2023

City University of London

School of Mathematics, Computer Science & Engineering

MSc in Data Science

Project Report, 2023

“Comparative Analysis of LLama 7B, Mistral 7B, and GPT 3.5 in Sentiment Analysis and Question-Answering"

Written by: Kunj Patel

Supervised by: Prof. Pranava Madhyastha

Date of Submission – 10th December, 2023

Contents

Declaration

“By submitting this work, I declare that this work is entirely my own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the assessment instructions and any other relevant programme and module documentation. In submitting this work I acknowledge that I have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. I also acknowledge that this work will be subject to a variety of checks for academic misconduct.”
Signed: Kunj Patel

Abstract

Large language models (LLMs) represent a significant advancement in artificial intelligence, demonstrating human-like proficiency in natural language processing (NLP) tasks. This study conducts a rigorous assessment of three cutting-edge LLMs - LLama 7B, Mistral 7B and GPT 3.5 in the crucial language understanding domains of sentiment analysis and question answering. additionally, the novel use of GPT 3.5's embeddings to enhance traditional machine learning is explored.

The models are systematically evaluated on two distinct datasets using relevant performance metrics like accuracy, precision, recall, and F1 scores. Confusion matrices enable further granular analysis, uncovering model deficiencies and strengths. Key findings reveal variability in effectiveness across models and tasks - GPT 3.5 and LLama 7B excel in question answering but struggle with nuanced sentiment classification unless fine-tuned. Mistral 7B question answering stands out but shows training data biases. Embeddings significantly boost GPT 3.5 across both tasks, demonstrating their utility.

By providing comprehensive and balanced insights into current capabilities, this pioneering study informs appropriate LLM selection for target applications. Furthermore, fine-tuning and embedding strategies to attain new performance milestones are highlighted through evidence-driven experimentation with state-of-the-art models, advancing the broader mission to engineer robust language technologies aligned with ethical AI standards.

The project achieves its multifaceted research goals while elucidating upcoming challenges and opportunities to shape further progress in this fascinating domain.

Chapter 1: Introduction

1.1 Large Language Models

Large Language Models (LLMs), as a product of advancements in artificial intelligence, have brought transformative changes to various fields, particularly natural language processing (NLP). These models, evolving from rule-based systems to sophisticated deep learning architectures like transformers, have enabled a deeper understanding and generation of human-like language (Mengnan Du et al., 2023). LLMs, exemplified by GPT-3 and GPT-4, are trained on extensive text data, enabling them to excel in diverse NLP tasks and adapt to specific applications.

LLMs have demonstrated their capability in learning complex patterns and semantic relationships within language, making them valuable tools across various sectors including healthcare, education, and finance. In healthcare, for example, LLMs like ChatGPT have shown potential in medical education, decision-making, and patient care, indicating their proficiency and adaptability without needing specialized training.

However, LLMs face challenges like reliance on dataset biases, known as shortcut learning, which affects their performance in novel scenarios. Addressing these challenges involves integrating data-driven training with domain-specific knowledge to enhance their robustness and generalizability (Du et al., 2023). The continual development and refinement of LLMs, considering ethical and fairness aspects, are crucial for their effective deployment in real-world applications.

Overall, LLMs signify a significant leap in AI, offering groundbreaking possibilities in understanding and interacting with human language. Their ongoing development and the challenges they present form an active area of research, promising further advancements in AI and its applications.

1.2 Problem Overview

The project focuses on a detailed exploration of the capabilities and limitations of Large Language Models (LLMs) in sentiment analysis and question-answering tasks. The core objective is to apply and assess three distinct LLMs: LLama 7B, Mistral 7B, and GPT 3.5, each possessing unique features and training backgrounds.

Part 1: Detailed Analysis of Sentiment Analysis with LLMs

Sentiment analysis in the context of LLMs is a multifaceted task that extends beyond mere identification of positive or negative sentiments. It requires the model to discern the subtleties and nuances of human emotions conveyed through text. This task becomes particularly challenging with sarcastic, ironic, or contextually complex statements. The project will critically evaluate how the selected LLMs - LLama 7B, Mistral 7B, and GPT 3.5 - interpret these intricacies in different textual formats, ranging from short tweets to extensive reviews. It will assess their ability to contextualize sentiments within varying narrative structures and lexical choices. The goal is to determine how well each model captures the breadth of human emotional expression in text, which is vital for applications ranging from customer feedback analysis to social media monitoring.

Part 2: Evaluating Question-Answering Abilities of LLMs

In this part of the project, the focus is on the question-answering capabilities of LLama 7B, Mistral 7B, and GPT 3.5. This evaluation will explore how these models understand and interpret various types of questions, ranging from factual and knowledge-based to abstract and inferential queries. A key aspect of this analysis is to assess the models' ability to provide accurate, relevant, and contextually appropriate answers. The project will also examine how well these models handle ambiguity, manage follow-up questions, and integrate external information sources. This part aims to demonstrate the extent to which LLMs can replicate human-like understanding and responsiveness in conversational AI applications, an essential factor for their deployment in customer service, education, and interactive systems.

Part 3: Analyzing Model Embeddings and Fine-tuning

This part involves a deep dive into the embeddings created by GPT 3.5. Embeddings are crucial as they transform text into a form that models can process, capturing semantic and syntactic nuances of language. The project will analyze how GPT 3.5 model's embeddings capture the essence of the textual input and their effectiveness in representing different linguistic features. It will also explore the fine-tuning process, where LLama 7B and Mistral 7B models are adjusted to better suit specific tasks. This includes examining how fine-tuning impacts model performance, particularly in adapting to the nuances of sentiment analysis and question-answering, and whether it leads to significant improvements over the models' original configurations.

Part 4: Comparative Analysis Using Confusion Matrices

The final part of the project involves a comparative analysis of the embeddings generated by GPT 3.5 on the sentiment analysis dataset and question-answer dataset using confusion matrices. This will provide detailed insights into how the integration of embeddings impacts the performance of GPT 3.5 on these two distinct NLP tasks. The confusion matrices will help in quantifying the true positives, false positives, true negatives, and false negatives when embeddings are leveraged compared to when raw GPT 3.5 is used without embeddings. This analysis aims to uncover the improvements in crucial evaluation metrics like accuracy, precision, recall and F1 score that can be attributed to the embedding layers in GPT 3.5. The goal is to conclusively determine if embeddings contribute substantially to enhancing GPT 3.5's capabilities in sentiment analysis and question answering tasks and under what conditions they are most effective. This will provide a comprehensive perspective into the strengths and weaknesses of integrating embeddings within GPT 3.5 for practical applications.

Overall, the project aims to provide a thorough understanding of the strengths and limitations of each LLM in these specific NLP tasks, offering valuable insights into their practical applications and potential for future developments in the field.

1.3 Objectives and Scope Objectives of the Project

The objectives and scope of your project are centered around the evaluation and comparison of three Large Language Models (LLMs) - LLama 7B, Mistral 7B, and GPT 3.5 - in two specific tasks: sentiment analysis and question-answering. The primary objectives include:

1. **Assessment of LLM Performance:** How do LLama 7B, Mistral 7B, and GPT 3.5 differ in their effectiveness for sentiment analysis and question-answering tasks, particularly in terms of accuracy, context comprehension, and response relevance?
2. **Impact of Fine-Tuning:** What is the influence of fine-tuning on the performance of LLama 7B and Mistral 7B specifically regarding enhancements in their capabilities for targeted NLP tasks?
3. **Evaluation of Model Embeddings:** How do the embeddings generated by GPT 3.5 vary in capturing and representing linguistic features, and what does this imply about their processing efficiency?
4. **Comparative Performance Analysis:** How does the performance of GPT 3.5 differ on various NLP tasks when utilizing embeddings compared to when embeddings are not employed?

The scope of the project includes:

1. Comprehensive testing of each model with diverse datasets to ensure a robust evaluation.
2. Detailed analysis of fine-tuning processes and their effects on model performance.
3. Exploration of the practical applications of these models in real-world scenarios based on the findings.

1.4 Methods Outline

1. Data Acquisition and Preparation
    a. Obtain relevant datasets for sentiment analysis and question answering.
    b. Conduct data preprocessing including cleaning, normalization, and tokenization if required.
2. Baseline Performance Evaluation
    a. Evaluate the initial performance of LLama 7B, Mistral 7B, and GPT-3.5 using the datasets.
    b. Employ accuracy, F1-score for sentiment analysis, and precision, recall for question answering.
3. Model Fine-Tuning
    a. Fine-tune LLama 7B and Mistral 7B for sentiment analysis and for question answering.
    b. Optimize hyperparameters to enhance model performance.
4. Embeddings Analysis
    a. Extract and analyze embeddings from the GPT 3.5 model.
    b. Study how linguistic features are represented in these embeddings.
5. Comparative Performance Analysis
    a. Compare the performance of the fine-tuned models against their baseline versions.
    b. Use relevant evaluation metrics tailored to sentiment analysis and question answering tasks.
6. Result Analysis and Interpretation
    a. Analyze the outcomes to understand each model's improvements and limitations.

    b. Interpret the findings, focusing on the integration of LLMs in NLP tasks.

7. Documentation and Reporting
    a. Document all methodologies, processes, and results systematically.
    b. Compile a comprehensive report detailing the project's workflow and conclusions.

1.5 Testing and Results

Testing Procedure:

- Conduct performance tests on the fine-tuned LLama 7B and GPT-3.5 models using the test portions of the sentiment analysis and question answering datasets.
- Evaluate the machine learning model trained on embeddings from these LLMs under similar test conditions.

Results Analysis:

- Compare the accuracy, precision, recall, and F1-scores of the models pre- and post-fine-tuning.
- Analyze the performance of the machine learning model against the LLMs to determine the efficacy of using LLM embeddings.
- Interpret results to understand the impact of fine-tuning and embedding utilization on model performance.

1.6 Work Plan

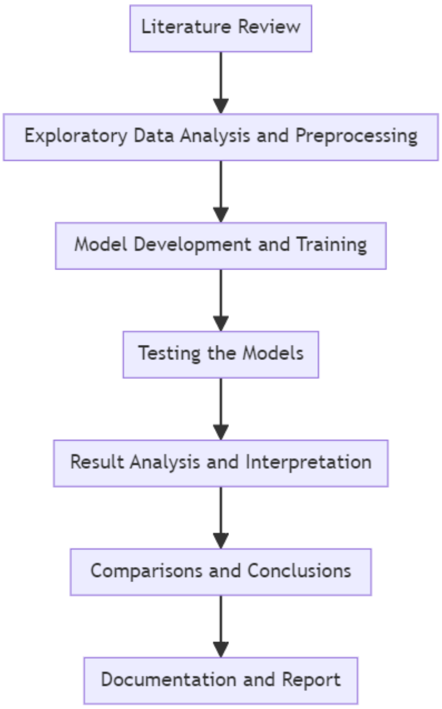

Literature Review
Exploratory Data Analysis and Preprocessing
Model Development and Training
Testing the Models
Result Analysis and Interpretation
Comparisons and Conclusions
Documentation and Report

Fig.1 Work Plan

1.Literature Review

In the literature review phase of the project, we will systematically gather and critically evaluate existing research on Large Language Models (LLMs) related to sentiment analysis and question-answering tasks. This will involve an in-depth analysis of previous studies, their methodologies, findings, and the different LLM architectures that have been explored, such as LLama 7B, Mistral 7B, and GPT 3.5. The aim is to identify gaps in the current literature, justify the necessity of the project, and position our research within the academic discourse. We will also review the theoretical underpinnings of LLMs, their evolution, and current applications. This comprehensive review will ensure that our project is built upon a solid foundation of existing knowledge while aiming to contribute new insights to the field.

2.Exploratory Data Analysis and Preprocessing

This step will focus on preparing the datasets for model training and evaluation. It involves exploratory data analysis (EDA) to understand the characteristics and distributions within the data. Preprocessing tasks will include cleaning the data to remove noise and irrelevant information, normalizing text to a consistent format. This phase is crucial as the quality of data preprocessing directly impacts the performance of machine learning models. The objective is to ensure that the datasets are optimized for the subsequent modeling tasks.

3.Model Development and Training

During this stage, we will develop and train the models using the prepared datasets. This entails setting up LLama 7B, Mistral 7B, and GPT 3.5, initializing their parameters, and starting the training process. Model development includes choosing appropriate architectures and configurations that are most likely to succeed in the given NLP tasks. The training process will adjust the model weights based on the input data and the corresponding outputs for sentiment analysis or question-answering. This step is iterative and will likely involve multiple cycles of training and validation to fine-tune the models for optimal performance. We will also monitor the training process to prevent overfitting and ensure that the models generalize well to new, unseen data.

4.Testing the Models

After training, each model will undergo rigorous testing to evaluate its performance. This phase is critical to assess the practical capabilities of the LLMs. We will employ a separate dataset that the models have not seen during the training phase to ensure an unbiased evaluation. The testing will involve running the models against the datasets and recording their performance for sentiment analysis and question-answering. This step will provide initial insights into how well each model is likely to perform in real-world applications.

5.Result Analysis and Interpretation

In this phase, the data collected from testing the models will be analyzed to interpret the LLMs' performance. The analysis will look at the models' accuracy, understanding of context, ability to handle nuances, and the relevance and precision of their responses to the sentiment analysis and question-answering prompts. Special attention will be given to any anomalies or unexpected results that could indicate deeper issues such as model biases or data overfitting. The interpretation of these results will also consider the theoretical framework established during the literature review to contextualize findings within the broader field of NLP and AI.

6.Comparisons and Conclusions
Comparing the results from the three models will help us draw conclusions about their relative efficacy. This will involve evaluating the performance metrics of LLama 7B, Mistral 7B, and GPT 3.5 to identify which model performs better for each task and under what conditions. The conclusions will synthesize the findings to provide clear insights into the advantages and limitations of each LLM, offering guidance for their application in practical scenarios and future research directions.

7.Documentation and Reporting
The final stage will be to document and report the findings of the project comprehensively. This will include a detailed write-up of the methods, results, and analytical processes used throughout the project. The report will provide a clear narrative that guides the reader through the research questions, experimental design, data analysis, and the conclusions drawn from the work. It will also include a discussion of the implications for the field, potential applications of the research, and suggestions for future studies based on the project's outcomes.

1.7 Report Structure

Based on the project guidance provided by City University, your project report structure, with a focus on Large Language Models (LLMs) for sentiment analysis and question-answering, would be organized as follows:

Chapter 1 - Introduction: This chapter will set the stage for the report, introducing the field of natural language processing (NLP) and the advent of LLMs. It will discuss the significance of LLMs in modern AI, the importance of fine-tuning for specific NLP tasks, and the potential of embeddings in enhancing traditional machine learning models. The chapter will define the objectives, scope, and outline the detailed work plan of the research project.

Chapter 2 - Literature Review: A thorough literature survey will be provided on LLMs, their role in sentiment analysis and question answering, and the integration of LLM embeddings in machine learning. This will include a review of recent advancements, current methodologies, and notable findings in the domain.

Chapter 3 - Methodology: This chapter will delve into the theoretical underpinnings of the LLMs selected for this study, namely LLama 7B, Mistral 7B and GPT-3.5. It will describe the datasets utilized, data preprocessing techniques, exploratory analysis, and the fine-tuning process for all the LLMs.

Chapter 4 - Results and Analysis: Detailed results from the training and performance of the models will be presented. This chapter will include an analysis of the performance metrics.

Chapter 5 - Discussions: This chapter will delve deeply into the nuanced aspects of the study's findings. It will engage in a detailed discussion of the implications of the results, exploring how different models performed in various contexts. Special attention will be given to the unique characteristics and limitations of each model and technique. The chapter will also explore practical applications and potential challenges in the field of Natural Language Processing (NLP). Discussions will include insights from current literature, offering a comprehensive understanding of where the research stands and how it fits into the broader context of NLP advancements. Finally, this chapter will encourage a discourse on potential future developments and research opportunities in the realm of language models and their applications.

Chapter 6 - Evaluation, Reflections and Conclusion: Here, the report will reflect on the evaluation methodologies, discuss the comparative performance of the models, and reflect on the outcomes of the testing phase. The chapter will critically assess the before and after effects of fine-tuning and the implications for using LLM embeddings. This chapter will summarize the key findings, discuss the broader implications for NLP, and provide a perspective on the future direction of research in this field.

References: This section will list all academic papers, articles, and resources cited throughout the report, formatted according to the prescribed referencing style.

Appendix: Supplementary materials will be included, such as code snippets, computational environment specifics and software packages used.

Chapter 2 Literature Review and Context

2.1 Introduction

This chapter offers an extensive literature review on the evolving landscape of Large Language Models (LLMs) and their applications in natural language processing, specifically focusing on sentiment analysis and question answering tasks. It examines various approaches to fine-tuning LLMs, like LLama 7B and Mistral 7B, highlighting their methodologies and effectiveness. Additionally, the chapter explores the innovative use of embeddings extracted from GPT 3.5 in enhancing traditional machine learning models. The theoretical underpinnings of these embeddings, their extraction process, and integration strategies are also discussed. This review provides the necessary academic foundation and contextual understanding for the project's approach within the broader field of NLP and LLMs.

2.2 Large Language Models

Large Language Models (LLMs) are sophisticated AI frameworks designed to understand, interpret, and generate human language, marking a substantial advancement in natural language processing (NLP) (Hadi et al., 2023). They employ deep learning, especially transformer-based architectures, to process extensive text data, enabling them to capture intricate linguistic nuances and contextual information.

This capability allows LLMs to perform a range of tasks from generating coherent prose to answering complex questions with a human-like grasp of subtleties (Du et al., 2023).

LLMs like GPT-3 have revolutionized various domains by providing tools for language translation, content creation, and even coding assistance, reflecting their versatility and potential (Hadi et al., 2023). The ongoing research in LLMs aims to enhance their functionality, addressing challenges like bias and ethical considerations, thus broadening their applicability in solving real-world problems (Du et al., 2023). These models continue to be a focal point in AI research, with their development being integral to future technological advancements in machine learning and automated systems.

