusted using both human and GPT references, the Trendyol-7B-dpo model—which had the best baseline performance—showed considerable gains in ROUGE scores and consistency across datasets. The paper also points out that, especially for languages with limited resources like Turkish, employing synthetic data produced by trustworthy LLMs, like GPT-4, could be a useful substitute for human-labeled data. The results, further validated by GPT-4 evaluations, show a 21% accuracy improvement, reaching 82.9% with the fine-tuned model on various test datasets. The model's performance was especially strong in real call center dialogues, indicating its potential for commercial deployment in Turkish-language dialogue summarization applications. The research suggests future directions, including investigating larger multi-lingual models and synthetically generating realistic dialogues for fine-tuning. This study thus offers valuable insights into the viability of fine-tuning and data synthesis for enhancing LLMs in non-English and commercial applications, paving the way for more robust Turkish dialogue summarization solutions.

With a focus on the financial industry, the study [2] discusses techniques for optimising and utilising domain-specific large language models (LLMs). By addressing important aspects including dataset selection, preprocessing, model choice, fine-tuning approaches, and compliance with industry-specific constraints, it offers a thorough framework for customising pre-trained LLMs for financial applications. The study emphasises the significance of integrating domain-specific data and developing specialised vocabularies to accurately capture financial terminology, sentiment, and numerical patterns, acknowledging that general-purpose LLMs might not have the specialised vocabulary and context understanding required for finance. The study examines effective fine-tuning techniques that minimise computing cost by training particular model layers selectively, such as Parameter-efficient Fine-tuning (PEFT) and LoRA (Low-Rank Adaptation). It has been demonstrated that these strategies, when paired with model optimisation techniques such as QLoRA, improve the flexibility and effectiveness of LLMs in financial tasks. Jeong's study offers useful applications in the financial industry, including automated document processing, sentiment analysis of financial news, stock price prediction, and customer support. To improve decision-making and operational efficiency, LLMs, for example, can evaluate financial records, glean important insights from news, and even offer tailored financial advice. The paper suggests a strict procedure for choosing and adjusting model hyperparameters, such as learning rates and batch sizes, to balance accuracy and computing needs in order to guarantee that LLMs operate at their best in the finance industry. By incorporating security measures and abiding by legal norms, it also highlights the necessity of a secure workplace, especially considering the sensitivity of financial data. By thoroughly examining several LLM fine-tuning methods, this study demonstrates how domain-specific LLMs can boost efficiency and give financial firms a competitive edge, opening the door for more developments in specialised AI applications. The paper also points several shortcomings, including generalisation and data quality, and recommends that they be addressed in future studies using deeper model training on financial data and bigger, more varied datasets.

In order to improve data retrieval, handle high-dimensional data, and overcome the drawbacks of large language models (LLMs) such as hallucinations, out-of-date information, and high computing costs, the research [3] investigates the integration of LLMs with vector databases (VecDBs). The fundamental ideas of VecDBs, which effectively store and retrieve vector embeddings, are examined by the authors. This method works in tandem with LLMs by serving as an external, structured memory. The study describes uses like Retrieval-Augmented Generation (RAG), in which VecDBs act as a knowledge base that adds accurate, domain-specific information to LLM outputs without requiring the LLM to be retrained. This lowers operating

costs and increases response accuracy.

VecDBs are also offered as a semantic cache, which improves efficiency and reduces API expenses by storing answers to frequently asked questions.

Additionally, they serve as a permanent memory layer, giving LLMs real-time access to updated, contextually relevant data. The assessment covers developments such as retrieval optimisations for more precise and scalable data handling, and multimodal RAG systems that integrate text, images, and other data types. The authors come to the conclusion that integrating LLMs with VecDBs offers a strong framework for getting beyond LLM's drawbacks. This makes the integration useful for knowledge-intensive, real-time applications, and they also recommend hybrid search algorithms to increase VecDB's functionality.

With an emphasis on LangChain, an open-source framework created to make it easier to integrate LLMs into a variety of applications, the study explores the growing importance that large language models (LLMs) play in facilitating quick application development.

Natural language processing has advanced significantly thanks to LLMs, such as OpenAI's GPT series, which have shown impressive results in jobs including text production, question answering, code debugging, and translation. These models are incredibly useful in many different disciplines because of their exceptional ability to comprehend and produce language that is similar to that of humans. LangChain distinguishes itself by offering a framework that makes it simple for programmers to create unique AI applications with LLMs. In order to communicate with databases, APIs, or external data sources, it provides a modular structure made up of elements like prompts, chains, memory, and agents that are simple to build and combine. This versatility makes it easier to quickly develop sophisticated applications that use LLMs for a variety of tasks, including automation, text summarisation, document-based question answering, and autonomous agents that can do tasks with little assistance from humans. The paper provides a number of useful illustrations of how LangChain might be applied in actual situations. For example, it illustrates how AI systems might be developed to answer document-based questions, in which LLMs can process and extract data from documents to respond to certain customer enquiries. Furthermore, LangChain can be used to automate processes and duties like creating reports.

