# From Words to Numbers: AI-Driven Analysis of Earnings Calls

A dissertation submitted as part of requirements for the award of

**Master of Science in Business Analytics**

in Bayes Business School, City University of London

September 2024

Michael Papageorgiou
ID: 230042127

Academic Supervisor: Ahmad Abu Khazneh

Industrial Partner: CitiBank

# Abstract

The integration of Large Language Models (LLMs) in financial analysis presents an opportunity to automate and enhance the interpretation of complex financial data, such as earnings call transcripts. This project investigates the capabilities and limitations of various LLMs, including ChatGPT3.5, ChatGPT4, FinBERT, and Gemini, in performing sentiment analysis on earnings calls of ExxonMobil Corporation across different quarters. By comparing the model outputs against human judgments, the research evaluates their accuracy, efficiency, and consistency in classifying sentiments as positive, neutral, or negative. The results indicate that while more advanced models like ChatGPT4 demonstrate superior performance in identifying neutral sentiments and handling mixed language, biases toward neutrality and positive sentiments are prevalent across models. Additionally, keyword extraction was explored to identify key topics and sentiments in the transcripts. The findings highlight the potential of LLMs as tools to support financial analysts but also underscore their current limitations in matching contextual understanding of human experts.

# Acknowledgement

I want to thank my supervisor Ahmad Abu-Kazneh for his continuous guidance throughout the project. I would also like to thank my fellow team members for their contributions with data collection and analysis. I also want to thank the industrial partner Jan Rock of CitiBank for his contribution.

# Contents

# Chapter 1

# Introduction

The integration of AI in financial analysis, particularly through Large Language Models (LLMs), has become more relevant than ever before. LLMs, trained on big datasets using machine learning techniques, demonstrate almost human-like text comprehension and production in a matter of seconds. An earnings call is a conference call between the management of a public company, analysts, investors, and the media to discuss financial results during a given reporting period, such as a quarter or a fiscal year (Investopedia, 2021). In the past up until now, interpreting earnings calls has been conducted by financial experts with extensive industry knowledge and years of experience, as the language used in these calls is often complex with financial jargon and carefully chosen words. Usually, in earnings calls the board of directors prettifies corporate challenges and downplays issues of companies. This kind of language necessitates a high degree of attentiveness and context understanding. With the recent development of more capable LLMs, the question arises if financial experts could use these models or potentially even be replaced entirely in the coming years. Earnings calls are often influenced by subjective factors such as tone, context, and implied forward-looking statements. The market reacts strongly to the tone and content of analysts' questions during earnings calls, as these interactions often reveal underlying sentiments and expectations about a company's future performance (Chen et al., 2018, p. 3). The automation of earning calls could bring numerous benefits for financial institutions such as deeper analysis, cost efficiencies and the capacity to process and interpret vast quantities of information in real time much faster. This is crucial in financial markets where decisions are made in a matter of minutes. Professional traders monitor closely earnings calls for stocks they own or trade, making decisions to buy, sell, or hold based on the general sentiment conveyed during these calls minutes after the call is ended. This is the reason why the price of a stock immediately changes after an earning call is released. An LLM could potentially provide traders with a time advantage of several minutes if it accurately captures and analyses the sentiment of the call, potentially allowing for quicker decision-making in response to the earnings information. J.P. Morgan already launched its own LLM called Suite, which will perform tasks traditionally executed by research analysts. The vision is that this tool will assist in writing, generating ideas, and summarizing documents within their asset and wealth management divisions. According to an internal memo, LLM Suite is positioned as a "research analyst that can offer information, solutions, and advice on a topic" (Financial Times, 2023). Similarly, Morgan Stanley has introduced its AI assistant, "Debrief", to support wealth advisors. This shows that LLMs are already establishing their way into financial institutions (Morgan Stanley, 2024). Nevertheless, this remains a largely undiscovered field, and it is uncertain whether LLMs can match the accuracy and contextual understanding of human experts. This project aims to investigate the abilities and limitations of these models by comparing the performance of LLMs against human experts in analyzing transcripts of earnings calls of Exxon Mobil Corporation. Exxon is one of the world's largest publicly traded international oil and gas companies, known for its significant role in the global energy sector (ExxonMobil, 2023). The models tested are ChatGPT 3.5 with an API pipeline, ChatGPT 4, FinBERT, and Gemini. Each model will be evaluated based on their ability to determine positive, negative or neutral sentiments compared to human judgement. This analysis will focus on accuracy, efficiency, consistency and timing. Another part of the project involves keyword extraction with the according sentiment from the earnings call transcripts. This step helps identify the most relevant topics and key points discussed during the calls. 1 ...

