ECC Analyzer: Extract Trading Signal from Earnings Conference Calls using Large Language Model for Stock Performance Prediction

Yupeng Cao[*,†] **Zhi Chen**[*,‡], **Qingyun Pei**[*,‡], **Prashant Kumar**[†]
K.P. Subbalakshmi[†], **Papa Momar Ndiaye**[‡]
[*]Equal Contribution
[†]Department of Electrical and Computer Engineering, Stevens Institute of Technology
[‡]School of Business, Stevens Institute of Technology
{ycao33,zchen100,qpei1,pkumar14,ksubbala,pndiaye}@stevens.edu

Abstract

In the realm of financial analytics, leveraging unstructured data, such as earnings conference calls (ECCs), to forecast stock performance is a critical challenge that has attracted both academics and investors. While previous studies have used deep learning-based models to obtain a general view of ECCs, they often fail to capture detailed, complex information. Our study introduces a novel framework: **ECC Analyzer**, combining Large Language Models (LLMs) and multi-modal techniques to extract richer, more predictive insights. The model begins by summarizing the transcript's structure and analyzing the speakers' mode and confidence level by detecting variations in tone and pitch for audio. This analysis helps investors form an overview perception of the ECCs. Moreover, this model uses the Retrieval-Augmented Generation (RAG) based methods to meticulously extract the focuses that have a significant impact on stock performance from an expert's perspective, providing a more targeted analysis. The model goes a step further by enriching these extracted focuses with additional layers of analysis, such as sentiment and audio segment features. By integrating these insights, the ECC Analyzer performs multi-task predictions of stock performance, including volatility, value-at-risk (VaR), and return for different intervals. The results show that our model outperforms traditional analytic benchmarks, confirming the effectiveness of using advanced LLM techniques in financial analytics.

1 Introduction

The integration of structured data such as stock prices and financial ratios with unstructured data including financial filings and company news is increasingly essential in investment decision-making (Fang and Zhang, 2016; Roeder et al., 2022). This trend stems from the recognition that unstructured data can provide insights not fully captured by structured data alone. For instance, financial reports elucidate the broader implications of numerical data through discussions of managerial decisions and corporate strategies (Wang and Hua, 2014), while company news can shed light on public sentiment, emerging market trends, and market perceptions (Kogan et al., 2009; Tetlock, 2007). These nuanced insights are vital for a comprehensive analysis of complex market dynamics.

Advancements in Natural Language Processing (NLP) have significantly enhanced the ability to analyze unstructured data within the financial sector. Traditional NLP applications initially used bag-of-words models, which simplify text into isolated words, thus ignoring syntactical structure and sequence (Zhang et al., 2010). Though effective in specific contexts like fraud detection (Purda and Skillicorn, 2015), these methods lack contextual understanding. More recent developments have introduced deep learning-based NLP algorithms, such as Embedding Language Models (ELMo) (Mikolov et al., 2013) and Long Short-Term Memory Networks (LSTM) (Hochreiter and Schmidhuber, 1997), which provide a more effective capture of textual context. Nevertheless, these supervised learning-based methods are highly task-specific and have limited adaptability to generalization (Singh et al., 2016).

The emergence of Large Language Models (LLMs) represents a paradigm shift in overcoming these limitations (Brown et al., 2020; Touvron et al., 2023; Achiam et al., 2023). Equipped with expansive knowledge bases and sophisticated zero-shot learning capabilities, LLMs are capable of performing a diverse array of text-related tasks—ranging from summarization (Zhang et al., 2024) and question-answering (Wei et al., 2022) to sentiment analysis (Zhang et al., 2023) and E-Commerce (Jia et al., 2023) without the need for specialized training in specific tasks or domains (Jia et al., 2022; Wang et al., 2023b; Yang et al., 2023; Gruver et al., 2024).

Drawing inspiration from recent advancements in LLMs, this work aims to explore the potential of LLMs to extract trading signals from earnings conference calls (ECCs) to improve predictions of stock performance. ECCs involve senior executives discussing quarterly results, providing a fertile ground for predicting stock movements through nuanced analysis of transcripts and audio. Despite the potential, existing research often overlooks finer details, lacks interpretability, or relies too heavily on sentiment analysis, leading to incomplete data interpretations.

Recent studies have increasingly leveraged domain-specific Large Language Models like FinBERT (Liu et al., 2021), which is adapted from Google's BERT algorithm for financial contexts. FinBERT evaluates the sentiment of ECCs by averaging the sentiment scores of each sentence, revealing a correlation between the sentiment and market reactions. Moreover, the integration of multimodal techniques has significantly improved the accuracy of financial risk predictions. For example, (Qin and Yang, 2019) and (Yang et al., 2020) utilize both textual and auditory data to generate embeddings that encapsulate semantic and auditory features. These models have proven effective in predicting market volatility at various intervals, demonstrating the potential of combining textual analysis with auditory data processing. While these studies may get the descent performance of prediction, another concern with using deep learning-based or LLM techniques is the challenge of explaining how the model arrives at its results. Several interpretability models have been proposed to explain the model's decision reasons. (Wang et al., 2023a; Tenney et al., 2020).

Existing research, however, reveals significant gaps that warrant further exploration: 1) the previous studies directly input entire texts or audio files into models, potentially missing important details and lacking interpretability, especially regarding which earnings call topics influence predictions; 2) While some studies focus on sentiment extraction from ECCs, they capture only a fraction of the available information, leading to potentially incomplete interpretations. The findings suggest that sentiment analysis alone offers limited explanatory power for predicting stock movements, highlighting the need for broader data utilization; 3) Additionally, integrating large language models (LLMs) into financial analysis while ensuring that investors understand the reasoning behind model outputs poses a distinct challenge, prompting ongoing research.

Given these gaps, in designing a framework to extract trading signals from ECCs to predict financial performance, this paper is interested in exploring the following research questions **(RQs)**:

- **RQ1**: How can large language models be used to provide investors with a more comprehensive understanding of a company's financial health and strategic direction?
- **RQ2**: Can a more comprehensive analysis provide additional predictive capability for stock performance?
- **RQ3**: Can LLMs be employed to generate interpretable content that aids investors in understanding the decision-making process?

To address the above **RQs**, this paper introduces the ECC Analyzer, a novel framework utilizing LLMs for in-depth analysis of ECC data. The framework initially provides an understanding of ECCs by segmenting transcripts into themes like financial performance and corporation project discussions. It summarizes these segments, distilling the essence of each thematic chunk, and then combines these summaries into a comprehensive overview. This hierarchical summarization enables stakeholders to grasp complex documents' main themes and insights. Regarding audio, the model analyzes speech features such as tone, pitch, and intensity to gauge the speaker's confidence level.

Furthermore, the ECC Analyzer simulates how investors examine key indicators and infer future market behavior. We begin by creating a database (**focus**) of key indicators with finance experts, such as financial metrics, management changes, operational costs, and strategic plans. Using the constructed database, the ECC Analyzer employs retrieval-augmented generation (RAG) to systematically examine ECCs and pinpoint factors critical to investment decisions. After thoroughly extracting and analyzing the ECC, the results are integrated with raw ECC data to conduct multi-task learning: predicting volatility as well as Value at Risk (VaR) and return for different time intervals (3, 7, 15, and 30 days).

