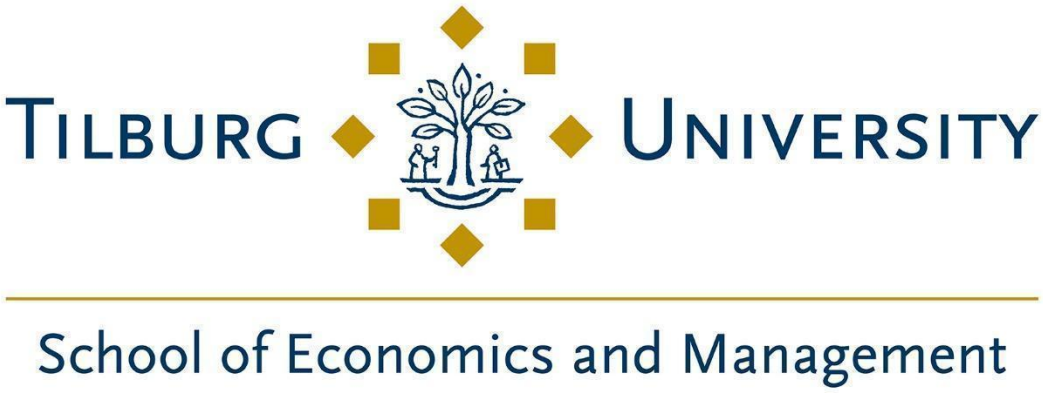


Market Reactions to Earnings Calls: Evidence from GPT-4o Derived Sentiment Analysis

Master Thesis Finance

Author	Tim van Mook
SNR	2108490
ANR	916802
Supervisor	Prof. Dr. Lieven Baele
Second reader	Dr. Giovanni Cocco
Date	29-07-2025

Key Words: Earnings Calls, Sentiment Analysis, Large Language Models, Event Study, Abnormal Returns.

Table of Contents

1. Introduction

Earnings conference calls are a key channel of communication between publicly traded companies and investors. During these calls, corporate management presents financial results through the prepared remarks and provides forward looking statements. This presentation segment is often followed by a question-and-answer session. Analyst raise questions during this session that may reveal additional insights into the firm's strategy, risk factors, or future expectations. This makes the Q&A segment a valuable aspect of the earnings call, as it helps reduce information asymmetry and enhances stock price efficiency (Palepu 2001). Investors and analysts are increasingly using the tone and sentiment of management disclosures, alongside the quantitative segment of the earnings calls, as a trading signal. In the academic field, this is known as text mining or sentiment analysis. Over the last decades, it has proven to be a valuable tool for extracting insights from corporate disclosures such as earnings calls, press releases, and other financial communications. The available techniques allow researchers and market practitioners to quantify positive and negative tone in these communications and offer an opportunity to examine the post-event market reactions.

While the sentiment of financial disclosures is clearly important, there remains a gap in understanding the most effective approach to extract and leverage financial sentiment with the latest technology. Traditional sentiment analysis in the financial field has primary relied on lexicon-based approaches. The Loughran-McDonald (LM) dictionary is a foundational lexicon and can be used for deriving quantitative results from positive and negative tone. These lexicon approaches have provided useful insights but may be limited in their ability to capture context, nuance and complex linguistic patterns. Recently, large language models (LLMs) like OpenAI's ChatGPT have become increasingly popular and may present a new alternative to perform more accurate sentiment analysis. However, it remains an open question whether such a general-purpose LLM can outperform domain-specific lexicon-approach.

This thesis investigates whether GPT4-o derived sentiment can improve the prediction of market reactions to earnings calls. The central question is whether sentiment extracted from these calls, quantified by GPT-4o, can predict short-term stock price movements in the days following the call. The study specifically examines whether the sentiment scores provide additional predictive value beyond both a traditional lexicon-based approach and quantitative earnings news. Leveraging recent advances in AI and its increasing accessibility, the analysis shows how modern techniques contribute to the field of financial sentiment analysis.

This study employs multiple regressions to analyse the relationship between financial sentiment in earnings calls and short-term abnormal returns. First, the study tests whether GPT-4o-derived sentiment provides significant explanatory power for short-term stock returns in addition to dictionary-based sentiment. The benchmark for comparison consists of Loughran-McDonald-derived sentiment scores, a widely used financial lexicon in the academic field. Second, the earnings call transcripts are split into

two sections: the presentation segment and the question-and-answer segment. The study tests which part of the earnings call exhibits a stronger correlation with returns. Third, the predictive power of the derived sentiment is evaluated alongside control variables to test the robustness of the model. The earnings surprise accounts for the deviation between expected and realised earnings, while size accounts for differences in market capitalisation. Finally, the study tests the economic application and relevance through a portfolio simulation. A long-short strategy is implemented based on the regression results, using GPT-derived investment recommendations. The portfolio applies (strong) buy, hold, or (strong) sell recommendations to test whether this strategy can generate abnormal returns. By following this framework, the study not only investigates the existence of sentiment effects on stock returns but also identifies which segment contains more relevant trading information. In addition, it tests whether sentiment is independent from quantitative factors, and evaluates its practical application through an investment simulation.

The empirical analysis is based on a panel of quarterly earnings call transcripts from publicly listed U.S. firms. The dataset consists of 3,360 transcripts from 60 different S&P-500 companies and covers the fiscal quarters from 2010 to 2023. Transcripts were obtained from the LSEG Workspace (Refinitiv) and span a diverse range of industry. Restricting the sample to a set of large-cap firms ensures comparability across different observations and reduces variation due to reporting quality. The 14-year times series data captures sufficient variation in market conditions to help generalise the results.

To quantify the sentiment in each call, two approaches were applied. First a dictionary-based approach, using the Loughran-McDonald master dictionary. This approach counts the frequency of positive and negative words in each transcript to generate a sentiment score, providing a commonly used and transparent benchmark for this study. The large language model approach employs GPT-4o to generate sentiment scores. The model evaluated each transcript based on a detailed prompt and assigns a continuous score from -1 (pessimistic) to +1 (optimistic). In addition, the model produces an investment recommendation based on the sentiment and relevant information. This LLM-based approach is implemented via the OpenAI API which enables batch processing. By comparing a domain-specific lexicon approach with an advanced AI model, the study assesses whether LLMs can capture nuances that dictionaries may fail to capture.

An event study framework is employed to measure the immediate market reactions around the earnings call date. Daily stock return data for each firm and the corresponding market index were collected from the CRSP database. These were used to compute cumulative abnormal returns (CARs) for each firm over multiple event windows ([0,+1], [0,+3], [0,+5]). The CAR represents the stock return over the event window, relative to the expected returns. This approach isolates the earnings call effects on stock prices, by adjusting for general market movements. The CARs serve as the dependent variable, while the sentiment scores serve as the independent variable. Control variables were included to ensure

robustness. The earnings surprise is collected from I/B/E/S and the market capitalisation is computed using data from CRSP.

The results of the study show the value of financial sentiment applications in earnings calls to understand short-term market reactions. In the event-study regressions, both the dictionary-based and GPT-4o derived sentiment exhibit positive coefficients and a statistically significant relationships with stock returns. Sentiment scores are significant across all the defined event windows, with the highest coefficient and explanatory power in the immediate window. The GPT-4o derived sentiment consistently outperforms the traditional approach in terms of adjusted R-squared. Moreover, when both the sentiment variables are included in the regression, the GPT-approach remains statistically significant, while the LM-approach loses significance. This suggests that the variables may be correlated and share explanatory power for the CARs. In addition, the incorporation of GPT-based demonstrates its effectiveness in identifying nuanced sentiment within the call, which is reflected in the market reactions.

The regressions also reveal that the sentiment, expressed during the Q&A segment, is a stronger predictor of market reactions than the presentation segment. This suggests that the unscripted and interactive discussion between analysts and management provides significant information for short-horizon investors. This finding is consistent with prior research and suggests that investors react more strongly to analyst tone and management responses (Price et al., 2012). Furthermore, the relationship between sentiment and returns remains statistically significant after controlling for earnings surprise and size. This robustness suggest that qualitive tone contributes to the market response beyond what is explained by financial earnings news.

Since the event study framework demonstrated statistical relevance, this study extends the analysis to a practical implementation. A portfolio strategy was simulated that takes long positions for events with a positive investment recommendation (based on the sentiment score) and short positions for events with negative recommendations. The strategy generated abnormal returns that were both statistically and economically significant relative to the market index. The GPT-based strategy yields a positive alpha, indicating that the derived sentiment scores produce excess returns relative to the market benchmark. The results from the simulation confirm the economic relevance of this sentiment approach and suggest that LLMs can serve as a valuable tool for investors to enhance their decision-making process.

This study contributes to the existing literature on financial sentiment analysis and market efficiency by providing new evidence on how LLM-derived sentiment analysis captures relevant information about market reactions. These results are consistent with earlier findings that qualitative textual information can enhance the prediction of market response beyond financial metrics. Henry (2006) showed that, compared to models relying solely on financial data, incorporating textual information improves the prediction of market returns. Building on this work, Price et al. (2012) examined earnings calls and

found that the sentiment in these calls, particularly during the Q&A segment, has significant predictive power for stock returns following an earnings call. The results of this thesis are consistent with these studies, confirming that earnings call sentiment correlates with short-term abnormal returns.

While the results support earlier findings on the role of sentiment in market reactions, this study offers a new contribution by showing that a general-purpose AI model can effectively capture financial sentiment. This study shows that GPT-4o outperforms a standard financial sentiment dictionary, suggesting that more advanced language models may provide meaningful improvements over traditional tools. By employing recent developments in AI with empirical methods in finance, this study shows that large language models can be practical and accessible tools to extract significant information from corporate communication. These findings highlight the potential for both academic research and practical implementations.

The remainder of this thesis layout is as follows. Chapter 2 reviews the relevant literature on financial sentiment analysis and the role of qualitative information in market returns. Chapter 3 describes the data collection and sample construction, introducing they key variables and preprocessing steps. Chapter 4 presents the development of research hypotheses. Chapter 5 presents the event study methodology, and Chapter 6 presents the empirical results of this event study analysis. Next, Chapter 7 introduces the design of the sentiment-based portfolio simulation, with Chapter 8 detailing the results. Finally, Chapter 9 concludes the thesis with a summary of the findings and suggestions for future research.

2. Literature Review

This chapter reviews the existing literature relevant to the research question. It focuses on the following key areas: traditional sentiment analysis in finance, the relationship between earnings calls and stock price prediction, and the application of deep learning and large language models.

2.1 Traditional Sentiment Analysis Techniques in Finance

Sentiment analysis, also known as opinion mining, refers to a field of research that analyses a corpus in order to extract opinions, sentiments, emotions and subjectivity from text (Liu, 2021). Research into sentiment analysis mainly started in the early 2000s, when researchers began to recognize the research challenges involved and the broad range of applications (Das & Chen, 2001; Pang et al., 2002; Tong, 2001; Turney, 2002; Yu & Hatzivassiloglou, 2003). To meet these challenges, researchers increasingly depended on Natural Language Processing (NLP), which can be seen as 'a theory-motivated range of computational techniques for the automatic analysis and representation of human language' (Cambria & White, 2014). The use of NLP methods to extract subjective opinions from source materials has become a widely used approach to measuring market sentiment in the stock market. This sentiment is driven by multiple factors such as global events, economic reports, historical and seasonal trends and various additional factors (Sohangir, Wang, Pomeranets, & Khoshgoftaar, 2018). The analysis can be conducted across various levels, including the document level, sentence level, phrase level, and aspect level. Document level analysis treats the entire text as a single entity and assigns an overall polarity score. Three main approaches include the lexicon-based approach, machine learning approach and hybrid approach (Wankhade, Rao, & Kulkarni, 2022).

The **lexicon-based approach** uses curated lists of words, each associated with a predefined sentiment score: positive, negative, or neutral. In the lexicon-based methodology, the text is segmented into individual tokens, and the sentiment of each token is evaluated. These individual assessments are then aggregated to arrive at an overall sentiment classification (Kiritchenko, Zhu, & Mohammad, 2014). Loughran and McDonald (2011) introduced the most commonly used lexicon in the Financial Sentiment Analysis (FSA) field. Their research pointed out that standard lexicons like the Harvard Dictionary misclassify many terms in a financial context. The Harvard word list includes terms such as cost, capital, tax and liability in its negative word count, despite being typically not negative in a financial setting. The authors manually built a financial lexicon by linking the word lists to 10-K filing returns, trading volume, return volatility, fraud, material weakness and unexpected earnings. This specialized financial lexicon has enhanced the precision of sentiment measurement in corporate disclosures (Kearney & Liu, 2014). They observed that, although the lexicon-based approach offers a user-friendly framework for financial analysts, reliance on general dictionaries often proves insufficient for capturing the nuances of financial contexts. Incorporating financial lexicons enhances the precision of analysis but it introduces a methodological challenge concerning the selection of an appropriate weighting scheme, which likely depends on the nature of the corpus being analysed and the hypotheses being tested. A weighting scheme

adjusts the influence of words based on their frequency and informational value within the corpus, making its selection critical for accurate sentiment analysis.

The **machine learning method** can be divided into two main approaches: supervised and unsupervised learning. Unsupervised learning approaches rely on curated resources such as knowledge bases, databases, ontologies, and lexicons. These have been carefully selected to identify sentiment-related information. Although these approaches can provide valuable domain-specific insights, supervised learning is more commonly employed, as it tends to achieve higher accuracy. Supervised methods require training algorithms on labelled datasets prior to application and typically include preprocessing of the text to extract relevant features (Wankhade, Rao, & Kulkarni, 2022). Tetlock (2007) represents a foundational study in the field of FSA by employing an unsupervised learning approach. The study analysed content from the Wall Street Journal, utilizing the General Inquirer, a lexicon-based tool for quantifying content analysis. The study suggests that high media pessimism can serve as a predictor of short-term decline in stock prices, followed by a reversion to fundamentals. These periods tend to align with increased market trading volumes, especially when sentiment is unusually high or low. His findings highlight the potential influence of media sentiment on market behaviour and the value of integrating analytical techniques in financial research to better understand market dynamics. Das and Chen (2007) proposed an innovative approach to FSA by applying supervised machine learning techniques. They employed classifiers such as Naive Bayes and vector-based models and integrated the output through a voting scheme. This methodology improved accuracy and reduced false positives. It proved effective in capturing investor sentiment from stock message boards and demonstrated its value in assessing reactions to events such as press releases, regulatory changes, and management announcements. Li (2010) found that machine learning approaches could achieve better performance than lexicon-based approaches. While these techniques have enhanced sentiment analysis, there are also limitations to the machine learning approaches. A primary limitation of supervised approaches is their sensitivity to both the quantity and quality of training data, which can lead to failures if the data is biased or insufficient. Additionally, detecting opinions at the sub-document level presents additional challenges for supervised approaches, as there is often limited information available for the classifier. In the case of unsupervised approaches, the main limitation is their requirement for a large volume of data to be trained accurately. Fully unsupervised models often generate incomprehensible topics, as their objective functions are not explicitly designed to align with human judgment. Despite its limitations, unsupervised learning remains a valuable approach for extracting insights from data without the need for annotated data (Madhoushi, Hamdan, & Zainudin, 2015).