# Chapter 2

# Literature Review

### 2.0.1 Overview of LLMs and Their Applications in Finance

The application of LLMs in finance is diverse, covering areas such as risk management, fraud detection, customer service, and predictions of market trends. LLMs provide financial firms with the potential to revolutionize their operational processes, from regulatory compliance to personalized customer interactions (Liu et al., 2023, p. 14). However, there is limited research on leveraging multiple LLMs to enhance the depth and accuracy of such analyses in the context of earnings calls" (Liu et al., 2023, p. 22). This gap presents an opportunity for further exploration, especially given the complexity and variability inherent in earnings call data.

### 2.0.2 FinBERT: Financial Sentiment Analysis Model

FinBERT is a domain-specific adaptation of the Bidirectional Encoder Representations from Transformers (BERT) model, specifically designed for sentiment analysis in the financial domain. The original BERT model employs bidirectional transformer architecture that allows for the understanding of context from both directions in a text sequence, leading to more nuanced understanding and better performance in NLP tasks (Devlin et al., 2018). FinBERT is trained only on financial data, thereby enhancing its ability to comprehend the terminology, idioms, and linguistic nuances of financial texts and jargon. The fine-tuning process involves training the model on domain-specific datasets like financial news, earnings calls, analyst reports, and other financial documents to distinguish sentiment with greater accuracy. This process ensures that FinBERT can effectively categorize text as positive, negative, or neutral. However, limitations are that its performance is depending on the domain specific data used for fine-tuning. If the data is not comprehensive or representative of the financial language diversity, the model's ability to generalize may be limited.

### 2.0.3 Generative Pre-trained Transformers (GPT): A Milestone in NLP

Generative Pre-trained Transformers (GPT) represent a groundbreaking development in the field of NLP and generally in the field of AI. The architecture of GPT employs self-attention mechanisms to process sequences of text, allowing the model to capture complex dependencies and contextual relationships within the data (Vaswani et al., 2017). The GPT models employ pre-training and fine-tuning. In the pre-training phase, the model learns to predict the next word in a sentence by leveraging a vast amount of unlabeled text data, enabling it to acquire a deep understanding of language structure, grammar, and even some elements of reasoning. The subsequent fine-tuning phase adapts this pre-trained model to specific tasks by training it on smaller, labelled datasets. This approach allows GPT models to generalize effectively across a wide range of NLP tasks, such as text generation, summarization, translation, and question answering (Radford et al., 2018; Brown et al., 2020). GPT3.5 has 175 billion parameters and can perform complex language tasks in a zero-shot or few-shot learning context. Its large-scale architecture and extensive pre-training dataset allow it to generate human-like text. However, GPT3.5 is resource-intensive in both training and deployment. Moreover, GPT models can sometimes generate plausible sounding but factually incorrect or biased outputs due to their reliance on patterns in the training data (Brown et al., 2020). To mitigate some of these limitations, ongoing research focuses on improving the factual accuracy and reliability of these models, including methods like fine-tuning on domain specific data, implementing fact-checking mechanisms, and introducing more robust evaluation metrics to assess generated content. GPT4 builds on GPT3.5 with advancements in model architecture, fine-tuning techniques, and data diversity, resulting in more refined language understanding and generation. It has a larger number of parameters than GPT3.5 and improved handling of complex instructions, greater factual accuracy, and the integration of more advanced safety techniques to reduce biases and harmful outputs (OpenAI, 2023). The data it is trained on is also more recent. ChatGPT4 has enhanced training methodologies, including Reinforcement Learning from Human Feedback (RLHF), which ensures alignment with human values and preferences (Bubeck et al., 2023).

### 2.0.4 Gemini: Google DeepMind Most Advanced LLM

Gemini, developed by Google DeepMind, is the main rival of ChatGPT. It processes text, code, audio, images, and video like the paid version of GPT. Gemini leverages Tensor Processing Units (TPUs), to deliver high performance. 2 ...

# Chapter 3

# Data Preparation

## 3.1 Data Collection

The base dataset consists of earnings call transcripts of ExxonMobil Corporation from three specific quarters: Q3 2020, Q2 2023, and Q1 2024. These transcripts were obtained from ExxonMobil's official website, excluding the third quarter of 2020, which is from Seeking Alpha. The selection of these periods allows for analysis across different time frames. Q3 2020 is of particular interest due to its release during the pandemic, which includes more negative sentiment and different language. I utilized the PyPDF2 library to extract textual content from PDF files of the earnings call transcripts, converting unstructured PDF data into a format suitable for analysis. Next, I eliminated repetitive elements such as page numbers and disclaimers, focusing the dataset on the main content. Regular expressions were used to split the cleaned text into individual sentences. Our team was not sure whether to parse into paragraphs or sentences, but after careful consideration, we concluded that we would generate much more data to compare with, when we parse into sentences. The LLM must be capable to determine the sentiment of each sentence without requiring the entire paragraph for context. When analyzing a full paragraph, the sentiment can become more generalized due to the presence of mixed sentiments. This approach provides more insights and a closer look at how the LLM interacts with each sentence. The segmented sentences were then organized into Data Frames and exported to separate Excel files for each quarter. 3.1 ...