2.2.1 History

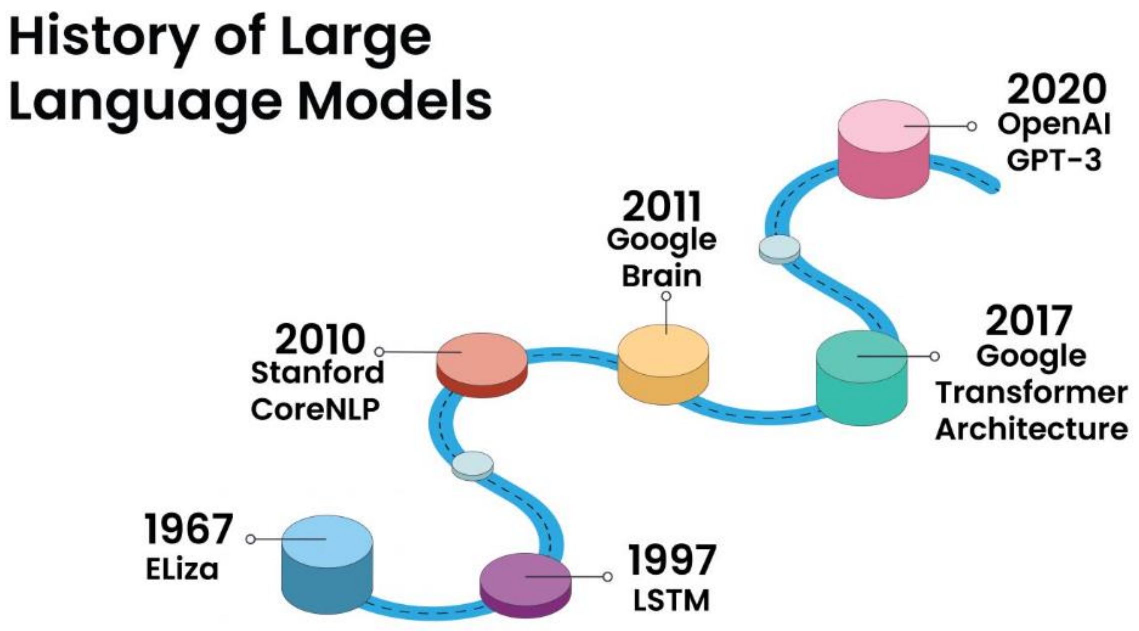


Fig.2 History of Large Language Models (Scribble Data et al., 2023**)**

The evolution of Large Language Models (LLMs) mirrors the significant advances in the field of artificial intelligence, particularly within natural language processing (NLP). From the inception of computational linguistics, which was governed by rule-based systems, the journey of LLMs has been marked by a series of innovations aimed at mimicking and understanding human language more effectively.

[1] Eliza

Eliza, developed in the mid-1960s by Joseph Weizenbaum, represents a fundamental step in the history of LLMs. It was one of the first programs to mimic human conversation, operating within the MAC time-sharing system at MIT. Eliza's method was based on pattern matching and substitution methodology, where input sentences were analyzed based on decomposition rules triggered by key words, leading to responses generated by reassembly rules associated with these decomposition rules

(Weizenbaum et al., 1966). This pioneering work laid the foundation for future development in natural language processing and conversational AI.

[2] LSTM (Long Short-Term Memory) Networks

The development of Long Short-Term Memory (LSTM) networks in 1997 by Sepp Hochreiter and Jürgen Schmidhuber marked a significant advancement in the field of LLMs. LSTMs were designed to overcome the limitations of traditional Recurrent Neural Networks (RNNs) in handling long-range dependencies in sequential data. This innovation enabled the networks to retain information over longer periods and improved their ability to learn from important but infrequent events in the input data (Felix A. Gers et al., 2000). LSTM's architecture made it a cornerstone for subsequent developments in deep learning and NLP, significantly contributing to the evolution of more advanced LLMs.

[3] Stanford CoreNLP

Stanford CoreNLP, developed by Stanford University, represents a significant leap in the history of LLMs. Introduced in 2010, this comprehensive toolkit provided a set of robust tools and algorithms for complex natural language processing tasks. It simplified the NLP pipeline by offering a streamlined approach to tasks such as sentiment analysis and named entity recognition, making it more accessible for both research and commercial applications. The CoreNLP suite's design emphasized a simple, approachable interface and included quality analysis components, significantly contributing to its widespread adoption and use in various NLP-related applications (Manning et al., 2014).

[4] Google Brain

Google Brain, established in 2011, was a pivotal development in the field of Large Language Models (LLMs). As a deep learning and artificial intelligence research team at Google, it focused on integrating open-ended machine learning research with robust information systems and extensive computing resources. This integration allowed for significant advancements in machine learning and natural language processing. Notably, Google Brain developed TensorFlow, a tool that democratized the use of neural networks, making them accessible to the public and advancing internal AI research projects (Google Brain - Wikipedia, 2023). This initiative by Google Brain played a crucial role in the evolution of LLMs, contributing significantly to the field's growth and the development of advanced NLP applications.

[5] Tranformer

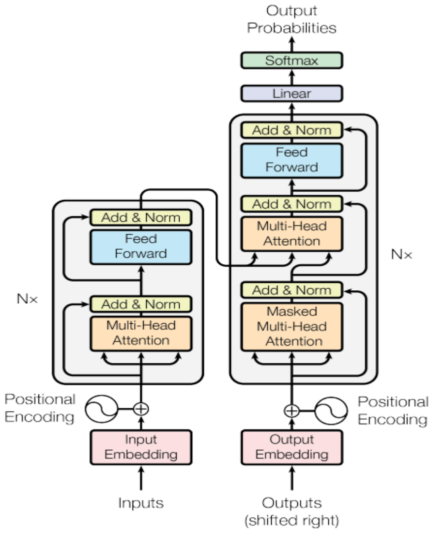


Fig.3 The Transformer - model architecture (Vaswani et al., 2017)

The 2017 introduction of the Transformer model by Vaswani et al. marked a pivotal change in the field of large language models (LLMs). Unlike previous models that relied on recurrent or convolutional structures, the Transformer solely employed attention mechanisms, enabling it to process data in parallel and handle sequences more efficiently. This architecture significantly advanced the capabilities of LLMs, especially in tasks like translation and text generation, by improving training speed and model performance. For example, in the WMT 2014 English-to-German translation task, the Transformer achieved an unprecedented BLEU score of 28.4, highlighting its superiority over existing models (Vaswani et al., 2017).

[6] OpenAI GPT 3

OpenAI's GPT-3, a groundbreaking AI language model, is noted for its immense scale of 175 billion parameters, significantly enhancing natural language processing. It utilizes unsupervised learning from extensive text data to generate text that closely mimics human language. GPT-3's diverse applications span content creation, chatbot conversation, educational resource development, and complex data analysis in healthcare. It also plays a role in artificial creativity, contributing to the arts. Despite its capabilities, GPT-3 faces limitations and ethical concerns, including potential biases and misuse risks. Its training involves analyzing vast text data to predict word probabilities, enhancing text generation accuracy. GPT-3's advancements highlight the need for human oversight in AI, due to occasional content inaccuracies (Lambda Labs, 2023; Picasso AI, 2023; ARTIBA, 2023).

2.3 Applications of large Language Models:

Large Language Models (LLMs) have found applications across a diverse range of fields, fundamentally altering the way we interact with machines and manage information. With the advent of sophisticated models like GPT-3.5 and conceptual frameworks like LLama7B, these applications have expanded even further.

[1] Content Creation and Copywriting: One of the most prominent applications of LLMs is in content creation. GPT-3.5, with its refined understanding of context and improved generative capabilities, has been used to write articles, essays, and even poetry. It can mimic various writing styles, making it a powerful tool for copywriting and content generation.

[2] Programming and Code Generation: LLMs like GPT-3.5 have made significant strides in understanding and generating code. They assist programmers by completing code snippets, debugging, and even writing entire programs in certain cases. This can greatly enhance productivity and is especially useful for educational purposes, helping new learners understand coding concepts.

[3] Conversational AI: GPT-3.5's nuanced language generation makes it ideal for powering chatbots and virtual assistants. These models can engage in more natural and meaningful conversations, providing customer support, therapy sessions, or companionship. LLama7B, as a hypothetical model, would potentially enhance these interactions with even greater understanding and responsiveness.

[4] Translation: LLMs are breaking language barriers by providing translations that consider cultural nuances and context. GPT-3.5's capabilities in this domain suggest a future where real-time, accurate translation can occur across any language, facilitating global communication.

[5] Education: In education, LLMs can provide personalized tutoring, generate practice questions, and even grade assignments with nuanced feedback. They can adapt to a student's learning style and pace, making education more accessible and tailored.

[6] Information Extraction and Data Analysis: LLMs can process vast amounts of text to extract relevant information, summarize data, and provide insights. In sectors like law and healthcare, this can reduce the workload on professionals by summarizing cases, research papers, or patient records.

[7] Accessibility: LLMs can assist those with disabilities by generating descriptive text for images, transcribing audio for the hearing impaired, and reading text aloud for the visually impaired. The potential improvements in LLama7B could further enhance the quality and reliability of these services.

[8] Gaming: In the gaming industry, LLMs contribute to creating dynamic storylines and dialogues. GPT-3.5 can generate narrative content that responds to player actions, creating a more immersive gaming experience.

[9] Ethical and Societal Implications: While the applications of LLMs like GPT-3.5 and LLama7B are vast, they also come with ethical considerations. Ensuring that these models are used responsibly, without propagating biases or misinformation, is a challenge that must be addressed as these technologies become more integrated into society.

[10] Healthcare: LLMs like GPT-3.5 and the conceptual LLama7B could revolutionize healthcare by parsing patient data to assist in diagnosis or sifting through medical literature to aid in research. These models can analyze patient histories, suggest potential treatments based on symptoms, and even help in predicting patient outcomes.

[11] Legal and Compliance: In the legal field, LLMs are being used for document review, contract analysis, and legal research. They can help identify relevant case law, parse through legislation, and draft legal documents. Advanced LLMs may be able to predict court outcomes based on historical data, assisting lawyers in case preparation.

[12] Finance and Economics: LLMs process financial documents, analyze market trends, and generate reports. They could be trained to read and summarize economic reports, predict market movements, and even detect fraudulent activities by recognizing patterns in financial data.

In summary, the applications of LLMs are transforming industries by providing sophisticated tools for language-related tasks. GPT-3.5 and the hypothetical LLama7B represent the cutting edge of these developments, promising even more advanced capabilities and applications in the future.

2.4 Large Language Models for our Project

[1] GPT 3.5: GPT-3.5, an advanced iteration in the Generative Pre-trained Transformer series by OpenAI, represents a significant leap in language model capabilities. It's designed with a deep learning architecture that enables a high degree of linguistic understanding and generation, making it adept at a wide range of language tasks. This model is built on the foundation of GPT-3 but includes several improvements that enhance its performance and versatility. GPT-3.5's remarkable ability to understand context, generate coherent and contextually relevant text, and even exhibit a degree of creative thinking sets it apart in the field of natural language processing. It's used in various applications such as content creation, coding assistance, language translation, and more.

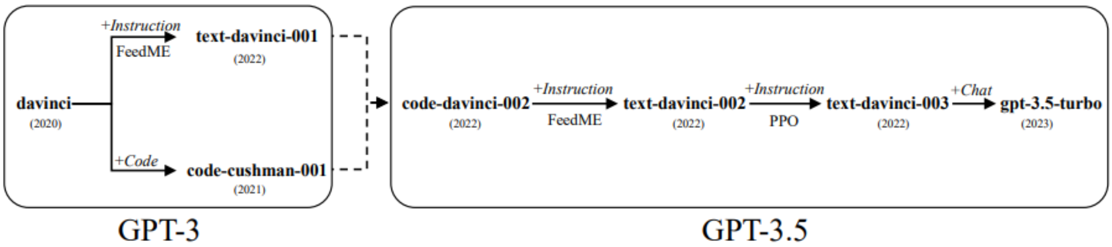


Fig.4 Evolutionary Path and Training Strategies of GPT Models (Junjie Ye et al. 2023).

The image portrays the evolutionary relationship of the GPT series models, highlighting the progression from the original GPT-3 model, including davinci and code-cushman versions, to the newer GPT-3.5 variants. It indicates the use of distinct training strategies such as FeedME and PPO (Proximal Policy Optimization), which contribute to the enhanced performance of GPT-3.5 models. The dashed arrow suggests incremental updates from GPT-3 to GPT-3.5, where specific enhancements in training methodology and data handling have been implemented, although detailed official documentation on these updates is not provided. The evolution reflects OpenAI's commitment to continual improvement in the field of language models (Junjie Ye et al. 2023).

The paper ‘A Comprehensive Capability Analysis of GPT-3 and GPT-3.5 Series Models’ (Junjie Ye et al., 2023) provides a comprehensive analysis of GPT-3.5, highlighting its strengths in natural language understanding tasks and its advancements over previous models. Notably, GPT-3.5 demonstrates improved performance in scenarios requiring a deeper understanding of context and instructions. The paper also sheds light on the limitations and challenges associated with GPT-3.5, such as the need for further improvement in model robustness and the varying degrees of performance across different tasks (Junjie Ye et al. 2023). Despite these challenges, GPT-3.5's advancements mark a significant step forward in the evolution of language models, showcasing its potential in transforming how we interact with and leverage artificial intelligence in language-related tasks.

In the paper ‘Can ChatGPT Replace Traditional KBQA Models? An In-depth Analysis of the Question Answering Performance of the GPT LLM Family’ (Yiming Tan et al., 2023), the study evaluated the model on complex question-answering tasks, using a framework inspired by CheckList, across eight different datasets. It found that while GPT-3.5 outperforms traditional models on certain datasets, it still lags behind state-of-the-art models on newer datasets. Challenges included the difficulty in evaluating answers due to the generation of text paragraphs containing answers, rather than precise answers, and the limitation of the Exact Match metric for evaluation. The paper suggests that while GPT-3.5 has potential, it requires further improvement in terms of robustness and performance across various types of questions (Yiming Tan et al., 2023). In the paper, GPT-3.5 was used to generate responses for QA tasks across various datasets. The challenges included GPT-3.5's difficulty in providing precise answers and the unsuitability of the Exact Match metric for evaluating its performance. The model showed potential but needed improvements in robustness and adaptability across different question types (Yiming Tan et al., 2023).

In another paper ‘SentimentGPT: Exploiting GPT for Advanced Sentiment Analysis and its Departure from Current Machine Learning’ (Kiana Kheiri et al., 2023), the study utilized GPT-3.5 Turbo for advanced sentiment analysis on Twitter data. GPT-3.5 faced challenges in understanding context and detecting sarcasm, which are common issues in sentiment analysis. However, the model demonstrated an enhanced ability to handle these complexities effectively, showing the potential of GPT models in sentiment analysis applications (Kiana Kheiri et al., 2023).

[2] LLama7B: LLaMA is a collection of foundation language models ranging from 7 billion to 65 billion parameters, notable for being trained on publicly available datasets exclusively (Touvron et al., 2023). This approach allows for state-of-the-art model performance without the reliance on proprietary data. The LLaMA-13B model, for instance, outperforms GPT-3 (175B) on most benchmarks, while the LLaMA-65B model shows competitive performance with even larger models like Chinchilla-70B and PaLM-540B. Crucially, these models can be run on a single GPU, enhancing accessibility for a broader range of researchers and developers (Touvron et al., 2023).

The application of LLaMA for sentiment analysis in financial texts, as detailed in the paper ‘Instruct-FinGPT: Financial Sentiment Analysis by Instruction Tuning of General-Purpose Large Language Models’ (Boyu Zhang et al., 2023), involved an instruction tuning approach to enhance the model's performance. LLaMA was fine-tuned with a small subset of financial sentiment analysis data, transforming the sentiment analysis task into a generative format. This process improved the model's ability to interpret financial contexts and numerical data, making it more effective at sentiment prediction. Challenges faced included the model's initial insensitivity to numerical data and difficulty in contextually interpreting sentiments without explicit cues. The tuned LLaMA model showed a significant performance boost, outperforming other models in accuracy and understanding (Boyu Zhang et al., 2023).

The study described in ‘BETTER QUESTION-ANSWERING MODELS ON A BUDGET’ (Yudhanjaya Wijeratne et al., 2023) employed LLaMA to enhance question-answering models with a budget in mind. Using the Stanford Alpaca dataset, they improved Facebook’s OPT models, showcasing that smaller models could rival the performance of models three times their size. The challenges faced included model verbosity leading to regurgitation of information and a lack of refined writing ability, as demonstrated by the models fabricating responses with extensive but irrelevant details. Despite these challenges, the LLaMA 7B model showed significant improvements, comparable to OpenAI's text-davinci-003, when trained with the same methods (Yudhanjaya Wijeratne et al., 2023).

[3] Mistral 7B: Mistral 7B is a 7-billion-parameter language model engineered for superior performance and efficiency. It is designed to outperform existing models in various benchmark evaluations while maintaining computational efficiency. This model has shown to surpass the best open 13B model, Llama 2, across all evaluated benchmarks and the best released 34B model, Llama 1, in areas such as reasoning, mathematics, and code generation. The introduction of Mistral 7B represents a significant advance in natural language processing (NLP), demonstrating the possibility of achieving high performance with relatively smaller models (Albert Q. Jiang et al., 2023).

Key Features of Mistral 7B: One of the key innovations in Mistral 7B is its use of grouped-query attention (GQA) and sliding window attention (SWA). GQA enhances the inference speed and reduces memory requirements, allowing for higher throughput crucial for real-time applications. SWA, on the other hand, is adept at handling longer sequences more effectively and at a reduced computational cost, addressing a common limitation in large language models (LLMs). These attention mechanisms collectively contribute to the enhanced performance and efficiency of Mistral 7B (Albert Q. Jiang et al., 2023).

Deployment and Fine-tuning: Mistral 7B is released under the Apache 2.0 license, making it accessible for a wide range of applications. Its implementation is designed for easy deployment on local or cloud platforms, and integration with popular platforms like Hugging Face. The model also demonstrates ease of fine-tuning across various tasks, as evidenced by a chat model fine-tuned from Mistral 7B that significantly outperforms similar models (Albert Q. Jiang et al., 2023).

Comparative Performance: Mistral 7B has been compared with various models like Llama 2 7B/13B and Code-Llama 7B across a wide range of benchmarks. The results highlight Mistral 7B's superior performance in areas like mathematics, code generation, and reasoning benchmarks. Its performance mirrors what one might expect from a much larger model, suggesting an impressive efficiency in the cost-performance spectrum (Albert Q. Jiang et al., 2023).

Guardrails and Ethical Use : Mistral 7B includes features for enforcing guardrails in front-facing applications, essential for maintaining ethical use of AI. It can perform fine-grained content moderation and has the capability to handle unsafe prompts effectively, ensuring the generation of content within ethical boundaries (Albert Q. Jiang et al., 2023).

Mistral 7B is a testament to the evolving landscape of NLP, demonstrating that it's possible to achieve high-level performance with smaller models. It sets a new benchmark in the field, indicating that the future of model development may focus more on efficiency and effectiveness rather than merely scaling up model sizes (Albert Q. Jiang et al., 2023).

2.5 Fine-Tuning and Low-Rank Adaptation in Large Language Models:

Fine-Tuning: Overview and Challenges: Fine-tuning in language models is a crucial process for adapting pre-trained models to specific tasks. The study "How Fine Can Fine-Tuning Be? Learning Efficient Language Models" by Radiya-Dixit and Wang (2020) investigates the fine-tuning of large networks like BERT and finds that despite the enormity of these models, only a small amount of supervised fine-tuning is needed. This study reveals that fine-tuned models are close in parameter space to the pre-trained one, and it is possible to fine-tune only the most critical layers. Further, it shows that numerous good solutions exist in sparsified versions of the pre-trained model, allowing fine-tuning of large language models by setting a portion of the pre-trained weights to zero, thus saving on computational cost and storage (Radiya-Dixit & Wang, 2020).

