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

Yupeng Cao*
ycao33@stevens.edu
Stevens Institute of Technology
Hoboken, NJ, USA

Zhi Chen*
zchen100@stevens.edu
Stevens Institute of Technology
Hoboken, NJ, USA

Qingyun Pei*
qpei1@stevens.edu
Stevens Institute of Technology
Hoboken, NJ, USA

Nathan Jinseok Lee
nathanlee@hunterschools.org
Hunter College High School
New York, USA

K.P. Subbalakshmi
ksubbala@stevens.edu
Stevens Institute of Technology
Hoboken, NJ, USA

Papa Momar Ndiaye
pndiaye@stevens.edu
Stevens Institute of Technology
Hoboken, NJ, USA

Abstract

In the realm of financial analytics, leveraging unstructured data, such as earnings conference calls (ECCs), to forecast stock volatility is a critical challenge that has attracted both academics and investors. While previous studies have used multimodal deep learning-based models to obtain a general view of ECCs for volatility predicting, they often fail to capture detailed, complex information. Our research introduces a novel framework: **ECC Analyzer**, which utilizes large language models (LLMs) to extract richer, more predictive content from ECCs to aid the model's prediction performance. We use the pre-trained large models to extract textual and audio features from ECCs and implement a hierarchical information extraction strategy to extract more fine-grained information. This strategy first extracts paragraph-level general information by summarizing the text and then extracts fine-grained focus sentences using Retrieval-Augmented Generation (RAG). These features are then fused through multimodal feature fusion to perform volatility prediction. Experimental results demonstrate that our model outperforms traditional analytical benchmarks, confirming the effectiveness of advanced LLM techniques in financial analysis.

CCS Concepts

• **Computing methodologies** → **Natural language processing**; • **Information systems** → **Multimedia information systems**.

Keywords

Large Language Model, Earnings Conference Call Analysis, Volatility forecasting, Retrieval-Augmented Generation

ACM Reference Format:

Yupeng Cao, Zhi Chen, Qingyun Pei, Nathan Jinseok Lee, K.P. Subbalakshmi, and Papa Momar Ndiaye. 2024. ECC Analyzer: Extract Trading Signal from Earnings Conference Calls using Large Language Model for Stock Volatility Prediction. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

*Equal Contribution

Conference acronym 'XX, June 03–05, 2024, Woodstock, NY

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

1 Introduction

Predicting the stock volatility over a certain period is a crucial task in financial analysis, aiding capital market participants in making better investment decisions. As such, developing effective techniques for predicting stock volatility has become increasingly important among academics and the financial industry. Previous studies in economics have shown that stock volatility can be predicted using publicly available information [8]. Substantial research in finance and computational linguistics has made significant progress in predicting stock volatility from various textual sources, such as company disclosed reports [14, 18, 30], the transcripts of earnings conference calls [17, 37] and social media [3, 7, 26].

Recently, advancements in multimodal learning have enabled the use of more unstructured multimedia data, such as audio recordings of earning conference calls [22, 28], Merge & Acquisition calls [31] and the video of CEO's speech [29], for stock volatility prediction. Multimodal learning is particularly valuable in this context as it allows for the integration of diverse data sources, providing richer, more nuanced insights into market dynamics and sentiment. This comprehensive analysis is essential for understanding complex market behaviors [28]. Our study focuses on multimodal earning conference calls (ECCs) data for two primary reasons: 1) transcripts and audio recordings are publicly available, and 2) ECCs are often associated with high volatility primarily due to the market reaction to the earnings announcement [9].

Existing multimodal approaches to volatility prediction using ECCs typically extract features from speech and text separately. Subsequently, commonly used Natural Language Processing (NLP) models, such as LSTM or Transformer-based models, are employed to jointly model the extracted speech and text features, ultimately performing volatility prediction [28, 36, 40]. While these methods have shown that the multimodal approach can extract complementary information from multiple modalities, enhancing financial modeling performance and model robustness, several challenges remain unaddressed: firstly, previous work directly feeds features extracted from text and audio into the model, potentially missing important contextual details; secondly, these approaches assign

equal importance to all sentences and audio segments which hard to reflect the impact of important information in the ECCs on the prediction.