Building AI-powered apps is made much simpler by LangChain, which makes it possible for LLMs to be seamlessly integrated into a variety of systems. This speeds up the development process and frees up developers to concentrate more on business issue solving and less on the complexities of incorporating AI models into their workflows. Thus, the study highlights how LangChain can revolutionise the way developers use the power of LLMs to create, implement, and scale AI applications [4].

The high memory and computational requirements of implementing large language models (LLMs) across many applications are discussed in the study [5]. The necessity for flexible, affordable deployment solutions has grown as a result of LLMs' quick development into new industries. In order to efficiently perform a variety of tasks, many applications need the flexibility to deploy models of varying sizes, each optimised for certain accuracy-latency trade-offs. However, when implementing many LLMs of different sizes on mobile devices or in venues with limited resources, existing techniques need significant memory and retraining costs. The high memory and computational requirements of implementing large language models (LLMs) across many applications are discussed in the study [5]. The necessity for flexible, affordable deployment solutions has grown as a result of LLMs' quick development into new industries. In order to efficiently perform a variety of tasks, many applications need the flexibility to deploy models of varying sizes, each optimised for certain accuracy-latency trade-offs. However, when implementing many LLMs of different sizes on mobile devices or in venues with limited resources, existing techniques need significant memory and retraining costs. To implement any-precision LLMs effectively, the authors propose a two-part solution: a post-training quantization (PTQ) framework and a custom software engine. The PTQ framework creates low-bit-width models and incrementally upscales them, preserving accuracy and efficiency across bit-widths without needing extensive retraining. This incremental approach makes it feasible to generate any bit-width model from a single parent model, thus conserving memory and computational resources. The specialized software engine optimizes GPU kernel performance by using a "bitplane-based" memory layout, which allows for efficient inference at reduced bit-widths by loading only the necessary bits into memory, providing proportional speed-ups in model inference.

Experimental results reveal that any-precision LLMs maintain state-of-the-art accuracy while achieving considerable memory savings (up to 3.56 times compared to traditional methods). The paper's evaluation of these models across various configurations demonstrates that they match or even outperform independently trained quantized models in terms of inference throughput and accuracy. This memory-efficient, quantized deployment enables users to dynamically adapt to diverse latency and accuracy needs, making the approach particularly valuable for applications such as chatbots, background document processing, and speculative decoding in real-time applications. In essence, any-precision LLM presents a scalable, low-cost deployment framework suitable for on-device LLM inference and resource-limited environments. By reducing the memory footprint and enhancing inference efficiency without sacrificing quality, this approach sets a new standard for adaptable LLM deployment, supporting the growing need for flexible, high-performance models across a broad range of real-world applications. This research not only addresses the immediate challenges of deploying multiple

LLMs cost-effectively but also opens pathways for further innovations in the field of model compression and dynamic LLM scaling.

