

# RISKLABS: Predicting Financial Risk Using Large Language Model based on Multimodal and Multi-Sources Data

Yupeng Cao<sup>+</sup>, Zhi Chen, Prashant Kumar, Qingyun Pei, Yangyang Yu, Haohang Li,  
Fabrizio Dimino, Lorenzo Ausiello, K.P. Subbalakshmi<sup>+</sup>, Papa Momar Ndiaye  
Stevens Institute of Technology  
ycao33@stevens.edu, ksubbala@stevens.edu

## Abstract

The integration of Artificial Intelligence (AI) techniques, particularly large language models (LLMs), in finance has garnered increasing academic attention. Despite progress, existing studies predominantly focus on tasks like financial text summarization, question-answering (Q&A), and stock movement prediction (binary classification), the application of LLMs to financial risk prediction remains underexplored. Addressing this gap, in this paper, we introduce **RiskLabs**, a novel framework that leverages LLMs to analyze and predict financial risks. RiskLabs uniquely integrates multimodal financial data—including textual and vocal information from Earnings Conference Calls (ECCs), market-related time series data, and contextual news data—to improve financial risk prediction. Empirical results demonstrate RiskLabs' effectiveness in forecasting both market volatility and variance. Through comparative experiments, we examine the contributions of different data sources to financial risk assessment and highlight the crucial role of LLMs in this process. We also discuss the challenges associated with using LLMs for financial risk prediction and explore the potential of combining them with multimodal data for this purpose.<sup>1234</sup>

## Keywords

Large Language Model, Financial Risk Forecasting, Multimodal Learning, Multi-source Data

## ACM Reference Format:

Yupeng Cao<sup>+</sup>, Zhi Chen, Prashant Kumar, Qingyun Pei, Yangyang Yu, Haohang Li., Fabrizio Dimino, Lorenzo Ausiello, K.P. Subbalakshmi<sup>+</sup>, Papa Momar Ndiaye . 2024. RiskLABS: Predicting Financial Risk Using Large Language Model based on Multimodal and Multi-Sources Data. In *Proceedings of International Workshop on Multimodal Financial Foundation Models (MFFMs) at 5th ACM International Conference on AI in Finance (MFFM Workshop @ ICAIF '24)*. ACM, New York, NY, USA, 12 pages. <https://doi.org/XXXXXXX.XXXXXXX>

<sup>1</sup>Position Paper

<sup>2</sup>Present at International Workshop on Multimodal Financial Foundation Models (MFFMs) @ ICAIF'24

<sup>3</sup><https://sites.google.com/view/iwmffm2024/accepted-papers?authuser=1>

<sup>4</sup>This position work was completed in April 2024.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

MFFM Workshop @ ICAIF '24, November 15, 2024, Brooklyn, NY.

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-XXXX-X/18/06

<https://doi.org/XXXXXXX.XXXXXXX>

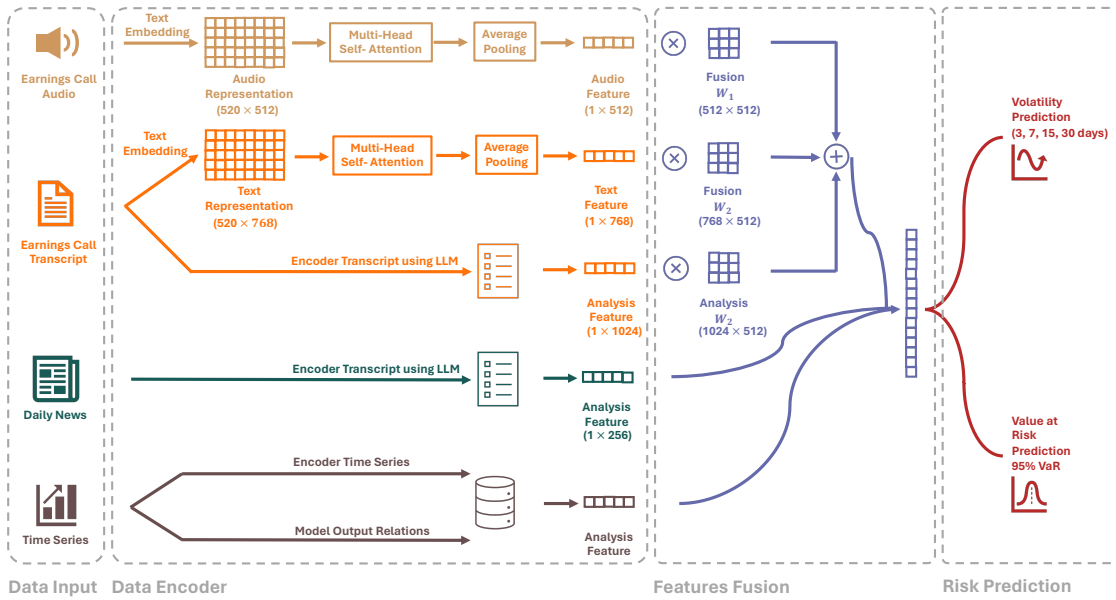
## 1 Introduction

The integration of artificial intelligence (AI) technology for financial risk prediction has been a longstanding focus of both industry and academia. Previously, researchers used machine learning techniques to analyze historical stock price data for risk prediction, such as using SVM to predict credit ratings based on financial statement ratios [10, 13] and using tree-based models to explore the correlation between economic indicators and the performance of various asset classes [9]. Recently, researchers have increasingly extracted trading signals from publicly available information for risk prediction. For example, financial reports and media news are used to analyze sentiment in news articles, capture market expectations for stock movements [14, 16, 21], risk factor analysis [29], and identify early warning signals from large amounts of textual data [1, 23]. With the rise of multimodal learning, more unstructured multimedia data, such as audio recordings of earning conference calls [17, 27] are utilized for stock volatility prediction.