By utilizing RAG, our method improves model interpretability by linking specific earnings call topics directly to stock performance, enhancing both the investor's understanding and the explainability of the analysis. To validate our approach, we

benchmarked our model against traditional methods, showing significant improvements in accuracy and predictive power. These results highlight our model's utility for financial analysts and investors in making informed decisions based on detailed data analysis.

2 Related Work

This paper integrates concepts from stock performance prediction and advancements in large language models (LLMs). In this section, we offer a concise review of key research in both areas relevant to our study.

2.1 Stock Performance Prediction

The use of NLP techniques to analyze unstructured data for predicting stock performance has attracted significant academic attention. A foundational study by (Kogan et al., 2009) shows that simple bag-of-words features from annual reports when combined with historical volatility, can outperform models based solely on historical data. Subsequent research, such as that by (Wang and Hua, 2014; Rekabsaz et al., 2017; Theil et al., 2018), proposed various document representation methods to predict stock price volatility. Drawing on multimodal technologies, (Qin and Yang, 2019) explored how audio features—such as tone, emotion, and speech rate—enhance stock movement predictions when combined with text analysis. Following by this, (Yang et al., 2020) further extends the idea of using multimodal data to improve risk prediction performance in multi-task learning, and the authors' experiments show that predicting multiple tasks at the same time can help the model further improve prediction performance. However, the aforementioned studies primarily input ECC data directly into models for prediction without conducting a thorough analysis of the ECC content.

2.2 Large Language Models in Finance

Numerous studies have explored the applications of LLMs in the financial sector. (Li et al., 2023) explore how LLMs have been adeptly applied to summarize and abstract complex financial documents such as 10-K, and 10-Q filings. (Yang et al., 2023; Yu et al., 2023) explores the usage of LLMs in mining media news for trading recommendations, showcasing the models' ability to discern subtle market indicators and sentiments. In the domain of customer service, the implementation of LLM-powered chatbots is spotlighted for offering context-aware interactions, serving as both assistants and consultants (Lakhani, 2023; Subagja et al., 2023; Soni, 2023). (Abdaljalil and Bouamor, 2021; Zmandar et al., 2021) explore the nuanced task of extracting financial and legal items from lengthy text documents, such as financial regulations and comprehensive policy manuals. However, these existing studies predominantly focus on tasks like financial text summarization, question-answering (Q&A), and stock movement prediction (binary classification), with a notable gap in the application of LLMs for comprehensive stock performance prediction.

3 Methodology

ECC Analyzer, illustrated in Figure (1) aims to comprehensively understand the multi-data types present in earnings conference calls, including both text and audio components. (3.1) audio encoding (3.2) transcript encoding. Furthermore, it focuses on extracting trading signals from the analyzed data to predict stock performance (3.3) earnings conference call focuses extraction. The model includes a component to optimally integrate different data sources (3.4) additive multi-model fusion. The model performs multi-task prediction: to predict the the volatility across different terms and the value at risk (VaR) at the same time (3.5). It also demonstrates strong predictive ability in forecasting stock returns.

3.1 Audio Encoding

Audio pre-trained models have achieved performing results in various downstream tasks (Pons and Serra, 2019; Cramer et al., 2019; Koh and Dubnov, 2021; Wang et al., 2021). We aim to leverage advanced audio pre-trained models like Wav2vec2, a transformer-based Large Language Model recognized for its effectiveness in processing raw audio (Baevski et al., 2020), to extract audio embeddings. After that, we employ a Multi-Head Self-Attention (MHSA) mechanism to distill specific audio features. This method is vital for integrating these features with other data modalities, facilitating a more detailed and comprehensive analysis.

To describe the Audio Encoding in more detail, we let the raw audio input data be represented by $A_c = \{a_c^1, a_c^2, \ldots, a_c^n\}$ where a_c^i represents the i^{th} audio frame in one data sample. Each audio frame will be converted into a vector representa-

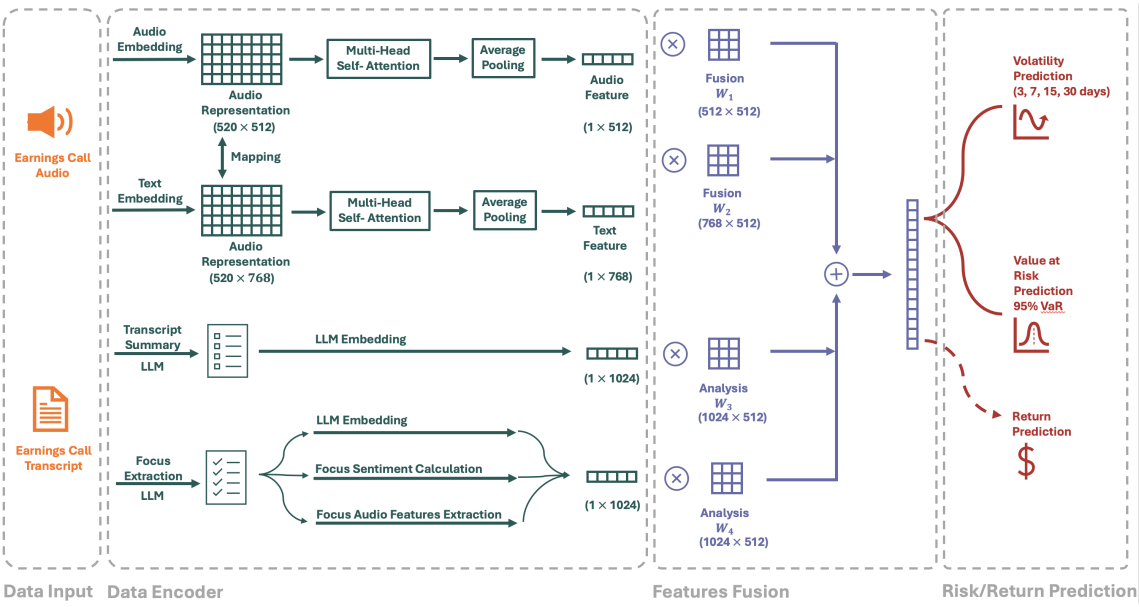

Figure 1: This figure illustrates the ECCs Analyzer's Framework. The model accepts multimodal inputs: ECC Audio and Transcript. The second area visualizes the model's pipeline to encode text and audio sources and illustrates how LLMs are applied for information analysis. This model analyzes text-audio pairs, summarizes ECCs, and extracts focal points of interest from a financial expert's perspective. The third area describes how the model consolidates outputs from both embeddings and LLM analysis for use in subsequent stages. The model will perform multi-task learning: our ECCs Analyzer will predict the Volatility of different terms and VaR in the meantime.

tion: $e^i_{ac} = \text{Wav2Vec2}(a^i_c)$. Therefore, we obtain the audio embeddings $E_{ac} = \{e^1_{ac}, e^2_{ac}, \ldots, e^n_{ac}\}$ which have dimensions of 520×512, representing the maximum number of audio files across companies and the transform dimensions for a single audio frame, respectively. Audio files with fewer than 520 frames ($n < 520$) are zero-padded for consistent matrix size.