In response to the aforementioned limitations and inspired by the significant performance improvements offered by large language models (LLMs) in various downstream NLP tasks, this paper introduces the ECC Analyzer — a novel framework that leverages LLMs for in-depth analysis of ECC data to enhance volatility prediction performance. The proposed ECC Analyzer employs a hierarchical information extraction strategy: 1) advanced pre-trained large models are used to extract embeddings from both audio and text, capturing the overall information content of ECCs.; 2) LLMs summarize ECC transcripts at the paragraph level, distilling important information from the transcript; 3) In collaboration with financial experts, we designed the "Question Bank" which contains a series of questions about ECC content that are of interest to investors. These questions are used as queries to extract fine-grained sentences containing key information from ECCs using Retrieval-Augmented Generation (RAG). We use these questions as queries and extract fine-grained sentences containing key information from ECC by RAG to improve the prediction performance of the model. Then the summarized text with extracted key focus sentences is transformed into text embedding and combined with the extracted audio and text features for volatility prediction.

Our extensive experiments on the real-world S&P 500 ECCs dataset demonstrate clear and significant improvements in prediction accuracy attributable to our proposed approach. **We claim that the contribution is two-fold: 1) Our framework innovatively integrates multi-modal information with fine-grained details extracted by Large Language Models (LLMs) to effectively distill and utilize rich information from financial documents, such as ECCs, for prediction tasks. 2) Our pipeline demonstrates a** 27.7% **reduction in average Mean Squared Error (MSE) compared to the current state-of-the-art (SOTA) model. In detail, The results show that our method achieves substantial performance improvements in forecasting short-term volatility (3-day, 7-day) compared to current SOTA methods. The results for medium-term volatility (15 days) and long-term volatility (30 days) are comparable to the best existing methods.**

2 Related Work

2.1 Stock Volatility Prediction

Volatility forecasting has long been of interest to researchers due to its practical applications. Conventional forecasting approaches rely on historical stock prices and use continuous and discrete time series models [5, 10, 13, 15, 16, 20]. Recently, the use of NLP techniques to analyze unstructured data for predicting stock performance has attracted significant academic attention. A foundational study by [18] 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 [30, 34, 37], proposed various document representation methods to predict stock price volatility. Drawing on multimodal technologies, [28] explored how audio features—such as tone, emotion, and speech rate—enhance stock movement predictions when combined with text analysis. Following by this, [40] 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. [36] addressed the reduction of gender bias in ECC predictions through adversarial training. 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 Financial Application

Numerous studies have explored the applications of LLMs in the financial sector. [23] explore how LLMs have been adeptly applied to summarize and abstract complex financial documents such as 10-K, and 10-Q filings. The FinBen [38] provides the benchmark performance of LLM for each task in the financial domain. FinGPT [39] and FinMem [41] explore 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 [21, 32, 33]. [1, 42] 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 volatility prediction.

3 Problem Formulation

Volatility can be expressed as the natural log of the standard deviation of return prices r in a window of τ days. We calculate the 3, 7, 15, and 30 trading days volatility using the equation:

$$v_{[d-\tau,d]} = \ln\left(\left(\frac{\sum_{i=0}^{\tau}(r_{d-i} - \bar{r})^2}{\tau}\right)^{\frac{1}{2}}\right) \tag{1}$$

The notations used across the paper are discussed herein. Let $s \in S$ denote a stock, $c \in C$ be an earnings call for stock s. For each stock, there exist multiple earnings calls c that are held periodically. Each call c can be segmented into a set of $a_c^i \in A_c$ audio clips, and corresponding $t_c^i \in T_c$ text sentences for $i \in [1, N]$, where is the N is the maximum number of audio clips in a call. For each stock, there exists a daily return denoted by $r_i = \frac{p_i - p_{i-1}}{p_{i-1}}$, where p_i is the adjusted close price at the end of the day and $\bar{r}$ is the average return over a period of τ days. We aims to develop a predictive regression function $f(c) \rightarrow v_{[d-\tau,d]}$