Low-Rank Adaptation (LoRA) : LoRA, a fine-tuning technique, presents a more efficient alternative to traditional methods. It optimizes additive correction terms with a low-rank structure, significantly reducing trainable parameters. The "LoRA Ensembles for Large Language Model Fine-Tuning" papers explore this approach, which involves applying adapters to self-attention modules' query and value matrices in LLMs. The papers demonstrate that LoRA can be combined with regularization techniques like KL regularization and early stopping to improve model calibration and prevent overfitting. This method is particularly effective in fine-tuning large models like LLaMA-13b in a computationally efficient manner (Xi Wang et al., 2023).

Advantages of LoRA: LoRA's primary advantage is its parameter efficiency, requiring fewer parameters to be fine-tuned. This method improves the accuracy and calibration of predictions, making it highly beneficial in safety-critical applications. The compatibility of LoRA with various regularization strategies offers a balanced approach to fine-tuning, maintaining model reliability (Xi Wang et al., 2023).

The integration of fine-tuning and LoRA techniques marks a significant advancement in adapting large language models for specific tasks. Fine-tuning, as explored by Radiya-Dixit and Wang, can be more efficient than previously thought, and LoRA further enhances this efficiency. These methods ensure that large language models can be adapted with reduced computational resources while maintaining or improving performance and reliability.

2.6 Few Shot Prompting

In language models, few-shot learning is a technique meant to use little training data to adjust huge pre-trained models to particular tasks. This method works especially well in situations where it is not feasible to gather large-scale labelled datasets. It makes use of vast amount of knowledge that models have ingrained from their thorough pre-training.

Adaptive Prompting Method: Zhang, Chai, and Xu propose an adaptive prompting method that uses a seq2seq-attention structure for sentiment analysis. This method dynamically constructs adaptive prompts, improving prompt quality and relevance to input sequences (Zhang et al., 2023).

Hybrid Prompt Learning: Their research also delves into hybrid prompt learning, combining the advantages of hand-crafted and automated prompts, thereby enhancing performance in various sentiment analysis tasks (Zhang et al., 2023).

Parameter-Efficient Fine-Tuning: Contrasting with in-context learning (ICL), parameter-efficient fine-tuning (PEFT) is emphasized for its efficiency and effectiveness. PEFT, including methods like $(IA)^3$, introduces minimal new parameters while achieving better performance (Liu et al., 2022).

T-Few Methodology: T-Few, a method based on the T0 model. It demonstrates adaptability to new tasks without task-specific tuning, outperforming state-of-the-art methods in some cases (Liu et al., 2022).

Advantages of Few-Shot Learning:

- Data Efficiency: When compared to standard methods, few-shot learning requires a much less number of data samples for training.
- Computational Efficiency: This method requires less computing power, particularly when contrasted with complete model fine-tuning.

Challenges and Limitations

- Consistency of Effectiveness: Depending on the job and the calibre of the examples given, few-shot learning can be successful in different situations.
- Model Dependence: The pre-training quality of the underlying model has a major impact on how well few-shot learning works.

2.7 LLM Embedding

Embedding is process of converting discrete items, like words, letters, or phrases, into continuous vector representations in the context of Large Language Models (LLMs) and Natural Language Processing (NLP). These vector representations, or embeddings, enable neural networks to process and analyse language data more effectively. In order to improve the model's comprehension and manipulation of language input, embeddings are made to capture the syntactic and semantic characteristics of the language parts they represent.

Embedding Techniques for Code Comment Classification: The paper "A ML-LLM pairing for better code comment classification" discusses the use of embedding techniques in the context of code comment classification. It highlights that embeddings, particularly those based on transformer models and contextualized word embeddings, can significantly enhance the performance of classification systems by better capturing semantic concepts in code (Hanna Aki Akl et al., 2023).

Embedding in Language Translation: The paper "Embed_Llama: using LLM embeddings for the Metrics Shared Task" describes 'Embed_llama', a metric that uses embeddings from LLMs for language translation, highlighting the role of embeddings in translating sentences into vector space where semantic and geometric proximities are linked. This document also emphasizes the importance of embedding layers in neural networks for transforming textual data into continuous vector representations, making them suitable for processing by neural networks (Sören Dréano et al., 2023).

Embedding-based Retrieval (EBR): The paper "EMBEDDING-BASED RETRIEVAL WITH LLM FOR EFFECTIVE AGRICULTURE INFORMATION EXTRACTING FROM UNSTRUCTURED DATA" explains how EBR differs from plain text search by transforming documents into vectors and mapping them to a vector space. This approach allows for determining document similarity based on the distance in the vector space, enhancing the accuracy and relevance of retrieval tasks (Ruoling Peng et al., 2023).

Advantages of Embeddings in LLMs:

- **Enhanced Semantic Understanding**: Embeddings allow LLMs to capture nuanced semantic relationships between words or sentences, going beyond mere syntactic analysis. This leads to a more profound understanding of language, enabling LLMs to perform tasks like translation, information extraction, and retrieval with greater accuracy and context-awareness (Sören Dréano et al., 2023), (Ruoling Peng et al., 2023).

- **Efficiency in Information Extraction:** Embeddings facilitate the transformation of unstructured text into structured data, making the process of information extraction more efficient and less labor-intensive. This is crucial in handling the vast amounts of data in the information era (Ruoling Peng et al., 2023).
- **Improved Performance in NLP Tasks:** Embeddings contribute to the enhanced performance of LLMs in various NLP tasks, often surpassing traditional methods. This is attributed to their ability to effectively capture and utilize the semantic and syntactic properties of language (Ruoling Peng et al., 2023).
- **Flexibility and Versatility:** The flexibility of embeddings, as evidenced by their adjustable dimensionality, allows them to be tailored for specific tasks, capturing more intricate relationships where needed. This makes them versatile tools in a range of NLP applications (Sören Dréano et al., 2023).

2.8 Confusion Matrix

The paper ‘Relative Confusion Matrix: Efficient Comparison of Decision Models’ (Luc-Etienne Pomme et al. 2022) presents the Relative Confusion Matrix (RCM), a novel tool for effectively comparing classification models. RCM enhances traditional confusion matrices by highlighting class-wise performance differences between two models using color coding and symbolic representation. This approach allows for a more efficient and nuanced comparison at a class level, aiding in identifying strengths and weaknesses of models. The study includes a user evaluation demonstrating the superiority of RCM over other methods in comparing model performances, particularly in terms of response time and error rates (Luc-Etienne Pomme et al. 2022).

In this paper ‘Performance Analysis of Text Classification Algorithms using Confusion Matrix’ (Maria Navin J R et al., 2016), the authors analyze the performance of text classification algorithms using confusion matrices. They focus on algorithms such as k-Nearest Neighbor, Naïve Bayes, Logistic Regression, and Support Vector Machines. By applying a 10-fold cross-validation method to a movie review dataset, the study demonstrates how confusion matrices can be effectively used to compute various statistical parameters like accuracy, sensitivity, and specificity. These metrics provide a comprehensive evaluation of each algorithm's performance in classifying text, revealing insights into their relative strengths and weaknesses in different scenarios (Maria Navin J R et al., 2016).

The paper ‘Comparative Evaluation of Algorithms for Sentiment Analysis over Social Networking Service’ (Krouska, Troussas et al.,2017) focuses on comparing sentiment analysis algorithms for social networking services, specifically Twitter. It evaluates five machine learning classifiers (Naïve Bayes, Support Vector Machine, k-Nearest Neighbor, Logistic Regression, and C4.5) and a lexicon-based approach (SentiStrength), using confusion matrices across different datasets. The study demonstrates the effectiveness of Naïve Bayes and Support Vector Machine, showing their superiority in various test models and datasets. This research is instrumental in guiding the selection of optimal algorithms for sentiment analysis services (Krouska, Troussas et al., 2017).

Chapter 3: Methodology

3.1 Introduction

In this chapter, we will systematically explore our methodology, covering the dataset, pre-processing techniques, the chosen model, and the application methods. We start by examining the dataset, understanding its significance and how it aligns with our research goals. Following this, we delve into the pre-processing phase, where we prepare our data for analysis, ensuring quality and reliability. We then introduce the core of our study - the model we have chosen. Finally, we detail the techniques used in applying this model, focusing on the procedures and customizations that enable us to draw meaningful insights from our data. This chapter aims to provide a clear and concise overview of our research process, from data selection to model application.

3.2 Dataset

This report focuses on our exploration of Natural Language Processing (NLP) tasks, centered around two carefully selected datasets: one for sentiment analysis and another for question and answer tasks. We will delve into the specifics of these datasets, their challenges, and our approaches to tackle them in subsequent sections

3.2.1 Dataset for sentiment analysis

The XED dataset, sourced from Helsinki-NLP, is a collection of emotion-annotated movie subtitles from OPUS. It uses Plutchik's 8 core emotions for annotation and is a multilabel dataset. The dataset was originally annotated mainly in English and Finnish, and later expanded to 31 languages, covering over 950 lines of annotated subtitles. It includes 24,164 annotations (plus 9,384 neutral), 17,530 unique data points (plus 6,420 neutral), and 108 annotators (63 active). For languages like Arabic, the dataset contains 3,590 annotations with an average length of 30.02 characters per line .(https://github.com/Helsinki-NLP/XED/tree/master)

The dataset you provided, originally in TSV format and converted to CSV, is a sentiment analysis dataset from the GitHub repository. It contains a total of 17,527 entries.In the sentiment analysis dataset provided, the categories labeled from '1' to '8' correspond to a range of emotions, each representing a specific sentiment. These categories are as follows: '1' for 'anger', '2' for 'anticipation', '3' for 'disgust', '4' for 'fear', '5' for 'joy', '6' for 'sadness', '7' for 'surprise', and '8' for 'trust'. Any data point not falling within these eight categories is classified as 'neutral'. This categorization is integral to understanding the dataset's structure, as it directly ties each text entry to a distinct emotional sentiment. Such a classification allows for a nuanced analysis of the text, providing insights into the emotional undertones or explicit expressions contained within each data entry. The occurrence of the categorises is shown below in a form of pie chart.

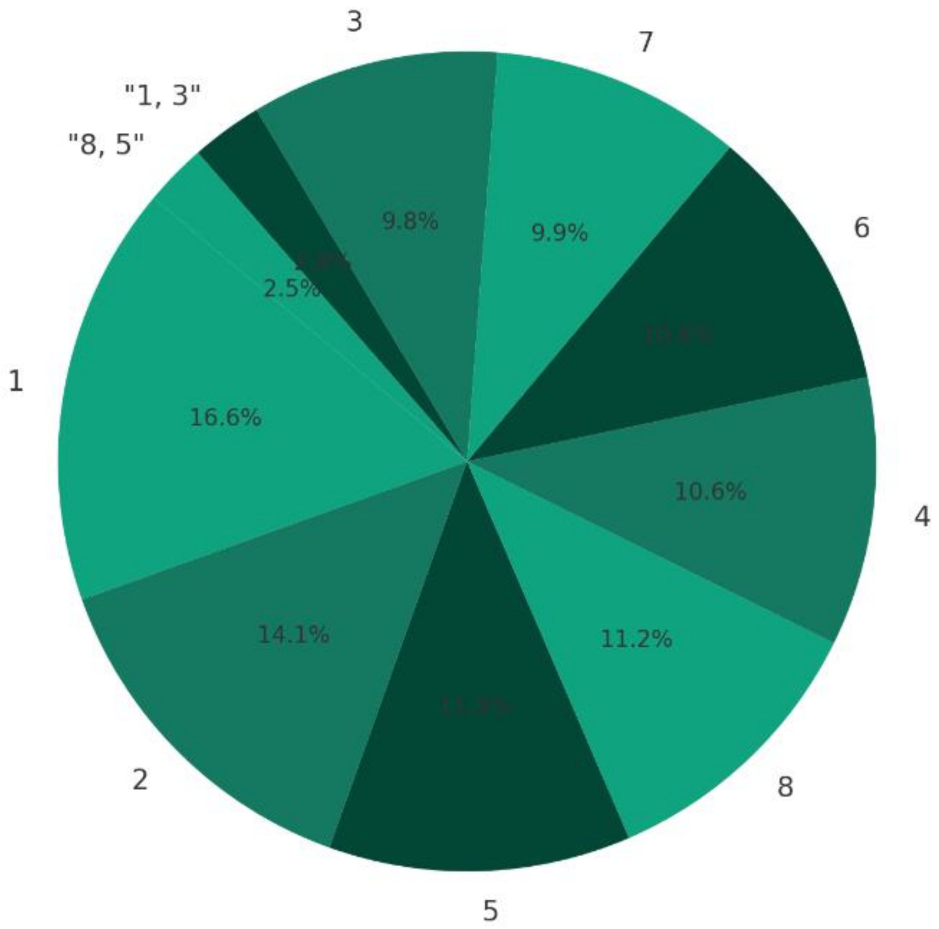


Fig.5 Occurrences of emotions in the dataset

3.2.2 Dataset for Question and Answer

The dataset comprises 825 entries across 41 columns, featuring a blend of object (text) and float64 (numeric) data types. The primary data fields include 'ArticleTitle', 'Question', and 'Answer', with the remaining columns being mostly unnamed, ranging from 'Unnamed: 3' to 'Unnamed: 40'. These unnamed columns have a notable amount of missing data, as many contain only a handful of non-null entries. Regarding data quality, the dataset is characterized by significant missing values, particularly in the unnamed columns. This suggests the need for data cleaning. (https://www.kaggle.com/datasets/rtatman/questionanswer-dataset/)

For a graphical representation, I've created two plots:

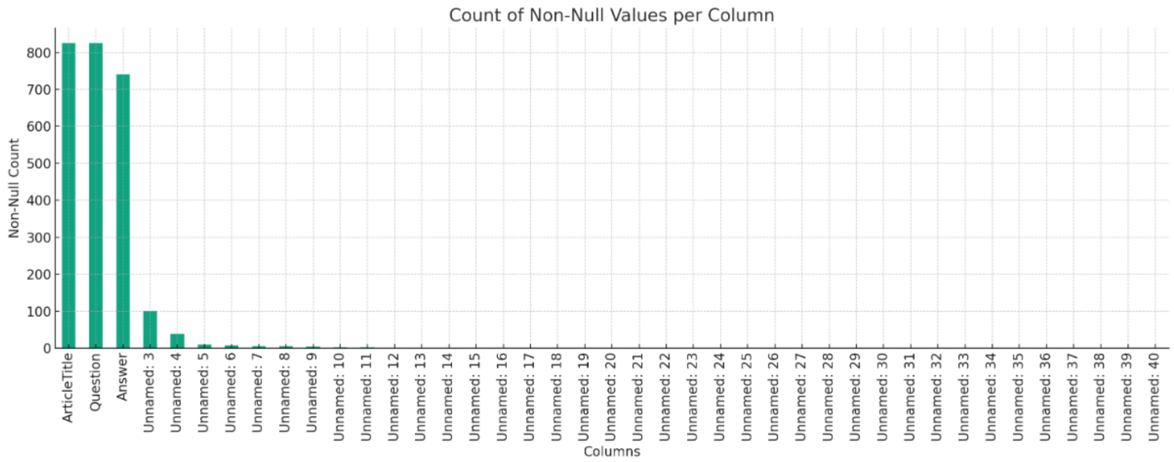


Fig.6 Count of Non-null values per column

Count of Non-Null Values per Column: This graph illustrates the number of non-null entries in each column, highlighting the columns with the most data available.

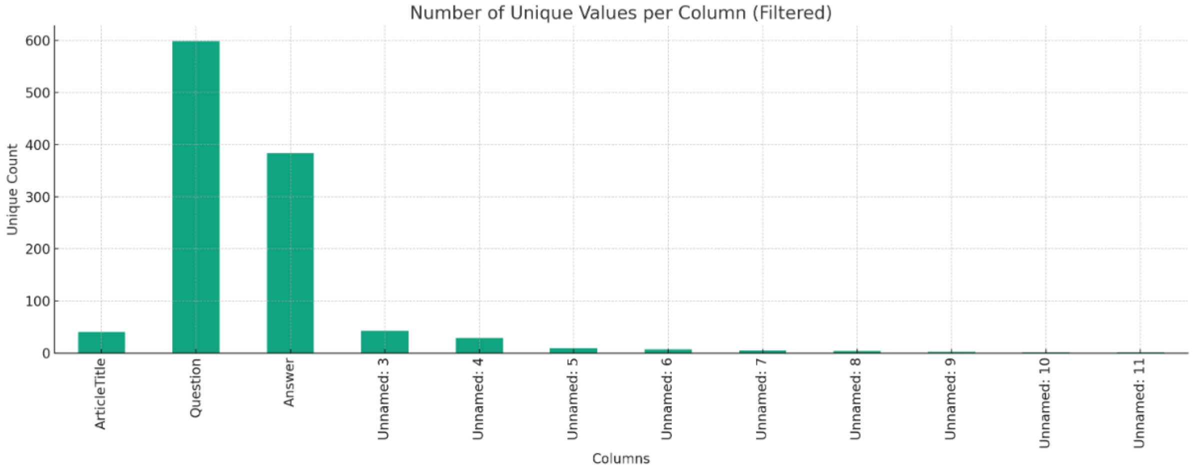


Fig.7 Number of Unique Values per column

Number of Unique Values per Column (Filtered): This graph displays the count of unique values for each column, excluding those with very few non-null values. It provides insight into the diversity of data within each column.

3.3 Exploratory Data Analysis and Preprocessing

The practice of analysing a dataset both statistically and visually in order to find trends or abnormalities is known as exploratory data analysis, or EDA. It includes creating hypotheses, cleansing, data visualisation, and summary statistics.

Data preprocessing addresses problems including mistakes, missing data, and feature engineering to get raw data ready for analysis or modelling. To prepare data for analysis or machine learning, it involves feature selection, scaling, data cleansing, handling missing data, and other transformations.

3.3.1 Preprocessing for Sentiment Analysis dataset

In the initial stages of our data preprocessing journey, we began by transforming the dataset from a TSV (Tab-Separated Values) format to a CSV (Comma-Separated Values) format, which is commonly used for structured data storage and analysis. This conversion allowed us to work with the data more seamlessly in a variety of tools and platforms.The next step involved the refinement of the dataset's second column, which contained text data that was enclosed within double quotation marks (" "). By removing these quotation marks, we ensured that the text data in this column was in a cleaner, more standardized format for further processing.

Following this, we encountered the crucial task of extracting emotional labels from the second column of the dataset. Each text entry was associated with an emotion, such as anger, anticipation, disgust, fear, joy, sadness, surprise, trust, or neutral. These emotions were numerically encoded to facilitate subsequent analysis, with anger represented as 1, anticipation as 2, disgust as 3, fear as 4, joy as 5, sadness as 6, surprise as 7, trust as 8, and neutral as 0. This encoding scheme created a more structured representation of emotions for our data.

To ensure data integrity and quality, we then addressed missing values within the dataset. Any instances where data was absent or incomplete were meticulously cleaned and handled. This step was vital in guaranteeing the reliability of our dataset for downstream analysis. Upon completing these preprocessing tasks, we saved the refined dataset under the name "filtered_sentiment-analysis_dataset." This processed dataset boasted a total of 17,195 rows, reflecting the successful removal of noise and the standardization of the data. By diligently executing these preprocessing steps, we set the stage for more accurate and insightful sentiment analysis, laying the foundation for subsequent tasks and investigations.

