STOCK MARKET ASSISTANCE
AI6127 - Deep Neural Networks For Natural Language Processing

Aradhya Dhruv
Matric: G2303518F
ar0001uv@e.ntu.edu.sg

Keerthivasan Krishnasamy Kumar
Matric: G2204686G
keerthiv001@e.ntu.edu.sg

Maheep
Matric:G2303665G
maheep001@e.ntu.edu.sg

Maheswaran Rohin Kumar
Matric:G2303513K
rohinkum001@e.ntu.edu.sg

Uwineza Joseph
Matric: G2303477F
joseph005@e.ntu.edu.sg

Abstract

In today's rapidly evolving financial landscape, vast amounts of unstructured data from news articles and social media significantly influence market trends. This project introduces an advanced stock market assistant powered by Natural Language Processing (NLP) and cutting-edge language models such as Mistral and Gemma. By synthesizing and summarizing complex financial data, the tool equips non-experts with accessible, accurate market insights derived from real-time and historical data sources[7]. It aims to democratize financial expertise by automating the extraction of actionable insights, thus enhancing individual investment strategies and broadening the understanding of market dynamics without the need for deep financial knowledge.

Keywords: Mistral, LLMs, LoRA

ACM Reference Format:
Aradhya Dhruv, Keerthivasan, Krishnasamy Kumar, Maheep, Maheswaran Rohin Kumar, and Uwineza Joseph. 2018. STOCK MARKET ASSISTANCE AI6127 - Deep Neural Networks For Natural Language Processing . In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX).* ACM, New York, NY, USA, 7 pages. https://doi.org/XXXXXXX.XXXXXXX

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

The overarching goal of this project is to harness the power of advanced NLP techniques and large language models to simplify the complex world of stock market analysis for the general public. Utilizing state-of-the-art models—Mistral[1] and Gemma[6]—the tool processes and summarizes significant volumes of financial news and market data, translating it into easily understandable formats. This process involves innovative methods such as sentiment analysis to predict market behavior and text summarization to condense information, providing users with reliable and timely insights.

Additionally, the project compares the performance of these models in real-world scenarios to ensure the highest accuracy and efficiency in delivering market summaries and sentiment analyses. By doing so, it enables users to make more informed decisions, fostering a more intuitive understanding of market trends and reducing the likelihood of investment errors caused by misinterpretation or lack of information

2 Related Works

2.1 Large Language Models(LLMs) on Tabular Data: Prediction, Generation, and Understanding - A Survey

The paper "Large Language Model Adaptation for Financial Sentiment Analysis" by Pau Rodriguez Inserte and Mariam Nakhlé [9] for financial sentiment analysis. It addresses challenges in using generalist LLMs for finance due to complex texts and specific terminology. The study presents adaptation methods for LLMs tailored to finance, demonstrating comparable performance of smaller models to larger ones. However, drawbacks include limited generative capabilities, focus on known tasks, and potential for improvement with larger models.

2.2 Pre-trained Large Language Models for Financial Sentiment Analysis

This [11] paper "Pre-trained Large Language Models for Financial Sentiment Analysis" by Wei Luo and Dihong Gong focuses on adapting pre-trained large language models (LLMs) for classifying financial news titles into sentiment categories. They propose to adapt the Llama2-7B model using supervised fine-tuning (SFT) technique, achieving significant performance improvements over previous state-of-the-art algorithms. However, the study focuses solely on the classification of financial news titles, limiting the generalizability of their findings to other types of financial texts, potentially overlooking variations in sentiment expressions across different financial contexts.