The **hybrid approach** combines machine learning and lexicon techniques to enhance sentiment analysis. Mudinas, Zhang, and Levene (2012) introduced a foundational hybrid model that integrates a lexicon-based framework with supervised learning. This method leverages the interpretability and structure of lexicons alongside the high accuracy of machine learning. Their approach demonstrates

strong adaptability to diverse writing styles and outperforms a leading state-of-the-art sentiment analysis system. Unlike single-purpose or domain-limited methods, the hybrid approach offers both precision and stability, producing structured, aspect-level sentiment insights and effectively combining the strengths of both techniques. By combining lexicons with machine learning, the hybrid approach uses predefined sentiment classifications to guide feature selection and interpretation, while machine learning models identify complex patterns and contextual nuances that lexicons may fail to capture.

2.2 Earnings Calls and Stock Price Prediction

Corporate disclosures, particularly earnings conference calls, serve as a key source of information for investors and have become increasingly important for capital market participants (Brown et al., 2019). During a conference earnings call, the management provides prepared comments on the performance and future outlook of the company, followed by a question-and-answer session with analysts. Healy and Palepu (2001) describe that "corporate disclosure is critical for the functioning of an efficient capital market" and indicate that this disclosure consists of both regulatory filings and voluntary communication including management forecasts and conference calls. This view is consistent with the theoretical foundation of the semi-strong Efficient Market Hypothesis, proposed by Fama (1970), which states that stock prices reflect all publicly available information. Brown et al. (2004) provides supporting evidence for this by showing that earnings calls reduce information asymmetry and enhance stock price efficiency. A significant number of financial economists and statisticians came to believe that stock prices might be at least partially predictable. Economists and psychologists in the field of behavioural finance suggest that short-term momentum in stock prices is often driven by investors' tendency to underreact to new information (Malkiel, 2003).

While much of the literature focuses on the importance of quantitative information, Henry (2006) extends the examination of market reaction by incorporating a verbal predictor variable. The study shows that the narrative elements such as tone, length and complexity contribute significantly to the predictive accuracy beyond the inclusion of financial information. Building on this line of research, Price et al. (2012) examined the additional informational value of earnings calls and the related market responses. Their findings suggest that the sentiment expressed during these disclosures significantly predicts abnormal stock returns and trading volume, with effects that exceed those of earning surprises over a 60-day window. Furthermore, the tone in the question-and-answer session offers significant explanatory power for the observed post-earnings-announcement drift. This is supported by a recent study by Medya et al. (2022), which finds that the sentiment and semantic content of earnings call transcripts significantly predict stock price movements, outperforming financial metrics. Their findings align with the fundamental research by Matsumoto et al. (2011). While the informational value of sentiment and tone in earnings calls is well established, extracting and interpreting this qualitative content presents a different challenge. Yeruva et al. (2020) observed that sentiment in text is often ambiguous, making it difficult for human annotators to interpret it consistently and accurately.

Recent research by Hassan et al. (2024) emphasizes the growing importance of earnings call transcripts, since they offer a direct insight from executives, analysts and market participants. Their findings indicate that FSA can be an essential tool for forecasting and they point out to the broad potential for innovative applications of textual analysis in the field of finance. These include improving economic surveillance, monitoring risk across firms, industries, and countries, and studying how markets react to shocks and adjust their expectations about the future. Jayaraman and Dennis (2020) classified sentiment from earnings conference calls to develop machine learning models that predict the direction of stock price movement following earnings announcements. Their lexicon-based approach generated four sentiment features, which achieved an accuracy of 73% in forecasting stock price direction. In addition, an OLS regression confirmed that earnings call sentiment is a significant predictor of the percentage change in stock prices after such announcements. While their method achieved strong predictive performance, more advanced NLP models may be able to capture sentiment more effectively and accurately.

2.3 Deep Learning and Large Language Models

Deep learning is a specialised area within machine learning that enables computers to learn from data and capture knowledge through multilayered structures. These layers enable the model to recognise complex patterns by integrating simpler concepts, reducing the reliance on explicitly programmed rules (Goodfellow et al., 2016). Deep learning methodologies have significantly advanced the field of natural language processing, particularly in the field of sentiment analysis. Techniques such as deep neural networks (DNN), convolutional neural networks (CNN), and recurrent neural networks (RNN) have demonstrated improved accuracy over traditional machine learning methods by automatically learning and extracting features from raw data. These techniques reduce the need for manual feature selection, thereby improving the overall accuracy and robustness of sentiment classification (Dang et al., 2020).

Minh et al. (2018) proposed a deep learning framework aimed at predicting stock trends using a Two-stream Gated Recurrent Unit (TGRU) model. Their approach analyses financial news from Bloomberg and Reuters, integrates sentiment-aware word embeddings (Stock2Vec) and includes financial indicators. The model shows its robustness and achieved an overall accuracy of 66.32% on the S&P-500 dataset. Additionally, trading simulations confirmed the effectiveness of the model, highlighting its practical value for supporting investment decisions. However, one notable limitation of the model is its complexity, which leads to high computational resources and a long training time.

Jing et al. (2021) introduce a hybrid deep learning framework that combines sentiment analysis with time series forecasting to predict the one-day-ahead closing prices. Their approach integrates a CNN to extract investor sentiment with a long short-term memory (LSTM) network for price prediction. The results indicate that this combined method achieves greater predictive accuracy compared to standalone models. Although CNNs and RNNs have contributed significantly to advancements in sentiment

analysis and financial forecasting, these methods often present challenges in terms of training efficiency and scalability (Seo et al., 2020).

The Transformer model, introduced by Vaswani et al. (2017), brought a fundamental change to natural language processing by relying entirely on a self-attention model to encode and decode syntactic and semantic information, achieving new state-of-the-art results with significantly reduced training time. A prominent transformer-based language model is the Bidirectional Encoder Representations from Transformers (BERT), which is pre-trained on a large corpus of text from Wikipedia and books. BERT employs a bidirectional attention mechanism that considers context from both directions and is trained using masked language modelling and next-sentence prediction. It has achieved state-of-the-art results on eleven NLP benchmarks, outperforming previous models, including OpenAI's GPT. Since BERT is empirically powerful and conceptually simple, it enables efficient fine-tuning across a wide range of NLP tasks (Devlin et al., 2019). Based on this framework, Araci (2019) pre-trained and fine-tuned the BERT base model on a financial corpus. The resulting model, known as FinBERT, is more effective at capturing financial sentiment and semantics, outperforming both traditional machine learning and other deep learning approaches on financial sentiment benchmarks. It achieved state-of-the-art performance, with an accuracy improvement of 15% over previous classification models. This demonstrates the effectiveness of domain-specific pre-training of language models since they require fewer labelled examples. Araci also points out that their work can be extended by using FinBERT directly with market return data on financial news. Recent work by Chin and Fan (2023) applied both dictionary-based approaches and FinBERT to extract sentiment from earnings call transcripts. Their findings indicate that FinBERT is consistently outperforming traditional methods, and incorporating FinBERT-based sentiment features can enhance investment strategy performance.

Guo et al. (2023) compared the performance of encoder-only large language models (LLMs), such as BERT and the financial variant FinBERT, with decoder-only models like ChatGPT. Their findings indicate that, while certain decoder-only LLMs demonstrate strong performance across most financial tasks through zero-shot prompting, they generally lag behind the level op performance achieved by fine-tuned expert models, particularly when it comes to proprietary datasets. This is supported by Shen and Zhang (2024) who investigate the use of LLMs and FinBERT for FSA of financial news articles and reports. They claim that FinBERT outperformed GPT-3.5 and GPT-4o in terms of accuracy, precision and F1 score. However, they also highlight that general-purpose LLMs show considerable promise for FSA when enhanced with prompt engineering, particularly in few-shots scenarios. In few-shot scenarios, the model is provided with a limited number of examples within the prompt to guide its understanding of the task. This contrasts with zero-shot prompting, where the model relies on its pretraining without any examples. In contrast, Fatouros et al. (2023) presented a more optimistic view of general-purpose LLMs, demonstrating that ChatGPT-3.5 can outperform FinBERT in short-term sentiment

classification. Their study employed a zero-shot prompting approach, resulting in an outperformance of 36% higher correlation with market returns and 35% in sentiment classification accuracy.

The existing literature shows a shift from traditional lexicon-based methods to more advanced deep learning models and LLMs in financial sentiment analysis. While traditional models remain valuable for their interpretability and transparency, the strong performance of FinBERT and GPT-4 highlights a significant potential. This study compares these models in extracting sentiment from earnings calls transcripts and assessing their effectiveness in predicting stock price movements.

3. Data

This chapter describes the data used in this study. It covers the sources of the datasets, the selection criteria of the sample, the preprocessing steps taken to prepare the data for analysis and the limitations of the final dataset. An overview of the data sources utilised in this study is provided in **Table 3.1**. In addition, a summary of the sample characteristics can be found in **Table 3.2**.

3.1 Data Collection

3.1.1 Earnings Call Transcripts

The primary data source for this study consists of quarterly earnings call transcripts, retrieved from LSEG Workspace (formerly Refinitiv Workspace). The dataset includes transcripts of 60 publicly traded companies listed in the Standard & Poor's 500 Index (S&P 500) as of 1 June 2025, covering the period from Fiscal Year 2010-Q1 to Fiscal Year 2023-Q4. To capture the market reaction, the overall sample period spans from January 1, 2009 to July 1, 2024. Due to the computational cost and processing time associated with financial sentiment analysis, a targeted sample of 60 firms was selected based on the availability of complete transcript records, industry representation and trading volume. This sampling strategy captures a broad representation across different sectors and time periods. Each earnings call typically consists two main components: the prepared remarks, usually presented by company executives, and the Question-and-Answer session with analysts. To enable a more detailed investigation into the predictive value of each section, the transcripts were segmented into the individual parts. This separation allows for a comparative analysis between the full earnings call, prepared remarks and Q&A, examining how sentiment in each segment relates to market reactions. In total, 3,360 full transcripts were collected. After preprocessing, the average number of sentences per transcript is approximately 530, resulting in a corpus of over 1.79 million individual sentences available for sentiment classification.

3.1.2 Loughran-McDonald Master Dictionary

As a benchmark for sentiment classification, this study employed the Loughran-McDonald Master Dictionary (version February 2024), a financial domain specific lexicon developed to improve the accuracy of textual sentiment analysis in financial contexts (2011). The dictionary was retrieved from the official source (The Notre Dame Software Repository for Accounting and Finance) and only words marked as *Positive* or *Negative* were retained for analysis.

3.1.3 Market and Earnings Data

In addition to the earnings call transcripts, this study incorporates key financial indicators that are used for the abnormal returns calculations and as control variables in the regression analysis.

Stock return data was retrieved from the CRSP (Center for Research in Security Prices) database. In addition, the daily return of the S&P 500 index was collected to serve as a market benchmark. These variables are used to compute the cumulative abnormal returns (CARs) in the event study framework. Furthermore, the risk-free rate, represented by the 1-month U.S. Treasury Bill rate was obtained from

the Kenneth R. French Data Library. The risk-free asset was included in the portfolio to complement stock positions whenever the model advised a partial allocation to stocks.

To control for factors that may influence stock returns independently of sentiment, two additional variables were included:

- **Earnings surprise**, obtained from the I/B/E/S database, which accounts for the degree to which reported earnings deviate from analysts expectations. This controls for the direct quantitative content of the earnings release.
- **Firm size**, calculated as the natural logarithmic function of market capitalization (closing price multiplied by shares outstanding), controls for differences in analyst and investors attention and information processing across firms of different sizes.

Table 3.1: Overview of Data Sources

Data Type	Description	Source
Transcripts	Earnings call transcripts	LSEG Workspace
Loughran-McDonald Master Dictionary	Financial sentiment lexicon	The Notre Dame Software Repository for Accounting and Finance
Stock returns	Daily returns for individual stocks to compute CARs	CRSP
S&P 500 Index	Benchmark index for the market return	CRSP
Earnings Surprise	Standardized Unexpected Earnings (SUE)	I/B/E/S
Firm Size	Log of market capitalization	CRSP
Risk-free rate	1-Month U.S. Treasury Bill Rate	Kenneth R. French Data Library

Table 3.2: Sample Characteristics

Characteristic	Details
Number of firms	60
Fiscal period covered	FY 2010-Q1 to FY 2023-Q4
Time period	January 2009 to July 2024
Number of transcripts	3,360
Transcript segments	Full Transcript/ Presentation/ Q&A
Total sentences after preprocessing	~ 1.79 million
Average sentences per transcript	~ 530

3.2 Preprocessing and Text Cleaning

Before performing the sentiment analysis, multiple cleaning and preprocessing steps were applied to the raw transcripts:

- **Removal of metadata and headers/footers:** All non-content blocks added by Refinitiv were removed. This included legal disclaimers, copyright statements, participant listings, and editorial annotations.
- **Exclusion of speaker and operator labels:** Identifiers such as "Operator," "Speaker," or names before each turn were stripped to reduce textual noise and avoid attribution bias in sentiment counts.
- **Split into Presentation and Questions-and-Answers:** The segments were split based on structural markers in the transcript. This enables separate sentiment analysis of these segments of the earning calls.
- **Parsing based on Loughran and McDonald's procedure:** Following a similar approach to Loughran and McDonald (2011), the remaining text was cleaned by removing all numbers, special characters, and isolated single-letter tokens to ensure a clean word-level analysis.
- **Tokenisation:** The cleaned text was finally tokenized into individual words, creating sentiment-ready structure that allowed accurate classification using the Loughran-McDonald sentiment dictionary.

This preprocessing ensured a standardized and noise-free input for sentiment analysis, enhancing the accuracy and consistency of results across all transcripts.

3.3 Dataset Summary and Limitations

The dataset represents a collection of both qualitative and quantitative financial data. It covers 14 fiscal years of data and includes high-quality transcripts and market data from widely used financial databases. However, the dataset is not without limitations:

- Transcripts vary in length and structure, especially across industries and time periods;
- Automated sentence segmentation and cleaning may introduce minor inconsistencies.

Despite these challenges, the dataset offers a robust foundation for the sentiment analysis measures and for evaluating their predictive power with respect to stock returns. A comprehensive overview of the descriptive statistics for the key variables is provided in **Table 3.3**. In addition, the distribution of the CARs is visualised in **Appendix Figure B.1**, providing further insight into the characteristics of the abnormal return data.

Table 3.3: Summary Statistics

Variable	N	Mean	Median	SD	Min	Max
CAR[-3,-1]	3,360	0.00	0.00	0.03	-0.34	0.45
CAR[0,+1]	3,360	0.00	0.00	0.05	-0.28	0.39
CAR[0,+3]	3,360	0.00	-0.00	0.06	-0.31	0.45
CAR[0,+5]	3,360	0.00	0.00	0.06	-0.29	0.43
Full Transcript LM-Sentiment Normalized	3,360	0.01	0.01	0.01	-0.02	0.03
Presentation Segment LM-Sentiment Normalized [1]	3,341	0.01	0.01	0.01	-0.02	0.04
Q&A Segment LM-Sentiment Normalized [2]	3,348	0.00	0.00	0.00	-0.02	0.03
Full Transcript GPT-Sentiment	3,360	0.38	0.50	0.25	-0.50	0.85
Presentation Segment GPT-Sentiment [1]	3,341	0.56	0.60	0.18	-0.50	0.90
Q&A Segment GPT-Sentiment [2]	3,348	0.31	0.40	0.20	-0.50	0.75
SUE score [3]	3,352	2.08	1.52	3.87	-33.96	89.81

[1] **Note:** 19 observations are missing due to the absence of a presentation segment in the earnings calls and transcripts.
[2] **Note:** 12 observations are missing due to the absence of a Q&A segment in the earnings calls and transcripts.
[3] **Note:** 8 observations are missing for the SUE score as the data was unavailable.