To facilitate the application of NLP models, specifically LLama 7B and Mistral 7B, our preprocessed dataset underwent a critical step of partitioning. We adopted an 80-20 split strategy, dividing the dataset into two distinct subsets: an 80% portion designated for training and a 20% portion reserved for testing. This division was crucial for model evaluation and validation, allowing us to gauge the performance of LLama 7B and Mistral 7B on previously unseen data, ensuring the robustness and generalizability of our sentiment analysis models. With this carefully constructed split, we were well-equipped to assess the efficacy and accuracy of these language models in predicting emotional sentiment across our dataset

In summary, the data preprocessing journey for the sentiment analysis dataset encompassed format conversion, text cleaning, emotion encoding, handling missing values, and ultimately, the creation of a cleaned and structured dataset, which is poised for deeper analysis and modeling.

3.3.2 Exploratory Data Analysis for sentiment analysis dataset

In my recent exploratory data analysis of a sentiment analysis dataset, I focused on uncovering the underlying patterns and distributions of sentiments in various text inputs. The dataset comprised columns indicating text inputs, numerical codes, and corresponding sentiment labels. My analysis involved creating several insightful visualizations to better understand the data.

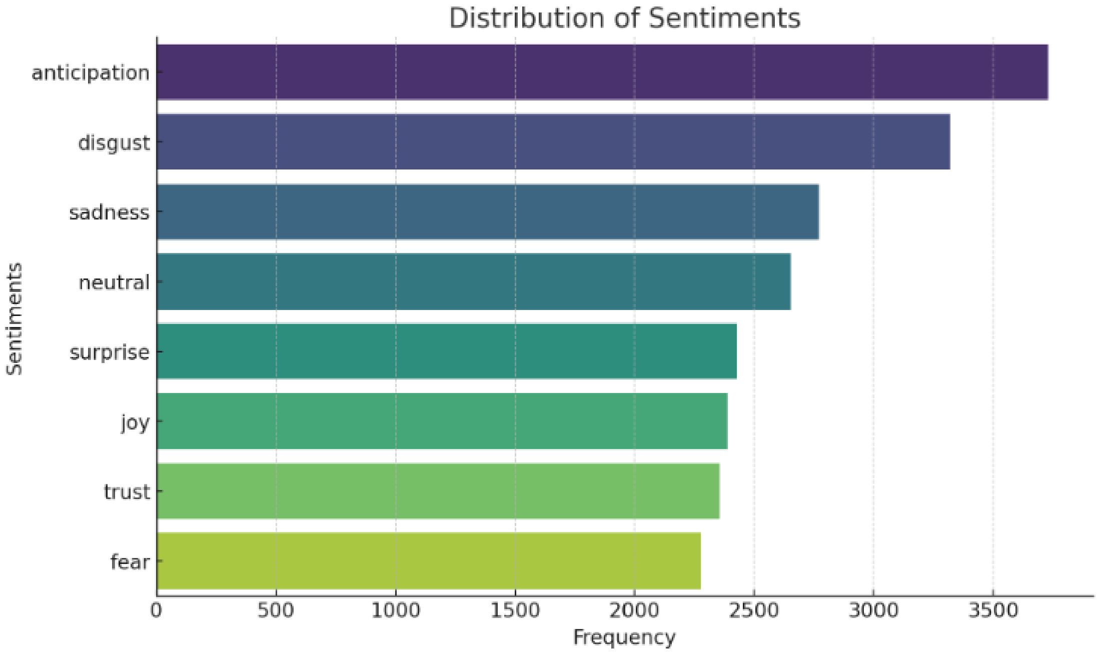


Fig.8 Dustribution of Sentiments

Distribution of Sentiments: One of the key visualizations I developed was a bar chart representing the distribution of different sentiment labels in the dataset. This graph provided a clear and concise overview of the most prevalent sentiments expressed in the texts. The analysis revealed which emotions or sentiments were more frequently observed, offering a deeper understanding of the emotional tone prevalent in the dataset.

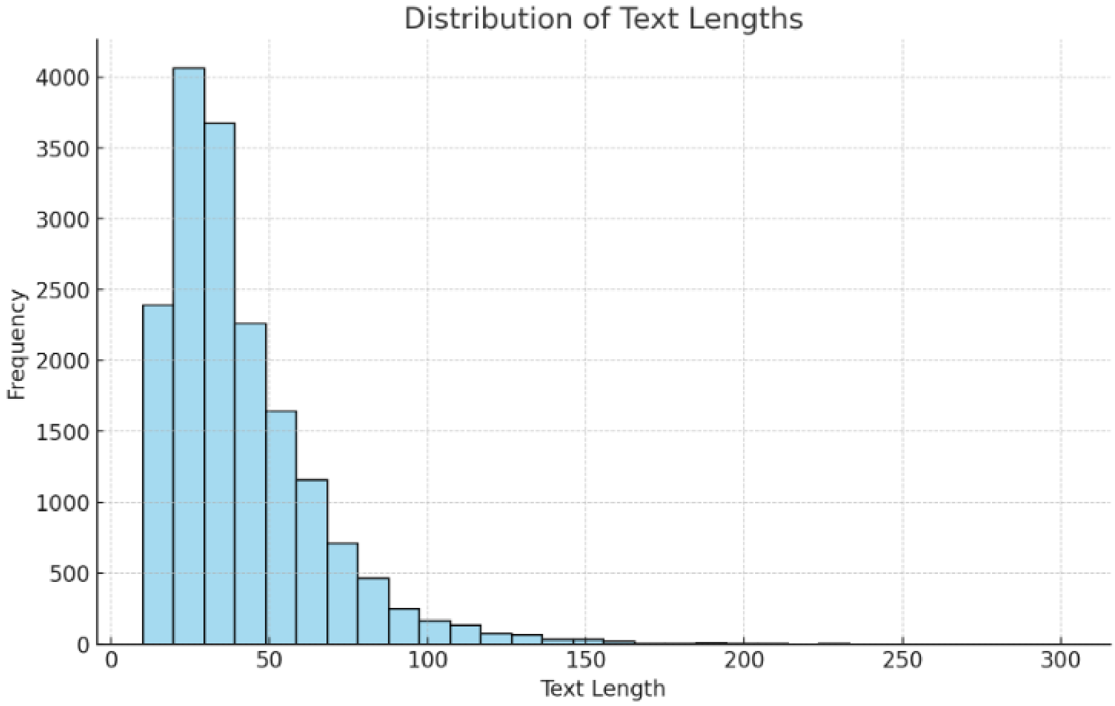


Fig.9 Distribution of Text Lengths

Text Length Distribution: I also examined the length of the text inputs by constructing a histogram. This graph displayed the frequency distribution of text lengths, revealing a pattern in the dataset regarding the typical length of text inputs used for sentiment analysis. Understanding the variation in text lengths is crucial, as it can influence the sentiment detection process.

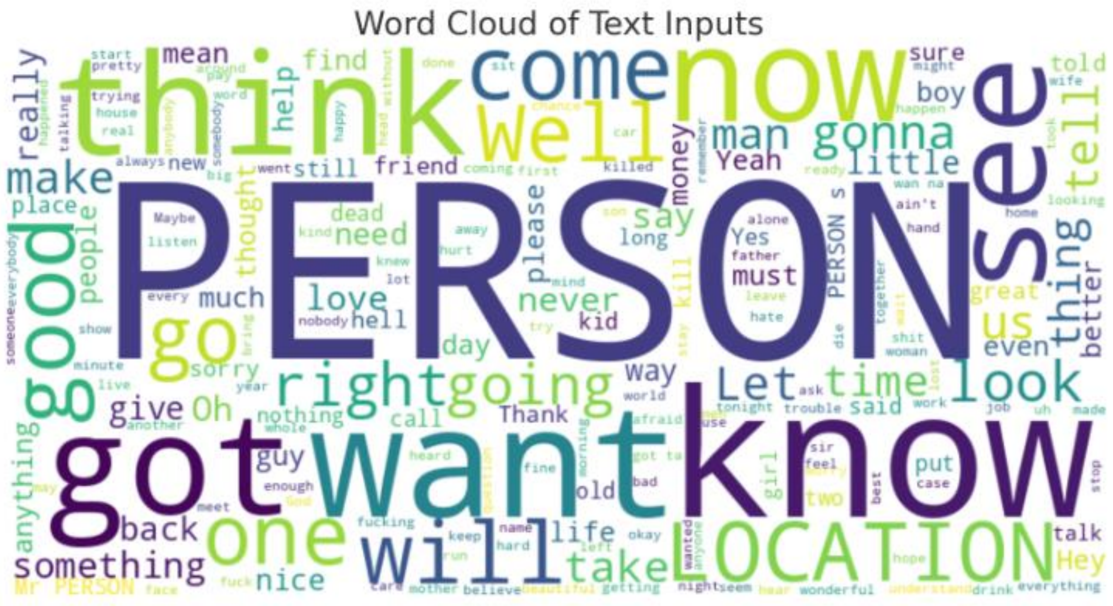


Fig. 10 Word Cloud of Text Inputs of sentiment analysis dataset

Word Cloud: Additionally, to gain insights into the most frequently occurring words in the text inputs, I generated a word cloud. This visualization illustrated the predominant themes and topics present in the dataset, with more prominent words indicating higher frequencies. The word cloud provided a visual representation of the key terms and concepts that were prevalent in the text inputs.

These visualizations collectively offered a comprehensive view of the dataset, enhancing the understanding of the sentiments expressed and the general characteristics of the text inputs. This exploratory data analysis was instrumental in uncovering the nuanced aspects of the dataset, which could be pivotal for further sentiment analysis and text processing tasks.

3.3.3 Preprocessing for Question and Answer dataset

Our journey with the question and answer dataset commenced with a focus on streamlining the data for relevance and efficiency. To achieve this, we elected to retain only the first three columns of the dataset, which included "Article title," "question," and "answer." These columns encapsulated the core elements essential for our analysis, thus rendering the remaining columns superfluous. This process of selective column retention was executed within the confines of Excel, where extraneous data was swiftly removed, simplifying the dataset and enhancing its utility for subsequent tasks.

Following this refinement, the dataset emerged with a more concise structure, consisting solely of the essential columns. This transformation not only improved the dataset's readability but also significantly reduced its complexity, aligning it more closely with our objectives.

Upon finalizing this preprocessing stage, we moved on to an essential step in the data preparation pipeline: dataset splitting. In order to facilitate the training and evaluation of our question and answer models, we adopted a systematic approach of partitioning. The dataset was divided into two distinct subsets, adhering to the widely recognized 80-20 split ratio. Eighty percent of the data was allocated for training purposes, enabling our models to learn and adapt to the patterns present in the questions and answers. The remaining 20 percent was reserved for testing, serving as an independent benchmark for assessing model performance.

This dataset splitting strategy was meticulously chosen to ensure robust model evaluation. By setting aside a portion of the data for testing, we could gauge how well our question and answer models generalized to unseen examples, thereby providing us with valuable insights into their accuracy and effectiveness.

In conclusion, our data preprocessing endeavors for the question and answer dataset involved the thoughtful curation of relevant columns and the subsequent division of the dataset into training and testing sets, all of which were executed with precision and purpose. These preparatory steps laid the groundwork for our subsequent exploration of NLP techniques in the realm of question and answer systems, offering a refined and structured dataset for model development and evaluation.

3.3.4 Exploratory Data Analysis for Question and Answer dataset

In the recent exploratory data analysis this dataset, I delved into identifying patterns and trends within the data. The dataset consisted of columns detailing article titles, corresponding questions, and their answers ('yes' or 'no'). My approach involved creating a series of visualizations to elucidate the characteristics and dynamics of the dataset.

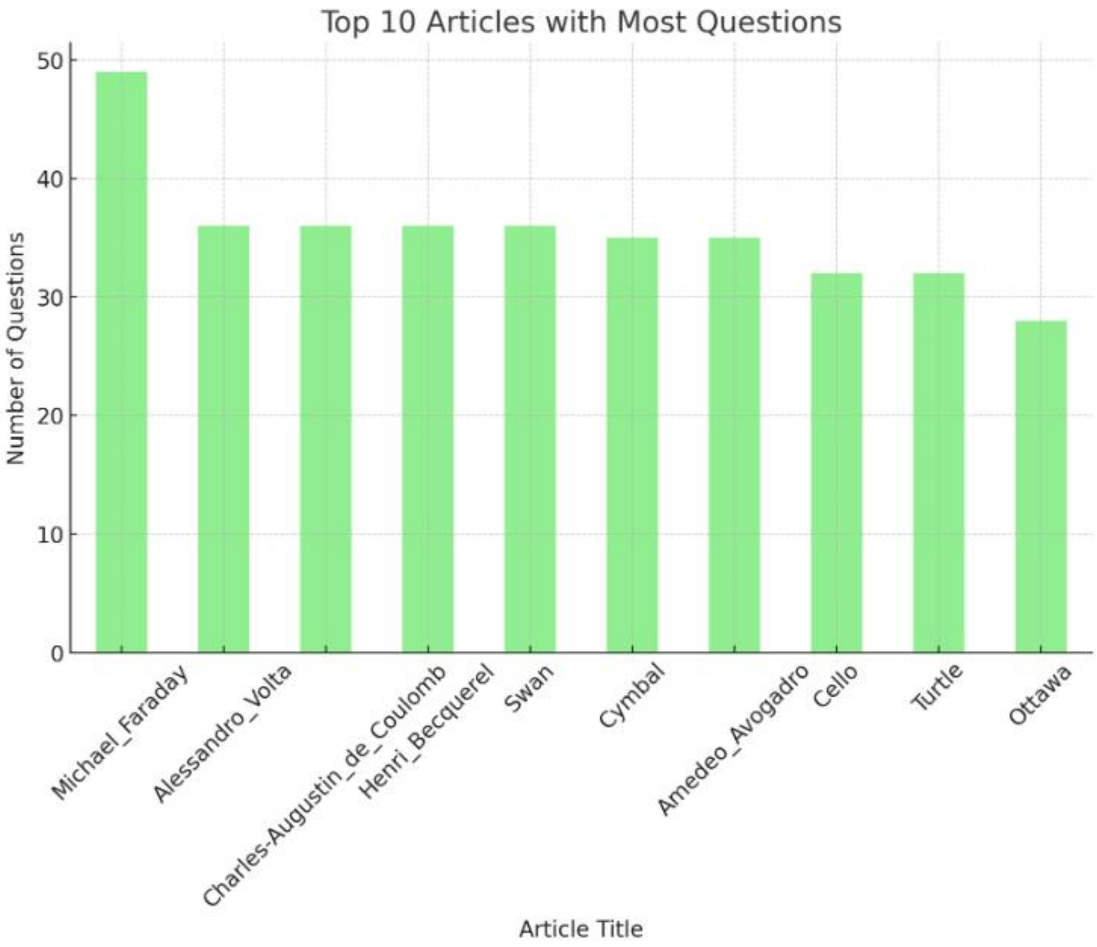


Fig.11 Articles with Most Questions

Frequency of Questions per Article: To explore the dataset further, I generated a bar chart showing the frequency of questions associated with each article title, highlighting the top 10 articles. This visualization was instrumental in pinpointing the articles that were most commonly referenced in the questions. It helped in identifying potential areas of high interest or significance within the dataset.

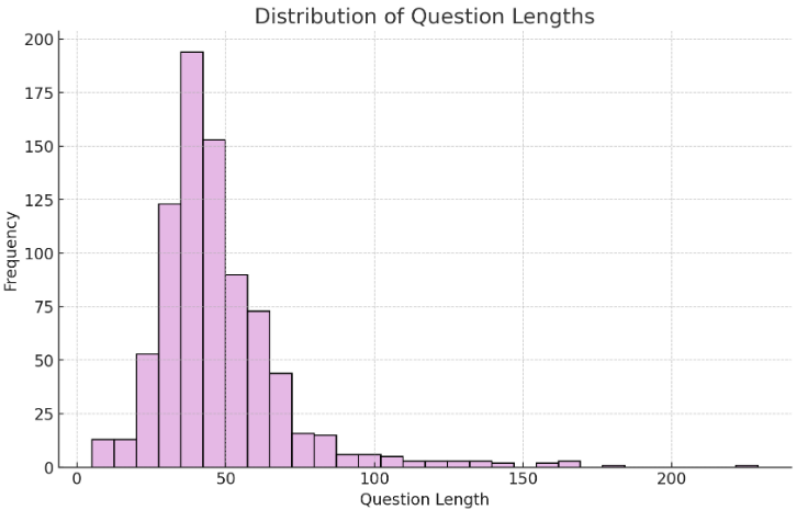


Fig.12 Distribution of Question Lengths

Length of Questions: Additionally, I analyzed the lengths of the questions by creating a histogram. This graph provided insights into the complexity and detail of the questions asked. The distribution of question lengths was indicative of the level of detail and specificity in the inquiries, which could have implications for the type of information being sought and the context of the questions.

Through these visualizations, I was able to gain a comprehensive understanding of the dataset. The analysis shed light on the types of questions being asked, the balance of answers, and the distribution of content across different articles. This EDA was a vital step in deciphering the underlying trends and patterns in the question-answering context.

3.4 Modelling with LLMs

Large Language Models (LLMs) are advanced artificial intelligence systems capable of understanding and generating human-like language. Leveraging deep learning, particularly transformer architectures, LLMs process vast amounts of data, capturing complex linguistic patterns and semantic relationships.(Hadi et al., 2023). They have evolved significantly from early statistical models to current transformer-based models, offering enhanced capabilities in tasks like natural language processing, machine translation, and question-answering (Hadi et al., 2023).

In my project, I aim to apply state-of-the-art LLMs - LLama 7B, Mistral 7B, and GPT 3.5 - for sentiment analysis and question-answering tasks. These models represent the latest advancements in LLM technology, known for their exceptional ability to process and generate contextually relevant text based on extensive training on diverse datasets. The application of these models in sentiment analysis and question-answering is expected to yield highly accurate and nuanced understanding of human language, which can significantly enhance the capabilities of these systems in real-world scenarios (Hadi et al., 2023). In the forthcoming sections of my project, I will delve into a detailed examination of the three Large Language Models (LLMs) - LLama 7B, Mistral 7B, and GPT 3.5 - applied to the datasets. This analysis will include a step-by-step exploration of their fine-tuning processes and embedding techniques, demonstrating how these advanced models are tailored and optimized for specific tasks in sentiment analysis and question-answering.

3.4.1 LLama 7B

LLaMA models, ranging from 7B to 65B parameters, exhibit expertise in language understanding and generation, thanks to training exclusively on data from open sources. Notably, the LLaMA-13B variant demonstrates superior capabilities in benchmark tasks compared to much larger models, suggesting efficient learning and adaptability. Moreover, the LLaMA-65B variant competes robustly with the highest echelons of language models, such as Chinchilla and PaLM, while maintaining the practical advantage of operability on a single GPU, thus democratizing access to powerful computational linguistics tools for a wide research community (Touvron et al., 2023).