2.3 Enhancing Financial Sentiment Analysis via Retrieval Augmented Large Language Models

This [3] paper "Enhancing Financial Sentiment Analysis via Retrieval Augmented Large Language Models" by Boyu Zhang and Hongyang (Bruce) Yang introduced a novel framework for financial sentiment analysis, combining instruction-tuned large language models (LLMs) with retrieval-augmentation techniques. While the proposed approach shows significant performance gains compared to traditional models and LLMs like ChatGPT and LLaMA, it has notable limitations. Firstly, the reliance on textual similarity for information retrieval may overlook crucial macroeconomic and microeconomic data, limiting the model's ability to make accurate judgments. Additionally, the effectiveness of the retrieval-augmented module may vary depending on the quality and relevance of the retrieved context, potentially affecting the model's predictive accuracy in certain scenarios. Thus, future work could explore integrating additional economic dimensions with textual data to enhance the precision and reliability of financial sentiment analysis performed by large language models.

3 Data Collection

We extracted historical stock price data from sources such as Yahoo Finance, ensuring a comprehensive dataset covering various time periods. For the news articles, we utilized the Google News API, which provided concise news summaries tailored to specific topics of interest. To demonstrate the efficacy of Language Models (LLMs) in sentiment analysis tasks, we narrowed our focus to prominent companies like Apple, Microsoft, Google, Meta, and Nvidia.

Once the data extraction phase was completed, we initiated the preprocessing stage to prepare the data for analysis.Specifically, we performed the following tasks:

- **Yahoo Finance API:** We extracted the closing prices of the above mentioned stocks using yahoo finance api
- **Google News API:** Utilized the Google News API to search for news articles related to the specified company or stock ticker within a defined timeframe. We collected relevant information such as publication date, media source, article title, full text, summary, and keywords.

Overall, the data extraction process involved a combination of web scraping techniques, API queries.

4 Concepts Utilized in the Project

4.0.1 LoRA. Fine-tuning large pre-trained models is resource-intensive due to the need to adjust millions of parameters, creating significant computational demands and time constraints. LoRA offers a promising alternative by decomposing the update matrix during fine-tuning, and to better understand its effectiveness, it's beneficial to first revisit the traditional fine-tuning methods [5].

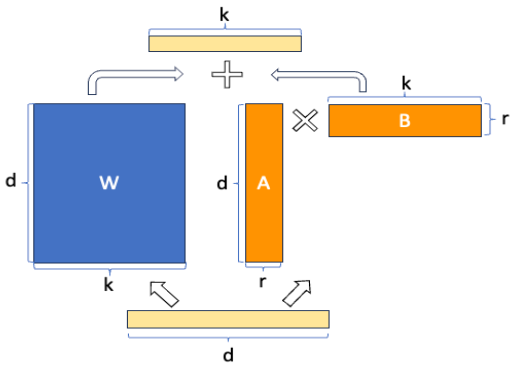

Figure 1. Supervised Finetuning Using LORA

LoRA optimizes fine-tuning by decomposing the weight adjustments $\Delta\mathbf{W}$ into two lower-rank matrices, $\mathbf{A}$ and $\mathbf{B}$, effectively reducing computational complexity. This method, based on the intrinsic rank hypothesis, suggests that significant modifications in a neural network can be represented with fewer parameters. Consequently, LoRA enables updating the weight matrix $\mathbf{W}'$ as $\mathbf{W} + \mathbf{BA}$, while keeping $\mathbf{W}$ unchanged, enhancing efficiency by decreasing memory usage and training time. This approach facilitates fine-tuning large models on less powerful hardware, making the management of extensive AI models more feasible across various domains.

4.0.2 QLoRA. Quantized LoRA[8], builds upon the foundational principles of LoRA. It aims to reduce the precision of numerical representations in the weight matrices, thereby lowering the computational overhead associated with matrix operations. This approach leverages the insight that many deep learning tasks do not require high precision for parameter updates, allowing for significant savings in memory usage and computational resources.