4. Hypotheses Development

This study examines the relationship between corporate communication in earnings call transcripts and short-term price predictions. Building on the foundational work of Loughran and McDonald (2011), who introduced a financial-domain sentiment dictionary, prior research has shown that sentiment in earning-related disclosures influences market returns. While the Loughran-McDonald dictionary remains a benchmark in financial textual analysis, recent advancements in natural language processing, particularly LLMs such as GPT-4o, allow for more nuanced sentiment interpretation. This study combines both traditional lexicon-based and modern transformer-based techniques to assess their effectiveness in explaining cumulative abnormal returns (CARs) following earning announcements. Furthermore, this study goes beyond the traditional event study framework by examining the practical application of sentiment extraction in earnings calls. This study aims to develop a sentiment-based long-short portfolio and evaluate its potential to generate a statistically and economically significant alpha, relative to the passive market index.

Based on these objectives, the following main hypothesis is formulated:

Hypothesis 1 (H1):
Financial sentiment extracted from earnings call transcripts is significantly related with short-term cumulative abnormal returns (CARs) after the earnings event.

The main hypothesis can be further dissected in five supporting hypotheses:

Hypothesis 2 (H2):
Net sentiment scores derived using the Loughran-McDonald dictionary are significant predictors of CARs in the post-earnings period.

Hypothesis 3 (H3):
GPT-based sentiment scores provide additional explanatory power for CARs beyond traditional Loughran-McDonald sentiment methods.

Hypothesis 4 (H4):
Sentiment in the Q&A section of the earnings call is more predictive of CARs than sentiment in the prepared remarks section.

Hypothesis 5 (H5):
Even after controlling for earnings surprises, sentiment variables remain significant in explaining post-earnings stock price movements.

Hypothesis 6 (H6):
A long-short portfolio strategy constructed using GPT sentiment-based investment recommendations generates a statistically and economically significant alpha relative to the passive market benchmark.

The hypotheses are tested using a combination of an event study framework, sentiment scores derived from the Loughran-McDonald dictionary and GPT-4o, traditional financial metrics and a long-short portfolio. The following chapter presents the methodological framework to conduct this analysis.

5. Methodology Event Study

This chapter outlines the methodological framework used to examine the relationship between earnings call sentiment and stock price movements. The methodology can be divided into three sections: financial sentiment analysis techniques, sentiment score and investment recommendation extraction and event study design.

5.1 Financial Sentiment Analysis (FSA) Techniques

This study uses two sentiment analysis methodologies to quantify the tone of quarterly earnings calls: a dictionary-based approach using the Loughran-McDonald Master Dictionary and a large language model-based approach leveraging GPT-4o. These sentiment scores serve as primary explanatory variables is the regression analysis.

5.1.1 Loughran-McDonald

To quantify sentiment in earnings call transcripts, this study first employs a lexicon-based technique using the Loughran-McDonald (2011) dictionary, which is a finance-domain specific dictionary, specifically designed for financial contexts. The dictionary provides predefined classifications of words as positive, negative, uncertainty, litigious and modal, allowing a straightforward approach for sentiment classification by matching words in the transcripts with the lexicon. This study focuses only on the *Positive* and *Negative* word categories defined in the dictionary. Words categorized under other categories were excluded, to ensure focus on the directional sentiment of the transcript. This classification aligns with the study's objective of examining whether positive or negative linguistic tone during earnings calls, can predict short-term market movements.

Each earnings call transcript was pre-processed according to the procedures described in Chapter 3, which included removing metadata, segmenting the presentation and Q&A, and tokenizing the cleaned text to individual word units. The tokenized words were cross-referenced against the Loughran-McDonald dictionary. A total count of positive and negative words for each transcript was collected, forming the basis for sentiment classification.

Following the above described classification of tokenized words, the transcript-level sentiment scores were computed to quantify the overall sentiment of each individual earnings call. Sentiment for each transcript was computed as:

$$Net\ Sentiment_i = Count_{positive,i} - Count_{Negative,i} \quad (1)$$

To account for transcript length differences and to ensure comparability across transcripts, the net sentiment score was normalized by the total number of words, resulting in a standardized sentiment score:

$$Normalized\ Net\ Sentiment_i = \frac{Net\ Sentiment_i}{Total\ Words_i} \quad (2)$$

These normalised net sentiment scores were calculated separately for the full transcript, the presentation section and the Q&A section. The split in segments allows for empirical testing of whether tone in the prepared statements or the analyst Q&A contributes more significantly to stock price movements. The resulting net sentiment scores serve as a key explanatory variable in the regression framework detailed in the third section of this chapter.

5.1.2 GPT-4o

To enhance and potentially improve the dictionary-based approach, this study employs GPT-4o for sentiment classification. GPT-4o is a transformer-based large language model capable of understanding financial tone, nuance and context beyond keyword matching. Sentiment is derived using a zero shot prompt-based approach in which GPT-4o is asked to analyse each earnings call (or segment) on a scale from -1 to +1. This continuous score reflects both the directional and intensity of the sentiment. In addition, a investment recommendation (strong sell, sell, hold, buy, strong buy) based on the sentiment score is generated for potential use in the portfolio simulation. The use of GPT-4o enables the incorporation of context and complex language patterns that would not be captured by dictionary methods.

5.2 Sentiment Score and Investment Recommendation Extraction

The following section describes the sentiment score and investment extraction with ChatGPT, which is the most innovative and experimental section of this study. The ability of LLMs to analyse large amounts of unstructured textual data will be leveraged to create a more comprehensive view of the financial information in earnings calls. A Python script was developed to send a call to the OpenAI API in order to extract the sentiment scores and investment recommendations from the transcripts. The pipeline begins with importing all the earnings call transcripts and the developed prompt. The script was developed to loop over the transcripts and to send the call with the set parameters and prompt to the OpenAI API. The JSON-structured output from the model was then parsed and saved as a CSV file type for further statistical analyses.

5.2.1 OpenAI Model

This study uses Open AI's API to implement the sentiment analysis at scale. The Python script allowed the batch processing of the 3,360 earnings call transcripts and was used to set the level of certain parameters. These parameters can be used to influence the quality and format of the model's output. The following section will discuss the main parameters of the OpenAI API that were essential in the GPT-extraction:

- **Model:** The model specifies which version of OpenAI's GPT is in use. There are multiple model versions which all differ in intelligence, speed and price. This study selected the "GPT-4o mini

model" to have a fast and affordable model for focused tasks. The *Chat Completions* endpoint is used to run the analysis. This model is selected for it's computational cost efficiency and fast response time.

- **Temperature:** The temperature parameter controls the randomness of text that is generated by the model. A lower temperature (close to 0.0) makes the model more consistent by prioritizing output tokens with the highest probability, which results in a more precise and accurate output. On the other hand, a higher temperature introduces more randomness by increasing the chances of selecting less probable output tokens, thereby generating a more creative and varied output. For this study, the temperature was set to 0.0 to ensure consistency and accuracy. Minimising the randomness of the model is important in financial sentiment analysis because this reduces the risk of inconsistent sentiment scores across different earnings call transcripts.
- **Max_Tokens:** The max tokens parameter sets the maximum length in total # of tokens that the model is allowed to generate. This parameter can be used to manage the length of the model's response and can prevent excessive long or irrelevant response. It could be used to control the computational cost and processing time. In this study, the max tokens parameter was not defined since the developed prompt has clear instructions on the preferred output.
- **Top_P:** The Top P parameter is also know as nucleus sampling. It is another technique to control for randomness of the model, but instead of selecting the output with the highest probability, it selects an output from an subset with the most probable output whose cumulative probability exceed a specified threshold (e.g. 95%). This allows the model to keep some consistency, while allowing for some degree of variation. In this study, Top P was not explicitly set and therefore defaulted to a value of 1.0. However, since the temperature was set to 0.0, the Top P setting had no practical effect on the model or the output.

Each transcript was submitted to the model through a structured API call consisting of the following two roles:

- **System message:** The system message can be used to give the model specified instructions. This study used the system message for context setting, behavioural guidelines, response style and operational constraints.
- **User message:** The user message can be used to instruct the model a specific task. This study used the user message for specific contextual information, in the form of the transcript and a response structure for the output of the model.

5.2.2 Prompt Engineering Techniques

Prompt engineering is the process of designing and optimising input prompts to effectively guide a language model's response. It is essential to the process of interacting with LLMs, since it directly influences the quality and accuracy of the model to generate relevant, precise and valuable output. This section outlines the key prompting techniques used in this study.

- **Zero-shot:** Zero-shot prompting relies on the pretraining of the LLM and its capability of generating suitable outputs. In contrast to few-shot prompting techniques, the model is not provided with examples of expected output. Key advantages of zero-shot prompting techniques include the simplicity, ease of use and flexibility. The technique can be used without requiring additional labelled training datasets, making it especially useful for the application of new domains and for capturing sentiment that varies across transcripts from different industries, company contexts or communication styles.
- **Role specification:** Role-based prompting involves instructing the model to "act as" or "be" a specified professional, character or personality. These instructions impact the model's tone, style and content to align with the expectations of that role, resulting in more effective and professional analysis. In this study, the model was prompted to act like a hedge fund analyst, ensuring that the outputs reflected the reasoning style and critical thinking of a professional analyst.
- **Chain of Thought (CoT):** Chain of Thought techniques enhance the performance of LLMs on complex tasks by guiding them through a step-by-step reasoning. This technique improves the accuracy, transparency and multistep-reasoning abilities of the model. In this study, CoT techniques were used to enable more comprehensible analysis of sentiment by simulating a human-like problem solving process. The prompt was structured to guide the model through different reasoning steps such as evaluating tone, assessing strategic clarity and applying judgment-based scoring.

5.2.3 Prompt Development

The final prompt used in the study is the following:

```
Act like a hedge fund analyst with expertise in financial NLP, investment analysis, and sentiment-driven quantitative modelling. You specialize in carefully reviewing corporate earnings call transcripts to extract forward-looking sentiment scores that reflect management's tone, strategic direction, and risk signals. These scores will later be used to support short-term trading and investment decisions.

Objective: Your task is to perform sentiment analysis on earnings call transcripts by closely examining how company executives communicate. Focus on what their language reveals about business outlook, financial performance, and confidence in strategy.

Instructions:

Step 1: Process each transcript by dividing it into the following sections:
- Full Transcript (the entire transcript)
- Presentation Section (prepared remarks from executives)
- Q&A Section (analyst questions and management responses)
```

If a section is not present in the transcript, mark both the Sentiment Score and Investment Recommendation as "NA" for that section.

Step 2: For each segment, perform a deep qualitative assessment, focusing on:
- Language indicating financial strength or weakness (e.g., revenue trends, margins, guidance)
- Tone of future outlook (confidence, certainty, hedging, vagueness)
- Clarity and credibility of strategic articulation (growth plans, innovation, restructuring)
- Presence of risk signals (macro concerns, regulatory issues, hesitation, defensive tone)
- In the Q&A section, also consider analyst tone: if analysts express doubt, highlight risks, or challenge management, treat this as a negative input; confident analyst tone or bullish questioning can support a more positive sentiment, especially when it reinforces management's narrative.[4]
- Be sceptical of overly polished or vague optimism.[5] Promotional language without clear backing should be treated as neutral or even a risk signal. Read between the lines like a buy-side analyst who discounts promotional language unless supported by fundamentals.[6]

Step 3: For each segment, assign two evaluations:

1. Sentiment Score:
A continuous value between -1.0 and +1.0 (rounded to two decimal places) that reflects the overall tone, conviction, and clarity of communication. This is a confidence-weighted and risk-adjusted signal:
- Use values close to -1.0 for strongly negative, defensive, or vague language that reveals concerns or lack of clarity.
- Use values near 0.0 when the tone is mixed, non-committal, promotional without substance, or uncertain.
- Use values close to +1.0 only when communication is explicit, confident, clearly forward-looking, and free from hedging or ambiguity - whether from management or analysts.

2. Investment Recommendation:
A categorical decision derived from the sentiment score, but based on judgment, not mechanical thresholds. Use the following scale:

[4] See Matsumoto et al. (2011) for evidence on the informational value of analyst questions in the Q&A.
[5] See De Amicis et al. (2021) who find that greater vagueness in conference calls reduces their informational value and weakens market reactions.
[6] See Huang et al. (2014) for evidence that promotional language not backed by fundamentals can mislead investors.

```
-1.0 (Strong Sell) - Overwhelmingly negative, poor outlook, or serious risk indicators
-0.5 (Sell) - Generally negative tone, weak positioning, or cautious language outweighing positive signs
0.0 (Hold) - Mixed, unclear, promotional-but-unsupported, or lacking directional conviction
0.5 (Buy) - Positive signs supported by credible, confident, and strategic forward-looking language
1.0 (Strong Buy) - Highly optimistic and unambiguous tone, strong evidence of execution strength and strategic clarity

Only assign a Buy or Strong Buy when positive sentiment is both clear and supported, not just implied through vague optimism.

Step 4: Do NOT use any external financial data, keyword dictionaries, or market context. Base all judgments solely on the transcript's internal language. Think like a buy-side analyst reading between the lines for implicit signals and subtext.

Step 5: Output format:
Use only numeric code, not labels (e.g., use 0.0, not "Hold").
Provide your response as a JSON object with the following fields:
{
  "Section": "[Full Transcript / Presentation / Q&A]",
  "Sentiment Score": -1.0 to 1.0 or "NA",
  "Investment Recommendation": -1.0 / -0.5 / 0.0 / 0.5 / 1.0 or "NA".
}

Maintain a professional and analytical tone. Focus on forward-looking language, management confidence, and real strategic signalling. Approach each section with methodical and risk-aware judgment.

Approach this task step-by-step with careful analysis.
```

5.3 Event Study Design

To evaluate the impact of earnings call sentiment on stock price movement, this study employs a short-horizon event study framework centred on each quarterly earnings calls. This study estimates abnormal returns, which are stock price movements that deviate from the expected performance based on a benchmark, and employs these as the dependent variable in the regression analysis (de Jong & de Goeij, 2011).

5.3.1 Event Definition and Window

The event date (t = 0) is defined as the trading day on which the earnings call took place, or the next available trading day if the call occurred outside the market hours. To capture short-term investor reactions, cumulative abnormal returns (CARs) were calculated over multiple post-event windows:

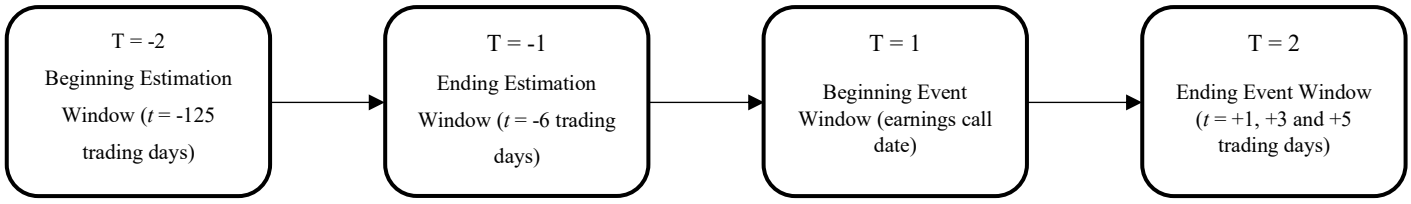
- **[-3,-1]:** The pre-event window capturing abnormal returns prior to the earnings call. This period serves as a benchmark to detect any potential information leakage or anticipatory price movements ahead of the earnings call.
- **[0, +1]:** The immediate two-day window (event day and the following trading day) capturing the immediate price reaction.
- **[0, +3]:** A short-term window (event day plus three trading days) to evaluate post-event drifts and information incorporation into stock prices.
- **[0, +5]:** An extended short-term window (event day plus five trading days) allowing a full week of trading to asses if sentiment effects continue to influence stock prices.