E_{ac} are then processed through a MHSA to distill specific audio features. The MHSA includes a multi-head attention block, a norm block, and a two-layer feed-forward network with ReLU activation, forming the basis for all subsequent architectures discussed. In detail, the MHSA calculation process is as follows:

$$\text{Multihead} = \text{Concat}(\text{head}_1, ..., \text{head}_h)W^o \quad (1)$$

$$\text{head}_i = \text{Attention}(QW^Q_i, KW^K_i, VW^V_i) \quad (2)$$

where Q (queries) and K (keys) of dimension d_k and V values of dimension d_v. The weights dimensions are: $W^Q_i, W^K_i, W^V_i \in R^{d_{model} \times d_k, d_k, d_v}$ respectively, and $W^o \in R^{d_v \times d_{model}}$. The dot product is then calculated for the query with all the keys. The attention scores are normalized using the softmax function:

$$\text{Attention}(Q, K, V) = softmax(\frac{KQ^T}{\sqrt{d_k}})V \quad (3)$$

The attention function on a set of queries is calculated simultaneously packed together in a matrix Q. The keys and values are also packed in the matrices K and V respectively. Combining (2)-(4), this results in a matrix:

$$T_{ac} = \text{MHSA}(E_{ac}) \quad (4)$$

where $T_{ac} = \{t^1_{ac}, t^2_{ac}, \ldots, t^n_{ac}\}$ with size 520×512. T_{ac} is then subjected to an average pooling layer to produce T_a, a condensed audio feature vector of size 512.

3.2 Transcript Encoding

The transcription process mirrors Audio Encoding, using SimCSE (Gao et al., 2021)to extract sentence-level vector representations from earnings conference transcripts. SimCSE is a Siamese neural network architecture that learns to embed pairs of sentences into a shared space where similar sentences are mapped close together and dissimilar sentences are mapped far apart. The raw transcripts are represented as $T_c = \{t^1_c, a^2_c, \ldots, t^n_c\}$, with each sentence t^i_c represents the i^{th} transformed into a vector representation: $e^i_{tc} = \text{SimCSE}(t^i_c)$.

We obtain the corresponding text embeddings given by $E_{tc} = \{e^1_{tc}, e^2_{tc}, \ldots, e^n_{tc}\}$ with size 520×768, where 520 is the maximum number of sen-

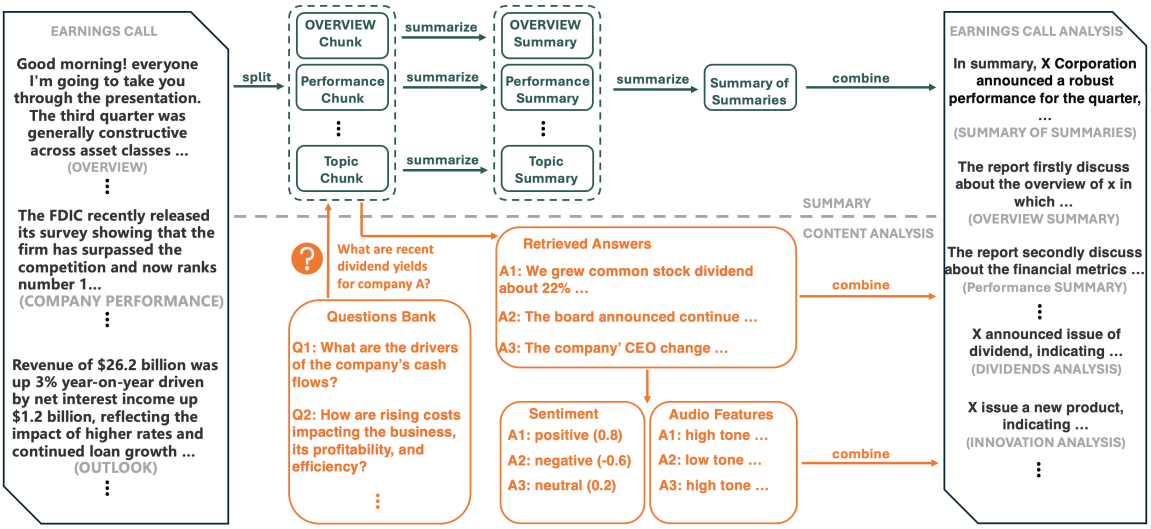

Figure 2: This figure visualizes the mechanism of the earnings conference call analyzer. This framework takes both earnings conference calls' audio and transcript as input. It starts by hierarchically summarizing the content of ECCs. Further, this model simulates the process through which an investor analyzes an ECC. It examines and extracts focuses on interest, such as financial indicators and business events, from a database designed by financial experts. Subsequently, this model calculates the sentiment for these focal points and extracts audio features, such as changes in tone and pitch, from the corresponding audio segments. All these analyses are combined to form a comprehensive encoder for further use.

tences amongst all data samples and 768 is the dimension of the output of SimCSE. Earnings conference calls with less than 520 sentences ($n < 520$) have been zero-padded for uniformity in input matrix size. Same with (1)-(4), the MHSA is applied to E_{tc} to get $T_{tc} = \{t_{tc}^1, t_{tc}^2, \ldots, t_{tc}^n\}$ with dimension 520×768. Then,T_{tc} is subjected to the average pooling layer to produce T_t, where T_t denotes the resultant extracted textual feature of size 768. This streamlined process effectively captures the textual nuances required for in-depth analysis.

3.3 Earnings Conference Call Focuses Extraction And Analysis

To obtain deep insights from an Earnings Conference Call on how it might influence future market performance, our approach encompasses several steps: (1) summarize ECCs, (2) extract investors' focus information (3) calculate focus sentiment and (4) extract audio's features of corresponding focus. Refer to the Figure 2 for additional details.

(1) Summarize ECCs. To accommodate the token limitations of LLMs, we start by dividing the entire document into several chunks and summarizing each one. We then summary these summaries to create a comprehensive overview of the entire document. This two-tiered approach ensures that the summary capture both the detailed and overall information. In further, we will use LLM to get the embedding with a size 1024: $T_s = \text{LLM}(summary + chunk\ summaries)$

(2) Extract the investor's focus using RAG. In this step, we attempt to use LLM to examine and extract focuses of interest to investors. Therefore, we start by identifying the core topics that analysts typically focus on. We consult with four finance experts, each possessing extensive experience in financial analysis. These experts are tasked with meticulously reviewing earnings conference calls across various sectors to pinpoint and summarize the key topics that are frequently discussed and also hold the greatest interest for investors. Their analysis aims to identify overlaps where common discussion topics align closely with investor concerns, highlighting areas of particular significance and interest in the investment community. Once we have established this focused database, our goal is to locate where these topics are mentioned during the earnings calls and systematically extract these segments. However, accurately extracting this information from the dense narrative of earnings conference calls can pose a challenge. Because LLM searches are based on similarity, comparing a single-word topic to an entire document presents a scale issue, and a single word may lack the necessary context, thus diminishing the extraction ca-

pability. To enhance the precision of information extraction, we formulate multiple questions related to each topic but phrased differently, ensuring a broad coverage and increasing the likelihood of accurately pinpointing relevant sections. For instance, regarding the dividends topic, we pose questions such as: (1) "Have there been any changes in the stock dividends, and if so, at what rate have the dividends increased or decreased?" (2) Does the company Board expect to increase the stock dividends in the next future? (3) How does the dividend yield compare to the peers?