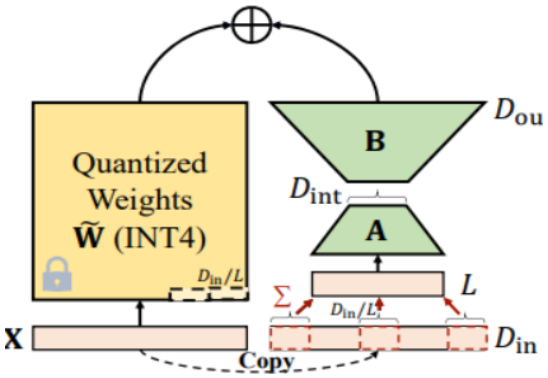

Figure 2. Quantized LoRA

- Quantization of weight matrices: QLoRA achieves this by quantizing the weight matrices into lower-precision formats while preserving model performance to a high degree.
- Efficient resource utilization: This innovative technique enables fine-tuning of large pre-trained models on resource-constrained devices with minimal loss in accuracy, making it particularly advantageous for edge computing and deployment scenarios.

Additionally, QLoRA's adaptability across a wide range of tasks ensures its relevance and applicability in various domains, further solidifying its position as a valuable optimization strategy in the realm of deep learning model adaptation and deployment.

4.0.3 Vector Database and LLM Embeddings. A vector database efficiently stores, indexes, and retrieves high-dimensional vector embeddings crucial for machine learning and AI applications[10]. These embeddings represent data in various forms, enabling quick retrieval of similar items based on their vector representations.

Large Language Models (LLMs) like GPT-3 or BERT produce embeddings that capture semantic meanings of texts. These embeddings, derived from neural network outputs, facilitate understanding, comparison, and retrieval of text-based content. In our project, vector databases will store and query embeddings generated from news articles and social media posts. These embeddings aid in identifying and summarizing relevant financial news and assessing sentiment (bullish or bearish) behind stock movements.

5 Methodology and Experiments

5.1 Prompt Construction for Finetuning:

Mistral Instruct 7B V0.2 requires inputs to adhere to a specific format to facilitate fine-tuning on a custom dataset. Typically, this format entails starting the prompt with a leading instruct tag [<INST>] and concluding it with a closing [/<INST>] tag. Additionally, context and questions are provided to the language model using "Context and Question". This approach aids in refining Mistral's behavior for the specified task.

5.2 Using Vector Database for Enabling Interaction with Gemini:

In order to enable conversational interactions with news articles, we have integrated Google's GEMINI model and utilized its text embeddings to grasp the context of news articles and respond appropriately to user queries. Additionally, we opted to employ Facebook AI Similarity Search (FAISS) as a vector database to house our embeddings. The sequential process detailing the integration with GEMINI is outlined below:

- News articles are initially extracted from our dataset and processed as raw texts. Subsequently, they are segmented into text chunks, and each chunk is assigned a vector embedding.
- Upon assigning vector embeddings, we store them in the FAISS vector database, facilitating semantic similarity search.
- When a user input prompt is received, it undergoes conversion into the format described in Section 4.1. Once transformed, it is converted into a vector embedding suitable for semantic search using FAISS to retrieve relevant prompts.
- The semantic search operation yields the top three ranked results, which are then forwarded to the Language Model (LLM) to generate prompt replies.

The above steps are described in the figure 3:

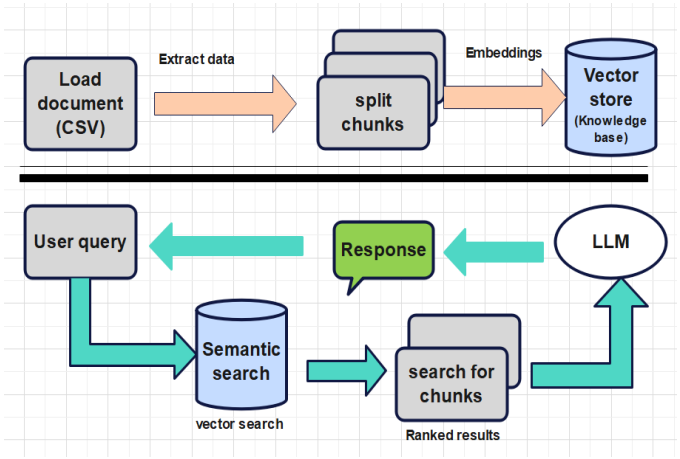

Figure 3. Architecture Followed for Conversation Interactions with Gemini Pro

5.3 Supervised Fine-Tuning (SFT) of Language Model:

To assess sentiment analysis capabilities, we utilized the **Mistral-7B-Instruct-v0.2** zero-shot model from Hugging