These windows enable the analysis of how strongly and quickly the market incorporates the sentiment from an earnings call.

5.3.2 Estimation Window

In line with the classical event study methodology proposed by MacKinlay (1997), an estimation window of 120 trading days is employed to calculate the "normal" return of each firm. These expected returns from the benchmark for identifying abnormal returns centred around the event. To mitigate the potential risk of information leakage or pre-event drift affecting the estimation, a 5-trading day buffer is implemented between the estimation window and the event window. As a result, the estimation window is defined as the period from t = -125 to t = -6, where t = 0 marks the event date. This buffer helps to ensure that early market reactions to anticipated announcement do not influence the estimation of normal stock returns.

Figure 5.1: Timeline of the Event Study Framework



5.3.3 Cumulative Abnormal Returns (CARs)

The market model is used to calculate the CARs around the earnings call date. This model assumes that individual firm stock returns are systematically related to returns of the overall market. The parameters α_i and β_i are estimated using the 120-day estimation window prior to the event. The formula to compute these parameters is the following:

$$R_{i,t} = \alpha_i + \beta_i R_{m,t} + \varepsilon_{i,t} \quad (3)$$

where:

- $R_{i,t}$ is the daily stock return of firm i on day t;
- α_i is the intercept term of firm i (alpha);
- β_i is the sensitivity to the market return of firm i (beta);
- $R_{m,t}$ is the return of the market index on day t;
- $\varepsilon_{i,t}$ is the error term of firm i on day t.

This OLS regression is run over the estimation window, where the S&P 500 index is used as the benchmark market in estimating these parameters. The resulting estimates are then used to predict the expected return on the event date and following dates, under the null hypothesis of no abnormal effects. The expected (normal) returns for stock *i* on day *t* in the event window is calculated using the following formula:

$$E[R_{i,t}] = \hat{\alpha}_i + \hat{\beta}_i R_{m,t} \quad (4)$$

where:

- $E(R_{i,t})$ is the expected return of firm i on day t;
- $\hat{\alpha}_i$ is the estimated intercept (alpha) for firm i;
- $\hat{\beta}_i$ is the estimated slope coefficient (beta) for firm i;
- $R_{m,t}$ is the actual return of the market index on day t.

Abnormal Returns (AR) are computed to measure the deviation between a stock's actual return and the expected return based on the market model. This isolates the effect of firm-specific events by removing the market-wide influences. For each firm i and each day t in the event window, the abnormal returns are defined as the actual return minus the expected return from the market model, by using the following formula:

$$AR_{i,t} = R_{i,t} - E(R_{i,t}) \quad (5)$$

where:

- $AR_{i,t}$ is the abnormal return of firm i on day t;
- $R_{i,t}$ is the actual return of firm i on day t;
- $E(R_{i,t})$ is the expected return of firm i on day t based on the market model.

Cumulative Abnormal Returns (CAR) is the sum of all abnormal returns over the event window, which captures the total impact of an event on the stock price. For each firm i and event window starting from day T_1 until day T_2, the cumulative abnormal returns were computed using the following formula:

$$CAR_i[T_1, T_2] = \sum_{t=T_1}^{T_2} AR_{i,t} \quad (6)$$

where:

- $CAR_i[T_1, T_2]$ is the cumulative abnormal return of firm i from day T_1 to day T_2;
- $\sum_{t=T_1}^{T_2} AR_{i,t}$ is the sum of abnormal returns from day T_1 to day T_2.

5.3.4 Regression Analysis

With the abnormal returns quantified through CARs, this study examines whether the sentiment in earnings calls predicts these abnormal returns. A cross-sectional regression model is employed with the CAR for each earnings call event as dependent variable and the key independent variables are the sentiment scores from the transcript (both from the Loughran-McDonald dictionary approach and GPT-4o analysis). The regression framework will test the explanatory power of the derived sentiment while controlling for standardized unexpected earnings, using the following key models:

$$CAR_i[T_1, T_2] = \beta_0 + \beta_1 Sentiment_i^{LM} + \gamma SUE_i + \varepsilon_i \quad (7)$$

$$CAR_i[T_1, T_2] = \beta_0 + \beta_1 Sentiment_i^{GPT} + \gamma SUE_i + \varepsilon_i \quad (8)$$

$$CAR_i[T_1, T_2] = \beta_0 + \beta_1 Sentiment_i^{LM} + \beta_2 Sentiment_i^{GPT} + \gamma SUE_i + \varepsilon_i \quad (9)$$

where:

- $CAR_i[T_1, T_2]$ is the cumulative abnormal return of firm i from day T_1 to day T_2;
- β_0 is the intercept term, capturing the baseline level of CAR_i;
- β_1 and β_2 are the sensitivity of CAR_i to the sentiment score;
- $Sentiment_i^{LM}$ is the LM-dictionary based sentiment score for firm i;
- $Sentiment_i^{GPT}$ is the GPT-4o based sentiment score for firm i;
- γSUE_i is the standardized unexpected earnings of firm i;
- ε_i is the error term of firm i.

5.3.5 Robustness Checks

To ensure the robustness of the regression models, multiple diagnostic tests were performed. These tests evaluate multicollinearity across explanatory variables and examine the relationship between sentiment measures and control variables. First, cross-correlations between the Loughran-McDonald sentiment scores, GPT-4o sentiment scores, and the standardized unexpected earnings (SUE) were calculated for each transcript segment. This analysis assesses whether the sentiment measures capture additional textual insights or whether they overlap with each other or with the quantitative earnings information.

Second, Variance Inflation Factors (VIFs) were computed for all explanatory variables, including sentiment scores, SUE, and firm size, to assess multicollinearity in the regression models. As long as all

VIF values remain below the commonly accepted threshold of 10, multicollinearity is unlikely to bias coefficient estimates or inflate standard errors. If the diagnostic tests indicate no evidence of variable collinearity, the statistical results can be interpreted with greater confidence.

The robustness of the event study results is supported by the analysis of alternative event windows and examining pre-event abnormal returns, which contribute to the reliability of the findings.

6. Results Event Study

This chapter presents the empirical results of the event study analysis. It examines the relationship between earnings call sentiment and short-term stock market reactions. The results are structured by different sentiment measures, transcript sections, and event windows, providing a comprehensive analysis of the research hypotheses.

6.1 Regression Results

6.1.1 Full Transcript Results

The first set of regression models examines the relationship between sentiment scores extracted from the full earnings call transcript and short-term cumulative abnormal returns (CARs). This analysis tests the first three hypotheses (H1, H2, and H3), assessing the predictive power of the Loughran-McDonald dictionary-based sentiment and the GPT-4o derived sentiment scores.

Table 6.1 (columns 1 and 3) presents the baseline regression results using the full transcript sentiment scores as the only explanatory variable for the CAR[0,+1] event window. Both LM-based and GPT-based sentiment scores demonstrate a statistically significant positive association with CARs. However, GPT-derived sentiment consistently demonstrates higher statistical significance levels and larger t-values compared to the LM-based scores. Specifically, the GPT sentiment coefficient for CAR[0,+1] is 0.0408 ($p < 0.001$), with a t-value of 11.31, whereas the corresponding LM coefficient is 1.054 with a t-value of 6.41. This suggests that while both models capture market reactions to sentiment, GPT-derived scores provide a more robust explanatory signal.

Furthermore, the explanatory power of the models, as captured by the Adjusted R-squared, is considerably higher for GPT-sentiment. For CAR[0,+1], the Adjusted R-squared of the GPT-model reaches 3.64%, while the LM-model achieves only 1.18%. This relative improvement is maintained across longer event windows, although overall explanatory power declines for the CAR[0,+5] window, consistent with information being quickly incorporated by the market.

To control for the quantitative information included in earnings announcements, the SUE score was added (columns 2 and 4 of **Table 6.1**). Incorporating SUE improves the model fit, with the Adjusted R-squared for the GPT-based sentiment rising to 6.99% for CAR[0,+1]. Both sentiment and SUE coefficients remain statistically significant, supporting the independent explanatory power of textual sentiment beyond quantitative earnings surprises. Importantly, GPT-sentiment continues to outperform the LM-sentiment model, both in terms of coefficient significance and model fit.

Extended results for longer event windows are provided in the Appendix. **Appendix Table A.1** presents the regression outcomes for CAR[0,+3], while **Appendix Table A.2** displays the results for CAR[0,+5].

Overall, these results provide strong support for the hypotheses H1, H2 and H3. First, the results confirm that financial sentiment extracted from earnings call transcripts is significantly related with short-term

CARs around the earnings event, thereby supporting H1. Second, the Loughran-McDonald sentiment scores are statistically significant which supports H2. Finally, the GPT-derived sentiment demonstrates more explanatory power for CARs relative to the traditional LM-approach. It consistently outperforms across all event windows and segments, which provides empirical support for H3. These results indicate that advanced language models capture relevant market information more effectively than traditional dictionary approaches.

Table 6.1: Regression Results - Full Transcript Sentiment for CAR[0,+1]

Dependent Variable: CAR[0,+1]				
Variables	**(1) LM-sentiment**	**(2) LM-sentiment with SUE**	**(3) GPT-4o Sentiment**	**(4) GPT-4o Sentiment with SUE**
Sentiment	1.0540***	0.7065***	0.0408***	0.0327***
	(0.164)	(0.1639)	(0.0036)	(0.0036)
SUE		0.0028***		0.0026***
		(0.0002)		(0.0002)
Intercept	-0.0066***	-0.0099***	-0.0144***	-0.0167***
	(0.0015)	(0.0015)	(0.0016)	(0.0016)
Adjusted R^2	0.0118	0.0526	0.0364	0.0699
Observations	3360	3352	3360	3352
Model Controls	None	SUE	None	SUE

Significance levels: *** p < 0.001, ** p <0.01, * p <0.05
Standard errors in parentheses

6.1.2 Sectional Analysis

To further investigate the sources of informative content within earnings calls, sentiment scores were analysed separately for the presentation and Q&A sections of the transcripts. This analysis aims to evaluate whether specific parts of the call show different explanatory power regarding short-term cumulative abnormal returns (H4).

Table 6.2 (columns 1 and 3) presents the regression results for the presentation segment sentiment, while **Table 6.3** (columns 1 and 3) reports the results for the Q&A segment. Consistent with earlier findings, both LM-based and GPT-based sentiment measures show a statistically significant positive relationship with CARs. However, the strength and explanatory power of the sentiment coefficients vary across sections.

For the presentation segment, GPT-derived sentiment continues to show statistical significance with a coefficient of 0.0411 (p < 0.001) for CAR[0,+1], accompanied by an Adjusted R-squared of 2.05%. The LM-based sentiment coefficient is also significant but with a lower explanatory power, yielding an Adjusted R-squared of only 0.65%. This suggests that the presentation section provides market-relevant information, but not as extensively as the full transcript analysis.

In contrast, the Q&A segment produces stronger results. As shown in **Table 6.3**, GPT-derived sentiment reaches a coefficient of 0.0547 ($p < 0.001$) for CAR[0,+1], with an Adjusted R-squared of 4.44%. Similarly, the Q&A segment LM-based sentiment yields the highest coefficient observed across all sections, amounting to 1.2633 ($p < 0.001$). These findings indicate that the Q&A portion of the call is the most informative in terms of predicting market reactions.

To account for quantitative information, the SUE score was added in columns 2 and 4 of both tables. The inclusion of SUE improves the model fit across both sections, yet the overall results remains consistent. The Q&A segment consistently shows higher levels of significance and greater explanatory power compared to the presentation section.

Extended results for longer event windows are provided in the Appendix. **Appendix Tables A.3 and A.4** report the presentation segment results for CAR[0,+3] and CAR[0,+5], respectively. Similarly, **Appendix Tables A.5 and A.6** present the corresponding results for the Q&A segment.

The sectional analysis supports hypothesis H4, confirming that the Q&A segment of earnings calls offers the highest informational value, as reflected in its superior predictive power for short-term CARs. Based on the results, the Q&A segment is even more informative than the full transcript, suggesting that the interaction between management and analysts provides unique insights.

Table 6.2: Regression Results - Presentation Segment Sentiment for CAR[0,+1]

Dependent Variable: CAR[0,+1]				
Variables	**(1) LM-sentiment**	**(2) LM-sentiment with SUE**	**(3) GPT-4o Sentiment**	**(4) GPT-4o Sentiment with SUE**
Sentiment	0.5335***	0.3233**	0.0411***	0.0296***
	(0.1116)	(0.1106)	(0.0049)	(0.0049)
SUE		0.0029***		0.0027***
		(0.0002)		(0.0002)
Intercept	-0.0057***	-0.0091***	-0.0221***	-0.0213***
	(0.0017)	(0.0017)	(0.0029)	(0.0028)
Adjusted R^2	0.0065	0.0506	0.0205	0.0586
Observations	3341	3333	3341	3333
Model Controls	None	SUE	None	SUE

Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$
Standard errors in parentheses

Table 6.3: Regression Results - Q&A Segment Sentiment for CAR[0,+1]

Dependent Variable: CAR[0,+1]				
Variables	**(1) LM-sentiment**	**(2) LM-sentiment with SUE**	**(3) GPT-4o Sentiment**	**(4) GPT-4o Sentiment with SUE**
Sentiment	1.2633***	0.9068***	0.0547***	0.0455***
	(0.1802)	(0.1793)	(0.0044)	(0.0044)
SUE		0.0027***		0.0025***
		(0.0002)		(0.0002)
Intercept	-0.0039***	-0.0082***	-0.0160***	-0.0183***
	(0.0012)	(0.0012)	(0.0016)	(0.0016)
Adjusted R^2	0.0142	0.0539	0.0444	0.0764
Observations	3348	3340	3348	3340
Model Controls	None	SUE	None	SUE

Significance levels: *** p < 0.001, ** p <0.01, * p <0.05
Standard errors in parentheses

6.1.3 Control Variables

To test the robustness of the sentiment-CAR relationship and account for potential influencing factors, additional control variables were introduced. Specifically, the logarithm of firm size (market capitalization) was added in addition to the SUE score. This adjustment aims to control for size-related effects that might influence stock price reactions to earnings calls.

The regression results with these controls are presented in **Table 6.4** (columns 1 and 3) for the LM-based and GPT-based sentiment scores derived from the full transcript. The inclusion of log(Size) alongside SUE improves the model fit, with the Adjusted R-squared increasing across all specifications. For the GPT-based model, the Adjusted R-squared for CAR[0,+1] increases to 7.47%, compared to 3.64% in the baseline model without controls. The LM-based model improved similarly, reaching an Adjusted R-squared of 5.50%.

Both SUE and log(Size) are statistically significant across all models, with SUE exhibiting a positive association with CARs, while firm size shows a negative relationship. The negative coefficient on size indicates that larger firms tend to experience smaller abnormal returns in response to sentiment and earnings information.

Importantly, the sentiment coefficients remain statistically significant after introducing these controls, for both LM and GPT-based models. This confirms the robustness of the sentiment effect on CARs and strengthens the argument that sentiment analysis contains unique information not captured by traditional quantitative variables.

Extended results for longer event windows are provided in the Appendix. **Appendix Table A.7** reports the results for CAR[0,+3], while **Appendix Table A.8** presents the results for CAR[0,+5]. In addition,

the corresponding presentation and Q&A segment results are reported in **Appendix Table A.9** to **Appendix Table A.14**.

Overall, the findings demonstrate that the predictive power of sentiment remains, even when controlling for firm size and earnings surprises, thereby providing support for hypothesis H5. This confirms the robustness of the results and highlights that the sentiment effect is not just driven by traditional quantitative factors like earnings surprises.