While supervised learning methods have proven effective, they are highly task-specific and lack adaptability [19], with their predictive performance constrained by the amount of input data and the size of parameters they can handle. However, the emergence of large language models (LLMs) represents a paradigm shift in overcoming these challenges. Equipped with vast knowledge bases and advanced zero-shot learning capabilities, LLMs can perform a wide range of text-related tasks, such as summarization [32], question answering [24], and sentiment analysis [30], without requiring task-specific training. In the financial domain, LLMs have been widely utilized for tasks such as financial text analysis [26] and financial report generation [25]. Furthermore, LLM-based agent systems have been employed for stock trading [28, 31, 33]. These studies highlight the versatility and impact of LLMs in finance, emphasizing their growing significance in enhancing business operations and decision-making processes.

Despite the expanding body of work on LLM applications in finance, there remains a significant gap in exploring their potential for financial risk prediction, particularly for metrics like stock volatility and Value at Risk (VaR). In addition, since LLMs are generative models, the challenge, of using them for precise numerical regression [20], also persists in financial risk prediction tasks. Therefore, our research aims to fill this gap by leveraging LLMs for financial risk prediction, contributing valuable insights to understanding market dynamics and ensuring sustainable success across various sectors. This position paper is interested in exploring the following research questions (RQs):

- **RQ1:** What role do large language models play in financial risk prediction?



**Figure 1:** This figure illustrates the RiskLABS Framework. The model accepts multiple-source inputs: Earnings Conference Call Audio & Transcript, Daily News, and Time Series Data. The second area visualizes the model’s pipeline to encode diverse sources and illustrates how LLMs are applied for data analysis. 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 RiskLABS will predict the Volatility of different terms and VaR in the meantime.

- **RQ2:** How does the predictive performance of Large Language Models (LLMs) compare with other artificial intelligence techniques in forecasting risk metrics?
- **RQ3:** How can multiple inputs of varying data types be effectively incorporated and balanced?

To address the issues previously mentioned, we propose the RiskLABS framework, designed to comprehensively understand the complexities of the investment environment. This framework ingests and analyzes multimodal data from multiple sources, providing a holistic view of the factors influencing financial markets. The multimodal data sources include: (1) earnings conference call transcripts, offering insights into corporate performance and future prospects; (2) audio from these calls, capturing tone and sentiment not evident in the text; (3) time series data for historical and real-time market trend analysis and (4) media news. Leveraging LLMs, RiskLABS extracts essential information from each modality and applies multimodal fusion techniques to integrate these features, enhancing financial risk prediction.

RiskLABS is equipped with four key modules to effectively process these diverse data streams: a) The Earnings Conference Call Encoder: leveraging LLM to handle data related to earnings conference call (1 and 2); b) The News-Market Reactions Encoder: establishing a pipeline through LLM to collect and interpret news data; c) The Time-Series Encoder: organizing and analyzing time-related data; and d) The Multi-Task Prediction: which amalgamates outputs from the aforementioned modules for multifaceted prediction. This synthesis of varied inputs allows the model to provide a nuanced understanding of the investment landscape, blending quantitative

market data with qualitative insights. Such an integrated methodology is crucial for precise forecasting and risk assessment, a key to navigating the often turbulent and complex terrain of financial markets. By incorporating a diverse range of inputs and using processing steps, RiskLABS solves a multitasking problem: forecasting volatility over various intervals (3 days, 7 days, 15 days, and 30 days), as well as the value at risk (VaR).

Our contributions are encapsulated as follows: 1) We utilize LLM to develop the RiskLABS framework, thereby addressing the shortfall in applying LLMs to financial risk prediction. 2) The versatility of our model lies in its capability to seamlessly integrate a variety of financial multimodal data from multiple sources, consequently bolstering risk prediction accuracy. 3) The efficacy of our framework is evidenced by experimental results, demonstrating its effectiveness in forecasting financial risks. 4) Furthermore, we provide a comprehensive analysis of the pivotal role LLM should assume in the realm of financial risk prediction.

## 2 RiskLABS Framework

Figure 1 illustrates the RiskLABS Framework, designed to handle multiple data types surrounding the financial information including audio, text, and time-series from different sources. The framework comprises four main modules: 1) Earnings Conference Call Encoder; 2) Time-Series Encoder; 3) Relevant News Encoder; and 4) Multimodal Fusion and Multi-Task Prediction. This section outlines the format of the data input and details each module. In brief, the Earnings Conference Call Encoder, Time-Series Data Encoder, and Relevant News Encoder are utilized to extract features from various

data types. These features are fused data that undergoes processing and modeling, after which it is fed into the Multi-Task Prediction Block. This block is responsible for forecasting both volatility for different intervals and VaR (Value at Risk) values.

## 2.1 Earnings Conference Call Encoder

The Earnings Conference Call Encoder module consists of three main components: Audio Encoding, Transcript Encoding, and Earnings Conference Call Analyzer.

**2.1.1 Audio Encoding.** Audio embeddings are extracted using Wav2vec2 [3] and then processed through a Multi-Head Self-Attention (MHSA) layer with pooling to obtain audio features  $T_a$ . To describe the Audio Encoding in more detail, we let the raw audio input data be represented by  $A_c = \{a_c^1, a_c^2, \dots, a_c^n\}$  where  $a_c^i$  represents the  $i^{th}$  audio frame in one data sample. Each audio frame will be converted into a vector representation:

$$e_{ac}^i = \text{Wav2Vec2}(a_c^i). \quad (1)$$

We obtain the audio embeddings  $E_{ac} = \{e_{ac}^1, e_{ac}^2, \dots, e_{ac}^n\}$ . The shape of the  $E_{ac}$  is  $520 \times 512$ , where 520 is the maximum number of audio files amongst all companies and 512 is the dimension of the resulting transform from the model for a single audio frame. Earnings conference calls with less than 520 audio frames ( $n < 520$ ) have been zero-padded for uniformity in input matrix size.