For the evaluation, following [28], we assess the accuracy of volatility predictions by comparing the predicted values y_i with the labeled volatility values $\hat{y_i}$. We use Mean Squared Error (MSE) as the evaluation metric:

$$MSE = \frac{\sum_i (y_i - \hat{y_i})^2}{n} \tag{2}$$

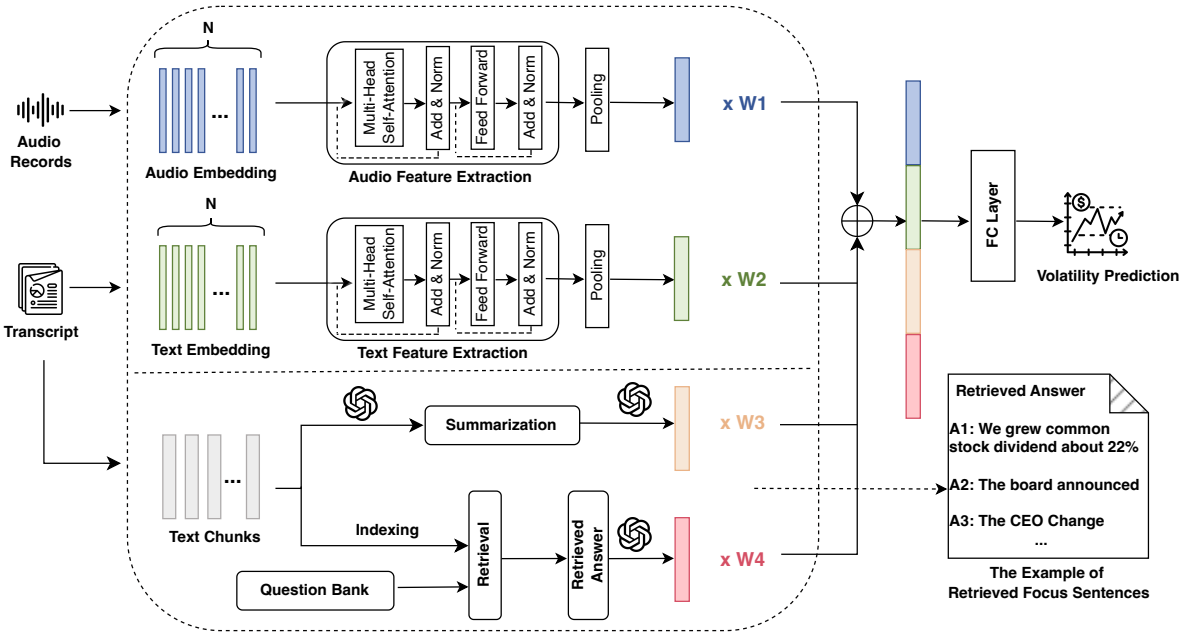

Figure 1: illustrates the ECC Analyzer Framework. The proposed method accepts multimodal inputs: audio record and transcript. The upper part of the box illustrates the feature extraction process for both the audio and text of the data. We use the pre-trained large models to generate embeddings from the audio and text, followed by a transformer encoder to extract the corresponding features. The lower part of the box represents a deeper analysis of the ECC. Here, the text is divided into chunks, which are then summarized into paragraphs using an LLM. Key sentences are extracted via RAG and converted into text features through text embedding. These text features are then fused with the features extracted from the upper part of the box to make the final volatility prediction.

This metric quantifies the average squared difference between the predicted and actual volatility values, providing a measure of prediction accuracy. Additionally, for constructing volatility labels, we collected stock price information via the Yahoo Finance API[1] and used equation 1 to calculate the labels for days 3, 7, 15, and 30.

4 Our proposed framework

ECC Analyzer (in Figure 1) aims to comprehensively understand the multi-data types present in earnings conference calls, including both text and audio components. In this section, we explain our methodology in 4 parts: (4.1) audio encoding, (4.2) transcript encoding,(4.3) fine-grained information extraction from the transcript by using LLM, and (4.4) multi-model fusion and model training.