Face. Initial tests involved its application to the Financial PhraseBank dataset, a collection of annotated financial texts. This dataset covers a wide range of financial phrases from various contexts like news articles and earnings reports. To refine the model's understanding, supervised fine-tuning (SFT)[4] was conducted using the SFTTrainer from the trl library. Training involved multiple epochs, with metrics such as loss and accuracy tracked to optimize the model's performance. We iteratively updated the model's weights to minimize the loss function during training.

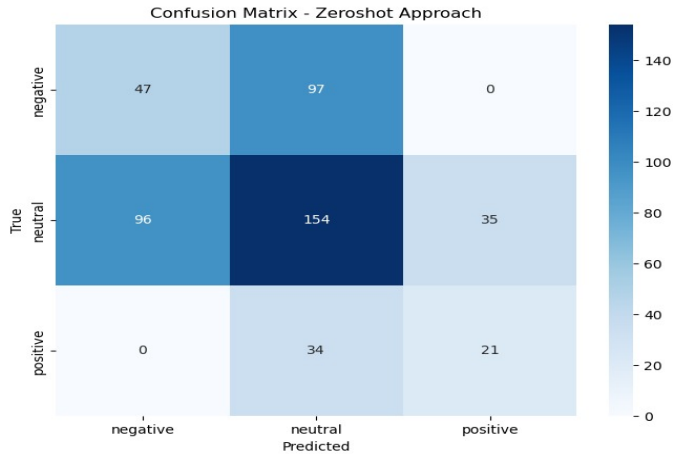

Figure 4. Confusion Matrix for Mistral 7B Instruct (Zero shot model) on Finanical Phrase Bank dataset

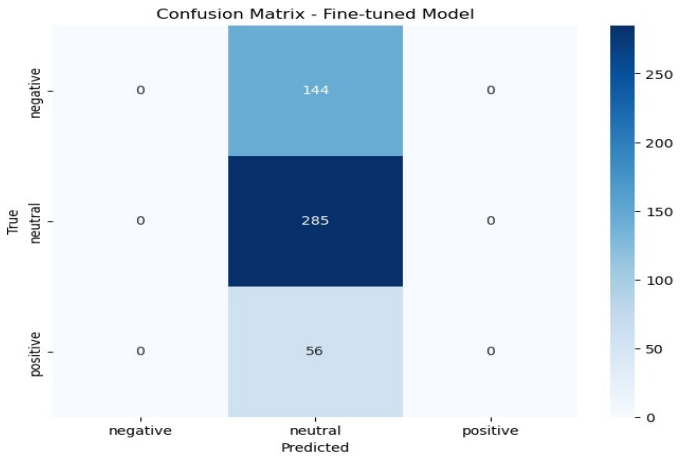

Figure 5. Confusion Matrix for Mistral 7B after supervised fine-tuning on Finanical Phrase Bank dataset

5.4 Interpretation of Confusion Matrix Results:

Using the zero-shot technique with Mistral to predict sentiments on the Financial Phrasebank dataset yielded an overall accuracy of 23%. Notably, accuracy for negative labels remained the highest (62%) compared to positive or neutral labels. The summarized results are presented in the table (1):

Sentiment Label	Accuracy
Positive	7%
Neutral	28.1%
Negative	62.5%

Table 1. Zeroshot Accuracy for Different Sentiment Labels

Given the 23% overall accuracy, we sought to rectify the class imbalance, notably skewed towards neutral and positive classes (see Figure 6).