Then, we fed audio embedding  $E_{ac}$  into the Multi-Head Self-Attention (MHSA) to extract the audio feature. The MHSA further comprises a multi-head attention block, followed by a norm block, and an MLP block. MLP denotes a two-layer feed-forward network (FNN) with a ReLU activation function. The MHSA block essentially forms the basis of all the architectures discussed later in the paper. In detail, the multi-head self-attention calculation process is as follows:

$$\text{Multihead} = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^o \quad (2)$$

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \quad (3)$$

where  $Q$  (queries) and  $K$  (keys) of dimension  $d_k$  and  $V$  values of dimension  $d_v$ . The dimensions of the weights are:  $W_i^Q \in R^{d_{\text{model}} \times d_k}$ ,  $W_i^K \in R^{d_{\text{model}} \times d_k}$ ,  $W_i^V \in R^{d_{\text{model}} \times d_v}$ , and  $W^o \in R^{d_v \times d_{\text{model}}}$ . The dot product is then calculated for the query with all the keys. These values are then normalized by dividing each value by  $\sqrt{d_k}$  and then we apply a softmax function to obtain the weights of the values:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{KQ^T}{\sqrt{d_k}}\right)V \quad (4)$$

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), we apply this process to  $E_{ac}$ :

$$T_{ac} = \text{MHSA}(E_{ac}) \quad (5)$$

where  $T_{ac} = \{t_{ac}^1, t_{ac}^2, \dots, t_{ac}^n\}$  with size  $520 \times 512$ . Following this, an average pooling layer is applied to  $T_{ac}$ :

$$T_a = \text{AveragePooling}(T_{ac}) \quad (6)$$

where  $T_a$  denotes the resultant extracted audio feature of size 512.

**2.1.2 Transcript Encoding.** Similar to Audio Encoding, this component first uses SimCSE [7] to generate vector representations for each sentence in the earnings call transcripts. These embeddings are then passed through an MHSA layer with pooling to capture text features  $T_t$ ; We let raw transcripts as  $T_c = \{t_c^1, a_c^2, \dots, t_c^n\}$  where  $t_c^i$  represents the  $i^{th}$  sentence in the transcript. Therefore, each sentence will be converted into a vector representation:

$$e_{tc}^i = \text{SimCSE}(t_c^i). \quad (7)$$

We obtain the corresponding text embeddings given by  $E_{tc} = \{e_{tc}^1, e_{tc}^2, \dots, e_{tc}^n\}$ . The shape of the  $E_{tc}$  is  $520 \times 768$ , 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 (2)-(5), the MHSA is applied to  $E_{tc}$  to get  $T_{tc} = \{t_{tc}^1, t_{tc}^2, \dots, t_{tc}^n\}$  with dimension  $520 \times 768$ . Then, the average pooling layer is applied to  $T_{tc}$ :

$$T_t = \text{AveragePooling}(T_{tc}) \quad (8)$$

where  $T_t$  denotes the resultant extracted textual feature of size 768.

**2.1.3 Earnings Conference Call Analyzer.** This component utilizes an LLM to summarize the earnings call, extracting essential and relevant information and converting it into a text feature. We show this process in Figure 2.

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:

$$T_s = \text{Embedding}(\text{Concatenated}[\text{chunk summaries}; \text{summary}]) \quad (9)$$

Next, 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 ??). 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:

$$T_f = \text{Embedding}(\text{Concatenated}[\text{selected sentences}; ]) \quad (10)$$

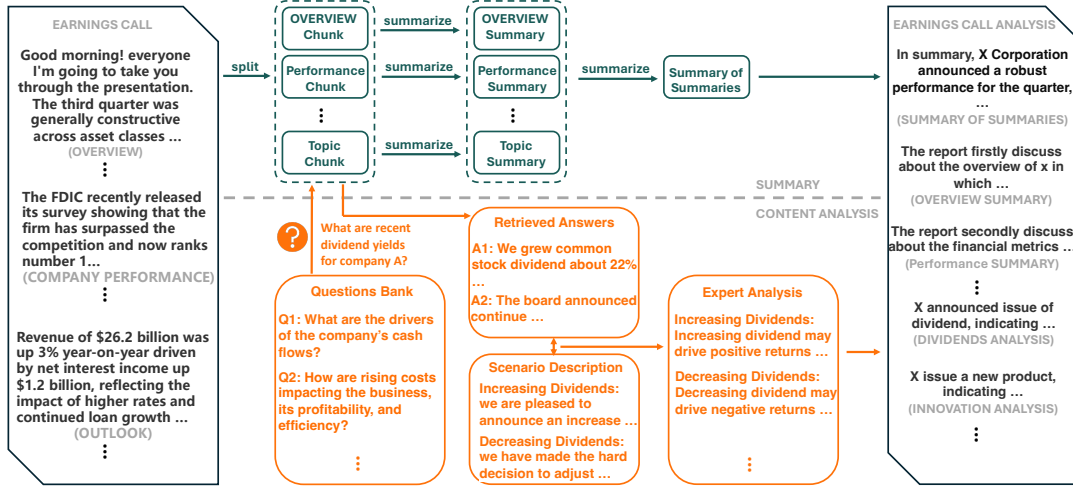


Figure 2: visualizes the process of fine-grained information extraction from ECC transcript.

## 2.2 Time-Series Encoder

For our time series input, we capture the VIX values from a 30-day period preceding the ECC release date. To distill meaningful features from this VIX series, we employ a Bidirectional Long Short-Term Memory (BiLSTM) network [18]:  $T_v = \text{BiLSTM}(\text{VIX})$ . In this setup, the BiLSTM is configured with 64 hidden states. Consequently, the output  $T_v$ , representing the extracted time series features, has a dimensionality of 128, reflecting the bidirectional nature of the BiLSTM.

## 2.3 Relevant News Encoder

News significantly influences stock movements and serves as a valuable indicator of market trends. It encompasses a wide range of information, including macroeconomic indicators, industry trends, and company-specific events such as earnings reports, mergers and acquisitions, regulatory changes, and management shifts. For each stock, we compile relevant news into a single text string, which is then processed by an LLM to generate a news feature  $T_n$ .

## 2.4 Multimodal Fusion and Multi-Task Learning

The additive fusion method can be seen as learning a new joint representation:

$$E = w_0 + w_1 \cdot T_a + w_2 \cdot T_t + w_3 \cdot T_f + T_v + T_n + \epsilon \quad (11)$$

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.