4.1 Audio Encoding

Audio pre-trained models have achieved performing results in various downstream tasks [6, 19, 27, 35]. 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 [2], 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.

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 representation: $e_{ac}^i = \text{Wav2Vec2}(a_c^i)$. Therefore, we obtain the audio embeddings $E_{ac} = \{e_{ac}^1, e_{ac}^2, \ldots, e_{ac}^n\}$ 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:

$$\text{Multihead} = \text{Concat}(\text{head}_1, ..., \text{head}_h)W^o \tag{3}$$

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{4}$$

where Q (queries) and K (keys) of dimension d_k and V values of dimension d_v. The weights dimensions are: $W_i^Q, W_i^K, W_i^V \in R^{d_{model} \times d_k, d_k, d_v}$ 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

[1] https://finance.yahoo.com/

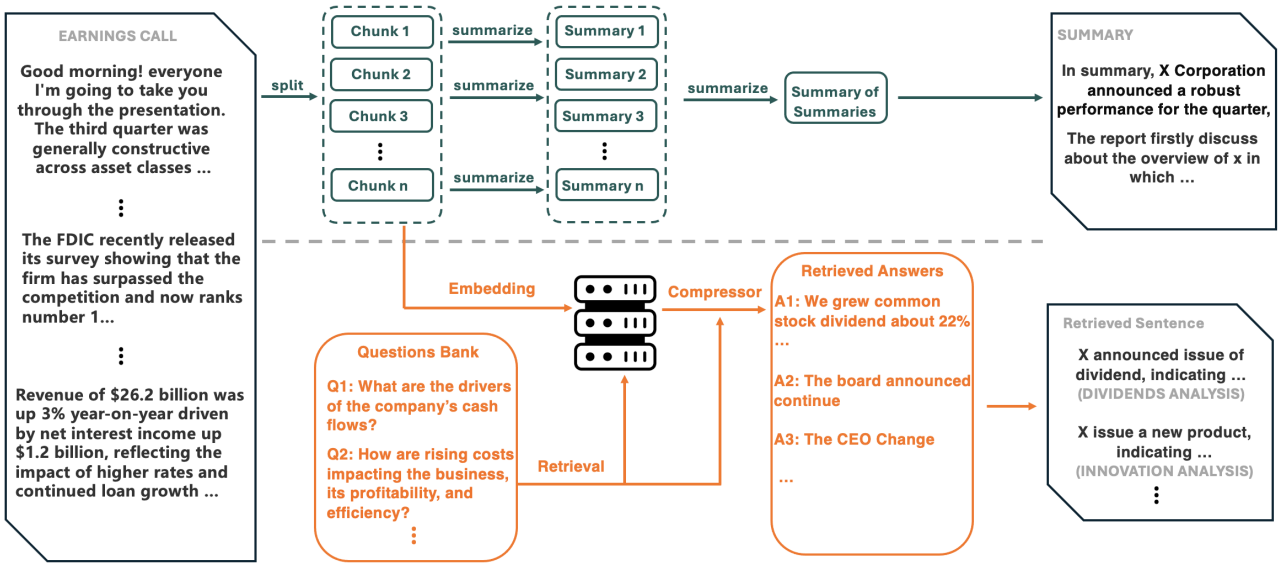

Figure 2: visualizes the process of fine-grained information extraction from ECC transcript.

scores are normalized using the softmax function:

$$\text{Attention}(Q, K, V) = softmax(\frac{KQ^T}{\sqrt{d_k}})V \tag{5}$$

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:

$$T_{ac} = \text{MHSA}(E_{ac}) \tag{6}$$

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.

4.2 Transcript Encoding

The transcript encoding process mirrors Audio Encoding, using SimCSE [11] 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 sentences 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 $T_{tc} = \{t^1_{tc}, t^2_{tc}, \ldots, t^n_{tc}\}$ 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.