After locating the segments where these topics are discussed, we extract the corresponding paragraphs. These paragraphs, however, may still contain irrelevant text such as preceding discussions. To address this, we use a contextual compression method from LLM, adept at distilling the paragraph down to its most relevant sentences, effectively obtaining the crucial data from the irrelevant text.

(3) Calculate focus sentiment. Sentiment is a strong indicator for reflecting market perception. We determined to calculate the sentiment score based on the focus. Because this offers a more targeted and insightful view, which may be overlooked in a broader context. It provides a clearer picture of how specific developments or concerns are influencing investor behavior and market movements.

(4) Extract audio's features using RAG. Once we identify the focal points of interest, we aim to enrich our analysis by locating and examining the audio segments that correspond to these specific topics. By analyzing the audio features related to a particular focus, such as the company's current projects, we can delve deeper into the emotional nuances and confidence levels exhibited by the speakers. This approach allows us to capture subtle cues in tone, pace, and emphasis that might indicate underlying sentiments or confidence about the discussed topics. The goal is to provide a more layered understanding of how speakers convey their messages and the potential impacts these emotional expressions have on the listeners' perceptions and reactions. We utilize Praat (Boersma and Van Heuven, 2001) to extract vocal features, such as pitch, intensity, jitter, HNR(Harmonic to Noise Ratio) and etc, from ECC audio files.

After obtaining the four key components, we then explore effective methods to merge these elements into a cohesive analysis. In further, we will use LLM to get the embedding with size 1024: $T_f = \text{LLM}(focus\ analysis + focus\ sentiment + focus\ audio\ features)$

3.4 Additive Multi-modal Fusion

Given the model's reliance on several inputs and diverse data types, we identify an effective fusion structure to integrate these features into the training process to ensure a balanced weighting among components. We use additive interactions to handle the representational fusion of different abstract representations. These operators can be viewed as differentiable building blocks that combine information from several different data streams and can be flexibly inserted into almost any unimodal pipeline (Liang et al., 2022). Given the audio feature T_a, textual feature T_t from the transcript, and T_s, T_f from ECC analyzed text, additive fusion can be seen as learning a new joint representation:

$$E = w_0 + w_1 \cdot T_a + w_2 \cdot T_t + w_3 \cdot T_s + w_4 \cdot T_f + \epsilon \quad (5)$$

where $w_1 \in R^{512\times512}$, $w_2 \in R^{768\times512}$ and $w_3, w_4 \in R^{1024\times512}$ are the weights learned for additive fusion, w_0 the bias term and ϵ the error term. E is a vector with 512 as the final feature from the Earning Conference Call Encoder.

3.5 Multi-Task Prediction

We begin our prediction process by aggregating features from various modules into a comprehensive feature representation. This unified feature set is fed into a two-layer neural network designed to perform the regression task. Integrating these diverse inputs into a cohesive output is crucial, as it harnesses the strengths of each module to enhance analysis and prediction accuracy.

Building on insights from previous research in multimodal financial risk prediction, which has demonstrated substantial improvements in prediction performance through multitask learning, we adopt a joint modeling approach. Here, we concurrently model volatility prediction and VaR prediction using a multi-task framework. The multi-task prediction module is comprised of two separate single-layer feedforward networks, each responsible for predicting volatility (vol) and Value at Risk (var) values individually. We train ECC Analyzer

by optimizing multitask loss:

$$\mathcal{L} = \mu(\sum_i (\hat{y}_i - y_i)^2) + (1-\mu)\max(q \times (v - \hat{v}), (1-q)(\hat{v} - v)) \quad (6)$$

multi-task learning allows us to optimize performance by accurately capturing and predicting multiple stock performance factors simultaneously.

4 Results and Discussions

4.1 Dataset

The dataset utilized in this study is sourced from the publicly available S&P 500 ECC dataset as constructed by (Qin and Yang, 2019). It includes both audio recordings and corresponding text transcripts from the 2017 earnings calls of 500 major companies listed on the S&P 500 and traded on U.S. stock exchanges. The dataset consists of 572 unique instances where the audio recordings were accurately and closely aligned with the text transcripts. Following the setup by (Qin and Yang, 2019), we partitioned the dataset into a training set and a test set with an 8:2 ratio, organized temporally to ensure that the data in the training set precedes those in the test set. This temporal division is crucial for maintaining the integrity of our predictive model, aligning the training process with the principle of using historical data to predict future risks—thus enhancing the accuracy and reliability of our forecasting approach.

4.2 Baseline

We compare our approach to several important baselines including 1) GARCH-based Classical Methods; 2) LSTM (Gers et al., 2000) model; 3) MT-LSTM+ATT (Luong et al., 2015) employing an attention-enhanced LSTM as the foundational model; 4) HAN (Glove) uses a Hierarchical Attention Network with dual-layered attention at the word and sentence levels for prediction; 5) MRDM (Qin and Yang, 2019) proposed a multimodal deep regression approach for volatility prediction tasks; 6) HTML (Yang et al., 2020) presented a state-of-the-art method; and 7) GPT-4-turbo-2024-04-09: directly utilize LLM to predict stock performance. We explain each baseline method in detail in the Appendix A.1

4.3 Implementation Details

We use GPT-4 for the ECC Analyzer experiment, utilizing it to analyze ECC data and build the analysis database. Throughout this process, we set the temperature parameter to 0. This ensures that the Large Language Models (LLMs) produce the most predictable responses, which aids in maintaining consistency in our experiments.

For the overall training of the ECC Analyer framework, we developed the code using PyTorch. Each Multi-Head Attention layer in the network comprises 6 layers and 8 individual heads in each layer. The training process utilized batch sizes $b \in \{2, 4, 8\}$. We use a grid search to determine the optimal parameters and select the learning rate λ for Adam optimizer among $\{1e-3, 1e-5, 1e-6, 1e-7\}$. The best hyper-parameters were kept consistent across all experiments, with the exception of the trade-off parameter μ which varied between the two tasks. We list the evaluation metrics in Appendix A.2.