RiskLabs is a multi-task training model that trains and predicts the volatility of different terms and VaR. The individual stock volatility can be expressed as the natural log of the standard deviation of

return prices  $r$  in a window of  $\tau$  days. We calculate the 3, 7, 15, and 30 trading days volatility using the equation:

$$v_{[d-\tau, d]} = \ln \left( \left( \frac{\sum_{i=0}^{\tau} (r_{d-i} - \bar{r})^2}{\tau} \right)^{\frac{1}{2}} \right) \quad (12)$$

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  $N$  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  $\tau$  days. We aim to develop a predictive regression function  $f(c) \rightarrow v_{[d-\tau, d]}$ .

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

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

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

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

We train RiskLabs by optimizing multitask loss:

$$\mathcal{L} = \mu \left( \sum_i (\hat{y}_i - y_i)^2 \right) + (1 - \mu) \max(q \times (v - \hat{v}), (1 - q)(\hat{v} - v)). \quad (13)$$

**Table 1: Performance results on our proposed framework RiskLabs 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	<b>0.158</b>	/	✓
GPT-3.5-Turbo	2.198	2.152	1.793	2.514	2.332	0.371	✓
<b>RiskLabs</b>	<b>0.324</b>	<b>0.585</b>	<b>0.317</b>	<b>0.233</b>	0.171	<b>0.049</b>	✓

### 3 Experiment and Discussion

#### 3.1 Experiment Setup

We compare our approach to volatility prediction to several important baselines as described in the Appendix A.2. We use GPT-4 for the Earnings Conference Call Analyzer and the News-Market Reactions Encoder. 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 framework, we developed the code using PyTorch. Each Multi-Head Attention layer in the network comprises 8 individual heads, and 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  $\lambda$  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  $\mu$  which varied between the two tasks. We list additional details in Appendix A.1.

#### 3.2 Performance Comparison (RQ1 & RQ2)

The performance of different models in predicting financial risks is detailed in Table 1. This comparison includes baseline models, our proposed RiskLabs framework, and a range of periods (3, 7, 15, and 30 days). Besides forecasting accuracy, we also consider factors like the predicted Value at Risk (VaR) and the multi-tasking capabilities of each model. Notably, the RiskLabs framework outperforms others in prediction accuracy, particularly in short-term and medium-term forecasts, evidenced by the lowest Mean Squared Error (MSE) values. This improvement becomes more pronounced when the RiskLabs framework integrates various data sources such as earnings conference calls, time series, and news feeds, outshining the current state-of-the-art HTML solutions. Moreover, the RiskLabs framework demonstrates superior performance in VaR prediction, underscoring the effectiveness of our proposed methodology in offering a more nuanced and comprehensive approach to financial risk prediction. This could be invaluable for investors seeking to make more informed decisions. Nevertheless, the framework’s performance in 30-day forecasts lags behind the HTML models, indicating potential areas for further enhancement in long-term risk forecasting with LLM-based solutions.

In addition to comparing the effectiveness of various AI techniques in risk prediction, our study also evaluates the differences

in Value at Risk (VaR) predictions made by Large Language Models (LLMs), traditional financial methodologies, and neural network frameworks. This comparison sheds light on each approach’s strengths and weaknesses in risk quantification, offering a clearer understanding of how modern LLMs are different from financial and neural network models. Table 3 displays the performance of value at risk prediction between the traditional finance model and the AI techniques.

The predefined VaR value is 0.05, meaning that the closer the model’s predictions are to 0.05, the better its performance. The result above shows that the prediction of VaR via applying the historical method is 0.016, which is significantly below the pre-defined percentile(5%). It indicates that the historical method overestimated the 95% VaR benchmark. Tracing back to 2016, we know there was a global financial crisis in 2015, and its effect lasted till the beginning of 2016. On January 20, 2016, the price of crude oil fell below \$27 a barrel; the DJIA index took a roller coaster from down 565 points to down 249 intraday. In February, the YTD (yield to Date) return came to -10.5%. These events, together with the sequelae of the 2015 stock market crisis, define 2016 as a risky year. Comparing 2016, 2017 will be much better. In January 2017, DJIA achieved a new historical height, landing above 20,000. The stock market experienced a boost with a 25% growth rate for DJIA, 19% for S&P 500, and 28% for Nasdaq. The market had strong confidence, and the VIX index in 2017 came to its historical lowest point. That explains the reason why the historical method may overestimate the 95% VaR benchmark, due to this method duplicating the extreme scenarios from 2016 to 2017, which leads to the extra estimation of financial risks. We show more results of the analysis and visualization in Appendix B.1.

Turning our focus to Table 1, we observe that the direct application of LLMs for financial risk prediction is markedly ineffective, akin to making random guesses. This underscores a crucial caution; if LLMs are not utilized appropriately, they might elevate investment risks. Consequently, in response to Research Questions 1 and 2 (RQ1 and RQ2), we conclude:

- Utilizing LLMs through simple prompt instructions for direct financial risk prediction is ineffectual and potentially hazardous, increasing investment risks.
- **LLM is a bad trader/predictor, but it’s a helpful assistant.** While LLMs alone may not be reliable for direct risk prediction, they can serve as valuable tools in collating and

**Table 2: Ablation Study: performance results of different modules.**

Module	$\overline{MSE}$	$MSE_3$	$MSE_7$	$MSE_{15}$	$MSE_{30}$	VaR
Audio + Text	0.373	0.645	0.362	0.280	0.204	0.131
Audio + Text + Analysis	0.357	0.627	0.335	0.267	0.199	0.057
Audio + Text + Analysis + VIX	<b>0.324</b>	<b>0.585</b>	<b>0.317</b>	<b>0.233</b>	<b>0.171</b>	<b>0.049</b>

**Table 3: Comparison of Value at Risk Predictions: AI Techniques vs. Traditional Financial Methods. The predefined VaR value is 0.05.**

Method	Prediction of VaR
Historical Method	0.016
Fully Connected Neural Network	0.044
LSTM	0.056
<b>RiskLabs</b>	<b>0.049</b>

analyzing diverse financial data. This processed information, when fed into sophisticated deep learning models, significantly enhances AI’s capability in forecasting financial risks, thus positioning LLMs as beneficial assistants rather than standalone predictors.