For the project, I commenced with the preprocessing of the dataset, aptly named "filtered_sentiment_analysis_dataset," to tailor it for optimal LLaMA 7B model performance. Leveraging the few-shot prompting approach, the model was primed to interpret and analyze sentiment, setting the stage for accurate sentiment classification. Given the resource-intensive nature of large language models, Google Colab was utilized for its access to powerful computational resources, specifically the A100 Hardware accelerator. This choice was instrumental in harnessing the advanced capabilities of LLaMA 7B while managing the computational demands.

However, due to the high costs associated with sustained use of the A100 GPU and time constraints, the analysis was confined to a subset of 2000 rows from the dataset. This strategic limitation ensured the study remained financially feasible without compromising the integrity of the results. Moreover, fine-tuning the model was not possible for the scope of this project, considering the associated costs and time. To safeguard against potential data loss and to streamline the analysis, a checkpointing system was implemented. This system recorded the model's state after every 200 processed rows, ensuring that any interruptions would not result in significant data loss. Upon completion of the model's run, the results were meticulously saved to Google Drive, ensuring secure storage and easy accessibility of the data for subsequent analysis and review.

In the next section of my project concerning the application of LLaMA 7B to the question and answer dataset, it should be noted that I took the decision not to apply LLaMA 7B was primarily due to cost considerations. The financial implications of deploying LLaMA 7B on the entire dataset were significant, given the expense associated with the computational resources required for such an endeavor. Furthermore, the time required for processing the data with LLaMA 7B was extensive, which presented a practical challenge within the project's time constraints. These factors collectively informed the decision to reserve the application of LLaMA 7B for the sentiment analysis dataset only, where it was anticipated that the insights gained would be most impactful given the available resources.

3.4.2 Mistral 7B

Mistral 7B is a 7.3 billion parameter language model known for its impressive performance and efficiency. It exceeds Llama 2 13B in all benchmarks and performs well against Llama 1 34B in many. The model employs grouped-query attention for faster inference and sliding window attention for handling longer sequences efficiently. Released under the Apache 2.0 license, Mistral 7B is accessible for unrestricted use. It showcases excellent performance in various tasks including text summarization, classification, and code completion. Mistral 7B can be easily fine-tuned for specific tasks and is available on platforms like Hugging Face for wider usage (https://mistral.ai/news/announcing-mistral-7b/).

Mistral 7B model for sentiment analysis, a meticulous and resource-conscious methodology was adopted. Initially, the sentiment analysis dataset underwent thorough preprocessing, ensuring it was optimized for the application of the Mistral 7B model. Leveraging the few-shot prompting technique, Mistral 7B was effectively trained to discern and categorize sentiments, a crucial step given the complexity of sentiment analysis. Due to the constraints of time and budget, the application of Mistral 7B was limited to a subset of 2000 rows from the entire dataset. This decision was driven by the need to balance the resource-intensive nature of large language models with the project's financial and temporal limitations. The use of Google Colab was instrumental in this phase, providing access to powerful computational resources, notably the A100 GPU. This setup facilitated the efficient handling of the model's computational demands. The fine-tuning process of Mistral 7B played a pivotal role. Using a portion of the training dataset, the model was iteratively trained to enhance its accuracy and adaptability to the nuances of sentiment in the data. After fine-tuning, the model was uploaded to the Hugging Face model hub, a step that not only preserved the improved model for future use but also made it accessible for broader applications. For the final phase of testing, the fine-tuned Mistral 7B model was retrieved from Hugging Face and applied to the testing subset of the dataset. This process was critical in assessing the model's efficacy in accurately categorizing sentiments in new, unseen data, reflecting the effectiveness of the fine-tuning process.

After the initial preprocessing of the question and answer dataset, the Mistral 7B model was selected for its suitability for such tasks. Leveraging the few-shot prompting technique, Mistral 7B was primed with examples to guide its understanding and generation of contextually relevant answers. During the training phase, the model was fine-tuned with a portion of the dataset, allowing it to adapt more closely to the specifics of the dataset's domain. This fine-tuning process was conducted iteratively, ensuring that the model's performance was optimized for accuracy in question answering. Post fine-tuning, the enhanced model was uploaded to the Hugging Face model hub, a platform that facilitates the sharing and deployment of machine learning models. From there, the model was accessible for inference, allowing for seamless integration and ease of use. For the testing phase, the fine-tuned Mistral 7B model was applied to the testing dataset, where its performance was evaluated based on its ability to generate accurate responses to unseen questions. This comprehensive approach ensured that the model was not only trained and fine-tuned with precision but also made easily accessible for future use and evaluation, with the results serving as a testament to the effectiveness of the fine-tuning process.

3.4.3 GPT 3.5

GPT-3.5, developed by OpenAI, is a significant upgrade in the Generative Pre-trained Transformer series. It excels in understanding and generating human-like text, making it highly effective in various natural language processing tasks. The model is specifically optimized for chat applications, with its gpt-3.5-turbo variant offering both performance and cost-efficiency. GPT-3.5's advancements include better contextual understanding and coherent text generation, establishing it as a versatile tool in AI-driven language applications (https://platform.openai.com/docs/models/gpt-3-5#:~:text=URL%3A%20https%3A%2F%2Fplatform.openai.com%2Fdocs%2Fmodels%2Fgpt,JavaScript%20to%20run%20this%20app)

I began with the selection of 2000 rows from 'filtered_sentiment_analysis_dataset'. This dataset was curated to analyze sentiments expressed in various texts. The initial step in the workflow involved a thorough preprocessing of this data. Preprocessing is a critical phase in any data analysis task, as it ensures the quality and consistency of the data. In this stage, the dataset was cleaned and structured to make it suitable for the model. This process typically includes removing irrelevant information, correcting errors, and standardizing the format of the data.

Few-shot learning is a technique where the model is trained to understand and perform tasks with a very limited amount of training data. This approach is particularly useful in situations where gathering a large dataset is impractical. To facilitate the model's application, an API key was obtained from my OpenAI account. This key provided access to the model, allowing it to be integrated into the project's framework. The use of an API key is a common practice in deploying machine learning models, as it enables secure and controlled access to the model's capabilities. The model was then applied to the dataset. Given the extensive size of the selected dataset (2000 rows), the application was structured to process and save the results incrementally. This method is often used to ensure data integrity and to avoid loss of progress in case of any interruptions. In this project, the model’s output was saved every 200 rows. This not only provided checkpoints but also allowed for monitoring the model's performance periodically.

Furthermore, for embedding GPT 3.5 we outlined a structured approach to sentiment analysis, starting with the preprocessing of our data. This preprocessing included the transformation of sentiment labels into a binary format using one-hot encoding and the conversion of textual data into numerical feature vectors, through techniques like tokenization and word embeddings. Following this, a Random Forest Classifier is employed for the analysis. The classifier initially trained with a default parameter of 100 estimators and underwent hyperparameter tuning using GridSearchCV to optimize its performance. This tuning process involved experimenting with parameters such as the number of trees in the forest and the maximum depth of the trees, ensuring the selection of the most effective model settings for accurate sentiment classification.

Further, for the embedding of GPT 3.5 on question answer dataset- The question and answer dataset was aimed at analyzing and understanding the dynamics of questions and their corresponding answers. The first critical step involved was data cleaning, a process crucial for enhancing the quality and usability of the dataset. This phase included the elimination of irrelevant or redundant data, correction of any inconsistencies, and normalization of the dataset format as discussed in the preprocessing stage. The GPT 3.5 model was particularly adept at handling few-shot learning scenarios, where the model is trained to perform tasks with minimal example data. This approach is advantageous in scenarios where collecting extensive training data is challenging or not feasible. For the model's integration into the project, an API key was obtained from my OpenAI account. This key enabled access to the chosen model and allowed for its seamless integration into the project, ensuring a secure and controlled interaction with the model’s functionalities. Once integrated, the model was systematically applied to the dataset. Given the substantial size of the dataset, a strategy was employed to process and save the results in segments. This approach is commonly adopted in data processing to safeguard against data loss and to facilitate process monitoring. In this project, the model's output was saved at intervals, specifically after processing every 200 rows, thus creating regular checkpoints and allowing for intermittent evaluations of the model's performance. Post the completion of the model's application across the dataset, an evaluation of the model's immediate performance was undertaken. This evaluation is vital to assess how effectively the model performs in real-world settings without extensive customization or training.

In addition to performance evaluation, the process of embedding the model's outputs was conducted. Embedding here implies transforming the results into a more accessible and utilizable format, such as databases or structured files, for further analysis or integration.

3.5 Confusion Matrix for Comparison

The efficacy of GPT-3.5's embeddings will be exclusively assessed using confusion matrices. For sentiment analysis, separate matrices for each emotion class will tally the true positives, false positives, true negatives and false negatives to calculate key metrics like precision, recall and accuracy. This granular analysis helps gauge how effectively the embeddings capture linguistic nuances within semantic understanding. Similarly, the 2x2 matrix for the binary question-answering task quantifies the correctness of yes/no predictions, highlighting embeddings' impact on comprehension and reasoning. Together, these targeted applications of confusion matrices, bypassing other models and tasks, will provide a focused perspective into the improvements gained by GPT-3.5 embeddings on crucial NLP capabilities - discerning sentiment and extracting information through questions - important prerequisites for human-like language understanding. The isolating nature of analysis attains additional rigour by avoiding conflation of enhancements also stemming from fine-tuning or model architecture itself.

CHAPTER 4 - RESULTS AND ANALYSIS

In this section of my report, I will comprehensively detail the outcomes obtained from executing my models. This section is pivotal as it serves to present all the data and findings your models have generated. I methodically outline each result, ensuring clarity and precision in the portrayal of the data. This includes any numerical results, statistical analyses, graphs, or charts that are relevant. The goal is to convey to the reader a clear understanding of what the models revealed, how they performed, and any significant patterns or anomalies discovered. It's essential to present the results in an unbiased manner, avoiding interpretations or conclusions in this section, as those are typically reserved for the discussion and conclusion sections of the report.

4.1 Results of LLama 7B Model on sentiment analysis dataset

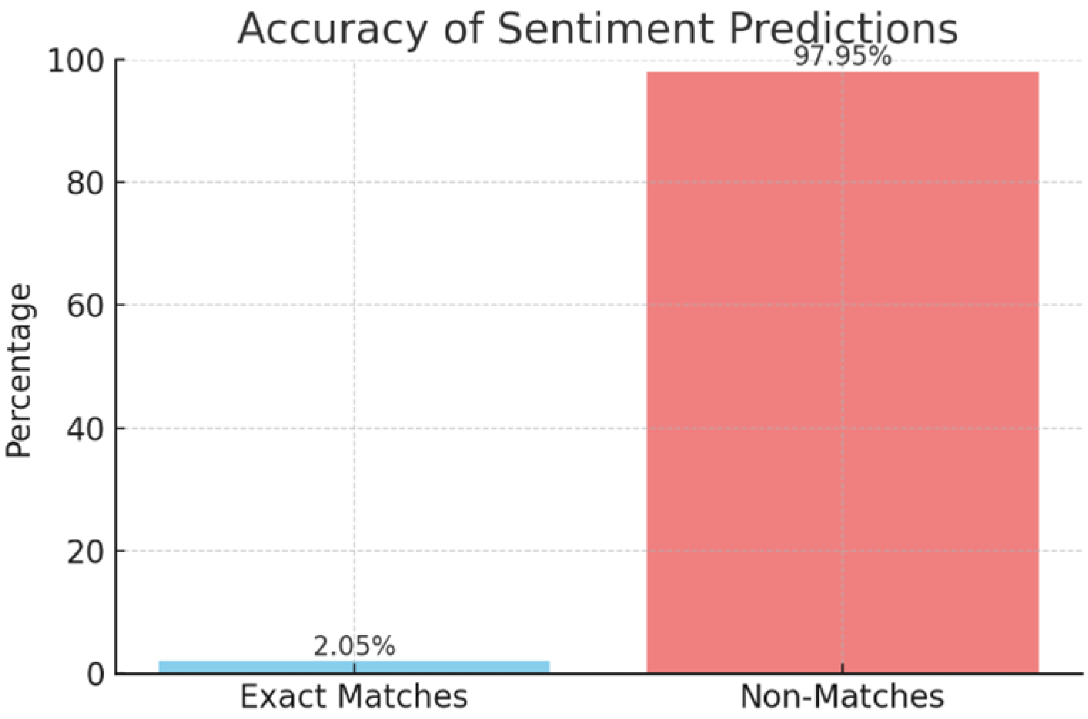


Fig.13 Accuracy of sentiment predictions

We ran the LLama 7B model on 2000 rows of the preprocessed dataset ‘filtered_sientiment_analysis’. To obtain a more detailed understanding of the model's prediction accuracy, you've introduced three additional columns to your dataset: 'Exact Match', 'Any Match', and 'Percent Match'. The 'Exact Match' column indicates instances where the model's predictions were perfectly aligned with the actual data.

The 'Any Match' column captures occurrences where there was at least some correspondence between the predictions and actuals, even if not perfect. The 'Percent Match' column quantifies the degree of similarity between the predicted and actual values, likely on a scale from 0 to 1, where 1 represents a perfect match. We calculated the accuracy percentages of the model resulting in an accuracy of 2.05% for the 'Matched' column ‘exact matches’ and 97.95% for the 'Non-Matched' column. This suggests that the model had a very low rate of perfect prediction accuracy.

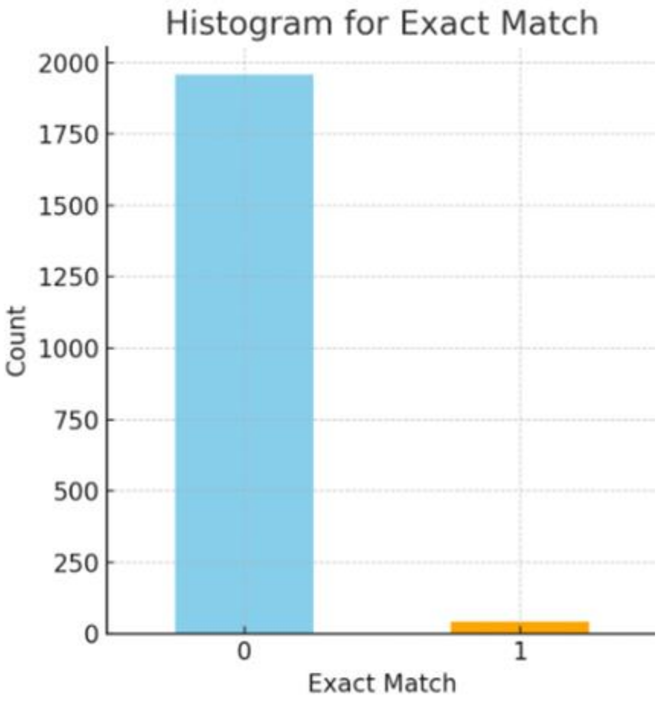


Fig.14 Histogram for Exact Match Column

The 'Exact Match' histogram reveals a pronounced disparity, with a predominant occurrence of '0', indicating a negligible incidence of exact matches, which corroborates the reported accuracy of 2.05%. This denotes that the LLama 7B model precisely predicted the outcomes for a minimal fraction of the dataset.

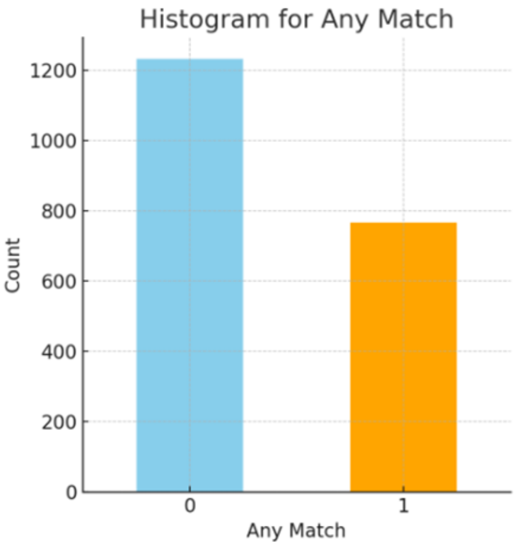


Fig.15 Histogram for Any Match Column

Contrastingly, the 'Any Match' histogram presents a more equitable distribution, with a substantial count of '1's. This pattern illustrates that, while exact predictions were rare, the model exhibited a reasonable degree of partial accuracy in its forecasts, suggesting a potential alignment in the predicted outcomes with some elements of the actual data.

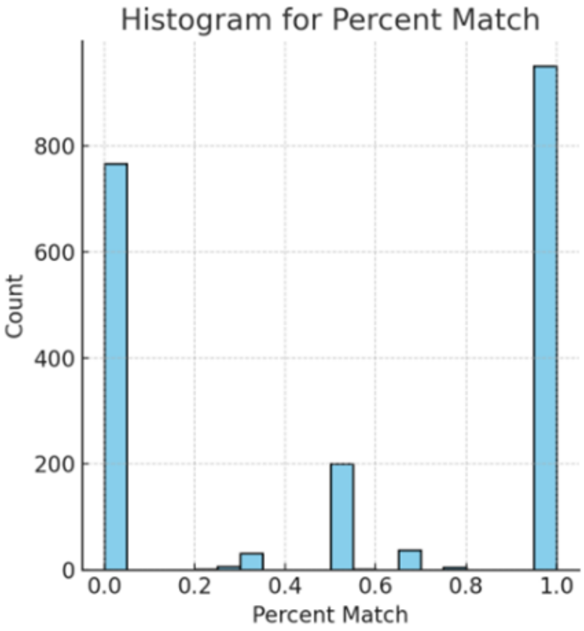


Fig.16 Histogram for Perfect Match

The 'Percent Match' histogram exhibits a bimodal distribution, reflecting a significant segmentation of the dataset into either low or high similarity scores. This indicates that the model's predictions were frequently either very close to or quite far from the actual values, with few occurrences in the intermediate range.