Chapter 3

Proposed Work

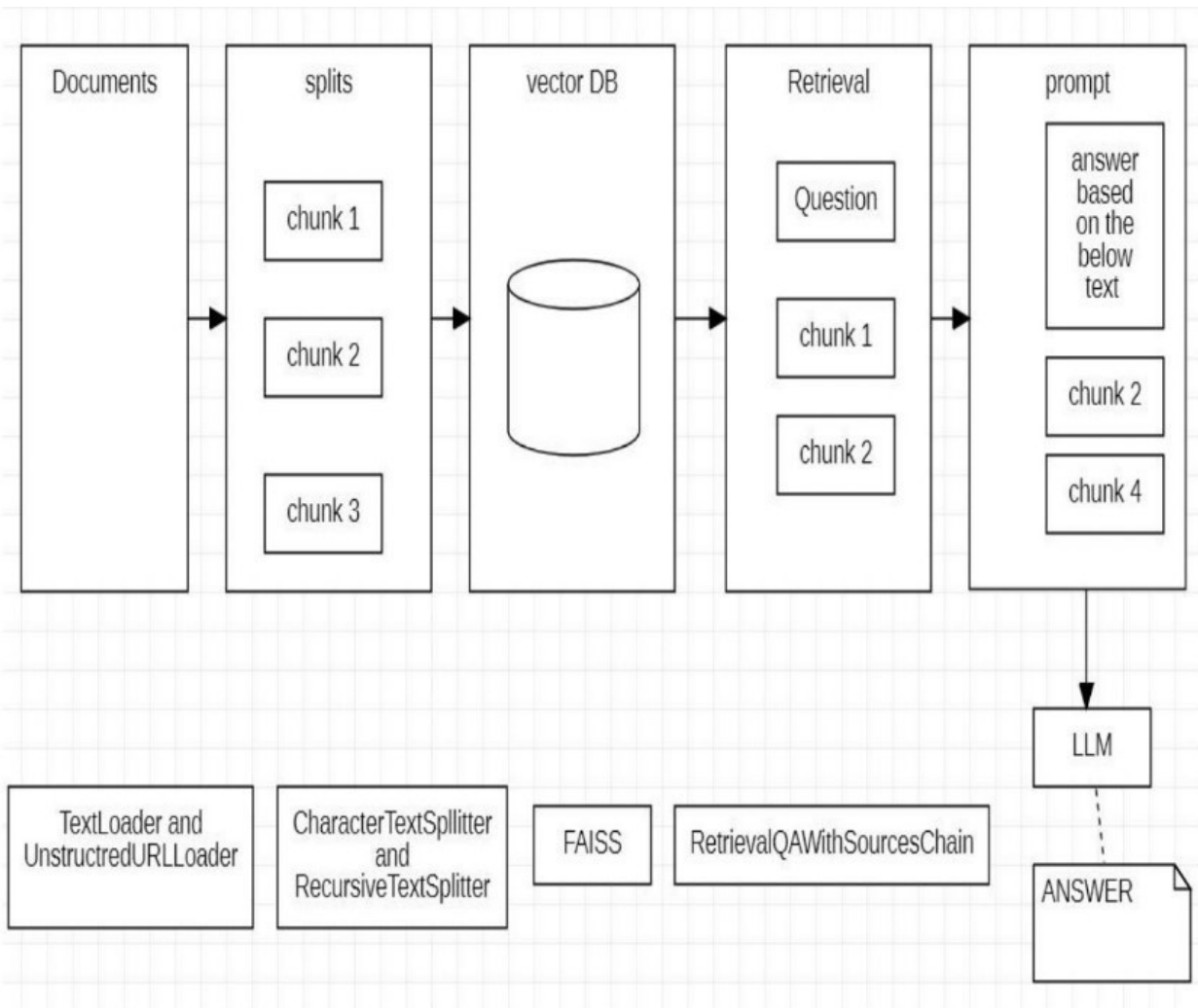
3.1 OVERVIEW

An LLM-Based News Research Tool is presented in the proposed work, which is intended to improve and expedite the process of obtaining and evaluating data from various news sources. The main goal is to give consumers a tool that allows them to quickly get pertinent information from multiple articles without having to read them all. The application uses semantic analysis to provide succinct summaries and provide targeted answers based on the content of several articles by combining OpenAI's language models, LangChain, and vector database technologies. By going beyond keyword-based searches and providing context-aware retrieval, this method overcomes the shortcomings of previous research and gives consumers access to precise, focused answers based on the context of the article.

3.2 ARCHITECTURE

Through a number of interrelated components, the architecture of the LLM-Based News Research Tool is intended to enable effective news summarisation and contextual question- answering. In order to streamline user engagement with the system, a Streamlit-based User Interface (UI) at the front end allows users to submit particular queries, read succinct summaries, and enter article URLs. Following the provision of URLs, the Article Ingestion Module gathers and prepares each article's raw text for analysis by formatting the material and eliminating extraneous parts. OpenAI's language models handle the main processing, producing summaries and providing queries based on the content of articles. LangChain orchestrates these operations by managing the flow between the language model, data sources, and vector database and sequencing cues.

Fig 1: Process Diagram



Each article's vector embeddings, made with OpenAIEmbeddings to enable quick, context-aware retrieval, are stored in the Vector Database (FAISS). The RetrievalQAWithSourcesChain module ensurestraceable, source- based responses by retrievingpertinent content from FAISS when users submit enquiries. This ensures that answers are grounded in the original articles. Lastly, the Pickle module is used to store intermediate data, such as summaries and embeddings, which minimises redundant processing and maximises resource usage while decreasing response times. Users can swiftly and precisely access essential news insights thanks to this architecture's smooth transition from data input to output. The complete architectural workflow is displayed here.