## 3.2 Data Manipulation

The parsed excel files were further processed by removing special characters, punctuation, elimination of single-character words, conversion to lowercase, and final whitespace trimming. LLMs achieve higher accuracy working with processed text. Missing values or ambiguous data were removed in crucial features for data quality. Feature engineering was performed to create additional context for the analysis, including sentence length and word count. To ensure data validation, the extracted text volume was compared against original PDF page counts. Last, random sampling of processed sentences to verify consistency. 3.2 ...

## 3.3 Human Evaluation Dataset for Validation

The human evaluation dataset serves as a validation benchmark for our sentiment analysis. For each earning call three people of my group independently assessed the sentiment of each sentence in the datasets as Positive (1) for Buy, Neutral (0) for Hold, or Negative (-1) for Sell. The results from all three evaluators were averaged, and the average was rounded up to the nearest integer to finalize the sentiment label for each sentence. Although we are not professional traders, this dataset serves as the benchmark for quantitatively comparing results to human judgment. It provides a solid reference point for evaluating the accuracy, reliability, timing, and consistency. 3.3 ...

# Chapter 4

# Methodology and Implementation

The key metrics were agreement rates and confusion matrix. The agreement rate and accuracy refer to the same metric. It measures the percentage of instances where the LLM predictions agree with human evaluations. 4 ...

$$\text{Accuracy} = \left(\frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}\right) \times 100$$

The confusion matrix provides a detailed breakdown of the true positives, false positives, and false negatives for each sentiment category. Other metrics used in the analysis were descriptive statistics and correlation coefficient, which helped to identify biases, relationships and variability in predictions.

### 4.0.1 ChatGPT3.5

The first model is ChatGPT3.5. For this approach I developed an NLP pipeline, which sends one sentence at a time to GPT3.5. Each sentence is individually processed. This was rather inefficient and took more than 15 minutes per earnings call. In addition, more API calls are made this way, which is cost intensive. Therefore, I used batch processing. This approach processes multiple sentences in a batch, sending them together in a JSON format to GPT3.5. For each batch, a tailored prompt was implemented to simulate the role of a financial analyst, instructing the model to classify each sentence as Positive (1), Neutral (0), or Negative (-1). The content of the prompt did not significantly impact the accuracy. The only important part in the prompt was to specifically command the model how to return the output. 50 sentences per batch were used across the quarters. The model needed around 5 minutes per earnings call, making it much more time efficient. The predicted sentiments were then merged with the original datasets for each quarter and concatenated with human evaluations to facilitate direct comparison and performance evaluation. The accuracy across all quarters is relatively low. The confusion matrix showed that the model tends to predict positive sentiments in cases where humans evaluated neutral or negative sentiments. The results suggest that GPT3.5 tends to over predict positive sentiments.

Table 4.1: Accuracy of GPT3.5 for Each Quarter

| **Quarter** | **Accuracy** |
|---|---|
| Q1 2024 | 0.41 |
| Q2 2023 | 0.42 |
| Q3 2020 | 0.57 |

### 4.0.2 FinBERT

The next model tested is FinBERT. For this approach, I implemented an NLP pipeline, where the model and tokenizer were loaded using the Hugging Face library. A sentiment analysis pipeline was then initialized to classify sentences into Positive (1), Neutral (0), or Negative (-1). Each earnings call transcript is processed in approximately 1.5 minutes, significantly less time required for analysis compared to other LLMs. There is a notable improvement in accuracy. In Q3 2020, the model achieved an accuracy of 0.71, which is the highest among all LLMs for this quarter. The confusion matrix showed a closer alignment with human evaluations across all sentiment categories. FinBERT performed well in identifying negative sentiment. However, the model still tends to over predict positive sentiments in some cases and could benefit from handling neutral sentiments more accurately.

Table 4.2: Accuracy of FinBERT for Each Quarter

| **Quarter** | **Accuracy** |
|---|---|
| Q1 2024 | 0.63 |
| Q2 2023 | 0.69 |
| Q3 2020 | 0.71 |

### 4.0.3 ChatGPT4

The next model analyzed was ChatGPT4. I used a conventional approach, as no API key was provided. ChatGPT4 was tested using a prompt-based approach. We uploaded the parsed sentences from each earnings call into an Excel file and provided the model with a custom prompt instructing it to predict the sentiment of each sentence as Positive (1), Neutral (0), or Negative (-1). Then we instructed it to create a new excel file with the according predicted sentiment. The resulting outputs were provided in a new Excel file, which allowed for direct comparison against human evaluations. The prompt is as follows:

> *"As a trader in the bank with 20 years of experience in the stock market, I want you to read the parsed earning call and score each sentence (negative: -1, neutral: 0, and positive: 1). How is each sentence important to make your decision after the earning call and to make your decision to either buy, sell, or hold?"*

Table 4.3: Accuracy of ChatGPT4 for Each Quarter

| **Quarter** | **Accuracy** |
|---|---|
| Q1 2024 | 0.67 |
| Q2 2023 | 0.76 |
| Q3 2020 | 0.66 |

The accuracies represent a significant improvement over its predecessors. In Q2 2023, it outperformed all other models. In terms of timing, it required approximately 2 minutes per earnings call, positioning it between the other models. ChatGPT4 demonstrated a more balanced approach and gave more neutral sentiment. This is probably because it could identify mixed sentiments in sentences, where positive as well as negative sentiments were included, and according to the logic, it evaluated it as neutral, which is also the safest choice if you are between two sentiments.

### 4.0.4 Gemini

The final model evaluated in this analysis was Gemini. Like ChatGPT4, Gemini was tested using a prompt-based approach. The prompt is as follows:

> *"You are a finance expert with 40 years of experience. I want you to grade each sentence of this earning transcript and give a sentiment score of either 1, 0 or -1 representing (positive sentiment, neutral and negative sentiment for investor). Your analysis should be based on the perspective of a senior stock trader who is analyzing Exxon earning call transcript. Please provide me with your answer in a excel file"*

Gemini misclassified positive sentiments as neutral, suggesting that it struggles with distinguishing between subtle sentiment differences when presented with mixed sentiment language. Overall, the performance across all three quarters indicates a consistent bias toward neutral sentiment. It displayed less ability to differentiate between negative and positive sentiments. In terms of timing it is close to ChatGPT4 needing around 2 minuites per earnings call. Nevertheless it is the second best performing model after ChatGPT4. This is also true for conventional users using the LLMs for everday tasks.

Table 4.4: Accuracy of Gemini for Each Quarter

| **Quarter** | **Accuracy** |
|---|---|
| Q1 2024 | 0.67 |
| Q2 2023 | 0.75 |
| Q3 2020 | 0.64 |

Figure 4.1 shows clearly that ChatGPT4 and Gemini are the best performing models, except for Q3 2020.

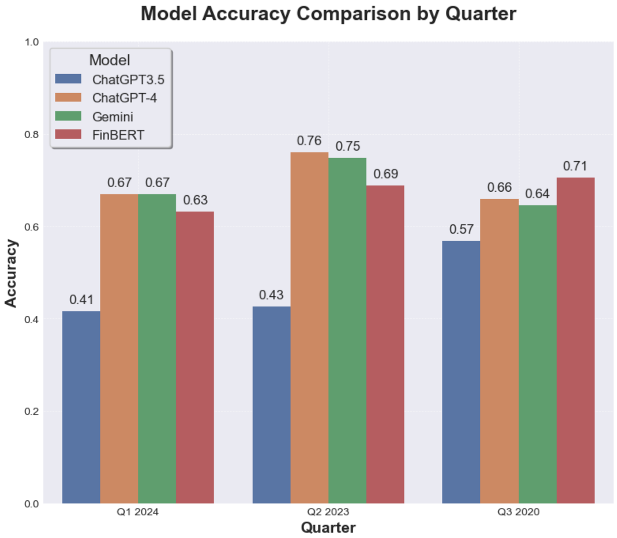


Figure 4.1: Comparison of accuracy across all quarters and all LLMs

## 4.1 Qualitative Analysis of Highlighted Discrepancies

To enhance the evaluation process, I created Excel files that specifically highlight sentences where there is a significant discrepancy between human judgment and LLM output, defined as an absolute difference of 2 in sentiment scores. This occurs, when a model assigns a sentiment score of 1 (positive) while human evaluators rate it as -1 (negative). This provides a clearer picture of the weaknesses and limitations of LLMs. I focused specifically on ChatGPT4, as it is the best performing model. The qualitative analysis revealed a consistent pattern where the model struggled with sentences containing mixed sentiments and complex financial language.

> *"Despite the considerable challenges associated with the pandemic, we have been able to achieve our best ever safety and best reliability performance in the Upstream in five years."*

This sentence contains a recognition of a challenge (negative sentiment) but with a positive outcome (positive sentiment).

Similarly, sentences involving detailed, domain-specific language, such as regulatory or technical terminology.