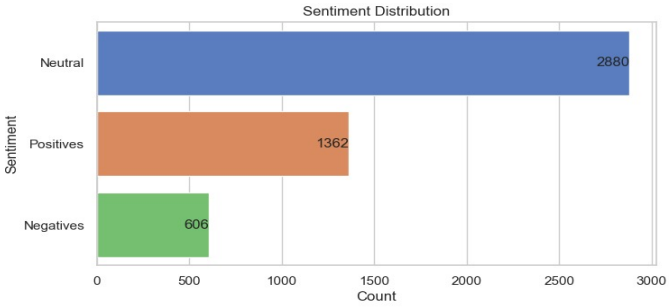

Figure 6. Class Imbalance In the Dataset

We applied both downsampling and upsampling techniques to tackle the class imbalance in our training dataset. The Mistral model was trained for 10 epochs each time to compare the effects of these approaches. Our observations revealed expected reductions in training loss across both sampling methods, with slight variations in each (see Figure 7).

Although we tried, the downsampling and upsampling techniques on the training dataset to balance the predictions.

Figure 7. Training Losses Observed with Different Approaches While Finetuning Mistral

They were still skewed towards neutral classes as observed in confusion matrix (Refer Figure(5)) and the accuracy after fine-tuning was 58%.

The figure below demonstrates the training time difference observed with different approaches. As evident, downsampling had the least training time whereas as Upsampling took the highest amount of training time.

The classification results after fine tuning reveal a mixed

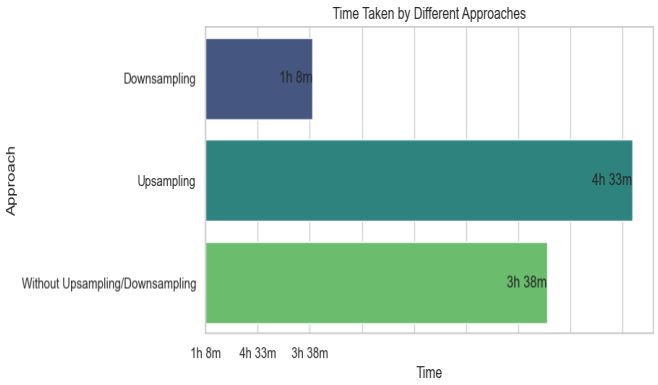

Figure 8. Training Time with Different Approaches

performance of the model across sentiment labels. While achieving a commendable accuracy of 1.0 for instances labeled as neutral, the model severely falters in accurately predicting positive and negative sentiments, with both precision and recall metrics at 0.00. This suggests a significant imbalance or lack of informative features for these classes. The overall accuracy of 0.588 indicates moderate success in sentiment prediction, but the low F1-scores for positive and negative sentiments (0.00) underscore the model's limitations in capturing the nuances of sentiment analysis.

5.5 Utilizing Streamlit for Market Sentiment Analysis and Summarization

Streamlit, a popular Python library, served as the foundation for developing an intuitive user interface (UI) to explore market sentiment and summarize financial news articles. The application is structured around several key functionalities, facilitated by defined functions within the codebase. Users can interact with the tool through a user-friendly sidebar, where they can input questions, select specific stocks, and choose time frames for analysis. Behind the scenes, the code orchestrates a series of operations, including data loading, preprocessing, sentiment analysis, and summarization. Utilizing Pandas, Matplotlib, and Streamlit, the tool dynamically generates visualizations, such as sentiment plots and stock data graphs, tailored to the user's selections. Additionally, the application incorporates a conversational chain for question answering, enriching the user experience with contextual insights. Summaries of pertinent information are displayed below the visualizations, providing concise yet informative narratives of the selected stock's performance and sentiment trends.