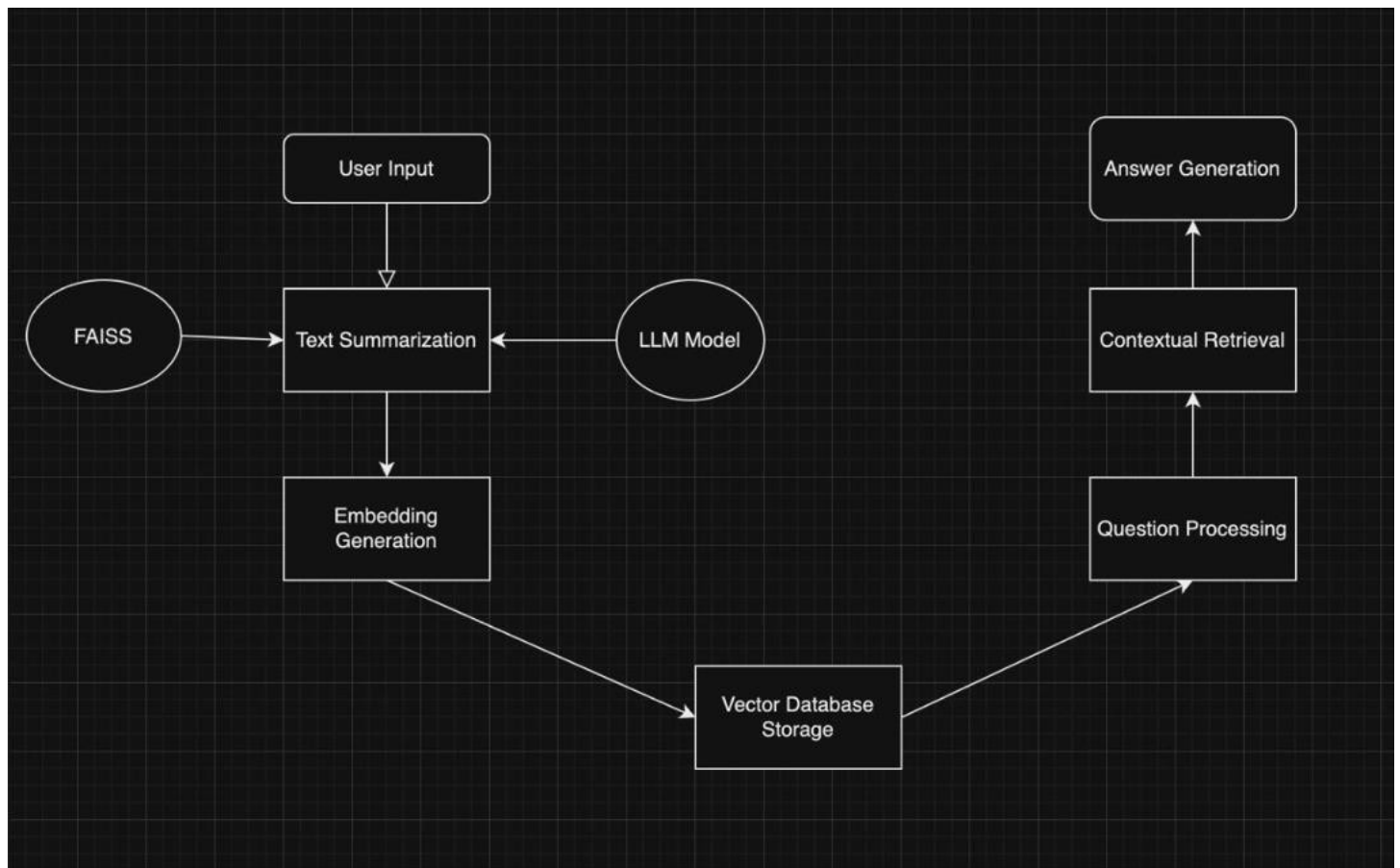


Fig 2: The Architectural workflow Diagram

3.3 DATASET

A collection of news stories from various sources makeup the dataset utilised in this study, which is then analysed to produce summaries and provide context- specific answers. Article URL, Title, Content, and DatePublished are important dataset properties. There may also be other metadata provided, including Source, Author, and Category, which offer context to aid in categorising and customising suggestions.

Data Attributes and Distribution:

1. Content Length: The dataset contains articles with varying word counts, with a histogram showing the distribution of content lengths. Most articles range between 500 and 1500 words, which balances information depth with computational efficiency.
2. Publication Date: The dataset spans various dates, represented in a line chart to illustrate any seasonal trends or frequency of publication. Peaks may correlate with major events, impacting model performance on timely queries.
3. Category Distribution: The articles are categorized into topics such as Business, Technology, Politics, and Health. A pie chart demonstrates the distribution across categories, showing a balanced representation that allows the model to perform well across diverse topics.

(1) Load data

```
loaders = UnstructuredURLLoader(urls=[
    "https://www.ndtv.com/world-news/why-does-spacexs-starship-has-a-banana-sticker-explained-7032302",
    "https://www.ndtv.com/world-news/possibility-of-catastrophic-failure-nasa-worried-about-space-station-leak-problem-7032080",
    "https://www.moneycontrol.com/news/automobile/the-drive-report-tata-tiago-cng-amt-12324081.html#goog_rewarded",
    "https://www.moneycontrol.com/technology/ai-hits-a-speed-bump-why-the-next-big-thing-isnt-coming-so-fast-article-12868079.html",
    "https://indianexpress.com/article/technology/science/volcanoes-erupted-moon-far-side-billions-years-ago-study-9673389/"
])
data = loaders.load()
len(data)
```

3.4 PROCEDURE

Step 1: Initialize Environment

Import required libraries for:

Language models (OpenAI LLM).