> ***"I mean, as I just commented here about the challenges with the regulation and the translation of IRA into regulation, you know, we've got a class six well permitting that's going to be required for sequestration,"***

In this case, the sentence contains complex references to regulatory processes. A possible reason could be that the model does not have access to the relevant data sources, which might be confidential. This raises the question about the type of data the models are primarily trained on and how this influences their decision-making processes. Further fine tuning on specific financial datasets could enhance performance. In addition, the LLMs make the same errors across the quarters. Gemini and ChatGPT4 made the exact same errors for Q1 2024, which shows similar weaknesses in their structure. FinBERT and ChatGPT3.5 make similar errors as well, likely due to their older architecture, different algorithm or number of parameters involved. The next step involved a reevaluation of the discrepant sentences. This was done by individually presenting the same sentences back to the models using the same prompt. Interestingly, this reassessment often resulted in correct sentiment classifications. This suggests that the model has the capacity to learn or respond more accurately when given a second chance to evaluate a specific context. This also shows that the models demonstrate higher accuracy with short, focused sentences. As sentences become longer or more, containing multiple clauses or mixed sentiments, the accuracy tends to decrease. More advanced LLMs rate the overall sentiment of earnings calls as neutral. Human Judgement is also tending to neutrality, which is due to the neutral language used in general during earnings calls. This pattern suggests a bias in the models towards safety, where they default to a neutral sentiment in mixed scenarios, potentially to avoid over-committing to positive or negative sentiment classifications that could be inaccurate. This can be also observed in Figure 4.2, where all evaluations are summed up across all quarters.

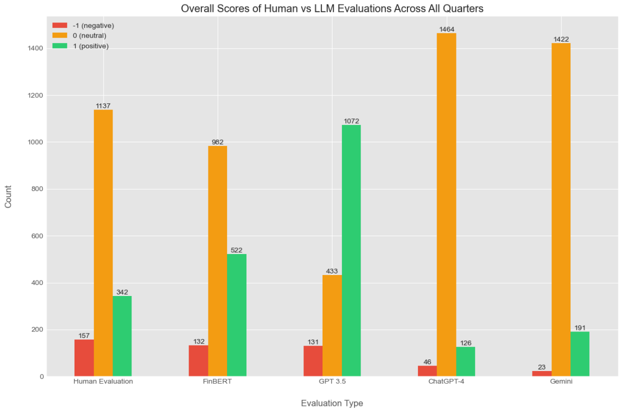
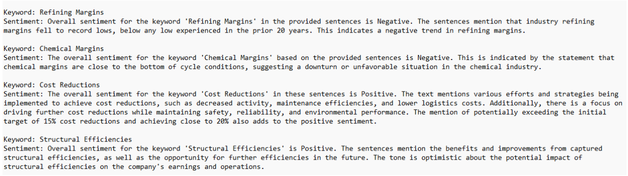


Figure 4.2: Sum of scores across all quarters of Human VS LLMs

## 4.2 Keyword Extraction and Sentiment Analysis

To further expand my project, I instructed the LLM to extract keywords and corresponding sentiment. This serves as an extension to my main analysis. I used GPT3.5-Turbo with an NLP pipeline. The goal was to provide a concise overview of the key topics discussed. For each extracted keyword, the LLM classified the overall sentiment. This offers a deeper understanding of the company's performance, market outlook, and strategic decisions as perceived during the calls. A benefit is that professionals can grasp the most important key topics and sentiments in less than a minute, giving them a time advantage in the markets. GPT 3.5 needed around 35 seconds. However, the LLM again tends to output neutral sentiment when keywords contain both positive and negative aspects. Personally, I think it is still useful to have such a concise summary of key topics and with further development and refinement, I think it might play a significant role in the future. LLMs are already creating transcripts in real time from interviews or video conferences etc. Figure 4.3 shows a snapshot from the text file GPT3.5 created with an API pipeline.



```
Keyword: Refining Margins
Sentiment: Overall sentiment for the keyword 'Refining Margins' in the provided sentences is Negative. The sentences mention that industry refining
margins fell to record lows, below any low experienced in the prior 20 years. This indicates a negative trend in refining margins.

Keyword: Chemical Margins
Sentiment: The overall sentiment for the keyword 'Chemical Margins' based on the provided sentences is Negative. This is indicated by the statement that
chemical margins are close to the bottom of cycle conditions, suggesting a downturn or unfavorable situation in the chemical industry.

Keyword: Cost Reductions
Sentiment: The overall sentiment for the keyword 'Cost Reductions' in these sentences is Positive. The text mentions various efforts and strategies being
implemented to achieve cost reductions, such as decreased activity, maintenance efficiencies, and lower logistics costs. Additionally, there is a focus on
driving further cost reductions while maintaining safety, reliability, and environmental performance. The mention of potentially exceeding the initial
target of 15% cost reductions and achieving close to 20% also adds to the positive sentiment.

Keyword: Structural Efficiencies
Sentiment: Overall sentiment for the keyword 'Structural Efficiencies' is Positive. The sentences mention the benefits and improvements from captured
structural efficiencies, as well as the opportunity for further efficiencies in the future. The tone is optimistic about the potential impact of
structural efficiencies on the company's earnings and operations.
```

Figure 4.3: Keywords and corresponding sentiment for Q3 2020

## 4.3 Data Limitations

The dataset used only includes earnings calls from ExxonMobil not taking into consideration different industries and market conditions. Expanding the dataset to include earnings calls from various sectors and companies could offer more insights of the capabilities and limitations. In addition, the human evaluation dataset is not evaluated by experts, but students such as me with no real applied financial knowledge. Another significant limitation is the lack of access to API keys for more advanced models like ChatGPT4 and Gemini. This restriction limits the ability to perform direct comparisons between models in terms of timing and efficiency. The use of an API would allow for more precise timing measurements and better control over prompt settings and batch processing, leading to more reliable insights.