### 3.3 Comparison Across Modules (RQ3)

In addition to overall performance, we evaluate the impact of different data combinations on predictive performance to better understand the relative contributions of data sources and each sub-module. We designed the ablation study as follows:

- **Audio + Text:** We design a comparison experiment starting with only using earnings conference calls. These calls are crucial in financial analysis as they are where executives of publicly traded companies discuss the company’s financial results for a specific period. They offer valuable insights into the company’s performance, strategic initiatives, and future projections. Investors highly regard these calls for the direct access they provide to the company’s leadership, offering in-depth financial data, operational updates, and forward-looking statements essential for informed investment decisions. In our approach, we initially concentrate on processing both the audio and textual content from these earnings conference calls using Multi-Head Self-Attention to extract salient features. Subsequently, we integrate these extracted features to directly predict financial outcomes.
- **Audio + Text + Analysis:** Then, we integrate the earnings conference calls analysis results into the prediction process. This is also the complete earnings conference call encoder in RiskLabs. This experiment will help us to verify whether the earnings conference calls analysis results are useful in helping to improve the model prediction performance.
- **Audio + Text + Analysis + VIX:** Next, we incorporate time series information into the model as well. We will determine whether the introduction of time series information

helps the model to make predictions by observing changes in predictive performance.

As shown in Table 2, the Audio + Text’ combination outperforms the baseline HTML model for 3-day forecasts. For longer periods—7, 15, and 30 days—RiskLabs achieves predictions closely aligned with HTML while still surpassing other baselines. Notably, with just Audio + Text’, our model is more streamlined than HTML, benefiting from a pre-trained model with extensive parameters that effectively map text and audio to vector representations. These vectors are processed through a multi-head self-attention mechanism for feature extraction, demonstrating the value of large-scale models in enhancing AI-driven risk prediction. Further improvements were observed upon integrating earnings conference call analysis and time-series data, leading to incremental gains in RiskLabs’ performance, especially for medium- to long-term forecasts (7, 15, and 30 days). These results suggest that earnings conference calls, as essential references for investors, have a more substantial impact on short-term risk volatility, while diverse information sources are crucial for improving accuracy in long-term forecasts. These findings underscore the contributions of each component within the RiskLabs framework, culminating in a robust predictive model.

### 3.4 Challenge and Opportunity

Our experimental findings demonstrate that leveraging large language models to integrate diverse information sources can significantly enhance the capability of our framework in predicting financial risks. With this in mind, we’ve expanded our data sources to include variables that could influence market volatility, such as news disseminated through social media. As detailed in Section 2.4, we utilize LLM to collect and analyze daily financial news. Initial small-scale experiments indicate that this addition can further improve the model’s performance. However, scaling up has introduced challenges: 1) the variable quality of news sources, with some containing misinformation, potentially introduces noise into our news database and affects audio model predictions; 2) there is a need to test the module’s effectiveness across a broader data spectrum. To address these challenges, we are implementing more nuanced steps to filter out low-quality news. Additionally, we are amassing newer data samples to enlarge our dataset, allowing for more comprehensive validation of our model’s efficacy. Furthermore, we will implement: a) Bayes-VaR method provides a probabilistic framework that allows for the incorporation of prior knowledge and the updating of beliefs as new evidence is presented. b) News-Market-Reactions Encoder c) "Dynamic Moving Time Window" and d) a "Time Decay Hyper-parameter" in RiskLabs. These features will enable more flexible training and forecasting on a daily basis, thus offering investors timelier and more accurate risk assessments. We described ongoing tasks as follows.

**3.4.1 Extract Relationship Among Multiple Response Variable Using Vector Auto-regression-based Method.** The task of RiskLabs is to predict multiple risk metrics in the meantime. Relationships will exist among them. Therefore, by identifying and modeling the relationships between them, we can significantly improve the accuracy of our predictions. This method shifts the focus from treating each metric as an isolated entity to understanding them as part of a complex system, where the dynamics between variables can provide critical insights for more precise risk assessment.

To achieve this, we attempt to find the relationship that links the two different stages: the  $m$ -day before and the  $m$ -day after. We will use the VAR(Vector Auto-Regression Model) to capture the linear relationship between the volatility of different terms. We set four different " $m$ -day volatility"(3-day volatility, 7-day volatility, 15-day volatility, and 30-day volatility). We define the  $\sigma_{-m,t}$  as the " $m$ -day future volatility" at time  $t - m$ ; the calculation method traces back  $m - \text{day}$  from today and calculates the  $m - \text{day}$ 's standard deviation of returns.

Volatility typically shows the cluster characteristics, which means that the change transmission pattern to the next stage tends to be similar to the last transmission pattern. These characteristics indicate that we could use the results from the close historical data as indicators or predictors to predict the next stage. We assume that the correlation between the volatility among different time scales will affect each other. We also measure the correlation by estimating the coefficient matrix and the information included in the historical data.

To estimate the coefficient of the model, we use the Bayesian Methodology by estimating the posterior distribution of the coefficient. If we have a general linear regression  $Y = X\theta$ , here  $\theta$  is the coefficient that we need to estimate. We can have the:

$$P(\theta|Y, X) = \frac{P(Y|\theta, X)P(\theta)}{P(Y|X)} = \frac{P(Y|\theta, X)P(\theta|X)}{\int_{-\infty}^{\infty} P(Y|X, \theta)p(\theta)d\theta} \quad (14)$$