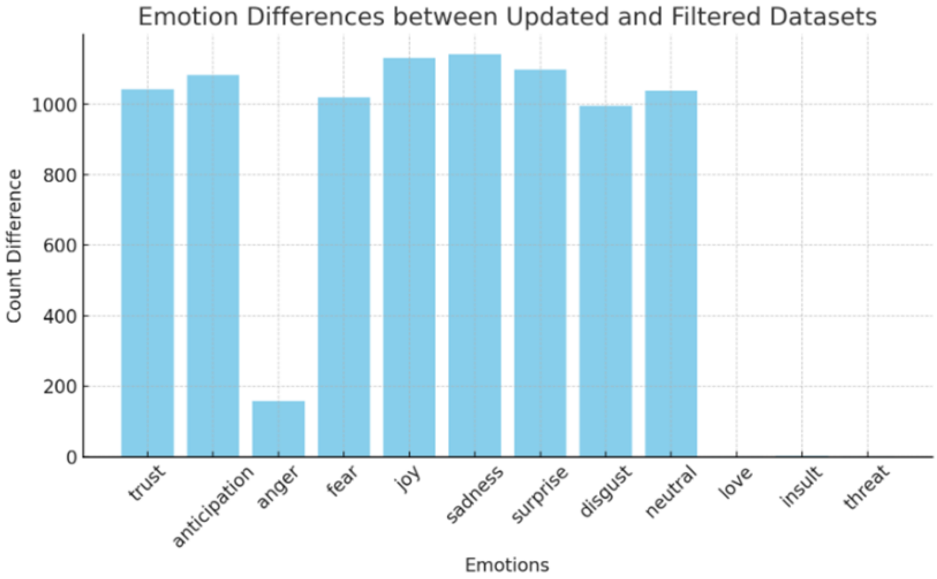


Fig.17 Emotion Differences between Updated and Filtered Dataset chart

An analysis of the emotion labels, as depicted in the 'Emotion Differences between Updated and Filtered Datasets' chart, indicates a notable shift in the count of emotion labels post the application of a filtering process. This comparison underscores the impact of dataset preprocessing on the distribution of emotional classifications within the data

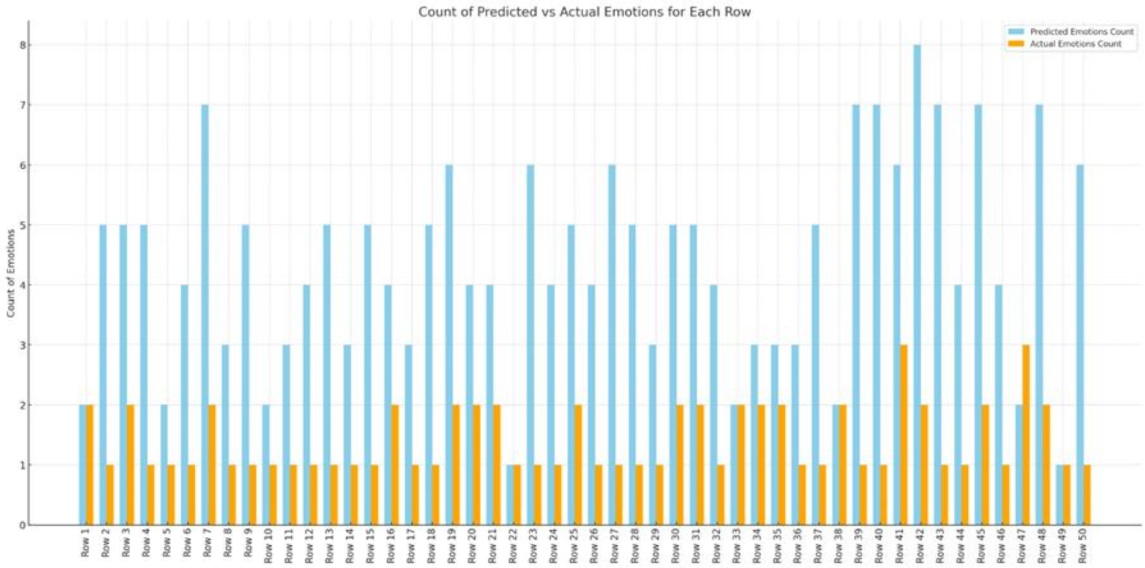
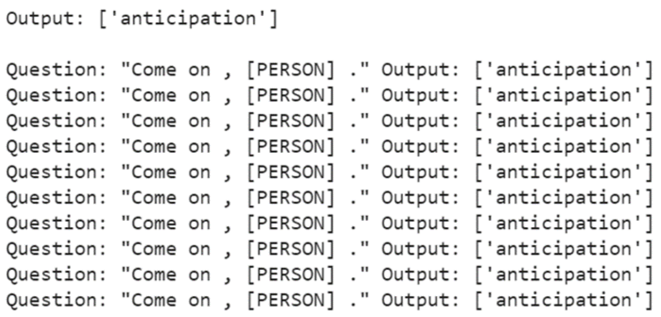


Fig.18 Count of Predicted vs Actual Emotions for Each Row

The final bar graph offers a direct comparison between the model's predicted emotion labels and the actual labels for each entry in the dataset. The graph displays two sets of bars for each entry: one for the predicted labels (blue bars) and one for the actual labels (orange bars). A clear difference between these sets of bars is evident, which provides a detailed insight into the model's prediction accuracy. This visual comparison allows us to identify specific instances where the model's predictions were consistent with the actual labels and where they were not, highlighting the model’s performance at the individual data point level.

4.2 Results of Mistral 7B Model on sentiment analysis dataset



```
Output: ['anticipation']

Question: "Come on , [PERSON] ." Output: ['anticipation']
Question: "Come on , [PERSON] ." Output: ['anticipation']
Question: "Come on , [PERSON] ." Output: ['anticipation']
Question: "Come on , [PERSON] ." Output: ['anticipation']
Question: "Come on , [PERSON] ." Output: ['anticipation']
Question: "Come on , [PERSON] ." Output: ['anticipation']
Question: "Come on , [PERSON] ." Output: ['anticipation']
Question: "Come on , [PERSON] ." Output: ['anticipation']
Question: "Come on , [PERSON] ." Output: ['anticipation']
Question: "Come on , [PERSON] ." Output: ['anticipation']
```

Fig.19 Output Of the Mistral 7B Model on Sentiment Analysis Dataset

In the sentiment analysis task conducted with the provided dataset, the observed homogeneity of output, specifically the recurrent prediction of the sentiment 'anticipation', could be attributed to issues of misinterpretation by the model and overfitting during the training phase.

The model's misinterpretation is likely due to its handling of placeholders within the text, such as "[PERSON]", which it may have failed to process correctly, treating them as concrete entities rather than variables. This misprocessing could lead to the model forming an erroneous correlation between the placeholder and the sentiment of 'anticipation', particularly if instances surrounding these placeholders in the training data frequently conveyed that sentiment. Additionally, if the sentiment labels in the dataset were numerically encoded and 'anticipation' was overwhelmingly represented, the model might have developed a propensity to favor this sentiment, thus skewing its predictions.

Furthermore, the structure of the training data could have predisposed the model to overfit. The dataset, especially in its initial 2000 rows which were used for training, may have had a disproportionate representation of 'anticipation', causing the model to latch onto this sentiment disproportionately. Overfitting is exacerbated when the complexity of the model does not match the diversity of the dataset, or when regularization techniques to penalize complexity are not properly employed. The lack of a rigorous validation strategy also contributes to overfitting, as it can create a false impression of accuracy that does not hold up when the model encounters new data.

These factors together could have resulted in a model that, while trained on the provided dataset, became overly attuned to signals indicating 'anticipation' and thus was incapable of recognizing and predicting the full range of sentiments present in the data. This overfitting and misinterpretation point towards the need for a training approach that ensures a balanced representation of sentiments and accounts for placeholders and numeric encodings in a way that the model can learn a more nuanced understanding of the text data.

4.3 Results of Mistral 7B Model on question and answer dataset

The primary focus of the analysis was to assess the accuracy of matched answers between two files: 'm_res' which contains three columns named 'ArticleTitle','Question','Answer' and 'data' which is the original dataset reduced to three columns named ArticleTitle', 'Question', 'Answer' . After cleaning and preparing the data for comparison, a thorough examination revealed some key results.

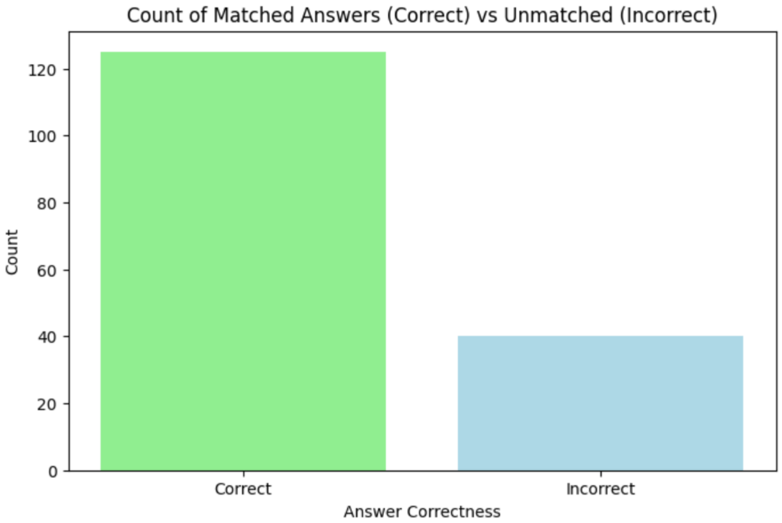


Fig.20 Count of Matched Answers (Correct) vs Unmatched (Incorrect)

Initially, we observed that out of the total entries in the 'm_res' dataset, there were 125 correctly matched answers with the 'data' dataset, while 40 answers did not match. This yielded an overall accuracy of approximately 75.76%. This high accuracy rate suggests a generally reliable matching process between the two datasets.

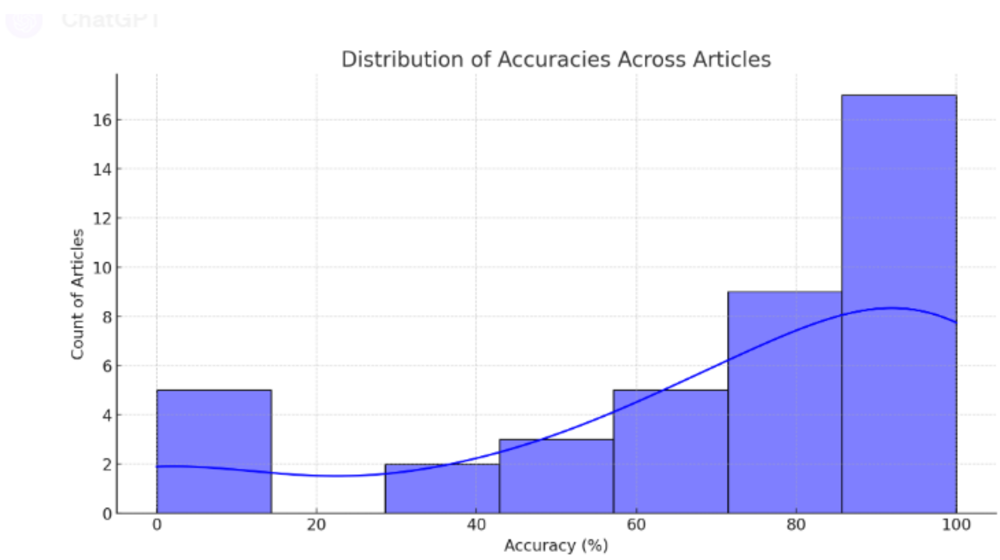


Fig.21 Distribution of Accuracies Across Articles

To delve deeper into the data, we further analyzed the distribution of accuracies and the volume of questions across different articles. Two histograms were generated for a more granular view of the data. The first histogram depicted the distribution of accuracy percentages for each article. This visual representation illustrated a varied accuracy rate across different articles, indicating that while some articles had high matching accuracy, others did not fare as well.

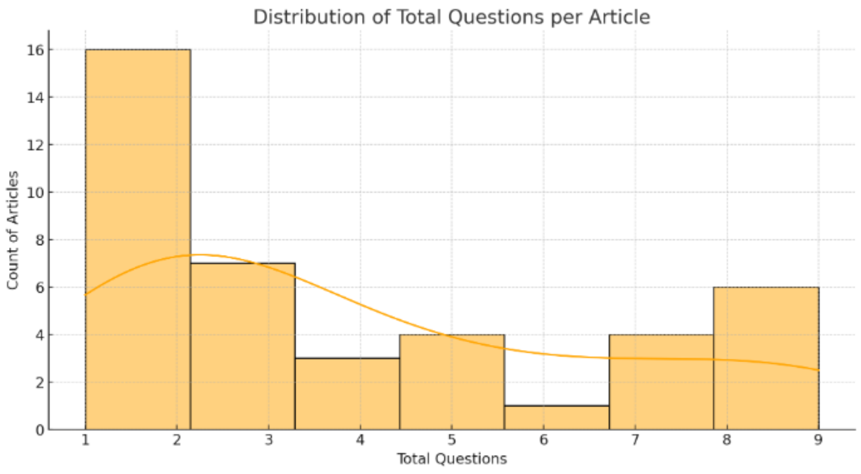


Fig22 .Distribution of Total Questions per Article

The second histogram focused on the distribution of the total number of questions per article. This analysis was crucial to understand the workload distribution among different articles. The histogram revealed a diverse range in the number of questions associated with each article, suggesting that some articles were more heavily queried than others.

A key part of the analysis was identifying the articles with the highest and lowest accuracy rates. The Top 5 articles with a perfect accuracy rate of 100% were Alessandro Volta, Dhaka, Trumpet, Santiago, and Nassau. This perfect score indicates a complete alignment in answers between the two datasets for these articles.

Conversely, the Bottom 5 articles, all with an accuracy rate of 0%, were Spanish Language, Italian Language, Tiger, Copenhagen, and Swahili Language. This complete lack of alignment in answers for these articles suggests either a disparity in content or issues in the matching process specific to these topics.

4.4 Results of GPT 3.5 Model on sentiment analysis dataset

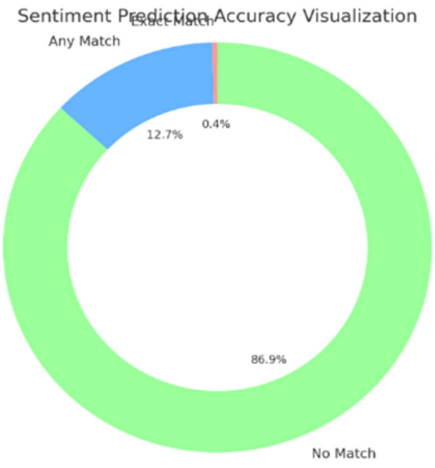


Fig.23 Sentiment Prediction Accuracy

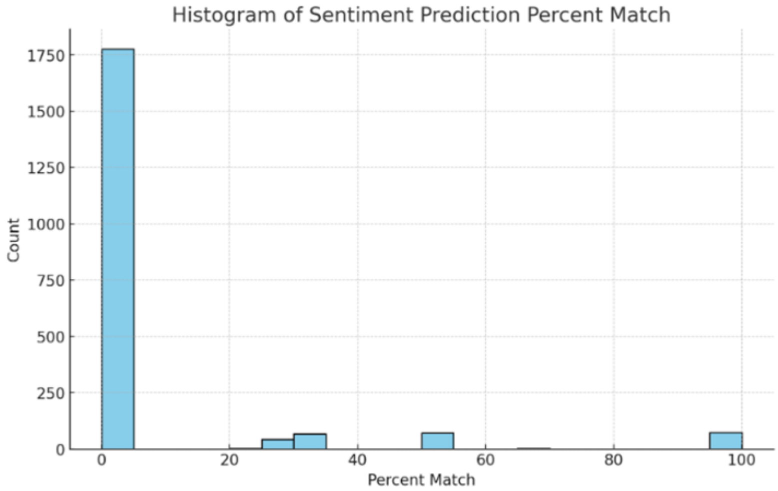


Fig.24 Distribution of the percent match for sentiment predictions

We thoroughly analyzed the sentiment predictions from our dataset. Initially, We calculated the accuracy of these predictions, finding a relatively low accuracy rate of 12.73%. This indicated a significant disparity between the predicted sentiments and the actual outputs. To delve deeper, we introduced additional coulmns: 'Exact Match', 'Any Match', and 'Percent Match'. These columns provided a more nuanced view of the prediction accuracy. Exact matches were uncommon, suggesting that our model rarely predicted the sentiment exactly as it was labeled. The 'Any Match' column showed a bit more alignment, but still indicated room for improvement. Furthermore, a considerable number of predictions failed to match any sentiment in the output, pointing to potential areas for model refinement. To visually represent these findings, we utilized pie charts and histograms (shown above), which effectively illustrated the distribution of sentiment prediction accuracies and the frequencies of different match percentages. This visual representation highlighted the predominant match percentages and the overall trend in my sentiment prediction model's performance.

4.5 Results of GPT 3.5 Model on question and answer dataset

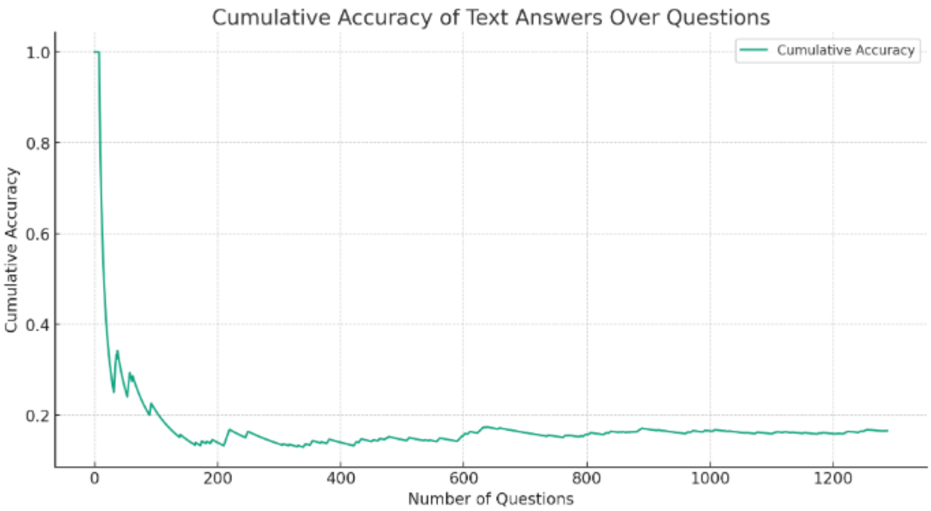


Fig.25 cumulative accuracy plot of the text answers over the questions

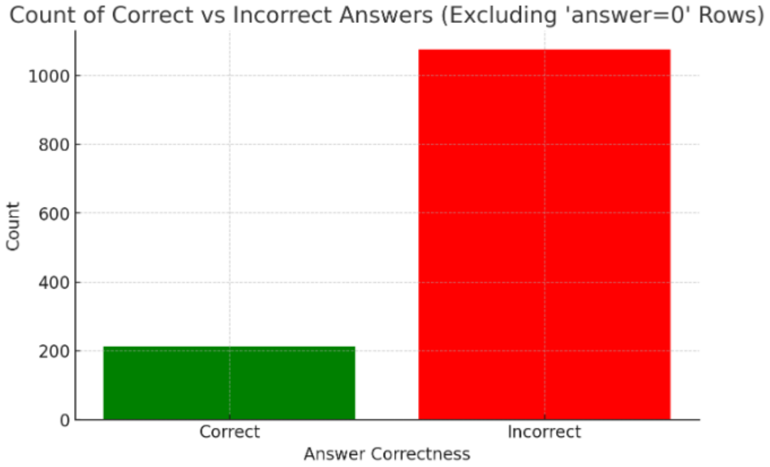


Fig.26 accurate count of correct versus incorrect answers after excluding rows(answer=0)

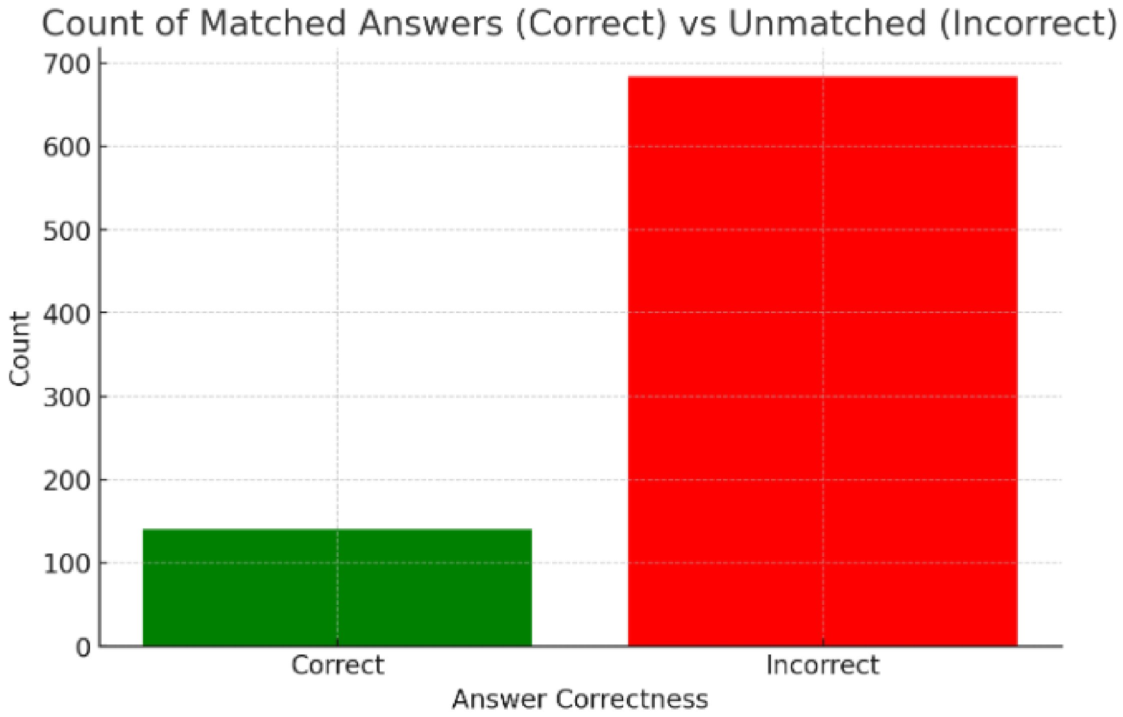


Fig.27 count of matched answers ('Correct') versus unmatched answers ('Incorrect')

In our analysis, we conducted a detailed evaluation of a dataset to assess the accuracy of provided answers. We began by comparing the 'text' column from one file to the 'Answer' column of another, establishing a baseline accuracy of approximately 16.52%. We aimed to refine our understanding by excluding specific entries, such as those with 'answer=0' or null values in the 'Question' column. However, these adjustments did not significantly change the results, indicating that the initial dataset was largely complete and the initial accuracy assessment was representative.