# Chapter 5

# Conclusion

Throughout this project, the aim is to assess the effectiveness of these models in analyzing earnings call transcripts from ExxonMobil. The results indicate that while LLMs can provide rapid and insightful sentiment analysis, they are not yet fully reliable substitutes for human expertise. They automatically default to neutral in scenarios with mixed ambiguous language. The findings of the keyword extraction and sentiment analysis further underscore these limitations. Although successful in identifying relevant keywords and providing a concise summary of the overall sentiment associated with each topic, the tendency for neutrality persisted again. Citibank could integrate LLMs in some of their activities but with caution. Incorporating an LLM into a legacy banking system would require significant training time and customization to ensure the model works effectively with proprietary banking data. In addition, there are a lot of legal and confidentiality issues when it comes to financial datasets of commercial banks. This would create a lot of bureaucratic work as well. If the LLMs become too sophisticated and decide on their own, they could pose a danger on how they manage confidential data. To sum up, LLMs should be viewed as complementary to human expertise, enhancing rather than replacing the judgment of experienced analysts. While LLMs excel at processing large datasets and identifying patterns quickly, they lack the contextual awareness that humans have. A hybrid approach, combining the data-driven insights and domain expertise of human analysts, offers a balanced strategy that maximizes both analytical accuracy and strategic depth.

# References

ExxonMobil. (2023). About Us. Retrieved from https://corporate.exxonmobil.com/About-us

Morgan Stanley. (2024). Morgan Stanley Wealth Management Announces Latest Game-Changing Addition to Suite of GenAI Tools. Morgan Stanley. Retrieved from https://www.morganstanley.com/press-releases/ai-at-morgan-stanley-debrief-launch.

Financial Times. (2024). JPMorgan pitches in-house chatbot as AI-based research analyst. Financial Times. Retrieved from https://www.ft.com/content/96dfec5f-4d5f-4c3e-8f66-ebd0dfc8392d.

Chen, J.V., Nagar, V., and Schoenfeld, J. (2018). Manager-analyst conversations in earnings conference calls. *Review of Accounting Studies*, 23(4), pp. 1315-1354.

Investopedia. (2021). Retrieved from https://www.investopedia.com/terms/e/earnings-call.asp

Liu, C., Arulappan, A., Naha, R., Mahanti, A., Kamruzzaman, J., & Ra, I-H. (2023). Large Language Models and Sentiment Analysis in Financial Markets: A Review, Datasets and Case Study. *IEEE*.

The Alan Turing Institute. (2024). The Impact of Large Language Models in Finance: Towards Trustworthy Adoption.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All You Need. *arXiv preprint arXiv:1706.03762*.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training. OpenAI.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language Models are Few-Shot Learners. *arXiv preprint arXiv:2005.14165*.

OpenAI. (2023). GPT-4 Technical Report. Retrieved from https://openai.com/index/gpt-4-research/