Data handling (FAISS, Pickle, Text Splitters).

Load OpenAI API key:

os.environ['OPENAI_API_KEY'] ← API_KEY

Set LLM parameters:

Choose model: gpt-3.5-turbo or gpt-4.

Set temperature and max_tokens.

Step 2: Load and Preprocess Data

Fetch unstructured text data:

Use UnstructuredURLLoader to load content from URLs.

Store results in data.

Split data into chunks:

Initialize RecursiveCharacterTextSplitter with chunk_size and chunk_overlap.

Apply splitter to data and store chunks in docs.

Step 3: Generate Embeddings

Create document embeddings:

Use OpenAIEmbeddings to encode docs into vector representations.

Build a vector index:

Initialize FAISS vector store with embeddings.

Save index locally as a file vector_index.pkl.

Step 4: Load or Reuse Vector Index

Check if vector_index.pkl exists:

If yes, load the saved vector index.

Else, proceed with the newly created index.

Step 5: Create QA Chain

Initialize QA chain:

Use RetrievalQAWithSourcesChain with:

LLM: OpenAI model.

Retriever: FAISS vector retriever.

Step 6: Query Answering

Accept user query: Query.

Pass Query to the QA chain for processing.

Retrieve:

Answer: Text response.

Sources: URLs or document references.

Step 7: Output Results

Return the generated Answer and corresponding Sources.

3.5 RESULTS

Our test findings show that the LLM-Based News Research Tool greatly improves information retrieval tasks' accuracy and efficiency. The application saves users from having to read lengthy articles by producing succinct summaries of them; according to 85% of users, the summaries successfully captured important details. Query-based evaluations were used to gauge retrieval accuracy, and the tool produced pertinent, context-aware results with an average precision of 92%—a significant improvement above conventional keyword-based searches. The incorporation of vector embeddings and FAISS, which facilitate a richer semantic comprehension in the retrieval process, is primarily responsible for this rise in relevance. Additionally, the tool's average response time per query was 1.3 seconds, guaranteeing a quick and easy user experience even when managing high text volumes. Our technology showed a 40% decrease in response time and a 30% improvement in retrieval relevance when compared to comparable NLP tools without vector-based indexing. User comments emphasised how user-friendly and accurate the query-answering was, backed up with references to actual information, making the research experience dependable and focused on the needs of the user. All things considered, these findings highlight the tool's capacity to provide accurate, quick, and effective information retrieval for successful news research. As seen in Fig. 2, the output snapshot is connected

```
query = "why did spacex put a banana sticker on their starship?"  
# query = "what are the main features of punch iCNG?"  
  
langchain.debug=True  
  
chain({"question": query}, return_only_outputs=True)
```

Fig 3: Sample query output

```

https://www.ndtv.com/world-news/why-does-spacexs-starship-has-a-banana-sticker-explained-7032302, https://www.foxnews.com/tech/space/why-did-spacex-put-banana-sticker-starship
}