4.4 Overall Results Analysis (RQ1)

Table 1 shows the performance of various methods in predicting stock performance. Notably, the ECC Analyzer framework excels, especially in short-term and medium-term forecasts, with the lowest Mean Squared Error (MSE) values. Its long-term prediction performance is comparable to the state-of-the-art method, HTML. This improvement highlights the benefits of deep analysis using LLMs extracted from ECCs. However, directly applying LLMs to stock performance prediction proves largely ineffective, similar to random guesses. This indicates that LLMs are more effective as tools to enhance investors' understanding of a company's financial health rather than direct predictors of financial metrics.

Additionally, the ECC Analyzer demonstrates outstanding performance in Value at Risk (VaR) prediction, confirming the efficacy of our methodology in providing a nuanced and comprehensive approach to financial risk prediction. We also report the results on Returns in Appendix B.

4.5 Ablation Study (RQ2)

In our research, we conducted an ablation study to assess how different combinations of ECC analysis results impact our model's performance. This systematic comparison helped us identify the individual contributions of each component.

According to Table 2, we can find that the audio and text features extracted by advanced large language models significantly improved short-term prediction accuracy compared to previous methods.

Table 1: Performance results on our proposed framework ECC Analyzer from different baseline models.

Model	$\overline{MSE}$	MSE_3	MSE_7	MSE_{15}	MSE_{30}	VaR	Multi-Task
Classical Method	0.713	1.710	0.526	0.330	0.284	/	⊗
LSTM	0.746	1.970	0.459	0.320	0.235	/	⊗
MT-LSTM-ATT	0.739	1.983	0.435	0.304	0.233	/	⊗
HAN	0.598	1.426	0.461	0.308	0.198	/	⊗
MRDM	0.577	1.371	0.420	0.300	0.217	/	⊗
HTML	0.401	0.845	0.349	0.251	**0.158**	/	✓
GPT-4-Turbo	2.198	3.187	5.059	7.959	11.824	0.371	✓
ECC Analyzer	**0.316**	**0.553**	**0.306**	**0.247**	0.159	**0.049**	✓

Table 2: Performance results of ablation study. We designed the ablation study as follows: 1) Audio+Text: uses raw audio and text data from ECCs; 2)Audio+Text+E_{os}: adds a comprehensive ECC summary generated by LLMs; 3) Audio+Text+$E_{os} + E_{cs}$: integrates both overall and chunk summaries for the ECC; 4) Audio+Text+E_{fo}: combines raw data with focused analytical results; 5) $E_{os} + E_{cs} + E_{fo}$: merges all LLM analyses for prediction without raw data; 6) Audio+Text+$E_{os} + E_{cs} + E_{fo}$: combines all data and analyses for enhanced prediction.

Module	$\overline{MSE}$	MSE_3	MSE_7	MSE_{15}	MSE_{30}
Audio+Text	0.373	0.645	0.362	0.28	0.204
Audio+Text+E_{os}	0.373	0.638	0.380	0.276	0.201
Audio+Text+$E_{os} + E_{cs}$	0.357	0.627	0.335	0.267	0.199
Audio+Text+E_{fo}	0.324	0.579	0.323	0.23	0.165
$E_{os} + E_{cs} + E_{fo}$	0.343	0.601	0.344	0.247	0.179
Audio+Text+$E_{os} + E_{cs} + E_{fo}$	**0.310**	**0.553**	**0.306**	**0.22**	**0.159**

Furthermore, incorporating summaries of the data slightly enhanced performance, but more notable improvements were observed when we added analysis of specific focus points. This indicates that our model effectively isolates and utilizes the most relevant information for predicting stock movements. Our best analytical results come from integrating the full spectrum of data and analytical outputs, underscoring the value of each component in our model. Notably, good predictive results were also can obtained using only analytics derived from LLMs, affirming the response to our RQ2: comprehensive analysis indeed enhances the predictive capability for stock performance.

Our findings also suggest that while earnings calls are information-rich, including every detail in the analysis can be counterproductive and may cloud essential insights. It is therefore critical to pinpoint and concentrate on the most predictive elements of the data, filtering out less relevant information to optimize the analysis process for stock performance prediction.

4.6 Interpretability Study (RQ3)

In our study, we initiate the analysis by collaborating with consultants to summarize and identify key focus points from earnings conference calls; the more comprehensive the summary, the better. Following this initial step, we combine these focus points and try different combinations to determine the optimal subset that most strongly predicts stock performance. (The finalized list of focus points is included in the appendix.) This method serves a dual purpose for investors. Firstly, it reveals which focus points are highly relevant to stock performance, allowing investors to prioritize their attention effectively. Secondly, by extracting and analyzing these key focuses, investors can understand the specific contents in the ECC that drive movements in the stock market.

5 Conclusions

This study confirms the effectiveness of using Large Language Models (LLMs) to extract and analyze key topics from earnings conference calls. Our approach not only pinpoints critical discus-

sions but also assesses their impact on stock performance, predicting metrics like volatility and Value at Risk (VaR). By integrating textual and audio data, our model offers a comprehensive view, capturing subtleties such as tone and pitch. Results from benchmark comparisons demonstrate our model's superior accuracy and predictive capabilities, highlighting LLMs' potential to improve interpretability and decision-making in investments.

References

Samir Abdaljalil and Houda Bouamor. 2021. An exploration of automatic text summarization of financial reports. In *Proceedings of the Third Workshop on Financial Technology and Natural Language Processing*, pages 1–7.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Paul Boersma and Vincent Van Heuven. 2001. Speak and unspeak with praat. *Glot International*, 5(9/10):341–347.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Aurora Linh Cramer, Ho-Hsiang Wu, Justin Salamon, and Juan Pablo Bello. 2019. Look, listen, and learn more: Design choices for deep audio embeddings. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3852–3856. IEEE.

Bin Fang and Peng Zhang. 2016. Big data in finance. *Big data concepts, theories, and applications*, pages 391–412.

Philip Hans Franses and Dick Van Dijk. 1996. Forecasting stock market volatility using (non-linear) garch models. *Journal of forecasting*, 15(3):229–235.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. 2000. Learning to forget: Continual prediction with lstm. *Neural computation*, 12(10):2451–2471.

Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew G Wilson. 2024. Large language models are zero-shot time series forecasters. *Advances in Neural Information Processing Systems*, 36.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Qinjin Jia, Yupeng Cao, and Edward Gehringer. 2022. Starting from "zero": An incremental zero-shot learning approach for assessing peer feedback comments. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 46–50.

Qinjin Jia, Yang Liu, Daoping Wu, Shaoyuan Xu, Huidong Liu, Jinmiao Fu, Roland Vollgraf, and Bryan Wang. 2023. Kg-flip: Knowledge-guided fashion-domain language-image pre-training for e-commerce. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 81–88.

Ha Young Kim and Chang Hyun Won. 2018. Forecasting the volatility of stock price index: A hybrid model integrating lstm with multiple garch-type models. *Expert Systems with Applications*, 103:25–37.

Shimon Kogan, Dimitry Levin, Bryan R Routledge, Jacob S Sagi, and Noah A Smith. 2009. Predicting risk from financial reports with regression. In *Proceedings of human language technologies: the 2009 annual conference of the North American Chapter of the Association for Computational Linguistics*, pages 272–280.