# Appendix A - FinBERT

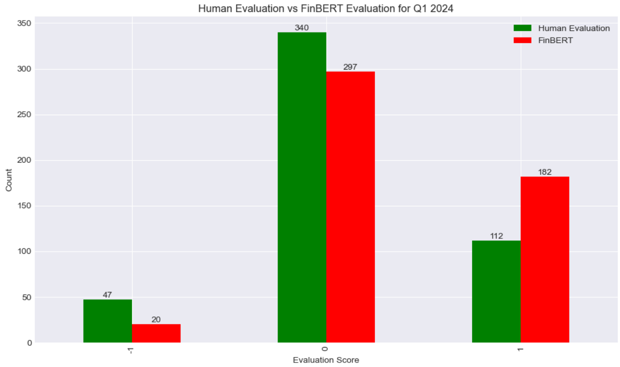


Figure 5.1: Human Evaluation vs FinBERT Evaluation for Q1 2024

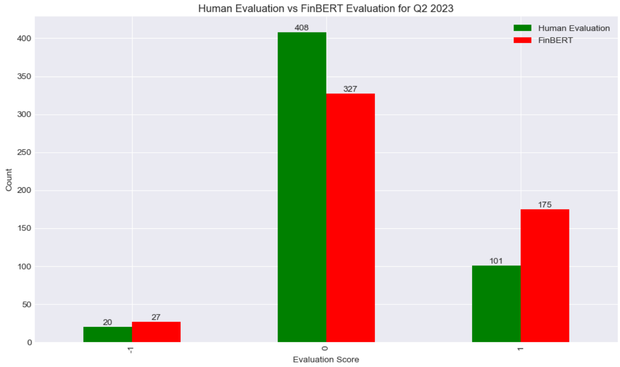


Figure 5.2: Human Evaluation vs FinBERT Evaluation for Q2 2023

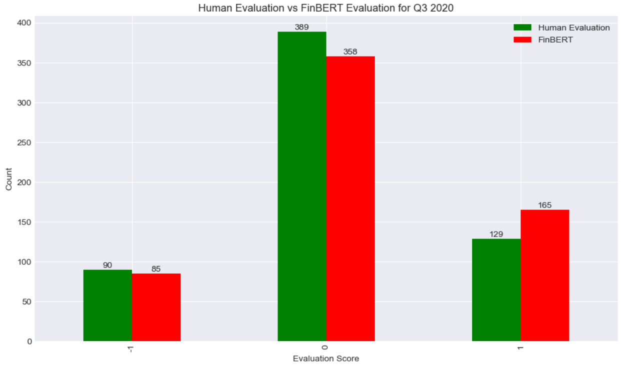


Figure 5.3: Human Evaluation vs FinBERT Evaluation for Q3 2020

# Appendix B - ChatGPT3.5-Turbo

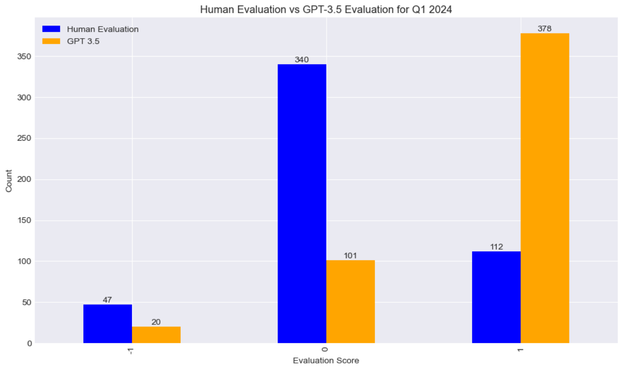


Figure 5.4: Human Evaluation vs GPT3.5 Evaluation for Q1 2024

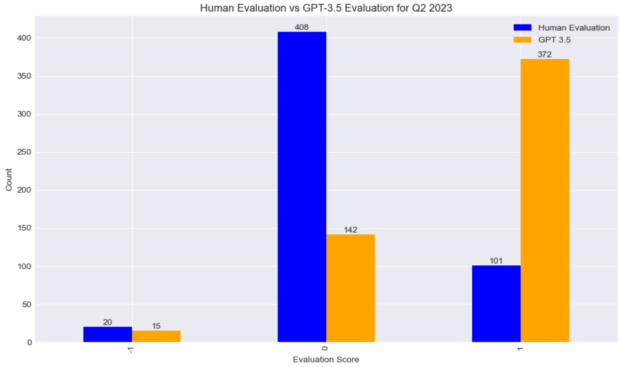


Figure 5.5: Human Evaluation vs GPT3.5 Evaluation for Q2 2023

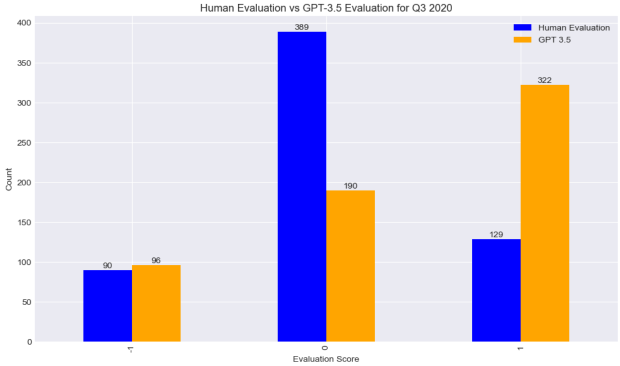


Figure 5.6: Human Evaluation vs GPT3.5 Evaluation for Q3 2020