4.3 Fine-grained Information Extraction from the Transcript by Using LLM

To obtain deep insights from an Earnings Conference Call on how it might predict future market performance, our approach encompasses two parts: (1) summarize the transcript, and (2) important sentences extracted by using RAG. We show this process in Figure 2. We list all Prompts in Appendix A.1.

4.3.1 Summarize the transcript. In order to effectively summarize key information in lengthy ECC records, we employ a hierarchical summarization strategy. First, the entire document is divided into chunks, and then we use LLM to summarize each chunk individually. These individual summaries are then further summarized through LLM, resulting in a comprehensive summarization paragraph of the entire document. This two-layer approach ensures that both detailed and aggregate information is captured. In further, we use the OpenAI 'text-embedding-3-small' text embedding model to generate embeddings T_s with size 1024 for summarized paragraph:

$$T_s = \text{Embedding}(summary + chunk\ summaries) \tag{7}$$

4.3.2 Important sentences extracted by using RAG. In this step, we aim to extract the most critical sentences containing relevant information from the entire transcript. To achieve this, we first convert the processed chunks into vector representations using the embedding model. Then, with input from financial experts, we design a list of questions about the ECC, which are stored in a 'Question Bank' as a query list (see Appendix A.2). For each question in the 'Question Bank', we retrieve the relevant chunks from the vectors individually. Next, we use the context compressor, in combination with the LLM chat, to quiz each extracted chunk and query on its relevance. This process filters out only the sentences useful for answering the questions. Finally, we synthesize the query-selected sentences into a coherent prompt, to which the large language model will generate responses as the final answers to the queries.

These generated answers are considered as the extracted important sentences. We then concatenate all the generated sentences and feed them into a text embedding model to generate the text embedding T_f with size 1024:

$$T_f = \text{Embedding}(\text{Concatenated}[selected\ sentences;]) \quad (8)$$

4.4 Multimodal Fusion and Prediction

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 [24]. 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:

$$E = w_0 + w_1 \cdot T_a + w_2 \cdot T_t + w_3 \cdot T_s + w_4 \cdot T_f + \epsilon \quad (9)$$

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. This unified feature set E is fed into two fully connected layers to perform the regression task. We train ECC Analyzer by optimizing the MSE loss:

$$\mathcal{L} = \mu(\sum_i (\hat{y}_i - y_i)^2) \quad (10)$$

5 Experiment

We describe the datasets used in the experiment and the baseline settings for performance comparisons. Our experiments aim to answer the following research questions (RQs): **RQ1**: Can our proposed pipeline, when integrated with LLMs, significantly improve volatility prediction performance compared to current state-of-the-art (SOTA) approaches? **RQ2**: Does the synergy between our model and the LLM result in better performance compared to raw LLM predictions? **RQ3**: Does each distinct component derived from ECC contribute to improving volatility prediction accuracy?

5.1 Dataset

The dataset utilized in this study is sourced from the publicly available S&P 500 ECC dataset as constructed by [28]. 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 [28], 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.

5.2 Baseline Setup

We compare our approach to several important baselines including:

- **Classical Method:** We incorporate the GARCH model and its derivatives, as described in [10, 16]. This model is 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 [12]**: The Long Short-Term Memory (LSTM) based method is a popular choice for financial time series prediction due to its efficacy in handling sequential data. We use a straightforward LSTM model as a benchmark for volatility prediction.
- **MT-LSTM-ATT [25]:** 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 gets word embeddings using pre-trained Glove vectors and then processed by a Bi-GRU [4] 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 [28]:** The MRDM model first introduced a multimodal deep regression approach to fuse the GloVe embeddings and hand-crafted acoustic features for volatility prediction tasks.
- **HTML [40]:** 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.
- **AMA-LSTM [36]:** The paper uses a strategy of adversarial training to minimize the effect of speaker gender in earnings conference calls.
- **GPT-4-turbo-2024-04-09:** We assessed the capability of LLMs in directly predicting volatility performance from ECCs. The model was set to generate a response with a zero temperature setting to ensure deterministic output.