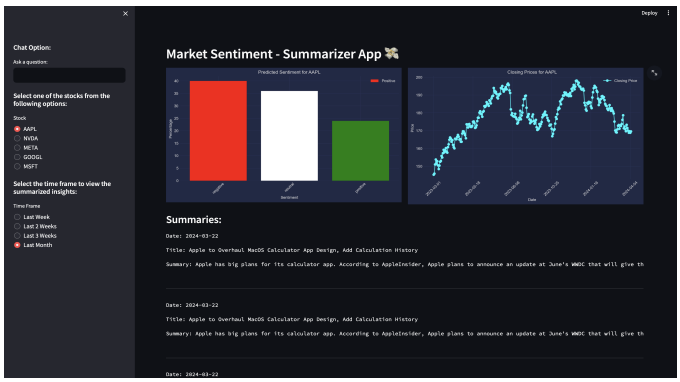

Figure 9. User Interface from Streamlit

NewsAPI[2] is used by the UI's News Summarizer feature to obtain online financial news articles. News Articles are retrieved from various websites for companies including Apple, Microsoft, Nvidia, Meta, and Google. The search query uses the company name and financial keywords like finance, financial, stock market etc to retrieve relevant financial news articles only.Newspaper3k is used for the web scraping and News data parsing

Web Scraping: Newspaper3k retrieval of news articles from various online sources by scraping web pages.

NEWS Parsing: Upon retrieval, it parses the content to extract data such as the article title, publication date, and News text and form a dictionary. This is then used to summarise.

Tokenization The News article is tokenized into individual words, and stop words are removed. This tokenization process prepares the text for further processing using the BERT model. This step helps prepare the text for further processing using the BERT model.

Extraction Bert-Extractive-Summarizer assists in extractive summarization by identifying the most important sentences and keywords from the original News that better represent the news.

Summarization

- **Text Preprocessing:** Before feeding the text into BERT, it's prepared by removing less relevant stop words and format it.

- **Representation:** BERT creates a detailed understanding of the text's meaning, capturing the important relationships between words. This helps with the better performance of the summarisation.

- **Extraction Mechanism:** The summarizer selects sentences and words from the original news based on their importance. This makes sure the summary contains the most important information that holds the original context.

6 Discussion

The supervised fine-tuning (SFT) of the Mistral-7B model aimed to enhance its sentiment analysis capabilities, particularly when applied to financial text data. However, the results obtained from the confusion matrices reveal some challenges and limitations encountered during this process.

Despite being fine-tuned, the model exhibited relatively low accuracy. Moreover, in the zero-shot setting on the Financial PhraseBank dataset, achieving an initial accuracy of just 23%. The breakdown of accuracy across different sentiment labels further highlighted the imbalance, notably with significantly higher accuracy for negative labels compared to positive or neutral ones. This skew in accuracy can be attributed to the inherent class imbalance within the dataset, as illustrated by the class distribution plot.

To address this class imbalance issue, we explored both downsampling and upsampling techniques during training. While these techniques led to reductions in training loss, the resulting confusion matrices still exhibited a significant skew towards neutral classes, indicating that the model struggled to generalize well across all sentiment categories.

One possible reason for the limited effectiveness of fine-tuning and class balancing techniques could be the complexity and diversity of financial language and sentiments. Financial texts often contain nuanced language and subtle contextual cues that may be challenging for the model to grasp, especially with a relatively small training dataset. Additionally, the pretrained Mistral-7B model may not have been sufficiently tailored to financial language, necessitating more extensive fine-tuning or domain-specific pretraining.

Furthermore, the effectiveness of downsampling and upsampling techniques may have been limited by the inherent noise and variability in the dataset, as well as the potential loss of information associated with data manipulation. Despite our efforts, achieving a balanced and accurate sentiment analysis model for financial texts remains a challenging task, warranting further investigation and experimentation with more sophisticated techniques and larger datasets.