# Appendix C - ChatGPT4

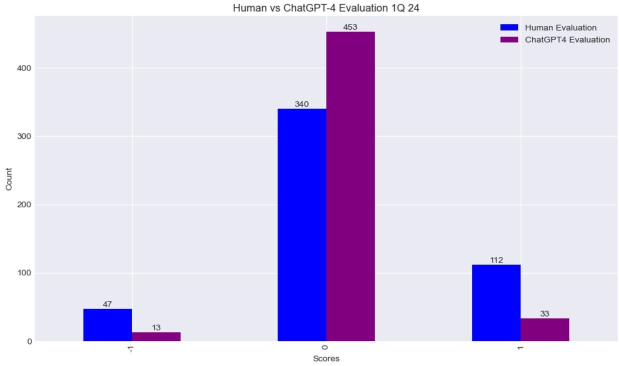


Figure 5.7: Human Evaluation vs GPT4 Evaluation for Q1 2024

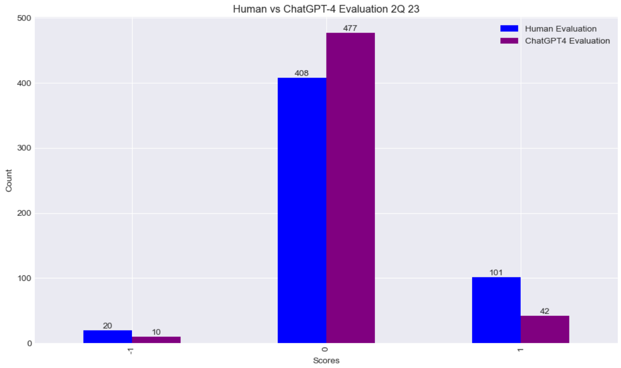


Figure 5.8: Human Evaluation vs GPT4 Evaluation for Q2 2023

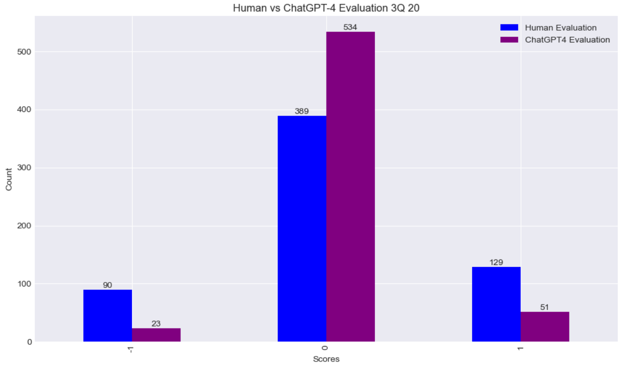


Figure 5.9: Human Evaluation vs GPT4 Evaluation for Q3 2020

# Appendix D - Gemini

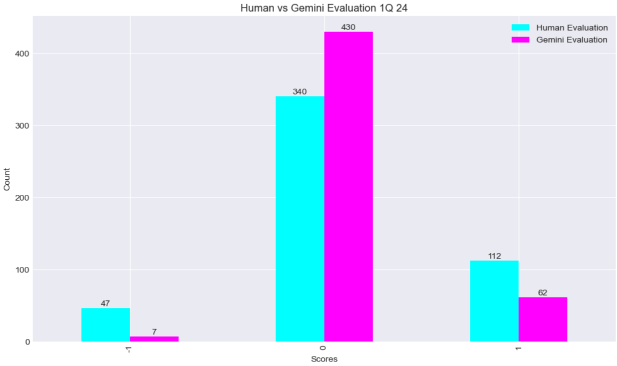


Figure 5.10: Human Evaluation vs Gemini Evaluation for Q1 2024

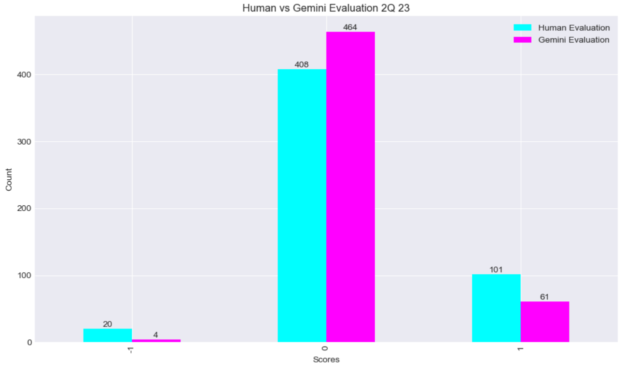


Figure 5.11: Human Evaluation vs Gemini Evaluation for Q2 2023

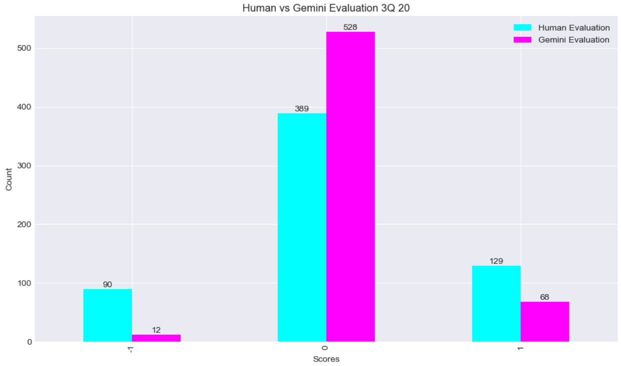


Figure 5.12: Human Evaluation vs Gemini Evaluation for Q3 2020