Table 6.4: Regression Results - Full Transcript Sentiment with Control Variables for CAR[0,+1]

Dependent Variable: CAR[0,+1]		
Variables	**(1) LM-sentiment with Controls**	**(2) GPT-sentiment with Controls**
Sentiment	0.6963***	0.0345***
	(0.1637)	(0.0036)
SUE	0.0028***	0.0026***
	(0.0002)	(0.0002)
log(Size)	-0.0024**	-0.0033***
	(0.0008)	(0.0008)
Intercept	0.0504**	0.0659***
	(0.0195)	(0.0193)
Adjusted R^2	0.0550	0.0747
Observations	3352	3352
Model Controls	SUE, log(Size)	SUE, log(Size)

Significance levels: *** p < 0.001, ** p <0.01, * p <0.05
Standard errors in parentheses

6.1.4 Combined Sentiment Analysis

To further examine the explanatory power of the sentiment measures, both the Loughran-McDonald (LM) dictionary-based sentiment and the GPT-4o sentiment scores were included jointly in the regression models. This analysis allows for an evaluation of whether combining the two sentiment measures provides complementary information in explaining cumulative abnormal returns (CARs).

The regression results for the CAR[0,+1] window are presented in **Table 6.5**, covering the full transcript, the presentation segment, and the Q&A segment. In all specifications, the GPT-derived sentiment scores remain statistically significant and positively associated with CARs. In contrast, the LM-sentiment scores are statistically insignificant across all segments, with large standard errors. The adjusted R-squared values are comparable to earlier models, with the Q&A segment showing the highest explanatory power.

Extended results for longer event windows are provided in the Appendix. **Appendix Table A.15** reports the outcomes for the CAR[0,+3] window, while **Appendix Table A.16** presents the results for CAR[0,+5]. Across these specifications, the GPT-sentiment maintains its significance, whereas LM-

sentiment remains insignificant. These findings suggest that the GPT-derived sentiment captures the relevant information presented during earnings calls more effectively than the traditional LM-based sentiment, even when both measures are included in the same model. After including the GPT-sentiment, the LM-derived sentiment no longer exhibits any explanatory power. This suggests that the relevant sentiment-related information is fully reflected in the GPT-based scores.

Table 6.5: Regression Results - Combined Sentiment Analysis (LM and GPT) for CAR[0,+1]

Dependent Variable: CAR[0,+1]			
Variables	**(1) Full Transcript**	**(2) Presentation Segment**	**(3) Q&A Segment**
GPT-Sentiment	0.0320***	0.0284***	0.0428***
	(0.0040)	(0.0053)	(0.0047)
LM-Sentiment	0.0686	0.0655	0.3191
	(0.1814)	(0.1203)	(0.1883)
SUE	0.0025***	0.0027***	0.0024***
	(0.0002)	(0.0002)	(0.0002)
Intercept	-0.0169***	-0.0214***	-0.0187***
	(0.0017)	(0.0029)	(0.0016)
Adjusted R^2	0.0697	0.0584	0.0769
Observations	3352	3333	3340
Model Controls	SUE	SUE	SUE

Significance levels: *** p < 0.001, ** p <0.01, * p <0.05
Standard errors in parentheses

6.1.5 Positive vs Negative Sentiment

To further explore the effect of sentiment on short-term abnormal returns, the GPT-derived sentiment scores were split into positive and negative sections. This analysis examines whether the predictive power of sentiment is primarily driven by positive or negative sentiment. The regression results are presented in **Table 6.6**, based on the GPT-derived sentiment scores from the full transcript, divided into positive and negative sentiment.

The results show that positive sentiment is significantly correlated with CARs, with a coefficient of 0.0354 ($p < 0.001$). In contrast, negative sentiment is not statistically significant, with a coefficient of 0.0341 and a large standard error. The Adjusted R-squared of the positive sentiment model is 4.02%, while for the negative sentiment model it is higher at 17.06%; however, this difference should be interpreted cautiously due to the very small number of observations in the negative sentiment model (74 compared to 2,637 for positive sentiment). The smaller sample size results from the exclusion of transcripts where either positive or negative sentiment is not present, resulting in a limited dataset for the negative sentiment analysis.

Overall, the findings indicate that positive sentiment in earnings calls is the main contributor to short-term market reactions, while negative sentiment does not demonstrate similar explanatory power. Additional results for extended event windows are included in the **Appendix**, with **Table A.17** presenting the outcomes for CAR[0,+3] and **Table A.18** for CAR[0,+5].

Table 6.6: Regression Results - Positive and Negative GPT-Sentiment CAR[0,+1]

Dependent Variable: CAR[0,+1]		
Variables	**(1) Positive Sentiment**	**(2) Negative Sentiment**
Sentiment	0.0354***	0.0341
	(0.0079)	(0.0545)
SUE	0.0022***	0.0073***
	(0.0002)	(0.0020)
Intercept	-0.0172***	-0.0128
	(0.0040)	(0.0187)
Adjusted R^2	0.0402	0.1706
Observations	2,637	74
Model Controls	SUE	SUE

Significance levels: *** p < 0.001, ** p <0.01, * p <0.05
Standard errors in parentheses

6.1.6 Robustness Checks

To ensure the robustness of the regression results, multiple diagnostic tests were performed. These tests assess the correlations between variables and examine whether multicollinearity may affect the reliability of the estimated coefficients.

First, the correlation between the two sentiment measures, Loughran-McDonald and GPT-4o, was evaluated across all transcript sections. As shown in **Panel A of Table 6.7**, the correlation is highest for the full transcript at 0.4645, followed by the presentation section (0.4208) and the Q&A section (0.3619). This indicates that while the two sentiment measures are related while also showing observable differences, this suggests that they capture unique information from the textual content.

Second, the relationship between sentiment and the Standardized Unexpected Earnings (SUE) was examined. **Panel B of Table 6.7** shows the correlations between each sentiment measure and SUE across the different transcript segments. The correlations are generally low, ranging from 0.1523 to 0.2019, suggesting that sentiment and earnings surprises capture separate aspects of information.

Lastly, Variance Inflation Factors (VIFs) were calculated to formally test for multicollinearity between the explanatory variables in the regression models. As shown in **Panel C of Table 6.7**, all VIF values are close to 1, with the highest being 1.0574 for GPT-4o sentiment in the full transcript. These values suggest that multicollinearity is not a concern in the regression models.

These diagnostic checks together indicate that the results are not affected by high correlations between explanatory variables and support the robustness of the findings.

Table 6.7: Summary of Diagnostic Statistics

Panel A: Cross-Correlation between Sentiment Techniques (LM vs GPT-4o)

Segment	Correlation
Full Transcript	0.4645
Presentation	0.4208
Q&A	0.3619

Panel B: Correlation Between Sentiment and SUE

Sentiment Technique	Segment	Correlation
Loughran-McDonald	Full Transcript	0.1777
	Presentation	0.1523
	Q&A	0.1700
GPT-4o	Full Transcript	0.2019
	Presentation	0.2015
	Q&A	0.1932

Panel C: Variance Inflation Factors (VIFs)

Sentiment Technique	Segment	Sentiment VIF	SUE VIF	log(Size) VIF
Loughran-McDonald	Full Transcript	1.0336	1.0380	1.0047
	Presentation	1.0237	1.0279	1.0042
	Q&A	1.0327	1.0351	1.0071
GPT-4o	Full Transcript	1.0574	1.0452	1.0178
	Presentation	1.0560	1.0442	1.0171
	Q&A	1.0486	1.0411	1.0133

6.2 Additional Regression Analysis

This section introduces two additional regression designs to extend the baseline analysis. The first specification adjusts the event window to CAR[-3,-1], allowing for an assessment of whether sentiment scores help explain return variation in the days preceding the earnings call. This serves as a direct test for potential information leakage or early market reactions prior to the formal announcement.

The second section adjusts the scale of LM-based sentiment scores by rescaling the normalised net sentiment to the [-1, 1] interval. The normalised score is calculated as the net positive word count divided by total word count, as described in the baseline methodology. To enable comparison of coefficient values across sentiment models, the following min-max transformation is applied:

$$LM\ Rescaled_i = 2 * \frac{LM_i^{norm} - \min(LM^{norm})}{\max(LM^{norm}) - \min(LM^{norm})} - 1 \quad (10)$$

Where:

- LM_i^{norm} is the normalised LM-score for event i.
- $LM\ Rescaled_i$ is the rescaled LM-score using a min-max transformation.

This transformation ensures the LM-based sentiment scores are placed on the same scale as the GPT-4o outputs, allowing for a direct comparison of their respective explanatory power in the regression models.

6.2.1 Pre-Event Window

To assess whether sentiment scores show predictive relevance prior to the earnings call, the regression window is adjusted to the CAR[-3,-1] interval. **Table 6.8** presents the results for the full transcript specification, and **Table 6.9** for the Q&A segment. Across all model variants, explanatory power remains limited, with adjusted R-squared values ranging from -0.0002 to 0.0023.

For the full transcript specification, neither the LM-based nor the GPT-4o sentiment coefficients are statistically significant, regardless of the inclusion of the SUE control. In the Q&A segment, GPT-4o sentiment shows a statistically significant association with CAR[-3,-1] in the baseline model (0.0055, $p < 0.05$). However, the effect becomes statistically insignificant once the earnings surprise variable is included as a control. The LM-based coefficients remain statistically insignificant throughout.

These results indicate that sentiment scores from both the full transcript and the Q&A segment do not systematically explain return variation prior to the earnings announcement. The explanatory power of GPT-4o sentiment appears concentrated in the immediate days around the event, as shown in the previous section, suggesting that the scores do not capture evidence of information leakage before the announcement.

Table 6.8: Regression Results - Full Transcript Sentiment for CAR[-3,-1]

Dependent Variable: CAR[-3,-1]				
Variables	**(1) LM-sentiment**	**(2) LM-sentiment with SUE**	**(3) GPT-4o Sentiment**	**(4) GPT-4o Sentiment with SUE**
Sentiment	0.0614	0.0207	0.0034	0.0024
	(0.0881)	(0.0895)	(0.0020)	(0.0020)
SUE		0.0003*		0.0003*
		(0.0001)		(0.0001)
Intercept	0.0003	-0.0001	-0.0005	-0.0008
	(0.0008)	(0.0008)	(0.0009)	(0.0009)
Adjusted R^2	-0.0002	0.0015	0.0006	0.0019
Observations	3360	3352	3360	3352
Model Controls	None	SUE	None	SUE

Significance levels: *** p < 0.001, ** p <0.01, * p <0.05
Standard errors in parentheses

Table 6.9: Regression Results - Q&A Segment Sentiment for CAR[-3,-1]

Dependent Variable: CAR[-3,-1]				
Variables	**(1) LM-sentiment**	**(2) LM-sentiment with SUE**	**(3) GPT-4o Sentiment**	**(4) GPT-4o Sentiment with SUE**
Sentiment	0.1740	0.1368	0.0055*	0.0045
	(0.0964)	(0.0977)	(0.0024)	(0.0024)
SUE		0.0003*		0.0003*
		(0.0001)		(0.0001)
Intercept	0.0001	-0.0003	-0.0009	-0.0012
	(0.0006)	(0.0006)	(0.0009)	(0.0009)
Adjusted R^2	0.0007	0.0019	0.0013	0.0023
Observations	3348	3340	3348	3340
Model Controls	None	SUE	None	SUE

Significance levels: *** p < 0.001, ** p <0.01, * p <0.05
Standard errors in parentheses

6.2.2 Rescaled LM-sentiment scores

To support the interpretation of the sentiment coefficients, LM-based scores are rescaled to the [-1, 1] range using a min-max transformation. The analysis focuses exclusively on the CAR[0,+1] window, as earlier results identified this event window as the most informative. The regression results are reported in **Table 6.10**. Across all sections, the sentiment coefficients are positive and statistically significant, with the strongest effect observed for the Q&A segment (0.0227, $p < 0.001$), followed by the full transcript (0.0149, $p < 0.001$) and the presentation section (0.0098, $p < 0.01$). Adjusted R-squared values remain in line with the baseline regressions, where the Q&A segment shows the highest explanatory

power (0.0540). The rescaling improves interpretability but does not replace the original regression, and is included in order to support the comparison of coefficient estimations across sentiment analysis approaches.

Compared to the GPT-4o sentiment regressions in **Tables 6.1** to **6.3**, the rescaled LM-coefficients are lower in absolute size across all transcript sections. In the Q&A segment regression, for instance, the LM coefficient (0.0227) is approximately half the size of the corresponding GPT-4o estimate (0.0455). These results indicate that GPT-4o sentiment incorporates a greater signal with respect to market-adjusted return variation. At the same time, the LM-based scores continue to capture statistically meaningful return effects, even when expressed on a standardised scale. This suggests that the LM-signal remains directionally consistent and robust. The observed differences in coefficients and explanatory power point out that traditional sentiment models are limited in their capacity to extract market-relevant content compared to LLM-based approach.

Table 6.10: Regression Results - Rescaled LM-Sentiment for CAR[0,+1]

Dependent Variable: CAR[0,+1]			
Variables	**(1) Full Transcript**	**(2) Presentation**	**(3) Q&A**
Sentiment	0.0149	0.0098	0.0227
	(0.0034)***	(0.0033)**	(0.0045)***
SUE	0.0028	0.0029	0.0027
	(0.0002)***	(0.0002)***	(0.0002)***
Intercept	-0.0057	-0.0045	-0.0029
	(0.0010)***	(0.0010)***	(0.0011)**
Adjusted R^2	0.0526	0.0507	0.0540
Observations	3352	3333	3340
Model Controls	SUE	SUE	SUE

7. Methodology Portfolio Simulation

To assess the economic application of GPT-derived sentiment scores from earnings calls, this study implements a simulated trading strategy. The primary objective is to examine whether the sentiment-informed investment recommendations can generate economically significant and risk-adjusted returns. The simulation constructs a portfolio strategy that follows a combined long-short approach, based on the investment recommendations derived from GPT-4o. The selection of a combined long-short strategy is based on the empirical findings from the event study analysis, where positive-only sentiment scores produced significant abnormal returns, while negative-only sentiment scores were insignificant. By implementing a combined long-short strategy, the simulation evaluates whether GPT-4o recommendations can serve as a profitable trading strategy. To simulate a realistic investment process, the strategy allocates capital on a daily basis based on the investment recommendations. Performance of the strategy is evaluated using standard financial metrics that assess both return and risk exposure.

7.1 Strategy Overview

The simulation includes all trading days on which at least one earnings call event occurred within the dataset. On each of these days, a fixed capital of 100,000 is allocated to the portfolio. This amount is initially divided equally across all events taking place that day, ensuring each event receives the same base capital allocation. Within each event, the allocated capital is further distributed according to the GPT-derived investment recommendation score, which ranges from -1 to +1, using the following categories:

Table 7.1: Mapping of GPT-4o Recommendations to Portfolio Allocation

GPT Investment Recommendation	Portfolio Position
- 1.0	Short 100% of the event's allocated capital
- 0.5	Short 50% and 50% allocated to the risk-free asset
0.0	100% allocated to the risk-free asset
+ 0.5	Long 50% and 50% allocated to the risk-free asset
+ 1.0	Long 100% of the event's allocated capital

Each position will be held for a period of two trading days, covering the event day ($t = 0$) and the following trading day ($t = 1$). This holding period is based on the event study analyses, which indicate that cumulative returns over the immediate two-day window (CAR [0,1]) captures the best impact of earnings call on stock prices.