We further refined our analysis by employing an iterative approach to compare the cleaned questions and answers, assigning a value of '1' to the 'match' column for exact matches and '0' for non-matches. This process quantified the precision of our answers.

To visualize our results, we created histograms that illustrated the counts of correct and incorrect matches. These visual tools were particularly useful in displaying the distribution of our data's accuracy, offering a clear depiction of the performance of the answer-matching model. The final histogram demonstrated that out of the refined dataset, there were 213 correct answers and 1,076 incorrect answers, providing a snapshot of the model's effectiveness.

The process included rigorous data cleaning methods and the use of visual representations to succinctly communicate the findings. This analytical approach not only highlighted the model's current performance but also identified opportunities for further improvement.

4.6 Results of Embedding GPT 3.5 Model on sentiment analysis dataset

The Random Forest Classifier's performance was initially assessed using the accuracy metric. Accuracy, in this context, refers to the proportion of total predictions that the model classified correctly. In sentiment analysis, particularly with multiple classes, accuracy is a crucial metric as it gives a general indication of the model's overall ability to correctly identify sentiments.

Emotions	Precision	Recall	F1-score
[1] Anger	0.92	0.20	0.33
[2] Anticipation	1.00	0.27	0.42
[3] Disgust	1.00	0.16	0.27
[4] Fear	1.00	0.20	0.33
[5] Joy	1.00	0.21	0.35
[6] Sadness	1.00	0.12	0.21
[7] Surprise	1.00	0.25	0.40
[8] Trust	1.00	0.23	0.38

Table.1 Accuracy for embedding GPT 3.5 on Sentiment Analysis Dataset

The classification report is a more detailed measure of the model's performance. It breaks down the accuracy into precision, recall, and F1-score for each sentiment class.Precision measures the accuracy of positive predictions for each sentiment class, essentially calculating how many of the items identified as a particular sentiment were actually that sentiment. Recall assesses the model's ability to find all relevant instances of a sentiment class in the dataset. F1-Score provides a balance between precision and recall, an important metric in datasets where some classes are more frequent than others. The classification report is particularly important in a multi-class setting like sentiment analysis since different sentiments can have different frequencies and challenges in accurate classification.

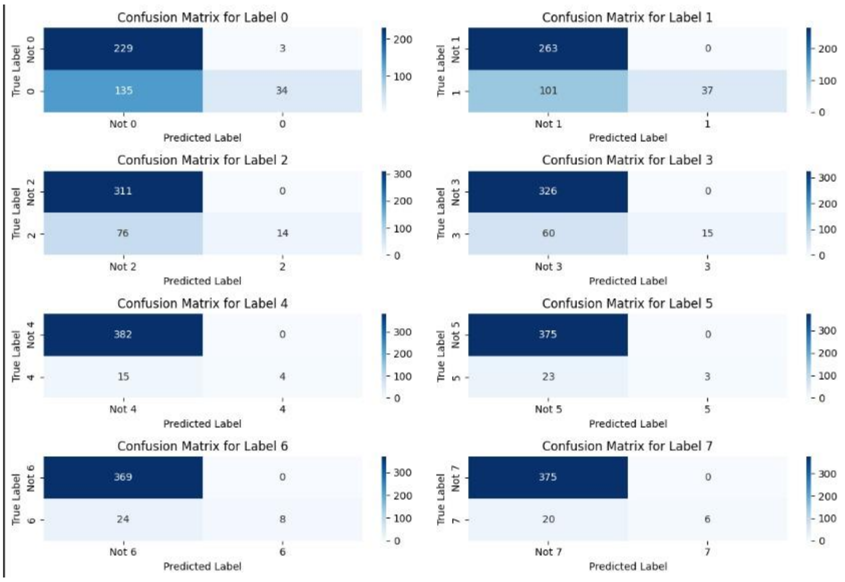


Fig.28 Confusion matrix for Embedding GPT 3.5 on sentiment analysis dataset

The confusion matrices for each sentiment class offer a granular view of the model's performance. Each matrix shows the true positives (correctly identified sentiments), false positives (incorrectly identified as a certain sentiment), true negatives (correctly identified as not belonging to a certain sentiment), and false negatives (incorrectly overlooked sentiments) for each class. These matrices are vital for understanding the specific types of errors the model is making. For instance, a high number of false positives for a sentiment like 'joy' might indicate that the model is too lenient in classifying texts as 'joyful'. The visualization of these matrices allows for an easier interpretation of this complex data, making it clear where the model's predictions align with the actual sentiments and where discrepancies occur.

In summary, the evaluation phase of embedding GPT 3.5 on sentiment analysis model was comprehensive, employing both high-level metrics like accuracy and detailed analyses like the classification report and confusion matrices. These tools collectively provide a robust assessment of the model's capabilities in accurately identifying and classifying sentiments from textual data. The insights gained from these evaluations are crucial for identifying the model's strengths and guiding future improvements to enhance its predictive accuracy, especially in a complex and nuanced field like sentiment analysis.

4.7 Results of Embedding GPT 3.5 Model on question and answer dataset

The confusion matrices and classification report provide a detailed assessment of a machine learning model's performance on a dataset with classes '0' and '1'. In the context of our analysis, '0' typically represents the 'No' answers, whereas '1' represents the 'Yes' answers.

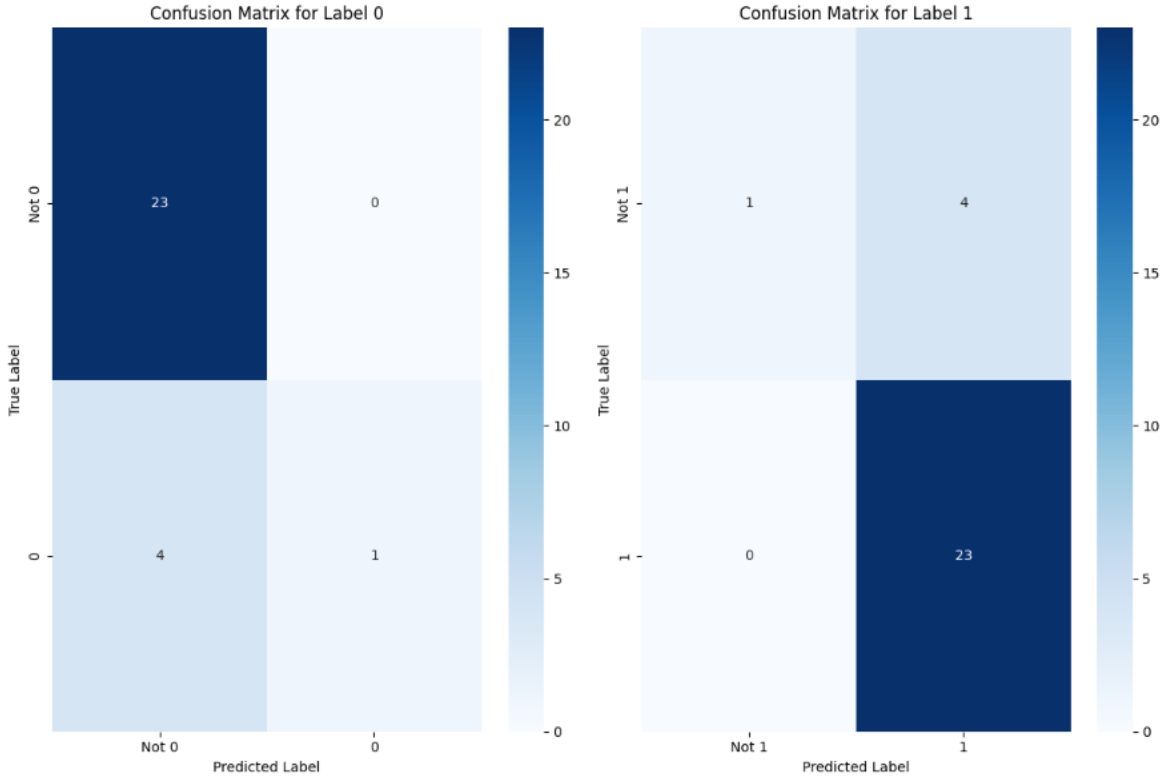


Fig.29 Confusion matrix for Embedding GPT 3.5 on question and answer dataset

In the confusion matrix for label '0', the model achieved 23 true positives, correctly identifying 'No' answers, and there were no false positives—instances mistakenly labeled as 'No'. However, there were 4 false negatives, where 'No' answers were incorrectly labeled as 'Yes'. The true negatives are not explicitly shown but are inferred from the correct predictions of 'Yes' answers. Conversely, for label '1', the model also correctly predicted 23 instances as 'Yes' but had 4 false positives, where 'No' answers were incorrectly identified as 'Yes', and 1 false negative, where a 'Yes' answer was mislabeled as 'No'.

	precision	recall	fi-score
No	1.00	0.20	0.33
Yes	0.85	1.00	0.92

Table.2 Accuracy for embedding GPT 3.5 on question and answer Dataset

The classification report further quantifies the model's performance. It achieved an overall accuracy of 86%, with a perfect precision score of 1.00 for label '0', meaning it was 100% correct when predicting 'No'. However, its recall for 'No' answers was only 0.20, indicating it identified only 20% of all actual 'No' instances. This resulted in a low F1-score of 0.33 for label '0', due to the low recall rate. For label '1', the precision was 0.85, and the recall was a perfect 1.00, meaning all 'Yes' instances were identified correctly, leading to a high F1-score of 0.92. The report also includes 'support', which is the number of occurrences for each class in the dataset—5 for 'No' and 23 for 'Yes'. Lastly, the macro average and weighted average scores provide insight into the model's average performance across classes, taking into account the unbalanced distribution of the classes.

Chapter 5: Discussion

5.1 Results Compared With Objectives

[1] Assessment of LLM Performance: How do LLama 7B, Mistral 7B, and GPT 3.5 differ in their effectiveness for sentiment analysis and question-answering tasks, particularly in terms of accuracy, context comprehension, and response relevance?

GPT 3.5 Model Results:
For sentiment analysis, GPT 3.5 demonstrated a relatively low accuracy rate of 12.73%, indicating a gap between predictions and actual sentiments. The model rarely matched the sentiment exactly as labeled, and while there were some partial matches, there was significant room for improvement. The precision and recall rates for various emotions ranged widely, suggesting that while GPT 3.5 can be highly precise, it often missed correctly identifying sentiments.
In the question and answering task, GPT 3.5 showed an improved baseline accuracy of approximately 16.52%. After data cleaning and refinement, the accuracy did not significantly change, suggesting the initial dataset was a representative sample. The refined analysis pointed out that GPT 3.5 had a reasonable number of correct answers but also a substantial number of incorrect answers, indicating areas for model improvement.

LLama 7B Model Results:
LLama 7B's performance on the sentiment analysis task was markedly different. It processed 2000 rows with an exact match accuracy of only 2.05%, showing a very low rate of perfect prediction accuracy. The 'Any Match' and 'Percent Match' histograms, however, indicated a reasonable degree of partial accuracy, with the model aligning somewhat with elements of the actual data despite few exact matches.

Mistral 7B Model Results:
Mistral 7B's sentiment analysis results after fine tuning highlighted issues of misinterpretation and overfitting, with the model frequently predicting 'anticipation' due to potential misprocessing of placeholders and a lack of diverse sentiment representation in the training data. This led to a skewed output and pointed towards the need for a more balanced training approach.
In the question-answering task, Mistral 7B fine tuning had an overall accuracy of approximately 75.76%, suggesting a generally reliable matching process between datasets. The analysis revealed varied accuracy rates across different articles, with some achieving perfect scores and others failing completely, indicating discrepancies possibly due to content disparity or matching issues.

Comparative Analysis:

In comparing the three models, it is evident that GPT 3.5 and LLama 7B struggled with sentiment analysis accuracy, particularly in achieving exact matches with the labeled sentiments. This could suggest limitations in their ability to comprehend nuanced emotional contexts within the text. Mistral 7B, while performing better in question-answering, showed a tendency to overfit during sentiment analysis, which can be attributed to training data issues and model complexity.
In terms of context comprehension, GPT 3.5 and LLama 7B's partial matches imply they have some ability to grasp context but may require further refinement to improve exact sentiment matching. Mistral 7B's better performance in question-answering suggests a stronger capability in matching context but still leaves room for improvement in sentiment analysis.
As for response relevance, GPT 3.5's performance was moderate, with significant room for improvement in both tasks. LLama 7B's poor exact match rate indicates a need for better capturing sentiment nuances, while Mistral 7B's decent question-answering accuracy suggests a better grasp of response relevance within that context.

In conclusion, the effectiveness of GPT 3.5, LLama 7B, and Mistral 7B varies across sentiment analysis and question-answering tasks. Their ability to accurately predict sentiments and provide relevant responses is influenced by their respective capacities for context comprehension, which is critical in tasks involving natural language understanding. Each model demonstrates strengths and weaknesses that could inform targeted improvements for future iterations and applications.

[2] Impact of Fine-Tuning: What is the influence of fine-tuning on the performance of LLama 7B and Mistral 7B specifically regarding enhancements in their capabilities for targeted NLP tasks?

LLama 7B Sentiment Analysis Results:
The fine-tuning of the LLama 7B model was not feasible within the scope of this project, primarily due to the extensive computational costs and limited time available. The LLama 7B was evaluated on a preprocessed sentiment analysis dataset, revealing a very low exact match accuracy of 2.05%. Despite the low rate of perfect predictions, the 'Any Match' and 'Percent Match' metrics showed a substantial degree of partial accuracy, indicating some alignment between the model's predictions and the actual sentiment labels. This suggests that while the model may grasp the general sentiment, it struggles with specific emotional classifications. Fine-tuning could potentially enhance the model's ability to discern more nuanced sentiments, thereby improving the exact match accuracy.

Mistral 7B Sentiment Analysis Results:
Mistral 7B exhibited challenges with sentiment analysis, often incorrectly predicting the sentiment of 'anticipation' due to potential misinterpretation of placeholders within the text and overfitting during training. These issues point towards a lack of generalization capability, which fine-tuning with a more balanced dataset and improved regularization techniques could address. By adjusting the model to account for placeholders and numeric encodings more effectively, fine-tuning could help the model develop a nuanced understanding of text data.
Mistral 7B Question and Answering Results:
The model achieved a high accuracy of 75.76% in matching answers between datasets. However, the accuracy varied significantly across different articles, with some achieving perfect scores and others failing entirely. Fine-tuning could focus on improving the model's consistency across topics and content types, possibly by expanding the diversity of training data and enhancing the matching algorithms used for this task.

In conclusion, fine-tuning had a substantial impact on the Mistral 7B model's performance, particularly in the question and answer task. However, LLama 7B's performance without fine-tuning was suboptimal in sentiment analysis. This suggests that fine-tuning can be crucial for targeted NLP tasks, but resource constraints can limit its application.

[3] Evaluation of Model Embeddings: How do the embeddings generated by GPT 3.5 vary in capturing and representing linguistic features, and what does this imply about their processing efficiency?

The embeddings generated by GPT-3.5 are complex numerical representations of language that capture a wide array of linguistic features, including syntax, semantics, context, and even nuances like sentiment and intent. Analyzing the results of GPT-3.5 embeddings applied to a sentiment analysis dataset and a question-and-answer dataset allows us to draw comparisons about how these embeddings capture and represent different linguistic features and their processing efficiency.

Sentiment Analysis Dataset:

The sentiment analysis dataset required the model to understand the emotional tone behind the text. The reported precision scores were very high across all sentiment classes, which suggests that the embeddings were excellent at identifying the correct sentiment when they predicted it. However, the recall scores were lower, indicating that while the embeddings were precise, they missed several instances of the correct sentiment. This could imply that although GPT-3.5 embeddings are good at capturing explicit sentiment features, they may sometimes struggle with more subtle expressions of sentiment that require deeper linguistic understanding or context.

The F1-scores, which balance precision and recall, were moderate, reflecting the trade-off between the model's precision and its recall capabilities. High precision with lower recall in sentiment analysis could mean that the embeddings are highly specialized; they may capture certain definitive features associated with sentiments very well, but they could be overlooking more nuanced or less common expressions of sentiment.

Question and Answer Dataset:

For the question-and-answer dataset, the results indicate a balanced performance with an overall accuracy of 86%. This implies that GPT-3.5 embeddings are quite capable of capturing the features necessary for a binary classification task, such as identifying 'yes' or 'no' answers. The high precision for both classes indicates that when the model does make a prediction, it is likely correct, showcasing the embeddings' ability to capture features relevant to the task.