Here,  $P(\theta)$ (or  $P(\theta|X)$ ) is the prior distribution that we could guess from the existing research or common knowledge. We also suppose that the independent variable is independent with the regression coefficient;  $P(Y|\theta, X)$  is the Likelihood Estimation from the data, and  $P(Y|X)$  is the marginal distribution. Usually, we could view  $P(Y|X) = \int_{-\infty}^{\infty} P(Y|X, \theta)p(\theta)d\theta$  as a constant because we could have the real value from the true dataset. From the Equation 14, we could have the following relation:

$$P(\theta|Y, X) \propto P(Y|\theta, X)P(\theta|X) \quad (15)$$

If we can determine the prior and the likelihood function, we can estimate the posterior distribution. We assume that the prior distribution  $P(\theta)$  satisfies the Normal distribution based on the simple linear regression. Subsequently, we consider the Likelihood  $P(Y|\theta, X)$ . In the empirical study from [2], the unconditional distribution of the realized historical distribution of the volatility is highly right-skewed. Besides, [5] tested the distribution historical volatility of S&P 500 stock from 1984 to 1996 and demonstrated that the pattern follows the log-normal distribution. Therefore, we took the log form of the original data. Given the preparation, we

build the VAR model as follows:

$$\begin{cases} \log(\sigma_{3,t}) = \alpha_3 + \beta_{1,1}\log(\sigma_{-3,t}) + \beta_{1,2}\log(\sigma_{-7,t}) + \beta_{1,3}\log(\sigma_{-15,t}) \\ \quad + \beta_{1,4}\log(\sigma_{-30,t}) + u_{3,t} \\ \log(\sigma_{7,t}) = \alpha_7 + \beta_{2,1}\log(\sigma_{-3,t}) + \beta_{2,2}\log(\sigma_{-7,t}) \\ \quad + \beta_{2,3}\log(\sigma_{-15,t}) + \beta_{2,4}\log(\sigma_{-30,t}) + u_{7,t} \\ \log(\sigma_{15,t}) = \alpha_{15} + \beta_{3,1}\log(\sigma_{-3,t}) + \beta_{3,2}\log(\sigma_{-7,t}) \\ \quad + \beta_{3,3}\log(\sigma_{-15,t}) + \beta_{3,4}\log(\sigma_{-30,t}) + u_{15,t} \\ \log(\sigma_{30,t}) = \alpha_{30} + \beta_{4,1}\log(\sigma_{-3,t}) + \beta_{4,2}\log(\sigma_{-7,t}) \\ \quad + \beta_{4,3}\log(\sigma_{-15,t}) + \beta_{4,4}\log(\sigma_{-30,t}) + u_{30,t} \end{cases}$$

Here,  $\sigma_{m,t}$  stands for the " $m$ -day" future volatility at time  $t$ ;  $u_{m,t}$  is a White noise term, which  $u_{m,t} \sim N(0, 1)$ ;  $\beta_{i,j}$  represents the linear relationship, which is the coefficient matrix of the VAR model,  $\alpha_m$ , which is the intercept term of the VAR model.

We apply the Monte Carlo Markov Chain(MCMC) algorithm to obtain the posterior distribution of the coefficient. MCMC is a method that is based on Bayesian Estimation. It is mainly used to predict the posterior distribution in a probability space by sampling. The Markov Chain follows the equation:

$$P(X_{t+1}|X_1, X_2, \dots, X_t) = P(X_{t+1}|X_t)$$

This could be explained by the fact that the probability of transforming only depends on the former statement. And based on the convergence theorem of Markov Chain, suppose the Markov Chain converges to one probability at  $n$ th step:  $\pi_n(x) \rightarrow \pi(x)$ , and  $X_n \sim \pi(x)$ ,  $X_{n+1} \sim \pi(x)$ , ..., we could say that  $X_n, X_{n+1}, \dots$  are i.i.d random variables.

We attempt to calculate the expectation of the posterior distribution by sampling. However, the posterior distribution is challenging to measure. To achieve this, we introduce the MCMC method to create a Markov Chain to achieve a stationary distribution and make it our desired posterior distribution. For MCMC, Metropolis-Hastings introduced a system to find a Markov Chain. For the model with many parameters to estimate, another alternative way is to use the Gibbs Sampling to update the estimated parameters one by one.

Regarding testing the MCMC process, we use several matrices, such as Monte Carlo standard errors (MCSE). There are two types, which measure the standard error of the mean and the standard error of the standard deviation of the chains. Also, because we did not generate the independent samples, the simulated samples are correlated. We want to know how many theoretically independent samples we drew and create the effective sample size (ESS), which measures the number of effectively independent samples we draw. Besides, we need to test if we achieve the stationary of our chains. Gelman-Rubin R-hat statistic( $\hat{R}$ ) gave us a measure. We measure  $\hat{R}$  by:

$$\hat{R} = \frac{\text{Variance between Chains}}{\text{Variance within Chains}}$$

Usually,  $\hat{R}$  should close to 1 and less than 1.01 otherwise it doesn't achieve the stationary with the Markov Chain.

In order to validate this methodology, we choose one of the stocks "TWTR," in our database as a case study, and then output the Bayesian regression results. Here, we used 250 days of data(from 2016-02-22 to 2017-02-15) as the model training period. Table 4 shows the training result. We observe that the  $\hat{R}$  of all parameters is equal to 1, and the  $ess\_tail$  value and  $ess\_bulk$  are all over 7000. Typically, there is no specific argument about the number of ESS.



**Table 4: Bayesian-VAR results by applying the MCMC method**

Dependent	Independent	N	Mean	Sd	Hdi_3%	Hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
3-Day	Intercept	250	-2.571	0.364	-3.275	-1.908	0.004	0.004	8342.0	7461.0	1.0
	3-Day	250	0.070	0.058	-0.044	0.174	0.001	0.000	9345.0	8076.0	1.0
	7-Day	250	0.118	0.116	-0.107	0.329	0.001	0.001	8512.0	7501.0	1.0
	15-Day	250	0.017	0.146	-0.249	0.297	0.002	0.001	7259.0	7407.0	1.0
	30-Day	250	0.203	0.143	-0.079	0.465	0.002	0.001	7072.0	7532.0	1.0
7-Day	Intercept	250	-2.748	0.251	-3.207	-2.265	0.003	0.002	8684.0	7809.0	1.0
	3-Day	250	0.042	0.040	-0.033	0.118	0.000	0.000	8600.0	7890.0	1.0
	7-Day	250	0.181	0.079	0.033	0.330	0.001	0.001	7281.0	7845.0	1.0
	15-Day	250	-0.078	0.100	-0.270	0.104	0.001	0.001	7428.0	7426.0	1.0
	30-Day	250	0.142	0.098	-0.035	0.333	0.001	0.001	7308.0	7386.0	1.0
15-Day	Intercept	250	-3.350	0.207	-3.729	-2.960	0.002	0.002	9410.0	8082.0	1.0
	3-Day	250	0.039	0.034	-0.025	0.103	0.000	0.000	10471.0	8203.0	1.0
	7-Day	250	0.068	0.067	-0.058	0.194	0.001	0.001	8018.0	7923.0	1.0
	15-Day	250	-0.026	0.085	-0.182	0.134	0.001	0.001	7053.0	7467.0	1.0
	30-Day	250	0.006	0.083	-0.149	0.165	0.001	0.001	7373.0	8124.0	1.0
30-Day	Intercept	250	-3.604	0.164	-3.904	-3.280	0.002	0.001	9388.0	8508.0	1.0
	3-Day	250	0.015	0.027	-0.034	0.067	0.000	0.000	9676.0	8515.0	1.0
	7-Day	250	0.108	0.052	0.006	0.200	0.001	0.000	7919.0	7510.0	1.0
	15-Day	250	0.005	0.065	-0.115	0.129	0.001	0.001	7004.0	7650.0	1.0
	30-Day	250	-0.140	0.065	-0.263	-0.020	0.001	0.001	7145.0	7494.0	1.0