For each event i on day t, the return is computed using the realised stock returns over the two day window, using the following formula:

$$R_{i,t}^{long} = \prod_{k=0}^{1}(1 + r_{i,t+k}) - 1 \quad (11)$$

where:

- $R_{i,t}^{long}$ is the realised return for event i over the two-day window starting on day t;
- Π is the product operator that accumulates the return over the two-day period;
- $r_{i,t+k}$ is the daily stock return for event i on day $t + k$. Where k indexes the holding period days, starting from the event ($k = 0$) to the next trading day ($k = 1$).

In case of a short position, the return is inverted using the following formula:

$$R_{i,t}^{short} = -R_{i,t}^{long} \quad (12)$$

where:

- $R_{i,t}^{short}$ is the realised return for event i on day t for a short position, computed as the negative of the long position return.

The allocated capital allocated to the risk-free asset earns the daily risk-free rate from the Fama-French dataset. The dataset provides the 1-month U.S. Treasury Bill rate expressed on a daily basis. Since the position is held over two trading days, the cumulative risk-free return is calculated as:

$$R_f^{(2)} = (1 + R_{f,t}) * (1 + R_{f,t+1}) - 1 \quad (13)$$

where:

- $R_f^{(2)}$ is the two-day compounded risk-free rate, calculated over the event window;
- $R_{f,t}$ is the daily risk-free rate on day t.

This framework enables the allocation of capital between stock positions and the risk-free asset, based on the GPT-4o generated investment recommendation.

7.2 Portfolio Construction

The portfolio construction methodology describes how capital is distributed across earnings events on each trading day. The total capital available per trading day is fixed at 100,000. This capital is initially distributed equally across all earnings events occurring on that trading day. Define N_t as the number of earnings events on day t, and C^{total} as the total daily capital. The allocated capital to each event i on day t is then defined as:

$$C_{i,t} = \frac{C^{total}}{N_t} \quad (14)$$

where:

- $C_{i,t}$ is the capital allocated to event i on day t;
- C^{total} is the total daily capital (100,00);

- N_t is the number of earnings events on day t.

Once the capital per event is defined, the allocation between stock position and risk-free asset is made based on the GPT-4o investment recommendation as defined in the previous section. The weight of the recommendation sets the proportion of allocated event capital invested in the stock position, while the residual capital is allocated to the risk-free asset. The realised return for each return is calculated by combining the realised stock return and the return on the risk-free asset over a two day holding period:

$$R_{i,t}^{event} = w_i^{stock} * R_{i,t}^{long/short} + w_i^{Rf} * R_f^{(2)} \tag{15}$$

where:

- $R_{i,t}^{event}$ is the total realised return for event i on day t, combining the returns from the stock position and risk-free asset;
- w_i^{stock} is the weight allocated to the stock position for event i;
- $R_{i,t}^{long/short}$ is the realised return of the stock for event i on day t, from either a long or short position;
- w_i^{Rf} is the weight allocated to the risk-free asset for event i;
- $R_f^{(2)}$ is the two-day compounded risk-free rate.

The daily portfolio return is then computed by accumulating the event-level returns, weighted by the capital allocated to each event relative to the portfolio capital:

$$R_t^{portfolio} = \sum_{i=1}^{N_t} \left(\frac{C_{i,t}}{C^{total}} * R_{i,t}^{event} \right) \tag{16}$$

This equal allocation approach leads to disproportionate risk exposure on days with a limited number of earnings events. When only one event occurs on a given day, the entire capital of 100,000 would otherwise be allocated to that single event, resulting in an undesirable concentration of risk. To mitigate this, a position cap is introduced to limit the maximum capital that can be allocated to any individual event, regardless of the number of events available on that day. The position cap is represented by C^{cap}, which redefines the capital allocation per event as:

$$C_{i,t} = \min \left(\frac{C^{total}}{N_t}, C^{cap} \right) \tag{17}$$

Any residual capital that remains unallocated due to the position cap is invested in the risk-free asset. For example, if $C^{cap} = 20{,}000$, and there is only one event, 20,000 is allocated to the stock, while the remaining 80,000 is invested in the risk-free asset.

Within each event, the capital allocated to the stock is proportional to the absolute value of the GPT-4o investment recommendation:

$$C_{i,t}^{stock} = C_{i,t} * \left| GPT\_inv_{i,t} \right| \tag{18}$$

The corresponding capital allocated to the risk-free asset at the event level is:

$$C_{i,t}^{rf} = C_{i,t} - C_{i,t}^{stock} \tag{19}$$

The realised return is then:

$$R_{i,t}^{stock} = C_{i,t}^{stock} * R_{i,t}^{long/short} \tag{20}$$

After all events on day t have been processed, the unallocated capital is:

$$C_t^{unallocated} = C^{total} - (C_{i,t} * N_t) \tag{21}$$

The total daily portfolio return is therefore:

$$R_t^{portfolio} = \frac{\sum_{i=1}^{N_t}\left(R_{i,t}^{stock} + C_{i,t}^{rf} * R_f^{(2)}\right) + C_t^{unallocated} * R_f^{(2)}}{C^{total}} \tag{22}$$

To assess the impact of different position caps, the portfolio strategy was simulated across a range of C^{cap} values, specifically from 20,000 to 100,000 in steps of 10,000. The performance at each position cap level was evaluated using the annualised daily ratio, CAPM alpha, beta, and R-squared. The Sharpe ratio served as the primary selection criterion, as it captures the trade-off between return and risk of the strategy. The implementation of a position cap provides a practical risk management mechanism by limiting the exposure to any single earnings event on a given day. At the same time, it ensures that the strategy remains adaptable to variation in the number of available events per trading day.

7.3 Performance Evaluation

The performance of the trading strategy is evaluated using standard financial metrics that capture both absolute returns and risk-adjusted performance. The first evaluation metric used is the daily Sharpe ratio, capturing the excess return per unit of risk, defined as the standard deviation of daily returns. The Sharpe ratio is calculated as:

$$Sharpe\ Ratio = \frac{\overline{R_t^{portfolio}} - \overline{R_{f,t}}}{\sigma(R_t^{portfolio})} \tag{23}$$

where:

- $\overline{R_t^{portfolio}}$ is the average return of the portfolio on day t;
- $\overline{R_{f,t}}$ is the average daily risk-free rate on day t;
- $\sigma(R_t^{portfolio})$ is the standard deviation of the daily portfolio returns;

In addition to the Sharpe ratio, the strategy's performance is evaluated using the Capital Asset Pricing Model (CAPM) regression. This model assesses the relationship between the strategy's excess returns and the market's excess returns. The CAPM regression is specified as:

$$R_t^{portfolio} - R_{f,t} = \alpha + \beta\left(R_{m,t} - R_{f,t}\right) + \varepsilon_t \quad (24)$$

where:

- $R_t^{portfolio}$ is the return of the portfolio on day t;
- $R_{f,t}$ is the daily risk-free rate on day t;
- α is the intercept, representing the portfolio's abnormal return, or alpha;
- β is captures the sensitivity of the strategy to the market;
- ε_t is the error term on day t.

The regression provides three key metrics: alpha, beta and R-squared. The alpha reflects the strategy's ability to generate returns beyond what can be explained by the market exposure. The beta measures the degree of systematic market risk captured by the strategy. The R-squared reflects the proportion of the variance in the strategy's excess returns that can be explained by the market's excess returns.

To validate the regression model, diagnostic tests are conducted for heteroskedasticity (using the Breusch-Pagan test) and autocorrelation (using the Durbin-Watson test). These checks help ensure that the regression results are statistically reliable and not affected by violations of the OLS assumptions. Together, these diagnostics and performance measures provide a comprehensive evaluation of the trading strategy, capturing both the return potential and associated risk.

8. Results Portfolio Simulation

This chapter evaluates the economic relevance of GPT-4o investment recommendation scores by implementing a long-short trading strategy. Capital is allocated across earnings events based on the model's investment recommendation, ranging from (strong) sell to (strong) buy. The simulation is conducted across a range of position caps, varying from 10,000 to 100,000, to evaluate how portfolio performance responds to different allocation constraints.

Performance is evaluated using standard financial metrics, including the daily Sharpe ratio, CAPM alpha and beta, and the R-squared, providing insight into both return characteristics and systematic risk exposure. To verify the robustness of the CAPM regression results, diagnostic tests for heteroskedasticity (Breusch-Pagan) and autocorrelation (Durbin-Watson) are applied.

This empirical analysis directly tests Hypothesis 6, which states that a long-short strategy based on GPT-4o investment recommendations generates a statistically and economically significant alpha relative to the passive market benchmark.

8.1 Position Cap Selection

Table 8.1 reports the risk-adjusted performance of the long-short portfolio constructed from GPT-4o investment recommendations based on the **full transcript**. The strategy is evaluated across ten capital position caps, ranging from 10,000 to 100,000 per earnings event. The results show that the daily Sharpe ratio reaches its maximum at the 30,000 cap level, with a value of 0.1966, after which performance declines steadily.

The Q&A-based strategy, presented in **Table 8.2**, shows a similar pattern with the highest Sharpe ratio observed at the 10,000 level (0.2011), followed by a decline in risk-adjusted returns. While **Q&A**-based recommendations result in slightly higher Sharpe ratios at lower position caps, their corresponding mean returns and alpha values remain consistently lower than those of the full transcript specification. The relation between position cap and risk-adjusted performance is visualised in **Appendix Figure B.2.**

Based on these results, the 30,000 cap is selected for further analysis for the full transcript recommendations, and the 10,000 cap for the Q&A segment strategy. These caps offer the most favourable trade-off between return and risk-adjusted performance, while maintaining Sharpe ratio efficiency relative to larger allocation levels. The following sections assess the financial characteristics and risk profile of the selected portfolio specifications in more detail.

Table 8.1: Portfolio Performance by Position Cap - Full Transcript Recommendations

Position Cap	Mean Return	Sharpe Ratio (Daily)	Alpha (α) (Daily)	Beta (β) (Daily)	R^2
10,000	0.0008	0.1952	0.0007***	0.0847***	0.0584
20,000	0.0015	0.1939	0.0014***	0.1661***	0.0597
30,000	0.0020	0.1966	0.0019***	0.2287***	0.0593
40,000	0.0025	0.1961	0.0023***	0.2757***	0.0573
50,000	0.0029	0.1939	0.0027***	0.3167***	0.0544
60,000	0.0031	0.1928	0.0029***	0.3365***	0.0529
70,000	0.0033	0.1906	0.0031***	0.3563***	0.0509
80,000	0.0035	0.1878	0.0033***	0.3760***	0.0487
90,000	0.0037	0.1847	0.0035***	0.3958***	0.0465
100,000	0.0039	0.1815	0.0037***	0.4156***	0.0445

Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

Table 8.2: Portfolio Performance by Position Cap - Q&A Segment Recommendations

Position Cap	Mean Return	Sharpe Ratio (Daily)	Alpha (α) (Daily)	Beta (β) (Daily)	R^2
10,000	0.0005	0.2011	0.0004***	0.0354***	0.0280
20,000	0.0009	0.1997	0.0008***	0.0697***	0.0290
30,000	0.0013	0.1984	0.0012***	0.0945***	0.0278
40,000	0.0015	0.1953	0.0014***	0.1126***	0.0260
50,000	0.0017	0.1912	0.0016***	0.1287***	0.0242
60,000	0.0019	0.1896	0.0018***	0.1359***	0.0233
70,000	0.0020	0.1870	0.0019***	0.1430***	0.0222
80,000	0.0021	0.1839	0.0020***	0.1501***	0.0211
90,000	0.0022	0.1805	0.0021***	0.1572***	0.0200
100,000	0.0023	0.1770	0.0022***	0.1643***	0.0189

Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

8.2 Performance Evaluation

Table 8.3 reports the key performance metrics of the selected long-short strategies based on the full transcript and Q&A segment recommendations, each evaluated at the position cap selected in the previous section. Both strategies yield positive Sharpe ratios on a daily basis, with the Q&A-based strategy showing a slightly higher value (0.2011) relative to the full transcript strategy (0.1966). This difference is driven by significantly lower volatility (0.0022 vs. 0.0100), rather than stronger return generation. The full transcript portfolio achieves a daily mean return of 0.0020, which is four times higher than the Q&A-based strategy (0.0005), indicating that the higher Sharpe ratio of the Q&A strategy primary reflects its lower volatility.

The CAPM regression results, reported in **Table 8.1** and **Table 8.2**, provide further insight into the sources of return. The full transcript portfolio yields a daily alpha of 0.0019 and a beta of 0.2287, both statistically significant at the 0.1% level. The corresponding R-squared of 0.0593 indicates that only a small share of return variation is attributable to market exposure. The Q&A-based strategy produces a lower alpha of 0.0004 and an even smaller beta of 0.0354, also significant at the 0.1% level, with an R-squared of 0.0280. While both strategies generate statistically significant abnormal returns, the full transcript strategy achieves a higher alpha in combination a limited degree of systematic risk exposure.

These results support hypothesis H6, which states that a long-short strategy based on GPT-4o investment recommendations generates a statistically and economically significant alpha relative to the passive market benchmark. The return performance, combined with low beta and limited explanatory power of the market factor, suggests that the observed performance cannot be attributed entirely to systematic risk. The following section examines the robustness of these regression outcomes by testing for heteroskedasticity and autocorrelation in the residuals.

Table 8.3: Strategy Performance Metrics

Strategy	Sharpe ratio (Daily)	Mean Return (Daily)	Standard Deviation (Daily)
Full Transcript	0.1966	0.0020	0.0100
Q&A Segment	0.2011	0.0005	0.0022

8.3 Robustness Checks

The robustness of the CAPM regressions is evaluated using standard diagnostic tests for heteroskedasticity and autocorrelation in the residuals. The Breusch-Pagan test is used to evaluate heteroskedasticity, while the Durbin-Watson statistic assesses serial correlation. Both diagnostics are performed on the full transcript and Q&A-based strategies at the selected position cap levels. The results are reported in **Table 8.4**.

For the full transcript strategy, the Breusch-Pagan test yields a test statistic of 0.2735 with a p-value of 0.6010, indicating no evidence of heteroskedasticity. The Durbin-Watson value is 2.0193 (p = 0.6485), suggesting no significant autocorrelation in the residuals. The Q&A-based portfolio produces a Breusch-Pagan statistic of 1.6512 (p = 0.1988) and a Durbin-Watson value of 2.0642 (p = 0.8957), indicating no violation of the residual assumptions.

These results confirm that the CAPM regressions are statistically robust across both strategies. The absence of heteroskedasticity and autocorrelation confirms the robustness of the estimated alpha coefficients and supports the reliability of the statistical conclusions.

Table 8.4: Diagnostic Tests for Portfolio Simulation CAPM Regressions

	Breusch-Pagan		Durbin-Watson	
Strategy	**BP-value**	**p-value**	**DW-value**	**p-value**
Full Transcript	0.2735	0.6010	2.0193	0.6485
Q&A Segment	1.6512	0.1988	2.0642	0.8957

9. Conclusion

This thesis investigates whether sentiment extracted from earnings call transcripts can predict short-term stock price movements. Using both a traditional dictionary-based approach (Loughran-McDonald) and LLM-based approach (GPT-4o), this study assesses their ability to predict cumulative abnormal returns (CARs) following an earnings call. The results are based on 3,360 transcripts from S&P-500 firms between 2009 and 2024, resulting in a corpus of over 1.79 million individual sentences. Using an event-study framework, this thesis examined how sentiment extracted by these two approaches relates to cumulative abnormal returns (CARs) around the earnings announcement.