7 Conclusion

The primary objective of this project is to provide access to stock market analysis, leveraging cutting-edge NLP techniques and LLMs to simplify complex financial insights for a broader audience. Through the integration of state-of-the-art models like Mistral and Gemma, our tool transforms extensive financial data into easily digestible formats, ensuring accessibility for all users. By employing sophisticated methods such as sentiment analysis and text summarization, we equip users with reliable and up-to-date insights, enabling them to make well-informed decisions and intuitively grasp market trends.

Our thorough evaluation of these models under real-world conditions guarantees the delivery of accurate and efficient market summaries and sentiment analyses to our users. Through this comparative analysis, we continually refine our approach, mitigating the risk of investment errors stemming from misinterpretation or insufficient information.

In essence, our project signifies a substantial advancement in the realm of accessible stock market analysis. Nevertheless, challenges such as class imbalance and model generalization persist, underscoring the ongoing need for research and improvement. By exploring advanced techniques and leveraging larger datasets, we strive to enhance the accuracy and efficacy of sentiment analysis in financial contexts, further bridging the gap between complex financial data and everyday users.

References

[1] Arthur Mensch Chris Bamford Devendra Singh Chaplot Diego de las Casas Florian Bressand Gianna Lengyel Guillaume Lample Lucile Saulnier Lélio Renard Lavaud Marie-Anne Lachaux Pierre Stock Teven Le Scao Thibaut Lavril Thomas Wang Timothée Lacroix William El Sayed Albert Q. Jiang, Alexandre Sablayrolles. 2023. *Mistral 7B.* Vol. v1. Cornell University. https://arxiv.org/abs/2310.06825

[2] Anonymous. 2023. *News in a word cloud.* DataCenter. https://www.pythonsherpa.com/static/files/html/NewsAPI.html

[3] Tianyu Zhou Boyu Zhang1, Hongyang (Bruce) Yang2. 2023. *Enhancing Financial Sentiment Analysis via Retrieval Augmented Large Language Models.* Vol. v2. Cornell University. https://arxiv.org/pdf/2310.04027.pdf

[4] Ph.D. Cameron R. Wolfe. 2024. *Supervised Fine-Tuning (SFT) with Large Language Models.* Towards Data Science. https://towardsdatascience.com/supervised-fine-tuning-sft-with-large-language-models-0c7d66a26788

[5] Phillip Wallis Zeyuan Allen-Zhu Yuanzhi Li Shean Wang Lu Wang Weizhu Chen Edward J. Hu, Yelong Shen. 2021. *LoRA: Low-Rank Adaptation of Large Language Models.* Vol. v2. Cornell University. https://arxiv.org/abs/2106.09685

[6] Google DeepMind Gemma Team. 2024. *Gemma: Open Models Based on Gemini Research and Technology.* Vol. v4. Google DeepMind. https://arxiv.org/pdf/2403.08295.pdf

[7] CBDA AAC CCA LN Mishra, CBAP. 2023. *Top 10 Tools used by Financial Analysts.* Adaptive US Logo 2024-min-2. https://www.adaptiveus.com/blog/top-10-financial-analyst-tools/

[8] Benjamin Marie. 2023. *QA-LoRA: Fine-Tune a Quantized Large Language Model on Your GPU.* Towards Data Science. https://towardsdatascience.com/qa-lora-fine-tune-a-quantized-large-language-model-on-your-gpu-c7291866706c

[9] Raheel Qader Gaëtan Caillaut Jingshu Liu1 Pau Rodriguez Inserte1, Mariam Nakhlé1. 2024. *Large Language Model Adaptation for Financial Sentiment Analysis.* Vol. v1. Grenoble Alpes University. https://arxiv.org/pdf/2401.14777.pdf

[10] Toni Taipalus. 2024. *Vector database management systems: Fundamental concepts, use-cases, and current challenges.* Vol. v2. https://arxiv.org/pdf/2309.11322.pdf

[11] Dihong Gong Wei Luo1. 2024. *Pre-trained Large Language Models for Financial Sentiment Analysis.* Vol. v1. https://arxiv.org/pdf/2401.05215.pdf