[12] in the book mentioned that we need more ESS samples for the interested parameters. In the meantime, we also need to balance the number of ESS data and model training costs. [22] in the paper recommends having the ESS number over 400, which should be dependent on practical experience and model, to get a stable Monte Carlo standard error.

**Table 5: Statistics Results about AEP(Absolute Error Percentage) of the Experiment**

Variable	N	Mean	SD	Skewness	Kurtosis	5% Percentile	25% Percentile	50% Percentile	75% Percentile	95% Percentile
3-Day	100	-0.13167	0.08377	-0.42741	-0.70133	-0.26854	-0.19993	-0.13011	-0.07004	-0.01447
7-Day	100	-0.08501	0.05202	-0.37466	-0.43682	-0.17094	-0.11859	-0.08499	-0.04545	-0.00960
15-Day	100	-0.07950	0.05310	-0.23159	-1.37432	-0.16238	-0.11840	-0.07607	-0.03119	-0.00952
30-Day	100	-0.04195	0.03775	-1.75213	3.12077	-0.11573	-0.05777	-0.03011	-0.01555	-0.00600

We used the rolling window(see Figure 6) to make a rough simulation only based on the historical data of different volatility terms. We chose 250 days(Approximately one year) as the training set window for the next-day prediction and iterated 100 times. The result is shown in Table 5. This table measures the distribution of AEP(Absolute Error Percentage), which is calculated by the following formula:

$$AEP = \frac{|\hat{y} - y|}{y}$$

Here, the  $\hat{y}$  is the estimated value from the Bayesian VAR model, and  $y$  is the true value. From Table 5, we can see that the average estimation bias will become smaller when we estimate the longer volatility term, the same as the standard deviation of AEP. The negative skewness tells us that the peak of the AEP distribution is close to the y-axis(Because the mean is negative, as we calculated). In general, this experiment shows that using the VAR model based on historical data to predict the future term should be possible and could maintain some accuracy.

**3.4.2 News-Market-Reactions Encoder.** News plays an important role in influencing stock movements, serving as a powerful indicator of market trends. It contains a wealth of information, ranging from macroeconomic indicators and industry trends to company-specific news such as earnings reports, mergers and acquisitions, regulatory changes, and management shifts.

News carries several key features that commonly impact its reception and influence on its audience. Firstly, news is most impactful when it is fresh. Information that is current and up-to-date is more

likely to influence decisions and perceptions than outdated news. Secondly, similar news tends to produce comparable impacts on market reactions. When news items share key characteristics—such as subject matter, sentiment, and relevance to investors—they often trigger similar responses among market participants. For instance, announcements of unexpected earnings surpassing market forecasts typically lead to positive stock price movements. Conversely, news of regulatory setbacks or legal challenges can prompt a downturn in stock values. This patterned response largely stems from investors collectively interpreting news through the lens of their past experiences and established market precedents, leading to the development of conditioned reactions to specific types of news. Consequently, understanding the similarities in news items can offer insights into potential market reactions. However, given the volume of news a company may receive in a day, and considering that each piece of news partially influences the future stock price, it's essential to assess the entire context. To account for the cumulative effect and interrelations among different news items, we analyze them collectively, treating all news related to a company within a single day as a unified analytical unit.

Initially, we collect news pertaining to a particular company from the three days prior to our target trading day, including the market responses for each day. This compilation allows a Large Language Model (LLM) to examine the data and deduce how the news might influence the stock's performance in the subsequent days. Subsequently, our strategy involves identifying historical dates on which the news profiles closely resemble those on the day before the intended trading day. This comparative analysis aims to understand the potential market reactions based on similar past events, thereby informing our trading strategies with a nuanced perspective on news impact.

Based on these news features, we propose a news-market reactions encoder composed of two main components. Firstly, we collect news for a specific company from the three days prior to our target trading day, along with the market reactions for each day. We then use an LLM to analyze the data and identify how the news might affect the stock's performance in the days that follow. Secondly, we attempt to identify the historical date whose news is similar to that of the intended trading date. In this scenario, we face the challenge of assessing similarity once more. Given that we are comparing two groups of news, the likelihood of identical combinations of news reoccurring is extremely small, which may lead to less accurate retrieval results.

To address this problem, we designed an enriched news pipeline to obtain attributes associated to the news and then attach this news to news group. Specifically, the enriched news pipeline comprises several key steps to analyze news content effectively.

After processing all news in our database through this pipeline, we assign these identified attributes to each news group as metadata. This metadata becomes instrumental in locating news similar to that of the target trading days. Rather than directly comparing two groups of news, we initially compare the metadata between them. This process helps us identify the top k news groups in the historical record that share similar metadata attributes. Subsequently, we assess the similarity between these top k groups and the news groups from the target trading date.





**Figure 3:** This figure illustrates the pipeline for enriching the news information. First, the pipeline will analyze the sentiments from the target news. Then, based on the binary questions bank designed for different topics, Pipeline can extract the information and answer these questions. Finally, if the pipeline could capture the signal of a specific topic, it would also give feedback on the potential market response.

This method significantly streamlines the process of finding similar news groups, enabling us to efficiently draw parallels. With these similarities established, we can make informed predictions about the potential impact of these news groups on the stock market in the days following the target trading date, based on the market reactions to historically similar news.