The findings show that both sentiment methods are individually positively related to CARs across all event windows. The shortest event window, CAR[0,+1] captures the most explanatory power of the different event windows. Suggesting that the short-term effects on stock price returns are most informative for short-horizon investors. In addition, the Q&A segment outperforms the presentation segment in terms of coefficient magnitude and Adjusted R-squared. This result is consistent with prior research and confirms that the interaction between analysts and management is more informative than the prepared remarks. This result is further extended by combining the sentiment analysis approaches to investigate the effects on abnormal returns. Across these regressions, the GPT-derived sentiment maintains significant, while the LM-derived sentiment becomes insignificant. These findings suggest that the GPT-derived sentiment captures the relevant information presented during earnings calls more effectively than the traditional LM-approach. The insignificance of the LM-derived sentiment suggest that all the relevant sentiment-related information is fully reflected in the GPT-derived scores.

To test the robustness of the sentiment-CAR relationship, quantitative factors that potentially influence the market reaction were included. The earnings surprise and market capitalisation both had a significant relationship with the CARs across the event windows. The earnings surprise show a positive correlation with abnormal returns, while firm size shows a negative relationship. Importantly, both the sentiment approaches remain statistically significant after including these control variables. This confirms the robustness of the sentiment effects on CARs and strengthens the hypothesis that sentiment analysis contains unique information which is not captured by quantitative data.

The simulated portfolio simulation, based on the results of the event study framework and GPT-derived investment recommendations, followed a long-short strategy. The strategy was based on the full transcript and Q&A segment recommendations and were evaluated using standard financial metrics. Both strategies yield a positive Sharpe ratio on a daily basis, with the Q&A segment (0.2011) outperforming the full transcript (0.1966) based strategy. However, the daily return of the Q&A based strategy shows a significantly smaller return mean return (0.0005) relative to the full transcript based strategy (0.0020). This difference is therefore driven by significantly lower volatility (0.0022 vs 0.0100), rather than stronger return generation. This suggest that the Sharpe ratio of the Q&A strategy primary

reflects its lower volatility. In addition, the CAPM results provide further insight into the sources of return. The full transcript strategy yields a daily alpha of 0.0019 and a beta of 0.2287, both statistically significant at the 0.1% level. The Q&A based strategy produces a lower alpha of 0.0004 and a smaller beta of 0.0354, also significant at the 0.1% level. While both strategies generate statistically significant abnormal returns, the full transcript strategy achieves a higher alpha in combination with a limited degree of systematic risk exposure. These results suggest that the GPT-based long-short strategy generates a statistically and economically significant return.

This thesis contributed to the literature by using a structured, accessible and replicable method to analyse financial sentiment in earnings conference calls. It shows the potential of analysis sentiment with general-purpose large language models and provides a new look on the application of AI-techniques compared to benchmark methods. Previous studies have rarely explored the predictive accuracy of LLM-driven sentiment analysis on earnings calls. Most studies have focused on news sentiment or social media, which may be unstructured and noisy. In contrast, this study employs a analysis on direct management disclosure, which is more accessible for general practitioners.

While the proposed methodology provides valuable insights, there remain several limitations that should be acknowledged. Although the GPT-derived sentiment scores offer advantages over the LM-derived approach, the limited research on long-horizon windows present a significant challenge. Since the portfolio simulation did not account for any form of trading costs, the real-world application relevance should be explored further. To address this, a more detailed portfolio simulation could be designed to test if the strategy is able to generate a positive alpha after trading costs. To further extend the portfolio simulation, variables such as volatility or trading volume could provide a more complete view on the strategy.

Another limitation is the assumption that short selling is always feasible, which is often not the case in real-word market conditions due to restrictions or costs associated with these positions. Additionally, the large language model may exhibit a sectoral or language bias, as some industries use more optimistic or technical language. This could distort the sentiment scores and create an alpha that does not generalise across sectors.

Future research could address these points and improve the proposed methodology. Give the resource constraints such as time, computational power and budget, a hybrid approach using machine learning and LLMs was not feasible, despite the potential benefits. In addition, the sample size could be extended across number of firms, sectors, countries and disclosure language. While this study employed the GPT-4o-mini model, more sophisticated models are available, potentially offering a more accurate analysis. In addition, instead of employing a zero-shot approach, the use of few-shot prompting could be tested. Finally, future research could examine different types of textual data such as press releases, analysts reports or other regulatory filings, to asses how their informational value and market impact deviate.

References

Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Brown, L. D., Call, A. C., Clement, M. B., & Sharp, N. Y. (2019). Managing the narrative: Investor relations officers and corporate disclosure☆. *Journal of Accounting and Economics*, *67*(1), 58-79.

Brown, S., Hillegeist, S. A., & Lo, K. (2004). Conference calls and information asymmetry. *Journal of Accounting and Economics*, *37*(3), 343-366.

Cambria, E., & White, B. (2014). Jumping NLP curves: A review of natural language processing research. *IEEE Computational intelligence magazine*, *9*(2), 48-57.

Chin, A., & Fan, Y. (2023). Leveraging text mining to extract insights from earnings call transcripts. *Journal of Investment Management*, *21*(1), 81-102.

Dang, N. C., Moreno-García, M. N., & De la Prieta, F. (2020). Sentiment analysis based on deep learning: A comparative study. *Electronics*, *9*(3), 483.

Das, S. R., & Chen, M. Y. (2007). Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management science*, *53*(9), 1375-1388.

Das, S., & Chen, M. (2001, July). Yahoo! for Amazon: Extracting market sentiment from stock message boards. In *Proceedings of the Asia Pacific finance association annual conference (APFA)* (Vol. 35, p. 43).

De Amicis, C., Falconieri, S., & Tastan, M. (2021). Sentiment analysis and gender differences in earnings conference calls. *Journal of Corporate Finance*, *71*, 101809.

de Jong, F., & de Goeij, P. (2011). Event studies methodology. Tilburg University.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171-4186).

Fama, E. F. (1970). Efficient capital markets. *Journal of finance*, *25*(2), 383-417.

Fatouros, G., Soldatos, J., Kouroumali, K., Makridis, G., & Kyriazis, D. (2023). Transforming sentiment analysis in the financial domain with ChatGPT. *Machine Learning with Applications*, *14*, 100508.

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning* (Vol. 1, No. 2). Cambridge: MIT press.

Guo, Y., Xu, Z., & Yang, Y. (2023). Is chatgpt a financial expert? evaluating language models on financial natural language processing. *arXiv preprint arXiv:2310.12664*.

Hassan, T. A., Hollander, S., Kalyani, A., van Lent, L., Schwedeler, M., & Tahoun, A. (2024). *Economic Surveillance using Corporate Text* (No. w33158). National Bureau of Economic Research.

Healy, P. M., & Palepu, K. G. (2001). Information asymmetry, corporate disclosure, and the capital markets: A review of the empirical disclosure literature. *Journal of accounting and economics*, *31*(1-3), 405-440.

Henry, E. (2006). Market reaction to verbal components of earnings press releases: Event study using a predictive algorithm. *Journal of Emerging Technologies in Accounting*, *3*(1), 1-19.

Huang, X., Teoh, S. H., & Zhang, Y. (2014). Tone management. *The accounting review*, *89*(3), 1083-1113.

Jayaraman, J. D., & Dennis, A. (2020). Can Earnings Call Sentiment Predict Stock Price Movement?. *Proceedings of the Northeast Business & Economics Association*.

Jing, N., Wu, Z., & Wang, H. (2021). A hybrid model integrating deep learning with investor sentiment analysis for stock price prediction. *Expert Systems with Applications*, *178*, 115019.

Kearney, C., & Liu, S. (2014). Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis*, *33*, 171-185.

Kiritchenko, S., Zhu, X., & Mohammad, S. M. (2014). Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, *50*, 723-762.

Li, F. (2010). The information content of forward-looking statements in corporate filings - A naïve Bayesian machine learning approach. *Journal of accounting research*, *48*(5), 1049-1102.

Liu, B. (2022). *Sentiment analysis and opinion mining*. Springer Nature.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of finance*, *66*(1), 35-65.

LSEG. (2025). LSEG Transcripts & Briefs [Dataset]. In *LSEG Workspace*. https://www.lseg.com/en/data-analytics/financial-data/company-data/events/earnings-transcripts-briefs/transcripts-database

MacKinlay, A. C. (1997). Event studies in economics and finance. *Journal of economic literature*, *35*(1), 13-39.

Madhoushi, Z., Hamdan, A. R., & Zainudin, S. (2015, July). Sentiment analysis techniques in recent works. In *2015 science and information conference (SAI)* (pp. 288-291). IEEE.

Malkiel, B. G. (2003). The efficient market hypothesis and its critics. *Journal of economic perspectives*, *17*(1), 59-82.

Matsumoto, D., Pronk, M., & Roelofsen, E. (2011). What makes conference calls useful? The information content of managers' presentations and analysts' discussion sessions. *The Accounting Review*, *86*(4), 1383-1414.

Medya, S., Rasoolinejad, M., Yang, Y., & Uzzi, B. (2022, April). An exploratory study of stock price movements from earnings calls. In *Companion Proceedings of the Web Conference 2022* (pp. 20-31).

Minh, D. L., Sadeghi-Niaraki, A., Huy, H. D., Min, K., & Moon, H. (2018). Deep learning approach for short-term stock trends prediction based on two-stream gated recurrent unit network. *Ieee Access*, *6*, 55392-55404.

Mudinas, A., Zhang, D., & Levene, M. (2012, August). Combining lexicon and learning based approaches for concept-level sentiment analysis. In *Proceedings of the first international workshop on issues of sentiment discovery and opinion mining* (pp. 1-8).

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, *2*(1-2), 1-135.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. *arXiv preprint cs/0205070*.

Price, S. M., Doran, J. S., Peterson, D. R., & Bliss, B. A. (2012). Earnings conference calls and stock returns: The incremental informativeness of textual tone. *Journal of Banking & Finance*, *36*(4), 992-1011.

Seo, S., Kim, C., Kim, H., Mo, K., & Kang, P. (2020). Comparative study of deep learning-based sentiment classification. *IEEE Access*, *8*, 6861-6875.

Shen, Y., & Zhang, P. K. (2024, July). Financial sentiment analysis on news and reports using large language models and finbert. In *2024 IEEE 6th International Conference on Power, Intelligent Computing and Systems (ICPICS)* (pp. 717-721). IEEE.

Sohangir, S., Wang, D., Pomeranets, A., & Khoshgoftaar, T. M. (2018). Big Data: Deep Learning for financial sentiment analysis. *Journal of Big Data*, *5*(1), 1-25.

Tong, R. M. (2001, September). An operational system for detecting and tracking opinions in on-line discussion. In *Working notes of the ACM SIGIR 2001 workshop on operational text classification* (Vol. 1, No. 6, pp. 1-6).

Turney, P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *arXiv preprint cs/0212032*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2020, October). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations* (pp. 38-45).

Yeruva, V. K., Chandrashekar, M., Lee, Y., Rydberg-Cox, J., Blanton, V., & Oyler, N. A. (2020, December). Interpretation of sentiment analysis with human-in-the-loop. In *2020 IEEE International Conference on Big Data (Big Data)* (pp. 3099-3108). IEEE.

Yu, H., & Hatzivassiloglou, V. (2003). Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing* (pp. 129-136).

Appendices

A. Tables

Table A.1: Regression Results - Full Transcript Sentiment for CAR[0,+3]

Dependent Variable: CAR[0,+3]				
Variables	**(1) LM-sentiment**	**(2) LM-sentiment with SUE**	**(3) GPT-4o Sentiment**	**(4) GPT-4o Sentiment with SUE**
Sentiment	1.0169***	0.6890***	0.0365***	0.0288***
	(0.1806)	(0.1809)	(0.0040)	(0.0040)
SUE		0.0027***		0.0025***
		(0.0003)		(0.0003)
Intercept	-0.0064***	-0.0095***	-0.0129***	-0.0151***
	(0.0016)	(0.0017)	(0.0018)	(0.0018)
Adjusted R^2	0.0091	0.0397	0.0242	0.0501
Observations	3360	3352	3360	3352
Model Controls	None	SUE	None	SUE

Significance levels: *** p < 0.001, ** p <0.01, * p <0.05
Standard errors in parentheses

Table A.2: Regression Results - Full Transcript Sentiment for CAR[0,+5]

Dependent Variable: CAR[0,+5]				
Variables	**(1) LM-sentiment**	**(2) LM-sentiment with SUE**	**(3) GPT-4o Sentiment**	**(4) GPT-4o Sentiment with SUE**
Sentiment	1.0293***	0.7038***	0.0338***	0.0261***
	(0.1872)	(0.1878)	(0.0041)	(0.0042)
SUE		0.0026***		0.0025***
		(0.0003)		(0.0003)
Intercept	-0.0067***	-0.0098***	-0.0121***	-0.0143***
	(0.0017)	(0.0017)	(0.0019)	(0.0019)
Adjusted R^2	0.0086	0.0366	0.0193	0.0437
Observations	3360	3352	3360	3352
Model Controls	None	SUE	None	SUE

Significance levels: *** p < 0.001, ** p <0.01, * p <0.05
Standard errors in parentheses

Table A.3: Regression Results - Presentation Segment Sentiment for CAR[0,+3]

Dependent Variable: CAR[0,+3]				
Variables	**(1) LM-sentiment**	**(2) LM-sentiment with SUE**	**(3) GPT-4o Sentiment**	**(4) GPT-4o Sentiment with SUE**
Sentiment	0.5375***	0.3411**	0.0382***	0.0272***
	(0.1224)	(0.1220)	(0.0054)	(0.0054)
SUE		0.0027***		0.0026***
		(0.0003)		(0.0003)
Intercept	-0.0058**	-0.0090***	-0.0206***	-0.0198***
	(0.0019)	(0.0018)	(0.0032)	(0.0031)
Adjusted R^2	0.0055	0.0385	0.0146	0.0435
Observations	3341	3333	3341	3333
Model Controls	None	SUE	None	SUE

Significance levels: *** p < 0.001, ** p <0.01, * p <0.05
Standard errors in parentheses

Table A.4: Regression Results - Presentation Segment Sentiment for CAR[0,+5]

Dependent Variable: CAR[0,+5]				
Variables	**(1) LM-sentiment**	**(2) LM-sentiment with SUE**	**(3) GPT-4o Sentiment**	**(4) GPT-4o Sentiment with SUE**
Sentiment	0.5454***	0.3513**	0.0366***	0.0256***
	(0.1268)	(0.1266)	(0.0056)	(0.0056)
SUE		0.0027***		0.0026***
		(0.0003)		(0.0003)
Intercept	-0.0061**	-0.0093***	-0.0199***	-0.0190***
	(0.0019)	(0.0019)	(0.0033)	(0.0033)
Adjusted R^2	0.0052	0.0354	0.0125	0.0392
Observations	3341	3333	3341	3333
Model Controls	None	SUE	None	SUE

Significance levels: *** p < 0.001, ** p <0.01, * p <0.05
Standard errors in parentheses

Table A.5: Regression Results - Q&A Segment Sentiment for CAR[0,+3]

Dependent Variable: CAR[0,+3]				
Variables	**(1) LM-sentiment**	**(2) LM-sentiment with SUE**	**(3) GPT-4o Sentiment**	**(4) GPT-4o Sentiment with SUE**
Sentiment	1.1810***	0.8406***	0.0496***	0.0409***
	(0.1977)	(0.1978)	(0.0048)	(0.0049)
SUE		0.0026***		0.0024***
		(0.0003)		(0.0003)
Intercept	-0.0037**	-0.0077***	-0.0145***	-0.0167***
	(0.0013)	(0.0013)	(0.0018)	(0.0018)
Adjusted R^2	0.0103	0.0401	0.0303	0.0549
Observations	3348	3340	3348	3340
Model Controls	None	SUE	None	SUE