5.3 Implementation Details

In the experiment, all interactions with LLMs were conducted using "GPT-4 Turbo-2024-04-09". We set the temperature parameter to 0 to ensure that the LLMs produce the most predictable responses, thereby maintaining consistency in our experiments.

For the overall training of the ECC Analyer framework, we developed the code using PyTorch. Each Multi-Head Self-Attention layer in the model comprises 6 layers and 8 individual heads in each layer. The training process utilized batch sizes $b \in \{2, 4, 8, 16\}$. We use a grid search to determine the optimal parameters and select the learning rate λ for Adam optimizer among $\{1e-3, 1e-5, 1e-6, 1e-7\}$.

5.4 Overall Results Analysis (RQ1 & RQ2)

Table 1: Performance results on our proposed framework ECC Analyzer with different baseline methods.

Model	$\overline{MSE}$	MSE_3	MSE_7	MSE_{15}	MSE_{30}
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 (Glove)	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**
AMA-LSTM	/	0.680	0.360	**0.230**	/
GPT-4-Turbo	2.198	3.187	5.059	7.959	11.824
ECC Analyzer	**0.314**	**0.553**	**0.306**	0.237	**0.158**

Table 1 shows the performance of various methods in predicting volatility performance. Notably, the ECC Analyzer framework excels, especially in short-term forecasts (day 3 and day 7), with the lowest Mean Squared Error (MSE) values of 0.553 and 0.306, respectively. Its long-term prediction performance is comparable to the state-of-the-art method, HTML. However, the ECC Analyzer performs slightly lower than AMA-LSTM in predicting medium-term (day 15) volatility. Nevertheless, ECC Analyzer's overall average MSE achieves superior performance. These encouraging experimental results illustrate that extracting fine-grained information from ECC data using LLMs can significantly enhance the model's volatility prediction accuracy. In particular, the extracted fine-grained information features are especially beneficial for improving short-term forecasting performance, which is crucial for investors. In summary, addressing **RQ1**, the prediction performance of our proposed method surpasses that of the current SOTA approaches.

In response to **RQ2**, directly applying LLMs to volatility prediction proves largely ineffective, akin to random guessing. These results also raise concerns about the potential misuse of LLMs, considering their impact on public safety. If investors use LLMs inappropriately for numerical outputs, they may increase financial risk. 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.

5.5 Ablation Study (RQ3)

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. Each component is represented as follows: E_{os} **represents the embedding of the overall ECC summary;** E_{cs} **represents the embedding of the summary for chunks;** E_{cs} **represents the embedding of the extracted fine-grained important sentence.**

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. 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. We also obtain good predictive results using only analytics derived from LLMs, affirming the response to our **RQ3**: comprehensive analysis indeed enhances the predictive capability for stock volatility prediction 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.

6 Conclusion

In this paper, we propose the ECC Analyzer, a novel framework that leverages large language models (LLMs) for in-depth analysis of ECC data to enhance volatility prediction performance. The ECC Analyzer extracts information from ECCs at multiple levels to assist in volatility prediction. Our experiments demonstrate that our proposed method improves the overall average Mean Squared Error (MSE). Specifically, we achieve better results in short-term forecasting, with medium-term and long-term forecasts also on par with SOTA methods. Additionally, we analyze the role of each component in the proposed framework for the forecasting process through ablation experiments, highlighting the contribution of each element to the overall performance.

References

[1] 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*. 1–7.

[2] 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 (2020), 12449–12460.

[3] Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of computational science* 2, 1 (2011), 1–8.

[4] 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* (2014).

[5] John C Cox and Stephen A Ross. 1976. The valuation of options for alternative stochastic processes. *Journal of financial economics* 3, 1-2 (1976), 145–166.

[6] 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)*. IEEE, 3852–3856.

[7] Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2015. Deep learning for event-driven stock prediction. In *Twenty-fourth international joint conference on artificial intelligence*.