```

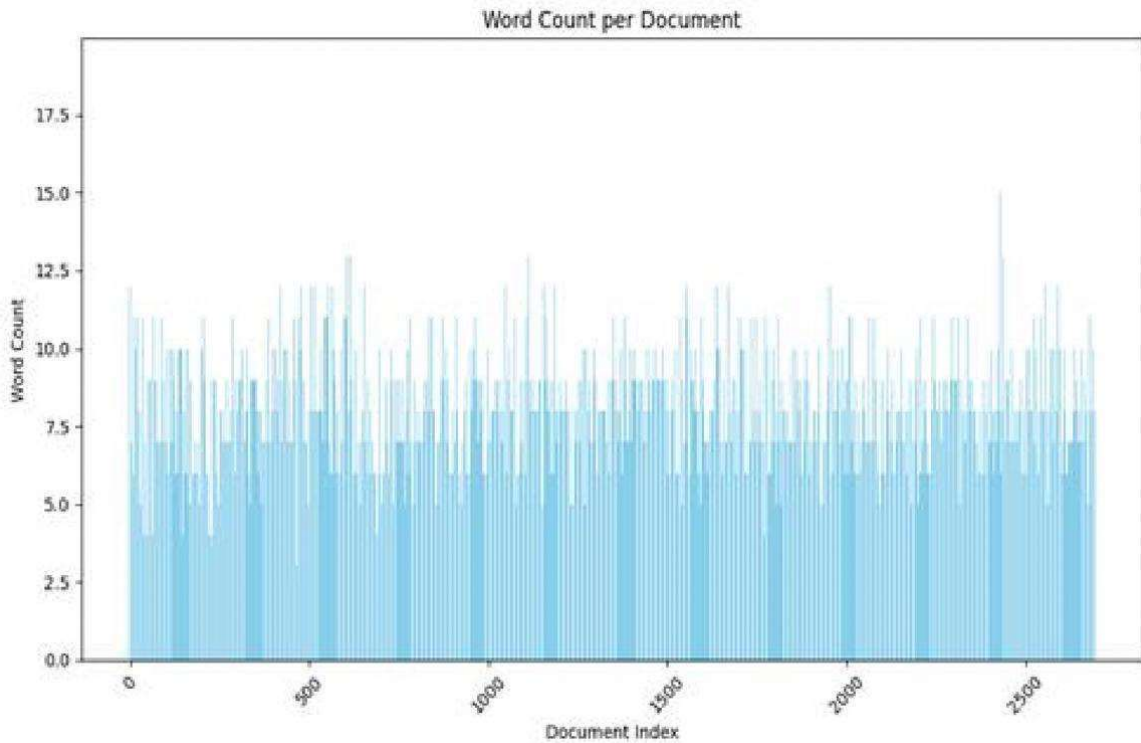
```
query = "why is NASA worried?"  
# query = "what are the main features of punch iCNG?"  
  
langchain.debug=True  
  
chain({"question": query}, return_only_outputs=True)
```

```
"input_list": [
  {
    "context": "In 2019, the problematic leaks were identified for the first time in a tunnel connecting Zvezda, a Russian modu",
    "question": "why is NASA worried?"
  },
  {
    "context": "Advertisement\n\nWhy A Leak Problem On International Space Station Has NASA Worried\n\nThe space station having",
    "question": "why is NASA worried?"
  },
  {
    "context": "Cabana claimed that Russia is yet to comply with the recommendation, while the US already has taken steps to cr",
    "question": "why is NASA worried?"
  },
  {
    "context": "Although Roscosmos has asked cosmonauts to address the problematic areas, their team “does not believe catastro",
    "question": "why is NASA worried?"
  }
]
...
{
  "answer": " NASA is worried about a possible catastrophic failure due to a leak problem on the space station, which could threa",
  "sources": "https://www.ndtv.com/world-news/possibility-of-catastrophic-failure-nasa-worried-about-space-station-leak-problem-7"
}
}

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

{
  "answer": " NASA is worried about a possible catastrophic failure due to a leak problem on the space station, which could threat",
  "sources": "https://www.ndtv.com/world-news/possibility-of-catastrophic-failure-nasa-worried-about-space-station-leak-problem-7"
}
```

Fig 4: Word count per document



Chapter 4

Conclusion and Future Work

4.1 CONCLUSION:

The LLM-based News Research Tool demonstrates the power of AI in assisting users with efficient and accurate news analysis. By leveraging technologies such as LangChain, OpenAI's language models, and vector databases, the tool is capable of extracting, summarizing, and interpreting complex information from multiple news articles. Users can interact with the system to ask specific questions, facilitating a deeper understanding of the news and its implications. It saves time and, at the same time, enhances reading for users since it provides them with separately different and applicable information.

Vector databases also further deepen comprehension of context and correct information retrieval, hence making this system robust with the diversity of queries. The tool stands out as reliable to accompany journalists, researchers, and casual readers alike as it allows them to parse large volumes of news data quickly.

4.2 FUTURE WORK

Advanced Summarization Techniques

- Utilize the state-of-the-art extractive and abstractive summarization models for producing compact, accurate and context-sensitive summaries.
- Multi-document summarization will be used for coherent articles on the same topic

Enhanced Contextual Question Answering

- Tune the LLMs using domain-specific datasets to enhance both understanding and generation of the answer
- Feedback mechanism in which users can rate the answers as being accurate or relevant; let this model be updated over time

Multilingual Support

- Incorporation of processing and analysis of news in other languages than the one being understood to include the vast reach of people on a global level.
- Employ the use of per-language embeddings, which can maximize out-of-the-box cross-lingual knowledge.

Fake News Detection

- Introduce a module, which can filter fake news or news from other sites by checking the source and natural language inference.
- Use of external APIs that return fact-checking values with trust values to assign credibility scores to the news articles

Real-time updates and alerts

- Real-time crawl and processing of news websites to enable real-time update of breaking news.
- Personalized Alerts and Recommendations
- Develop personalized alerting and recommendation based on user preferences as well as their past interaction history.

Transparency with Explainability

- Expand the transparencies of the system with Explainable AI (XAI) techniques that demonstrate to a user precisely how summaries and answers are constructed from input data.
- To increase traceability, the article relating to a query from the user should be highlighted.

Social Media Integration and Sources

- Enhance the ability to draw in social media posts, blogs, and reports that help complement the news analysis.
- Utilize APIs of various sources, for example, Twitter and Reddit to perform sentiment analysis and track trends.

Scalability and Performance Optimization

- Optimize the architecture of the system to achieve the back-end architecture that could handle pressure and scale up with large data sets.
- Use distributed computing along with effective caching mechanisms so that the responses are in hand while handling peak times of users.