Significance levels: *** p < 0.001, ** p <0.01, * p <0.05
Standard errors in parentheses

Table A.6: Regression Results - Q&A Segment Sentiment for CAR[0,+5]

Dependent Variable: CAR[0,+5]				
Variables	**(1) LM-sentiment**	**(2) LM-sentiment with SUE**	**(3) GPT-4o Sentiment**	**(4) GPT-4o Sentiment with SUE**
Sentiment	1.1892***	0.8512***	0.0457***	0.0369***
	(0.2050)	(0.2054)	(0.0050)	(0.0051)
SUE	0.0026***		0.0024***	0.0026***
	(0.0003)		(0.0003)	(0.0003)
Intercept	-0.0040**	-0.0080***	-0.0136***	-0.0158***
	(0.0013)	(0.0014)	(0.0019)	(0.0019)
Adjusted R^2	0.0097	0.0368	0.0239	0.0470
Observations	3348	3340	3348	3340
Model Controls	None	SUE	None	SUE

Significance levels: *** p < 0.001, ** p <0.01, * p <0.05
Standard errors in parentheses

Table A.7: Regression Results - Full Transcript Sentiment with Control Variables for CAR[0,+3]

Dependent Variable: CAR[0,+3]		
Variables	**(1) LM-sentiment with Controls**	**(2) GPT-sentiment with Controls**
Sentiment	0.6767***	0.0309***
	(0.1807)	(0.0040)
SUE	0.0027***	0.0025***
	(0.0003)	(0.0003)
log(Size)	-0.0029***	-0.0037***
	(0.0009)	(0.0009)
Intercept	0.0629**	0.0772***
	(0.0215)	(0.0214)
Adjusted R^2	0.0426	0.0551
Observations	3352	3352
Model Controls	SUE, log(Size)	SUE, log(Size)

Significance levels: *** p < 0.001, ** p <0.01, * p <0.05
Standard errors in parentheses

Table A.8: Regression Results - Full Transcript Sentiment with Control Variables for CAR[0,+5]

Dependent Variable: CAR[0,+5]		
Variables	**(1) LM-sentiment with Controls**	**(2) GPT-sentiment with Controls**
Sentiment	0.6912***	0.0281***
	(0.1876)	(0.0042)
SUE	0.0027***	0.0025***
	(0.0003)	(0.0003)
log(Size)	-0.0030***	-0.0037***
	(0.0009)	(0.0009)
Intercept	0.0643**	0.0780***
	(0.0224)	(0.0222)
Adjusted R^2	0.0395	0.0484
Observations	3352	3352
Model Controls	SUE, log(Size)	SUE, log(Size)

Significance levels: *** p < 0.001, ** p <0.01, * p <0.05
Standard errors in parentheses

Table A.9: Regression Results - Presentation Segment Sentiment with Control Variables for CAR[0,+1]

Dependent Variable: CAR[0,+1]		
Variables	**(1) LM-sentiment with Controls**	**(2) GPT-sentiment with Controls**
Sentiment	0.3226**	0.00478**
	(0.1104)	(0.0049)
SUE	0.0029***	0.0027***
	(0.0002)	(0.0002)
log(Size)	-0.0025**	-0.0031***
	(0.0008)	(0.0008)
Intercept	0.0538**	0.0547**
	(0.0195)	(0.0194)
Adjusted R^2	0.0533	0.0627
Observations	3333	3333
Model Controls	SUE, log(Size)	SUE, log(Size)

Significance levels: *** p < 0.001, ** p <0.01, * p <0.05
Standard errors in parentheses

Table A.10: Regression Results - Presentation Segment Sentiment with Control Variables for CAR[0,+3]

Dependent Variable: CAR[0,+3]		
Variables	**(1) LM-sentiment with Controls**	**(2) GPT-sentiment with Controls**
Sentiment	0.3402**	0.00151**
	(0.1218)	(0.0054)
SUE	0.0028***	0.0026***
	(0.0003)	(0.0003)
log(Size)	-0.0030***	-0.0036***
	(0.0009)	(0.0009)
Intercept	0.0666**	0.0680**
	(0.0215)	(0.0214)
Adjusted R^2	0.0418	0.0481
Observations	3333	3333
Model Controls	SUE, log(Size)	SUE, log(Size)

Significance levels: *** p < 0.001, ** p <0.01, * p <0.05
Standard errors in parentheses

Table A.11: Regression Results - Presentation Segment Sentiment with Control Variables for CAR[0,+5]

Dependent Variable: CAR[0,+5]		
Variables	**(1) LM-sentiment with Controls**	**(2) GPT-sentiment with Controls**
Sentiment	0.6912***	0.0281***
	(0.1876)	(0.0042)
SUE	0.0027***	0.0025***
	(0.0003)	(0.0003)
log(Size)	-0.0030***	-0.0037***
	(0.0009)	(0.0009)
Intercept	0.0643**	0.0780***
	(0.0224)	(0.0222)
Adjusted R^2	0.0395	0.0484
Observations	3333	3333
Model Controls	SUE, log(Size)	SUE, log(Size)

Significance levels: *** p < 0.001, ** p <0.01, * p <0.05
Standard errors in parentheses

Table A.12: Regression Results - Q&A Segment Sentiment with Control Variables for CAR[0,+1]

Dependent Variable: CAR[0,+1]		
Variables	**(1) LM-sentiment with Controls**	**(2) GPT-sentiment with Controls**
Sentiment	0.3226 **	0.00478 **
	(0.1104)	(0.0049)
SUE	0.0029 ***	0.0027 ***
	(0.0002)	(0.0002)
log(Size)	-0.0025 **	-0.0031 ***
	(0.0008)	(0.0008)
Intercept	0.0538 **	0.0547 **
	(0.0195)	(0.0194)
Adjusted R^2	0.0533	0.0627
Observations	3340	3340
Model Controls	SUE, log(Size)	SUE, log(Size)

Significance levels: *** p < 0.001, ** p <0.01, * p <0.05
Standard errors in parentheses

Table A.13: Regression Results - Q&A Segment Sentiment with Control Variables for CAR[0,+3]

Dependent Variable: CAR[0,+3]		
Variables	**(1) LM-sentiment with Controls**	**(2) GPT-sentiment with Controls**
Sentiment	0.3402 **	0.00151 **
	(0.1218)	(0.0054)
SUE	0.0028 ***	0.0026 ***
	(0.0003)	(0.0003)
log(Size)	-0.0030 ***	-0.0036 ***
	(0.0009)	(0.0009)
Intercept	0.0666 **	0.0680 **
	(0.0215)	(0.0214)
Adjusted R^2	0.0418	0.0481
Observations	3340	3340
Model Controls	SUE, log(Size)	SUE, log(Size)

Significance levels: *** p < 0.001, ** p <0.01, * p <0.05
Standard errors in parentheses

Table A.14: Regression Results - Q&A Segment Sentiment with Control Variables for CAR[0,+5]

Dependent Variable: CAR[0,+5]		
Variables	**(1) LM-sentiment with Controls**	**(2) GPT-sentiment with Controls**
Sentiment	0.3504 **	0.0014 **
	(0.1264)	(0.0056)
SUE	0.0028 ***	0.0026 ***
	(0.0003)	(0.0003)
log(Size)	-0.0030 ***	-0.0037 ***
	(0.0009)	(0.0009)
Intercept	0.0694 **	0.0710 **
	(0.0223)	(0.0222)
Adjusted R^2	0.0387	0.0437
Observations	3340	3340
Model Controls	SUE, log(Size)	SUE, log(Size)

Significance levels: *** p < 0.001, ** p <0.01, * p <0.05
Standard errors in parentheses

Table A.15: Regression Results - Combined Sentiment Analysis (LM and GPT) for CAR[0,+3]

Dependent Variable: CAR[0,+3]			
Variables	**(1) Full Transcript**	**(2) Presentation Segment**	**(3) Q&A Segment**
GPT-Sentiment	0.0274***	0.0252***	0.0382***
	(0.0045)	(0.0059)	(0.0052)
LM-Sentiment	0.1428	0.1127	0.3165
	(0.2009)	(0.1329)	(0.2087)
SUE	0.0024***	0.0026***	0.0023***
	(0.0003)	(0.0003)	(0.0003)
Intercept	-0.0155***	-0.0200***	-0.0171***
	(0.0019)	(0.0032)	(0.0018)
Adjusted R^2	0.0500	0.0434	0.0552
Observations	3352	3333	3340
Model Controls	SUE	SUE	SUE

Significance levels: *** p < 0.001, ** p <0.01, * p <0.05
Standard errors in parentheses

Table A.16: Regression Results - Combined Sentiment Analysis (LM and GPT) for CAR[0,+5]

Dependent Variable: CAR[0,+5]			
Variables	**(1) Full Transcript**	**(2) Presentation Segment**	**(3) Q&A Segment**
GPT-Sentiment	0.0238***	0.0231***	0.0337***
	(0.0047)	(0.0061)	(0.0054)
LM-Sentiment	0.2292	0.1417	0.3895
	(0.2089)	(0.1380)	(0.2172)
SUE	0.0025***	0.0026***	0.0023***
	(0.0003)	(0.0003)	(0.0003)
Intercept	-0.0151***	-0.0194***	-0.0162***
	(0.0020)	(0.0033)	(0.0019)
Adjusted R^2	0.0438	0.0392	0.0476
Observations	3352	3333	3340
Model Controls	SUE	SUE	SUE

Significance levels: *** p < 0.001, ** p <0.01, * p <0.05
Standard errors in parentheses

Table A.17: Regression Results - Positive and Negative GPT-Sentiment CAR[0,+3]

Dependent Variable: CAR[0,+3]		
Variables	**(1) Positive Sentiment**	**(2) Negative Sentiment**
Sentiment	0.0327***	0.0188
	(0.0086)	(0.0627)
SUE	0.0020***	0.0093***
	(0.0003)	(0.0023)
Intercept	0.0020***	0.0093***
	(0.0003)	(0.0023)
Adjusted R^2	0.0285	0.1910
Observations	2,637	74
Model Controls	SUE	SUE

Significance levels: *** p < 0.001, ** p <0.01, * p <0.05
Standard errors in parentheses

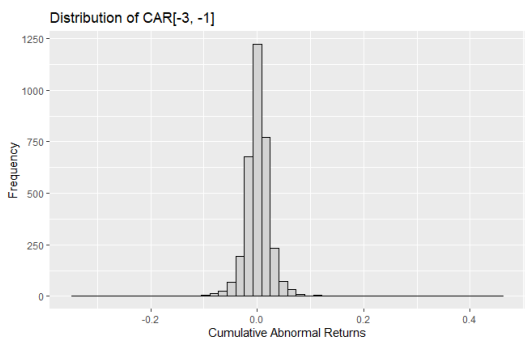
Table A.18: Regression Results - Positive and Negative GPT-Sentiment CAR[0,+5]

Dependent Variable: CAR[0,+5]		
Variables	**(1) Positive Sentiment**	**(2) Negative Sentiment**
Sentiment	0.0303***	-0.0142
	(0.0089)	(0.0648)
SUE	0.0019***	0.0104***
	(0.0003)	(0.0023)
Intercept	-0.0151***	-0.0235
	(0.0045)	(0.0222)
Adjusted R^2	0.0239	0.2071
Observations	2,637	74
Model Controls	SUE	SUE

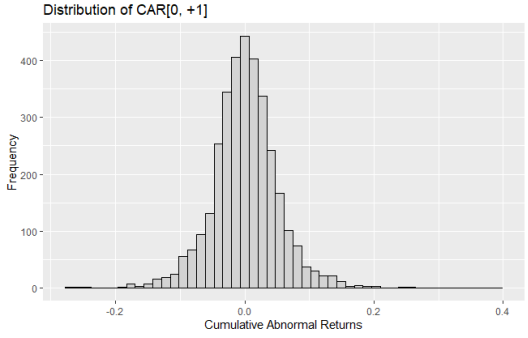
Significance levels: *** p < 0.001, ** p <0.01, * p <0.05
Standard errors in parentheses

B. Figures

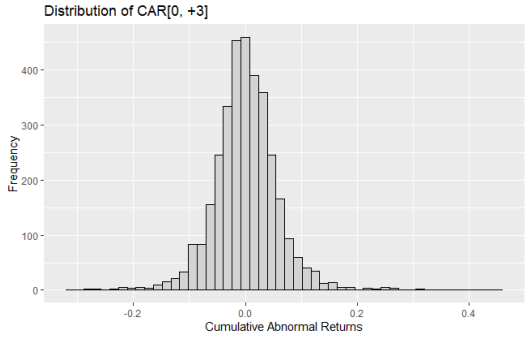
Figure B.1: CARs Distribution



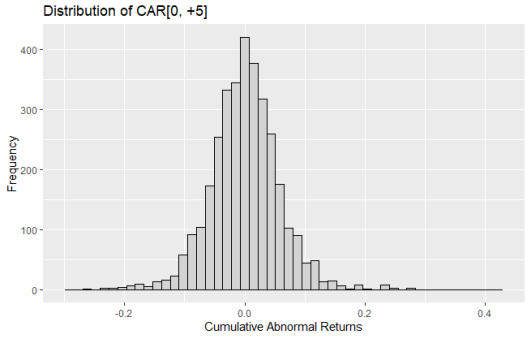
(a) CAR[-3,-1]



(b) CAR[0,+1]

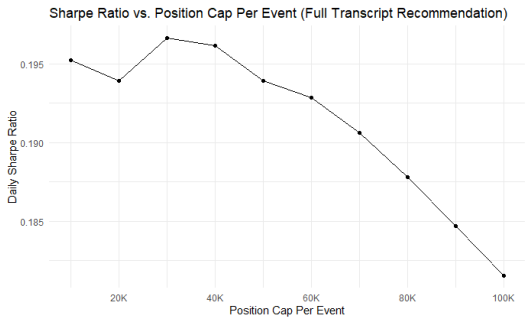


(c) CAR[0,+3]

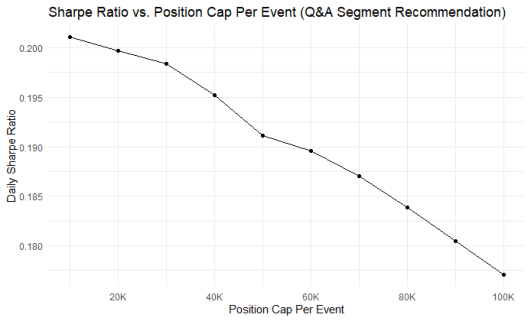


(d) CAR[0,+5]

Figure B.2: Sharpe Ratio vs. Position Cap



(a) Full Transcript Recommendations



(b) Q&A Segment Recommendations

C. Transparency of AI use

This section will document the (responsible) use of AI tools in this thesis. The primary AI tool used, employed as a sentiment analysis tool, was OpenAI's ChatGPT. For further information about the analysis, please refer to Chapter 5 (Methodology Event Study) and Chapter 7 (Methodology Portfolio Simulation). In addition, ChatGPT was employed for coding support in both Python and Rstudio. This included assistance with debugging, understanding errors and improving code structure. Finally, Grammarly was used as a grammar tool, which can be described as a co-reader tool that helps identify unnoticed mistakes in writing. Similar to the grammar functions of Microsoft Word, it uses a built-in AI to suggest synonyms, correct grammatical errors and improve the overall readability of a written text.