Eunjeong Koh and Shlomo Dubnov. 2021. Comparison and analysis of deep audio embeddings for music emotion recognition. *arXiv preprint arXiv:2104.06517*.

Akbar Lakhani. 2023. Enhancing customer service with chatgpt transforming the way businesses interact with customers.

Seoki Lee and Daniel J Connolly. 2010. The impact of it news on hospitality firm value using cumulative abnormal returns (cars). *International Journal of Hospitality Management*, 29(3):354–362.

Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. 2023. Large language models in finance: A survey. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pages 374–382.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2022. Foundations and trends in multimodal machine learning: Principles, challenges, and open questions. *arXiv preprint arXiv:2209.03430*.

Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. 2021. Finbert: A pre-trained financial language representation model for financial text mining. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence*, pages 4513–4519.

Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2015. Multi-task sequence to sequence learning. *arXiv preprint arXiv:1511.06114*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Jordi Pons and Xavier Serra. 2019. musicnn: Pre-trained convolutional neural networks for music audio tagging. *arXiv preprint arXiv:1909.06654*.

Lynnette Purda and David Skillicorn. 2015. Accounting variables, deception, and a bag of words: Assessing the tools of fraud detection. *Contemporary Accounting Research*, 32(3):1193–1223.

Yu Qin and Yi Yang. 2019. What you say and how you say it matters: Predicting stock volatility using verbal and vocal cues. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 390–401.

Navid Rekabsaz, Mihai Lupu, Artem Baklanov, Allan Hanbury, Alexander Dür, and Linda Anderson. 2017. Volatility prediction using financial disclosures sentiments with word embedding-based ir models. *arXiv preprint arXiv:1702.01978*.

Jan Roeder, Matthias Palmer, and Jan Muntermann. 2022. Data-driven decision-making in credit risk management: The information value of analyst reports. *Decision Support Systems*, 158:113770.

Amanpreet Singh, Narina Thakur, and Aakanksha Sharma. 2016. A review of supervised machine learning algorithms. In *2016 3rd international conference on computing for sustainable global development (INDIACom)*, pages 1310–1315. Ieee.

Vishvesh Soni. 2023. Large language models for enhancing customer lifecycle management. *Journal of Empirical Social Science Studies*, 7(1):67–89.

Agus Dedi Subagja, Abu Muna Almaududi Ausat, Ade Risna Sari, M Indre Wanof, and Suherlan Suherlan. 2023. Improving customer service quality in msmes through the use of chatgpt. *Jurnal Minfo Polgan*, 12(1):380–386.

Ian Tenney, James Wexler, Jasmijn Bastings, Tolga Bolukbasi, Andy Coenen, Sebastian Gehrmann, Ellen Jiang, Mahima Pushkarna, Carey Radebaugh, Emily Reif, et al. 2020. The language interpretability tool: Extensible, interactive visualizations and analysis for nlp models. *arXiv preprint arXiv:2008.05122*.

Paul C Tetlock. 2007. Giving content to investor sentiment: The role of media in the stock market. *The Journal of finance*, 62(3):1139–1168.

Christoph Kilian Theil, Sanja Štajner, and Heiner Stuckenschmidt. 2018. Word embeddings-based uncertainty detection in financial disclosures. In *Proceedings of the first workshop on economics and natural language processing*, pages 32–37.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Dan Wang, Zhi Chen, Ionuţ Florescu, and Bingyang Wen. 2023a. A sparsity algorithm for finding optimal counterfactual explanations: Application to corporate credit rating. *Research in International Business and Finance*, 64:101869.

Ning Wang, Yupeng Cao, Shuai Hao, Zongru Shao, and KP Subbalakshmi. 2021. Modular multi-modal attention network for alzheimer's disease detection using patient audio and language data. In *Interspeech*, pages 3835–3839.

William Yang Wang and Zhenhao Hua. 2014. A semiparametric gaussian copula regression model for predicting financial risks from earnings calls. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1155–1165.

Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, et al. 2023b. Factcheck-gpt: End-to-end fine-grained document-level fact-checking and correction of llm output. *arXiv preprint arXiv:2311.09000*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. Fingpt: Open-source financial large language models. *arXiv preprint arXiv:2306.06031*.

Linyi Yang, Tin Lok James Ng, Barry Smyth, and Riuhai Dong. 2020. Html: Hierarchical transformer-based multi-task learning for volatility prediction. In *Proceedings of The Web Conference 2020*, pages 441–451.

Yangyang Yu, Haohang Li, Zhi Chen, Yuechen Jiang, Yang Li, Denghui Zhang, Rong Liu, Jordan W Suchow, and Khaldoun Khashanah. 2023. Finmem: A performance-enhanced llm trading agent with layered memory and character design. *arXiv preprint arXiv:2311.13743*.

Boyu Zhang, Hongyang Yang, Tianyu Zhou, Muhammad Ali Babar, and Xiao-Yang Liu. 2023. Enhancing financial sentiment analysis via retrieval augmented large language models. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pages 349–356.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Yin Zhang, Rong Jin, and Zhi-Hua Zhou. 2010. Understanding bag-of-words model: a statistical framework. *International journal of machine learning and cybernetics*, 1:43–52.

Nadhem Zmandar, Abhishek Singh, Mahmoud El-Haj, and Paul Rayson. 2021. Joint abstractive and extractive method for long financial document summarization. In *Proceedings of the 3rd Financial Narrative Processing Workshop*, pages 99–105.

A Supplement Material for Experiment Setup

A.1 Baseline Setup

- **Classical Methods:** We incorporate the GARCH model and its derivatives, as described in (Franses and Van Dijk, 1996; Kim and Won, 2018). These models are well-recognized for short-term volatility prediction but may not be as effective for forecasting average volatility over longer periods, such as n-day volatility.

- **LSTM (Gers et al., 2000)**: Long Short-Term Memory Networks (LSTMs) are a popular choice for financial time series prediction due to their efficacy in handling sequential data. We use a straightforward LSTM model as a benchmark for volatility prediction.

- **MT-LSTM+ATT (Luong et al., 2015):** combines the prediction of average n-day volatility with the forecasting of single-day volatility, employing an attention-enhanced LSTM as the foundational model.

- **HAN (Glove):** uses a Hierarchical Attention Network with dual-layered attention at the word and sentence levels. HAN first get word embeddings using pre-trained Glove vectors and then processed by a Bi-GRU (Chung et al., 2014) encoder, while another Bi-GRU encoder simultaneously forms a sentence-level representation of each document. The resulting document representation is input into a regression layer to produce predictions.

- **MRDM (Qin and Yang, 2019):** The MRDM model first introduced a multi-modal deep regression approach to fuse the GloVe embeddings and hand-crafted acoustic features for volatility prediction tasks.

- **HTML (Yang et al., 2020):** This work presented a state-of-the-art model that employs WWM-BERT for text token encoding. Similar to MDRM, HTML also leverages the same audio features. These unimodal features are then combined and processed through a sentence-level transformer, resulting in multimodal representations for each call.