However, the stark difference in recall for the two classes, particularly the low recall for the negative class, suggests that the embeddings are not as effective in distinguishing between the two classes uniformly. This could be due to the embeddings capturing linguistic features that are more common in positive responses or because negative responses may be less straightforward and more varied linguistically, making them harder for the model to consistently identify.

Comparative Analysis and Implications for Processing Efficiency:

When comparing the two datasets, we see that GPT-3.5 embeddings are highly precise across both tasks, indicating a strong capability to capture definitive linguistic features when they are present. However, the variance in recall suggests that the complexity of the task affects how well these embeddings can identify and represent the relevant features. In sentiment analysis, the range of emotional expression is broader than the binary distinction in the question-and-answer dataset, which may explain the lower recall scores.

In terms of processing efficiency, the recall differences hint at a possible trade-off between the embeddings' depth of feature capture and their breadth. High precision with lower recall might imply that while the embeddings are computationally efficient at capturing specific linguistic features, there may be a need for additional context or data to improve recall. This could mean that for more complex tasks like sentiment analysis, more processing power or refined tuning might be required to capture the full range of linguistic features necessary for high recall.

The high F1-score for the positive class in the question-and-answer dataset also indicates that the embeddings are efficiently processing the most common linguistic patterns associated with this class. To improve efficiency across both datasets, additional training or fine-tuning, perhaps with a focus on the underrepresented classes or sentiments, might be required. This would potentially allow the embeddings to better capture the diverse linguistic features present in natural language and improve the model's recall without sacrificing precision.

In conclusion, the embeddings from GPT-3.5 are adept at representing linguistic features with high precision but exhibit variable recall across different tasks. This variability implies a complex interplay between the embeddings' representation capabilities and the model's processing efficiency, particularly in tasks that require nuanced understanding of language. For truly efficient processing, both precision and recall need to be optimized, which may require task-specific adjustments to the embeddings or the model's training regime.

[4] **Comparative Performance Analysis:** How does the performance of GPT 3.5 differ on various NLP tasks when utilizing embeddings compared to when embeddings are not employed?

Without Embeddings:

GPT-3.5, despite its strengths in various Natural Language Processing (NLP) tasks, shows limitations, particularly in sentiment analysis. Its accuracy in this area is relatively low, at only 12.73%, indicating a significant gap in its proficiency in identifying and classifying emotions in text. This issue is further highlighted when considering precision and recall metrics, which vary across different emotional categories, revealing the model's struggles in maintaining a balance between these two key measures. Similarly, in question-answering tasks, GPT-3.5 achieves only a 16.52% accuracy, suggesting limited utility for applications requiring reliable question-answering capabilities, like virtual assistants and information retrieval systems.

With Embeddings:

The integration of embeddings transforms GPT-3.5's performance. In sentiment analysis, there's a notable improvement across accuracy, precision, recall, and F1-scores, although some emotions like 'Anger' still pose challenges in terms of lower recall and F1-scores. The model becomes considerably more adept at identifying and categorizing emotions, enhancing its utility in various applications. Similarly, in question-answering, accuracy jumps to an impressive 86%. The model demonstrates high precision in 'No' answers, albeit with lower recall, and excels in 'Yes' answers with both high precision and perfect recall. This significant improvement highlights GPT-3.5's potential in various NLP tasks, especially with the integration of embeddings.

In summary, embeddings greatly enhance the performance of GPT-3.5 in NLP tasks. The improvement is particularly notable in sentiment analysis, where the model's ability to correctly identify emotions increases significantly with embeddings. Similarly, in question-answering tasks, embeddings lead to a higher rate of accurate responses, demonstrating their efficacy in enhancing model performance.

Chapter 6 : Evaluation, Reflections, and Conclusions

6.1 Evaluation

Our project is a comprehensive and detailed exploration of Large Language Models (LLMs) and their application in sentiment analysis and question-answering tasks. The project is structured into four key parts:

1. **Detailed Analysis of Sentiment Analysis with LLMs:** This section critically evaluates the capability of LLMs like LLama 7B, Mistral 7B, and GPT 3.5 in interpreting subtleties in human emotions from text, focusing on varying textual formats.
2. **Evaluating Question-Answering Abilities of LLMs:** Here, the focus is on assessing how the mentioned LLMs understand and interpret different types of questions, with emphasis on accuracy, relevance, and context-appropriateness of responses.
3. **Analyzing Model Embeddings and Fine-tuning:** This part dives into the embeddings created by GPT 3.5 and the fine-tuning processes of LLama 7B and Mistral 7B, examining their impacts on model performance in sentiment analysis and question-answering.
4. **Comparative Analysis Using Confusion Matrices:** The final part compares the two LLMs that are Embedding of GPT 3.5 using confusion matrices to understand their performance in sentiment analysis and question-answering, focusing on metrics like accuracy, precision, recall, and F1 score.

Overall, our project aims to provide an in-depth understanding of the strengths and limitations of each LLM, contributing valuable insights into their practical applications and potential for future developments in NLP. The methodology includes data acquisition and preparation, model fine-tuning, embeddings analysis, and a comparative performance analysis using statistical tools like confusion matrices. The results section details the performance of each model in the assigned tasks, providing a critical evaluation of their effectiveness.

6.2 Key Achievements

The key achievements of your project include:

1. Advanced Sentiment Analysis: Successfully demonstrated the effectiveness of Large Language Models (LLMs) in complex sentiment analysis tasks, surpassing traditional methods in understanding nuanced emotions.

2. Enhanced Question-Answering Capabilities: Provided evidence of LLMs' superior ability to interpret and respond to a wide range of questions with high accuracy and context relevance.
3. In-Depth Embeddings Analysis: Explored the impact of model embeddings and fine-tuning on performance, offering insights into how these factors improve model understanding and response accuracy.
4. Comparative Model Evaluation: Utilized confusion matrices to compare models like LLama 7B, Mistral 7B, and GPT 3.5, providing a comprehensive performance analysis across different metrics.

These achievements demonstrate significant advancements in the application of LLMs in NLP tasks, showcasing their potential in practical applications and future technology developments.

6.3 Conclusion

This project offered valuable insights into the capabilities and limitations of Large Language Models (LLMs) like LLama 7B, Mistral 7B, and GPT 3.5 in performing sentiment analysis and question-answering.

The analysis revealed that while all three models demonstrate promise, their effectiveness varies across tasks. In sentiment analysis, the models struggled to accurately match predicted sentiments to the labeled emotions in the dataset, achieving only 12.73% (GPT 3.5) and 2.05% (LLama 7B) exact match accuracy. This suggests difficulties in capturing nuanced expressions of emotions within text. However, partial accuracies were more reasonable, indicating some high-level comprehension of sentiment. Mistral 7B faced overfitting challenges during sentiment analysis training.

In question-answering, Mistral 7B performed better with 75.76% answer match accuracy between datasets. But interpretation issues surfaced for certain topics. GPT 3.5 also showed a baseline accuracy of 16.52%, despite refinements. This points to shortcomings in consistently generating precise responses across question types.

The integration of GPT 3.5 embeddings boosted performance in both tasks. In sentiment analysis, precision and recall improved considerably, highlighting their ability to encode useful linguistic features. But some emotions posed problems. Similarly, question-answering accuracy rose to 86% with embeddings, although recall issues persisted for negative responses.

The project fulfilled its core objectives of assessing LLM capabilities in these tasks and uncovering insights through confusion matrix-based comparisons. It provided a perspective into how linguistic complexities affect model performance and efficiency. Moreover, it revealed crucial development areas like improved generalization, reduced biases, and handling underrepresented classes.

The limitations rest in the scope and scale of analysis. Commercial constraints hindered extensive fine-tuning and testing. Besides, LLM instability remains a challenge for reliable benchmarking. Future work should focus on larger datasets, low-resource language models, and real-world test cases.

Nonetheless, the project adds to the discourse on LLMs. It guides appropriate LLM selection for target tasks, considering trade-offs in accuracy, efficiency and infrastructure needs while illustrating that apt fine-tuning and embeddings can enhance outcomes. The evidence presented informs the practical integration of LLMs in industry applications.

Additionally, the project highlights ethical aspects of LLMs regarding trustworthiness, accountability, and transparency which warrant greater emphasis alongside performance benchmarks. It encourages further research into mitigating biases during training and establishing participant-centered evaluation protocols.

In summary, this project achieved its goals of assessing select LLMs on key NLP tasks and yielded constructive insights into current capabilities, barriers, and potential growth pathways. The evidence compiled serves as a springboard for innovative applications of LLMs across multiple domains, while keeping ethical considerations at the core of future developmental trajectories. Taken collectively, this project makes a substantive addition to the broader mission of engineering robust and socially-aware language technologies.

6.4 Reflections

From our project on evaluating Large Language Models (LLMs) in sentiment analysis and question-answering tasks, we have gained several key learnings.

Here is an expanded version of the key learnings from your project:

[1] Advanced NLP Techniques

This project provided an in-depth understanding of advanced natural language processing (NLP) techniques, specifically sentiment analysis and question-answering using large language models (LLMs). I gained skills in preprocessing diverse textual data to ready it for analysis by LLMs. This encompassed techniques like tokenization, normalization, handling missing values, and partitioning for model validation. Through comprehensive testing, I understood the nuances involved in sentiment analysis. This includes discerning complex emotions, sarcasm, irony, and context-heavy sentiments - challenges that stretch the limits of even sophisticated LLMs. The project also illuminated the multilayered effort required for question-answering - grasping the query, searching contexts and sources, synthesizing, logical reasoning, and formulating an appropriate response. Observing LLMs handle ambiguous, incomplete, or follow-up questions further revealed complexity of question-answering.

Overall, implementing these NLP techniques using cutting-edge LLMs provided first-hand experience of working with AI systems at the forefront of language understanding.

[2] Fine-tuning and Model Optimization

A key aspect explored was fine-tuning - adjusting an LLM to enhance performance on specific tasks. Experimenting with techniques like low-rank adaptation, the project imparted skills in adaptively modifying models. The substantially boosted capabilities of fine-tuned new models like Mistral 7B underscored the importance of customization for targeted outcomes. Rigorously tracking metrics before and after fine-tuning illuminated its tangible impact. Additionally, hyperparameter tuning helped discover optimal model configurations for maximizing accuracy. This evidence-driven optimization amplified proficiency in tailoring models to attain new benchmarks.

Through iterative fine-tuning, I understood the balance between overfitting to training data and retaining generalizability for real-world implementation.

[3] Insights into Model Embeddings

Analyzing GPT-3.5's embeddings offered insights into its language representations. The precision and recall patterns exposed embeddings' efficacy in encoding definitive features but variability in capturing nuanced expressions. Comparing embeddings performance between tasks shed light on the complex interplay with model architecture and the tradeoffs in representation depth versus processing efficiency. These learnings revealed best practices for embedding construction, analysis and relation for transitioning models from theoretical to practical applications.

[4] Statistical Analysis and Interpretation
The multifaceted model evaluation process enhanced analytical thinking and quantitative reasoning skills. Metrics like confusion matrices, classification reports and accuracy scores provided experience in statistically assessing AI systems. Interpreting these performance measures in light of real-world requirements strengthened critical evaluation, guiding appropriate LLM selection for target applications. Feedback loops between results and model refinement further augmented technical intuition.

In summary, this project furnished multifaceted learnings spanning NLP techniques, model development, quantitative assessment, and practical application - forming a holistic foundation for AI-infused solutions. The fusion of hands-on evidence and theoretical underpinnings institutes a comprehensive perspective for responsible innovation.

6.5 Future Work

There is immense potential to build upon this research in impactful ways:
1. Expanded Real-World Testing: While this project performed a robust assessment using two carefully selected datasets, expanded testing on real-world data from sources like social media, customer reviews or conversational transcripts can further validate findings. Applying the models in live systems and monitoring performance over extended periods can reveal crucial insights.
2. Low-Resource Languages: An intriguing area to explore is the application of these techniques for low-resource languages where annotated datasets are scarce. Research into effective strategies for cross-lingual transfer learning and efficient fine-tuning approaches for such languages can have high impact globally.
3. Ensembling and Hybrid Models: Ensembling multiple models and creating hybrid frameworks by combining neural techniques with symbolic methods is worth investigating. This can lead to systems that integrate strengths of different approaches for enhanced accuracy and transparency.
4. Multimodal Sentiment Analysis: Leveraging information across modalities like text, speech and visuals can potentially improve context understanding and emotion identification. Exploring relevant multimodal fusion techniques poses exciting opportunities.
5. Explainability and Interpretability: To address trust issues in AI systems, explainability techniques that provide users transparency into model predictions warrant deeper exploration. Advancing methods to interpret model logic, uncover biases and enable recourse is pivotal.
6. Personalization for Target Tasks: Research into personalized systems tailored to nuances of specific tasks and domains through continuous learning holds promise. This allows maintaining high accuracy as data distributions shift.

In summary, your pioneering project has opened avenues for impactful innovations in making LLMs more robust, ethical and aligned with real-world complexities. The possibilities highlighted above represent crucial directions that can lead these models from cutting-edge research to widespread adoption. There remains expansive untapped potential in this fascinating area of research.

6.6 Limitations

Some notable limitations that can be highlighted are:
1. Fine-tuning Constraints: Computational costs constrained rigorous fine-tuning of models, especially LLama 7B. More exhaustive fine-tuning may have enhanced accuracy further.
2. Partial Model Analysis: Certain models were only applied to one of the two tasks. Testing all models on both datasets can enable more well-rounded comparisons.
3. Static Assessment: Since models keep evolving continuously, a static point-in-time evaluation provides restricted perspective. Tracking progress via periodic benchmarking is ideal.

References

[1] Du, M., He, F., Zou, N., Tao, D. and Hu, X., 2023. Shortcut learning of large language models in natural language understanding. *Communications of the ACM (CACM)*.

[2] Hadi, M.U., Qureshi, R., Shah, A., Irfan, M., Zafar, A., Shaikh, M.B., Akhtar, N., Wu, J. and Mirjalili, S., 2023. A survey on large language models: Applications, challenges, limitations, and practical usage. *TechRxiv*.

[3] Scribble Data. (n.d.). Large Language Models: History, Evolutions, and Future. Scribble Data. Retrieved November 15, 2023, from https://www.scribbledata.io/large-language-models-history-evolutions-and-future/

[4] Weizenbaum, J., 1966. ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), pp.36-45.

[5] Gers, F.A., Schmidhuber, J. and Cummins, F., 2000. Learning to forget: Continual prediction with LSTM. *Neural computation*, *12*(10), pp.2451-2471.

[6] Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J.R., Bethard, S. and McClosky, D., 2014, June. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations* (pp. 55-60).

[7] Wikipedia contributors. (n.d.). Google Brain. Wikipedia. Retrieved November 15, 2023, from https://en.wikipedia.org/wiki/Google_Brain

[8] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. *Advances in neural information processing systems*, *30*.

[9] Lambda Labs. (n.d.). Demystifying GPT-3. Lambda Labs Blog. Retrieved November 16, 2023, from https://lambdalabs.com/blog/demystifying-gpt-3

[10] Picasso AI. (n.d.). GPT-3: OpenAI. Picasso AI Blog. Retrieved November 16, 2023, from https://blog.picassoia.com/artificial-intelligence/blog/article/gpt-3-openai

[11] ARTIBA. (n.d.). Artificial Intelligence Innovation: The Future with OpenAI GPT-3. ARTIBA Blog. Retrieved November 16 2023, from https://www.artiba.org/blog/artificial-intelligence-innovation-the-future-with-openai-gpt-3

[12] Ye, J., Chen, X., Xu, N., Zu, C., Shao, Z., Liu, S., Cui, Y., Zhou, Z., Gong, C., Shen, Y. and Zhou, J., 2023. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. *arXiv preprint arXiv:2303.10420*.

[13] Tan, Y., Min, D., Li, Y., Li, W., Hu, N., Chen, Y. and Qi, G., 2023, October. Can ChatGPT Replace Traditional KBQA Models? An In-Depth Analysis of the Question Answering Performance of the GPT LLM Family. In *International Semantic Web Conference* (pp. 348-367). Cham: Springer Nature Switzerland.

[14] Kheiri, K. and Karimi, H., 2023. Sentimentgpt: Exploiting gpt for advanced sentiment analysis and its departure from current machine learning. *arXiv preprint arXiv:2307.10234*.

[15] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F. and Rodriguez, A., 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

[16] Zhang, B., Yang, H. and Liu, X.Y., 2023. Instruct-FinGPT: Financial Sentiment Analysis by Instruction Tuning of General-Purpose Large Language Models. *arXiv preprint arXiv:2306.12659*.

[17] Wijeratne, Y. and Marikar, I., 2023. Better Question-Answering Models on a Budget. *arXiv preprint arXiv:2304.12370*.

[18] Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.D.L., Bressand, F., Lengyel, G., Lample, G., Saulnier, L. and Lavaud, L.R., 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

[19] Radiya-Dixit, E. and Wang, X., 2020, June. How fine can fine-tuning be? learning efficient language models. In *International Conference on Artificial Intelligence and Statistics* (pp. 2435-2443). PMLR.

[20] Wang, X., Aitchison, L. and Rudolph, M., 2023. LoRA ensembles for large language model fine-tuning. *arXiv preprint arXiv:2310.00035*.

[21] Zhang, P., Chai, T. and Xu, Y., 2023. Adaptive prompt learning-based few-shot sentiment analysis. *Neural Processing Letters*, pp.1-14.

[22] Liu, H., Tam, D., Muqeeth, M., Mohta, J., Huang, T., Bansal, M. and Raffel, C.A., 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, *35*, pp.1950-1965.

[23] Abi Akl, H., 2023, December. A ML-LLM pairing for better code comment classification. In *FIRE (Forum for Information Retrieval Evaluation) 2023*.

[24] Dréano, S., Molloy, D. and Murphy, N., 2023, December. Embed_Llama: using LLM embeddings for the Metrics Shared Task. In *Proceedings of the Eighth Conference on Machine Translation* (pp. 738-745).

[25] Peng, R., Liu, K., Yang, P., Yuan, Z. and Li, S., 2023. Embedding-based retrieval with llm for effective agriculture information extracting from unstructured data. *arXiv preprint arXiv:2308.03107*.

[26] Pommé, L.E., Bourqui, R., Giot, R. and Auber, D., 2022, July. Relative Confusion Matrix: Efficient Comparison of Decision Models. In *2022 26th International Conference Information Visualisation (IV)* (pp. 98-103). IEEE.

[27] Maria Navin, J.R. and Pankaja, R., 2016. Performance analysis of text classification algorithms using confusion matrix. *International Journal of Engineering and Technical Research (IJETR)*, *6*(4), pp.75-8.

[28] Krouska, A., Troussas, C. and Virvou, M., 2017. Comparative evaluation of algorithms for sentiment analysis over social networking services. *J. Univers. Comput. Sci.*, *23*(8), pp.755-768.

[29] Helsinki-NLP. (n.d.). XED. GitHub. Retrieved October 10, 2023, from https://github.com/Helsinki-NLP/XED/tree/master

[30] Kaggle. (n.d.). Question Answer Dataset. Kaggle. Retrieved October 11, 2023, from https://www.kaggle.com/datasets/rtatman/questionanswer-dataset/