[8] Bernard Dumas, Alexander Kurshev, and Raman Uppal. 2009. Equilibrium portfolio strategies in the presence of sentiment risk and excess volatility. *The Journal of Finance* 64, 2 (2009), 579–629.

[9] George Foster, Chris Olsen, and Terry Shevlin. 1984. Earnings releases, anomalies, and the behavior of security returns. *Accounting Review* (1984), 574–603.

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

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

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

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 an overall ECC summary generated by LLMs; 3)Audio+Text+E_{cs}: adds the chucks summary generated by LLMs; 4) Audio+Text+$E_{os} + E_{cs}$: integrates both overall and chunk summaries for the ECC; 5) Audio+Text+E_{fo}: combines raw data with focused analytical results; 6) $E_{os} + E_{cs} + E_{fo}$: 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	0.280	0.204
Audio+Text+E_{os}	0.373	0.638	0.380	0.276	0.201
Audio+Text+E_{cs}	0.375	0.640	0.385	0.275	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	0.230	0.165
$E_{os} + E_{cs} + E_{fo}$	0.343	0.601	0.344	0.247	0.179
Audio+Text+$E_{os} + E_{cs} + E_{fo}$	**0.314**	**0.553**	**0.306**	**0.237**	**0.158**

[13] Steven L Heston. 1993. A closed-form solution for options with stochastic volatility with applications to bond and currency options. *The review of financial studies* 6, 2 (1993), 327–343.

[14] Gerard Hoberg and Gordon Phillips. 2016. Text-based network industries and endogenous product differentiation. *Journal of political economy* 124, 5 (2016), 1423–1465.

[15] Ying-Lin Hsu, TI Lin, and CF Lee. 2008. Constant elasticity of variance (CEV) option pricing model: Integration and detailed derivation. *Mathematics and Computers in Simulation* 79, 1 (2008), 60–71.

[16] 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 (2018), 25–37.

[17] Michael D Kimbrough. 2005. The effect of conference calls on analyst and market underreaction to earnings announcements. *The Accounting Review* 80, 1 (2005), 189–219.

[18] 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*. 272–280.

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

[20] Werner Kristjanpoller, Anton Fadic, and Marcel C Minutolo. 2014. Volatility forecast using hybrid neural network models. *Expert Systems with Applications* 41, 5 (2014), 2437–2442.

[21] Akbar Lakhani. 2023. Enhancing Customer Service with ChatGPT Transforming the Way Businesses Interact with Customers. (2023).

[22] Jiazheng Li, Linyi Yang, Barry Smyth, and Ruihai Dong. 2020. Maec: A multimodal aligned earnings conference call dataset for financial risk prediction. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 3063–3070.

[23] 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*. 374–382.

[24] 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* (2022).

[25] 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* (2015).

[26] Nuno Oliveira, Paulo Cortez, and Nelson Areal. 2017. The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices. *Expert Systems with applications* 73 (2017), 125–144.

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

[28] 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*. 390–401.

[29] Kumaran Rajandran. 2021. Interdiscursivity in corporate financial communication: an analysis of earnings videos. *Corporate Communications: An International Journal* 26, 2 (2021), 328–347.

[30] 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* (2017).

[31] Ramit Sawhney, Mihir Goyal, Prakhar Goel, Puneet Mathur, and Rajiv Shah. 2021. Multimodal multi-speaker merger & acquisition financial modeling: A new task, dataset, and neural baselines. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 6751–6762.

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

[33] 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 (2023), 380–386.

[34] 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*. 32–37.

[35] 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*. 3835–3839.

[36] Shengkun Wang, Taoran Ji, Jianfeng He, Mariam Almutairi, Dan Wang, Linhan Wang, Min Zhang, and Chang-Tien Lu. 2024. AMA-LSTM: Pioneering Robust and Fair Financial Audio Analysis for Stock Volatility Prediction. *arXiv preprint arXiv:2407.18324* (2024).

[37] 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)*. 1155–1165.

[38] Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, et al. 2024. The finben: An holistic financial benchmark for large language models. *arXiv preprint arXiv:2402.12659* (2024).