- **GPT-4-turbo-2024-04-09:** We assessed the capability of LLMs in directly predicting stock performance from ECCs. The model was set to generate response with a zero temperature setting to ensure deterministic output.

A.2 Evalation Metrics

A.3 Volatility of Different Future Intervals

We used the following formula to calculate the Volatility of Different Terms:

$$\sigma_{ijk}^2 = \frac{1}{k-1}\sum_{m=1}^{k}[S_{i,(j-m)} - \frac{1}{k}\sum_{l=1}^{k} S_{i,(j-l)}]^2$$

Here, we define the σ_{ijk} as the $k-days$ volatility of Stock i at time j, which is calculated as the sample standard deviation of the past $k-days$ stock closed price of company i. But in the real application, we took the log form of historical volatility.

A.4 VaR of Different Future Intervals

Our second task is predicting the 1-day VaR of the target stock based on the multi-source inputs. The definition of VaR is:

$$x = F^{-1}(p)$$

The $F(\cdot)$ is the cumulative loss distribution, p is the percentile we set, and x is the VaR. From the idea of Quantile Regression, we can have:

$$L_\tau(y, \hat{y}) = \begin{cases} \tau \cdot (y - \hat{y}) & \text{if } y \geq \hat{y} \\ (1-\tau) \cdot (\hat{y} - y) & \text{if } y < \hat{y} \end{cases}$$

Calculating and estimating VaR can help the company better deal with financial risks and avoid extreme scenarios in the future.

A.5 Returns with Different Future Intervals

We used the following formula to calculate the Returns with Different Gaps:

$$R_{ijk} = \frac{S_{i,(j+k)} - S_{i,j}}{S_{i,j}}$$

Here, we denote $S_{i,j}$ as the closed price of company i's stock at time j, and R_{ijK} as the $k-day$ return of stock price i at time j. We calculated this value as the difference between the closed price of company i at time j and time $j+k$ divided by the closed price of company i at time j.

B The Results on Returns

To evaluate the predictive capability of our model, we have applied it to forecast stock returns. The results of these forecasts are systematically compiled and displayed in a table 3.

Table 3: ECC Analyzer's Prediction Performance On Stock Returns

	Return_3d	Return_7d	Return_15d	Return_30d
Predited Return	0.0024	0.0068	0.0131	0.0185
True Return	0.0007	0.0017	0.0035	0.0139
Error	0.0018	0.0051	0.0096	0.0046
Percentage Error	2.6923	3.0964	2.7216	0.3348

C Regression and Sentiment Analysis

In the main chapter, we used the various returns and volatility of different terms as our predicted target. This part will introduce how we calculated these values and further results about the relationships between sentiments calculated by LLM and these target values.

C.1 Cumulative Abnormal Returns and Regression Analysis

We used both Abnormal Returns collected from the market and the returns of different terms to capture the sentiment delivered by ECCs. For the abnormal Returns, we used the following regression model (Lee and Connolly, 2010):

$$R_{it} = \alpha + \beta R_{mt} + \varepsilon_{it}$$

R_{it} is the return of the security at time t, also R_{mt} is the market portfolio return at time t. The Abnormal Returns we defined here are the difference between the real value of Returns and estimated ones, which is ε_{it}. For other dependent variables, we used the R_{ijk} for $k = 1, 3, 7, 15, 30$. Results are shown in Table 4.

For the volatility, we used the difference between the two closed periods of volatility. The dependent variable for volatility could be calculated as $log(\sigma_{i(j+1)k}) - log(\sigma_{ijk})$ for $k = 3, 7, 15, 30$. Results are shown in Table 5.

From Table 4 and Table 5, we can find that the Sentiment generated from the LLM is always statistically significant in the linear regression model results. Comparing the coefficient of different terms of Returns, we found that the coefficient gets larger when there are more significant term gaps. This indicates that the impact of the ECC's sentiment will appear as time progresses. There is a higher probability that the impact of the ECCs occurs after a certain period because, besides the financial data, most information, like future projects mentioned in ECCs, will not be reflated instantly. That could explain why the coefficient will rise as the period gets larger. However, the volatility change will have a smaller impact when the period gets larger(because the absolute value of Sentiments' coefficient in Table 5 is smaller when the period gets more extensive), which means that the same ECCs sentiments effect will have a more significant impact in short terms rather than the longer terms.

D Prompt Design

D.1 Prompt for Summarizing Earnings Conference Call Segments

- Identify Key Points
  For each segment, identify the key topics covered. Note any significant financial figures, strategic decisions, performance metrics, or forward-looking statements.

- Summarize Succinctly
  Write a concise summary for each segment, capturing the essence of the discussion. Aim to condense the information into a few sentences that clearly convey the main points and outcomes discussed.

- Highlight Relevant Details
  Include any specific details that are critical for understanding the segment's context or implications, such as notable quotes from the company's executives or specific data points that illustrate trends or changes.

- Connect the Dots
  If applicable, relate the segment's content to broader company objectives or industry trends

Table 4: This table shows the linear regression results of different kinds of returns. We used different types of returns as dependent variables. $CARs$ is the cumulative abnormal returns, calculated as the difference between the actual value of the predicted valve from factor models. The rest of the dependent variables are calculated as the difference between the closed price of company i at time j and time $j + k$ divided by the closed price of company i at time j. The independent variable is the Sentiment score, which is calculated by LLM. We extracted all the events in one ECC and evaluated their sentiment score. The percentage of events with positive sentiment minus the percentage of events with negative sentiments calculates the score.

	Dependent Variable					
	CARs	Return_1d	Return_3d	Return_7d	Return_15d	Return_30d
Constant	-0.0105^{***}	-0.0105^{***}	-0.0126^{***}	-0.0182^{***}	-0.0286^{***}	-0.0244^{***}
	(0.004)	(0.004)	(0.004)	(0.005)	(0.007	(0.008)
Sentiment	0.0223^{***}	0.024^{***}	0.0295^{***}	0.0442^{***}	0.0716^{***}	0.0852^{***}
	(0.008)	(0.008)	(0.009)	(0.010)	(0.015)	(0.016)
Adjusted R^2	0.016	0.018	0.022	0.038	0.048	0.055
Observation	572	572	572	572	572	572

Table 5: This table shows the linear regression results of different terms of Volatility. We used four different terms of volatility. Volatility here is defined by the log form of the sample standard deviation of the past $k - days$ stock closed price of company i. The independent variable is the Sentiment score, which is calculated by LLM. We extracted all the events in one ECC and evaluated their sentiment score. The percentage of events with positive sentiment minus the percentage of events with negative sentiments calculates the score.

	Dependent Variable			
	Volatility_3d	Volatility_7d	Volatility_15d	Volatility_30d
Constant	-0.0254^{***}	-0.017^{***}	-0.0071^{***}	-0.005^{***}
	(0.013)	(0.009)	(0.006)	(0.004)
Sentiment	-0.0681^{***}	-0.0556^{***}	-0.0438^{***}	-0.0266^{***}
	(0.026)	(0.018)	(0.013)	(0.009)
Adjusted R^2	0.013	0.019	0.022	0.019
Observation	572	572	572	572

to provide context and show how the segment fits into the bigger picture.