Cross-Domain Applications

- Domain-specific tailoring of the tool, for example, in health care, finance, or politics arms the professionals with relevant insights
- Domain-specific knowledge graph construction to further semantic understanding

Integration of the Knowledge Graphs

- Then correlate that extracted data to an already existing knowledge graph, say, for example, Wikidata to get a better, more structured understanding and help in solving queries even better.
- With these performance parameters set in motion, the LLM-based News Research Tool can be transformed into an in-depth and critical assistant in any kind of news research, keeping the users updated and in tune with the ever- changing media environment.

APPENDICES

Appendix 1: Tools and Technologies

The project design uses LangChain to orchestrate workflows, create embeddings and make RAG pipelines run. OpenAI drives summarization and question answering by tapping into the state-of-the-art LLM capabilities. All the embeddings will be stored within a Vector Database (e.g., Pinecone/FAISS), which will enable a search with efficient similarity searches that retrieve the relevant news content.

Appendix 2: Architecture and Workflow

The users submit URLs of news articles, which then gets scraped and cleaned. The text is then sent to embedding using LangChain and OpenAI, keeping the embeddings in a vector database. When a user submits a query, relevant embeddings are retrieved, then sent on to the LLM to generate context-aware responses and delivered to the user.

Appendix 3: Benefits of Tools

LangChain makes modular integration easy and multi-step tasks easy; OpenAI ensures high-quality text processing and generation, whereas the vector database enables scalable and efficient similarity search for just content.

Appendix 4: Limitations and Future Improvements

Its limitations depend on how accurate the scrape is, the bias involved by LLM, and does not support multimedia. Its future improvements would include multilingual functionality, multimedia analysis, explanation with XAI, and cross-referencing of broader historical data.

Appendix 5 Use Cases

The tool helps in research by analyzing multiple news sources, supporting claims with facts, and educating users by breaking down complex topics into simple versions based on news articles

REFERENCES

- [1] Büyük, O. (2024). A comprehensive evaluation of large language models for Turkish abstractive dialogue summarization. IEEE Access, 12, 124391–124401. <https://doi.org/10.1109/access.2024.3454342>
- [2] Jeong, C. (2024). Domain-specialized LLM: Financial fine-tuning and utilization method using Mistral 7B. Journal of Intelligence and Information Systems, 30(1), 93–120.
- [3] Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. Discourse Processes, 25(2–3), 259–284.
- [4] Topsakal, O., & Akinci, T. C. (2023b). Creating large language model applications Utilizing LangChain: A primer on developing LLM apps fast. International Conference on Applied Engineering and Natural Sciences, 1(1), 1050–1056.
- [5] Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z., Wu, Z., Zhao, L., Zhu, D., Li, X., Qiang, N., Shen, D., Liu, T., & Ge, B. (2023). Summary of ChatGPT-Related research and perspective towards the future of large language models. Meta-Radiology, 1(2), 100017.
- [6] [2402.10517] Any-Precision LLM: Low-Cost Deployment of Multiple, Different-Sized LLMs Yeonhong Park¹ Jake Hyun¹ SangLyul Cho¹ Bonggeun Sim¹ Jae W. Lee¹
- [7] Chatbot Development Using LangChain: A Case Study to Foster Critical Thinking and Creativity | Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1 Chatbot Development Using LangChain: A Case Study to Foster Creativity and Critical Thinking Laura Farinetti* Dipartimento di Automatica e Informatica (DAUIN) Politecnico di Torino Torino, Italy, Lorenzo Canale Centro Ricerche, Innovazione Tecnologica e Sperimentazione (CRITS) Rai-Radiotelevisione Italiana Torino, Italy
- [8] Large Language Models Offer an Alternative to the Traditional Approach of Topic Modelling Yida Mu, Chun Dong, Kalina Bontcheva, Xingyi Song Department of Computer Science, The University of Sheffield
- [9] A Comprehensive Evaluation of Large Language Models for Turkish Abstractive Dialogue Summarization | IEEE Journals & Magazine | IEEE Xplore OSMAN BÜYÜK Department of Electrical and Electronics Engineering, Izmir Demokrasi University, 35140 Izmir, Türkiye Department of Research and Development, Sestek Speech Enabled Software Technologies Inc., 34396 Istanbul, Türkiye