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

[40] 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*. 441–451.

[41] 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* (2023).

[42] 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*. 99–105.

A Appendix

A.1 Prompt Design

A.1.1 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 to provide context and show how the segment fits into the bigger picture.

A.1.2 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.

A.1.3 Prompt for Extracting Important Sentences.

- Check 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 prices.
- 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.

A.2 Design of Question Bank

Table 3: The designed Question Bank

Focus Category	Focus Item	Questions
Financial Indicator	Dividend	Q1: Did this company pay the investors dividend? Q2: Have there been any increases or decreases in the stock dividends? If yes, what is the rate at which the dividends have been increasing? Q3: What type of dividend did the company pay?
	Revenue	Q1: What was the company's reported revenue for the past quarter, and how does it compare to the same quarter in the previous year? Q2: What factors influenced the company's revenue performance this quarter? Q3: What are the company's revenue forecasts for the upcoming quarters, and what strategies are in place to achieve these targets?
	Return	Q1: What was the company's net profit margin for this quarter, and how has it changed from the previous quarter or year? Q2: What is the company's Return on Equity (ROE) and Return on Assets (ROA) for this period? Q3: How does the company plan to enhance shareholder value in the upcoming periods? Are there any dividends or buybacks planned?
	Earnings	Q1: 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? Q2: What is the outlook provided by the executives of this company in relation to the future earnings growth? Q3: Are the earnings above or below compared to the expectations? Are they attractive compared to the peers?
Employee Manager	Salary	Q1: 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? Q2: How does your company's compensation structure compare with industry standards, particularly in terms of salary, benefits, and bonuses? Q3: What were the average salary increases or decreases across the company this year compared to last year?
	Pension	Q1: 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? Q2: How does the company manage its pension liabilities, and what strategies are in place to address any underfunded positions? Q3: What are the expected impacts of current pension commitments on the company's future financial performance?
	Management Change	Q1: Have there been any recent changes in the company's key management positions, including the CEO, etc. Q2: What were the reasons behind any recent management changes, particularly in the CEO position? Q3: What impact are the recent management changes expected to have on the company's strategy in the near term?
Cost	Operating Costs	Q1: What were the total operating costs this quarter compared to the previous quarter? Q2: Which factors contributed to any significant changes in operating costs? Q3: How are you managing operating costs in light of current economic conditions?
	Cost of Goods Sold	Q1: How did the Cost of Goods Sold change this quarter, and what were the driving factors behind these changes? Q2: What percentage of revenue does the Cost of Goods Sold represent, and how does this compare to industry norms? Q3: Are there any initiatives in place to reduce Cost of Goods Sold without compromising quality?
	Marketing and Sales Costs	Q1: How much did the company spend on marketing and sales this quarter? Q2: What specific marketing or sales strategies contributed to these costs? Q3: Are there plans to adjust these strategies in the upcoming quarters based on performance?
Expansion	Geographic Expansion	Q1: What specific regions or markets is the company expanding into, and what factors influenced these selections? Q2: What are the initial costs associated with the geographic expansion and strategies are in place to support it? Q3: What is the projected timeline for the new regional operations to reach profitability?
	Product Line Expansion	Q1: What new products is the company introducing, and what consumer or market needs do they aim to address? Q2: How will the introduction of these products impact the company's production costs and overall financial performance? Q3: Are there any expected synergies between the new products and existing products or services?
	Market Segmentation Expansion	Q1: Which new customer segments is the company targeting, and what research supports this strategic direction? Q2: What marketing strategies will be employed to reach these new segments, and what are the anticipated costs? Q3: What are the growth expectations for these new market segments over the next few years?
Business	Business	Q1: What are the key projects or initiatives currently being undertaken by the company, and provide a brief overview of what each project entails? Q2: 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? Q3: 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	Q1: What are the company's primary strategic goals for the upcoming year, and what key initiatives are planned to achieve these objectives? Q2: How do these future plans align with current industry trends and market demands? Q3: What are the expected financial impacts of these plans on the company's performance in the short and long term?