D.2 Prompt for Creating an Overview Summary from Earnings Conference Call Segments

- Gather Segment Summaries
  Start by reviewing the summaries of each segment from the earnings conference call. Ensure that you have all the segment summaries available to reference.
- Identify Common Themes
  Look for common themes, recurring issues, or consistent messages across the segments. Note any overarching strategies, goals, or concerns expressed by the company executives.

D.3 Prompt for Extracting Focus And Explore Its Impact

- Listen to the Call
  Begin by thoroughly listening to the entire earnings conference call. Pay attention to both the prepared remarks and the question-and-answer session.
- Identify Focus Points
  Identify statements or discussions that involve significant financial metrics, strategic initiatives, new products or markets, regulatory impacts, or any notable shifts in operations. These are potential focus points that could influence investor perceptions and stock price.
- Document Evidence
  For each identified focus point, document the exact wording used, the context in which it was discussed, and who discussed it (e.g., CEO, CFO). This will be crucial for accurate interpretation and analysis.
- Analyze Impact on Stock Movement
  Pre and Post Analysis: Examine stock price movements immediately before and after the call to capture initial reactions.
- Longer-term Impact
  Review stock performance in the days or weeks following the call to assess sustained impacts.
- Compare with Market Trends
  Ensure to factor in overall market conditions and sector movements to isolate the impact of the earnings call from broader market trends.

D.4 Prompt for Analyzing Sentiment of Focus Points and Supporting Evidence from an Earnings Conference Call

- Analyze Sentiment for Each Focus Point
  Apply the sentiment analysis to the text surrounding each focus point. Pay attention to the language used, such as positive, negative, or neutral descriptors, and the intensity of the language. Please assign a continuous decimal sentiment score to each event, ranging from -1 to 1. A score of -1 represents a highly negative sentiment, 0 indicates neutrality, and 1 signifies a highly positive sentiment.
- Context Consideration
  Consider the context in which each point was discussed. Assess whether the sentiment is directly related to the focus point or influenced by broader discussion themes.
- Document Supporting Quotes
  For each focus point, document specific quotes or statements from the call that illustrate the sentiment. Note the speaker and their role to add credibility to the sentiment analysis.

D.5 Prompt for Analyzing Audio Features of Focus Points from an Earnings Conference Call

- Tone and Pitch Analysis
  Examine variations in tone and pitch within each audio segment. Look for patterns that might indicate emphasis, uncertainty, confidence, or stress.
- Volume and Speech Rate
  Measure changes in volume and variations in speech rate. High volume and rapid speech may indicate areas of strong emotion or importance.
- Pause Patterns
  Identify the frequency and duration of pauses, which can provide insights into the speaker's thought process or hesitation.

Figure 3: We asked four experts to help us analyze the possible scenarios of ECCs. Our experts focus on the following general Categories of the ECCs' transcript: Financial Indicators, Employee & Managers, Cost, Project Expansion, Business, and Future Look. Each general category is separated into some minor focusing points. Our experts put forward three questions for each small focusing point. We used these questions to ask LLM and wished for a directional response in the thinking pattern hidden in the questions that we wanted to teach the LLM.

Focus Category	Focus Item	Question 1	Question 2	Question 3
financial indicator	dividend	Did this company paid the investors dividend?	Have there been any increases or decreases in the stock dividends? If yes, what is the rate at which the dividends have been increasing?	What type of the dividend did the company pay?
	revenue	What was the company's reported revenue for the past quarter, and how does it compare to the same quarter in the previous year?	What factors influenced the company's revenue performance this quarter?	What are the company's revenue forecasts for the upcoming quarters, and what strategies are in place to achieve these targets?
	return	What was the company's net profit margin for this quarter, and how has it changed from the previous quarter or year?	What is the company's Return on Equity (ROE) and Return on Assets (ROA) for this period?	How does the company plan to enhance shareholder value in the upcoming periods? Are there any dividends or buybacks planned?
	earnings	Have there been any increases or decreases in the earnings? If yes, what is the rate at which the earnings have been increasing or decreasing?	What is the outlook provided by the executives of this company in relation to the future earnings growth?	Are the earnings above or below compared to the expectations? Are they attractive compared to the peers?
employee manager	salary	What percentage of the company's total expenses is currently allocated to employee salaries, and how has this changed in response to recent business developments?	How does your company's compensation structure compare with industry standards, particularly in terms of salary, benefits, and bonuses?	What were the average salary increases or decreases across the company this year compared to last year?
	pension	What is the current status of the company's pension fund, and what were the major changes to its funding status over the past year?	How does the company manage its pension liabilities, and what strategies are in place to address any underfunded positions?	What are the expected impacts of current pension commitments on the company's future financial performance?
	management change	Have there been any recent changes in the company's key management positions, including the CEO, etc.	What were the reasons behind any recent management changes, particularly in the CEO position?	What impact are the recent management changes expected to have on the company's strategy and operations in the near term?
cost	Operating Costs	What were the total operating costs this quarter compared to the previous quarter?	Which factors contributed to any significant changes in operating costs?	How are you managing operating costs in light of current economic conditions?
	Cost of Goods Sold	How did the Cost of Goods Sold change this quarter, and what were the driving factors behind these changes?	What percentage of revenue does the Cost of Goods Sold represent, and how does this compare to industry norms?	Are there any initiatives in place to reduce Cost of Goods Sold without compromising quality?
	Marketing and Sales Costs	How much did the company spend on marketing and sales this quarter?	What specific marketing or sales strategies contributed to these costs?	Are there plans to adjust these strategies in the upcoming quarters based on performance?
expansion	Geographic Expansion	What specific regions or markets is the company expanding into, and what factors influenced these selections?	What are the initial costs associated with the geographic expansion, and what financial strategies are in place to support it?	What is the projected timeline for the new regional operations to reach profitability?
	Product Line Expansion	What new products is the company introducing, and what consumer or market needs do they aim to address?	How will the introduction of these products impact the company's production costs and overall financial performance?	Are there any expected synergies between the new products and existing products or services?
	Market Segmentation Expansion	Which new customer segments is the company targeting, and what research supports this strategic direction?	What marketing strategies will be employed to reach these new segments, and what are the anticipated costs?	What are the growth expectations for these new market segments over the next few years?
business	business	What are the key projects or initiatives currently being undertaken by the company, and provide a brief overview of what each project entails?	How has the performance of these projects compared to the previous quarter, and how do they stand in relation to competitors in the same sector? What has been the market's response to these initiatives?	What are the generated revenues from these projects for the current reporting period, and what potential risks could impact their future performance?
future outlook	future outlook	What are the company's primary strategic goals for the upcoming year, and what key initiatives are planned to achieve these objectives?	How do these future plans align with current industry trends and market demands?	What are the expected financial and operational impacts of these plans on the company's performance in the short and long term?