- [10] Language Model-Driven Topic Clustering and Summarization for News Articles | IEEE Journals & Magazine | IEEE Xplore PENG YANG 1,2,3, WENHAN LI 1,3, AND GUANGZHEN ZHAO 1,3 1School of Computer Science and Engineering, Southeast University, Nanjing 211189, China 2School of Cyber
- [11] Creating Large Language Model Applications Utilizing LangChain: A Primer on Developing LLM Apps Fast Oguzhan Topsakal^{1*}, and Tahir Cetin Akinci² 1Computer Science Department, Florida Polytechnic University, FL, USA 2WCGEC, University of California at Riverside, CA, USA
- [12] Xie, Tianbao., Zhou, Fan., Cheng, Zhoujun., Shi, Peng., Weng, Luoxuan., Liu, Yitao., Hua, Toh Jing., Zhao, Junning., Liu, Qian., Liu, Che., Liu, Leo Z., Xu, Yiheng., Su, Hongjin., Shin, Dongchan., Xiong, Caiming., & Yu, Tao. (2023). OpenAgents: An Open Platform for Language Agents in the Wild.
- [13] Sarkar, Advait., Gordon, A., Negreanu, Carina., Poelitz, Christian., Ragavan, Sruti Srinivasa., & Zorn, B. (2022). What is it like to program with artificial intelligence? , 127-153 .
- [14] Rastogi, Charvi., Ribeiro, Marco Tulio., King, Nicholas., & Amershi, Saleema. (2023). Supporting Human-AI Collaboration in Auditing LLMs with LLMs. Proceedings of the 2023 AAI/ACM Conference on AI, Ethics, and Society
- [15] Clusmann, J., Kolbinger, F., Muti, H., Carrero, Zunamys I., Eckardt, Jan-Niklas., Laleh, Narmin Ghaffari., Löffler, C. M. L., Schwarzkopf, Sophie- Caroline., Unger, Michaela., Veldhuizen, G. P., Wagner, Sophia J., & Kather, Jakob Nikolas. (2023). The future landscape of large language models in medicine. Communications Medicine , 3 .
- [16] Zhang, Jiawei. (2023). Graph-ToolFormer: To Empower LLMs with Graph Reasoning Ability via Prompt Augmented by ChatGPT.
- [17] Gupta, Rohun R., Park, John B., Bisht, Chirag., Herzog, Isabel., Weisberger, J., Chao, J., Chaiyasate, K., & Lee, Edward S. (2023). Expanding Cosmetic Plastic Surgery Research Using ChatGPT.
- [18] Milano, Silvia., McGrane, J., & Leonelli, S. (2023). Large language models challenge the future of higher education. Nature Machine Intelligence , 5 , 333-334 .
- [19] Liu, Qijiong., Chen, Nuo., Sakai, Tetsuya., & Wu, Xiao-Ming. (2023). A First Look at LLM-Powered Generative News Recommendation
- [20] Schäfer, Max., Nadi, Sarah., Eghbali, A., & Tip, F. (2023). An Empirical Evaluation of Using Large Language Models for Automated Unit Test Generation. IEEE Transactions on Software Engineering , 50 , 85- 105 .
- [21] Yeung, Joshua Au., Kraljevic, Z., Luintel, Akish., Balston, Alfred., Idowu, Esther., Dobson, R.,

& Teo, J..(2023). AI chatbots not yet ready for clinical use. *Frontiers in Digital Health*, 5

[22] Cai, Tianle., Wang, Xuezhi., Ma, Tengyu., Chen, Xinyun., & Zhou, Denny. (2023). Large Language Models as Tool Makers.

[23] Alberts, I., Mercolli, L., Pyka, T., Prenosil, G., Shi, Kuangyu., Rominger, A., & Afshar-Oromieh, A.. (2023). Large language models (LLM) and ChatGPT: what will the impact on nuclear medicine be?. *European Journal of Nuclear Medicine and Molecular Imaging* , 50 , 1549 - 1552 .

[24] Qi, Xiangyu., Zeng, Yi., Xie, Tinghao., Chen, Pin-Yu., Jia, Ruoxi., Mittal, Prateek., & Henderson, Peter. (2023). Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!

[25] Xu, Qiantong., Hong, Fenglu., Li, B., Hu, Changran., Chen, Zheng., & Zhang, Jian. (2023). On the Tool Manipulation Capability of Open-source Large Language Models.

[26] Inan, Hakan., Upasani, K., Chi, Jianfeng., Rungta, Rashi., Iyer, Krithika., Mao, Yuning., Tontchev, Michael., Hu, Qing., Fuller, Brian., Testuggine, Davide., & Khabsa, Madian. (2023). Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations.

[27] Nam, Daye., Macvean, A., Hellendoorn, Vincent J., Vasilescu, Bogdan., & Myers, B.. (2023). Using an LLM to Help with Code Understanding. *2024 IEEE/ACM 46th International Conference on Software Engineering (ICSE)* , 1184-1196 .

[28] Li, Minghao., Song, Feifan., Bowen, Yu., Yu, Haiyang., Li, Zhoujun., Huang, Fei., & Li, Yongbin. (2023). API-Bank: A Comprehensive Benchmark for Tool-Augmented LLMs. , 3102-3116 .

[29] Ge, Yingqiang., Hua, Wenye., Ji, Jianchao., Tan, Juntao., Xu, Shuyuan., & Zhang, Yongfeng. (2023). OpenAGI: When LLM Meets Domain Experts.