In this way, it is efficient to identify the two groups that are similar, then we can make an inference for the target trading date about how these results could impact the following stock market based on the historically similar news and its corresponding market reactions.

**3.4.3 Time Decay Hyper-parameter and Dynamic Moving Time Window Training.** In previous sections, we examined the model inputs: (1) earnings conference call transcripts, (2) historical time series data, and (3) news articles. However, the availability frequency of these inputs varies. News and historical time series data are available on a daily basis, whereas earnings conference calls are released on specific dates when a company presents its earnings report.

Consequently, the input from earnings conference calls may be absent on some training days. Ignoring this input entirely would be imprecise, as the information from an earnings conference call can continue to affect stock price movements after its publication. To incorporate the continuous influence of earnings conference call information even on days without new input, we will introduce a

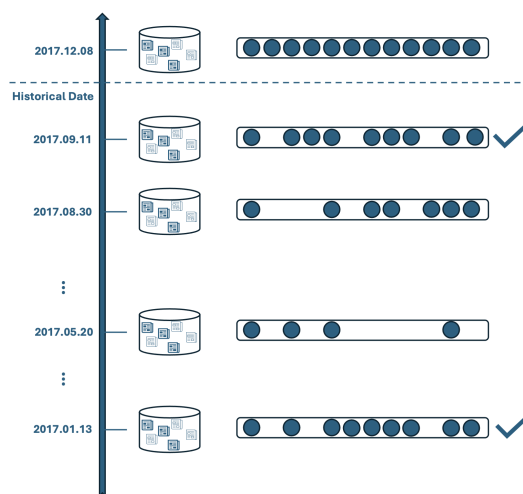
hyper-parameter to measure the decaying speed of its relevance. We will apply the Exponential Decay function:

$$I(t) = I(0)e^{-\lambda t}$$

In the equation,  $\lambda$  is the hyper-parameter that measures the decay rate.  $t$  is the time to the last earnings conference call release date. When the  $\lambda$  is fixed, the impact of the earnings conference call will still decrease as  $t$  gets larger, which satisfies our common sense. Figure 5 shows the daily data inputs, for days lacking a new earnings conference call, we still incorporate the information from the most recent call, adjusted by  $\lambda$  to reflect its diminishing impact speed over time.

Another consideration related to timing is the impact of input data on the market across time. RiskLabs takes time series data and news as inputs. The relevance of these inputs to the response variables is strongly dependent on the timing of the data. If we solely train a model without further updates and then use it to make predictions for dates far away the training period, the model's predictive performance may decline. This decrease in accuracy arises because the model becomes less relevant of later dates.

Hence, updating the model at a high frequency can continuously adjust the model's parameters to maintain its sensitivity to the latest market trends and fluctuations. To achieve this, we use a dynamic moving time window method for training. Specifically, we use a fixed window of historical data up to the target trading day to train our model, which is then used only to predict the



**Figure 4:** This diagram illustrates the process by which the News Analyzer assesses similarities across various news collections. As news items pass through the enrichment pipeline, they are tagged with multiple attributes. Each circle in the figure represents one of these attributes. To identify similar news collections from historical data, the analyzer starts by comparing these attributes and filtering out certain ones. Subsequently, it evaluates the similarity among the remaining attributes to determine the connections between different news collections.



**Figure 5:** This figure illustrates the methodology for modeling the impact of earnings conference calls on the stock market when earnings conference calls are not available. It uses a curve to represent the time decay effect of an earnings conference call's influence over time. When an earnings conference call is initially issued, its impact on the market is at its peak. Over time, this influence gradually diminishes.

response variable for that particular day. Once the day has passed, we shift the window to include the newly added day and proceed to train a new model for the next target trading day (See Figure 6). This iterative process ensures our model remains adaptive and up-to-date with the latest market data.

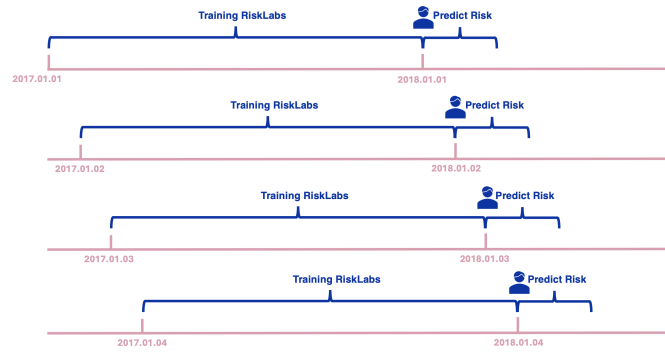
## 4 Conclusion

In this study, we explored the utilization of LLMs in predicting financial risks and introduced the RiskLabs framework. This innovative framework employs LLMs to systematically organize and analyze diverse financial data types and sources, augmenting deep learning models in financial risk prediction. Central to RiskLabs are specialized modules: the Earnings Conference Call Encoder, the Time-Series Encoder, and the News-Market Reactions Encoder, each designed to process specific financial data. These encoders collectively facilitate the merging of various data features for robust multi-task financial risk forecasting. The empirical findings from our study led to several key insights: 1) The RiskLabs framework

demonstrates a high efficacy in predicting financial risks, confirming its potential as a valuable tool in this domain. 2) While LLMs in isolation may not yield effective financial risk predictions, their strategic application in processing relevant financial data significantly enhances the predictive power of deep learning models. 3) Ablation studies further affirm that each individual module of RiskLabs meaningfully contributes to the accuracy of the final risk predictions. Overall, our research underscores the transformative potential of LLMs in financial risk assessment, marking a significant step forward in the application of AI in finance.

## References

- [1] Noujoud Ahbali, Xinyuan Liu, Albert Nanda, Jamie Stark, Ashit Talukder, and Rupinder Paul Khandpur. 2022. Identifying corporate credit risk sentiments from financial news. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*. 362–370.
- [2] Torben G Andersen, Tim Bollerslev, Francis X Diebold, and Heiko Ebens. 2001. The distribution of realized stock return volatility. *Journal of financial economics* 61, 1 (2001), 43–76.
- [3] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations.



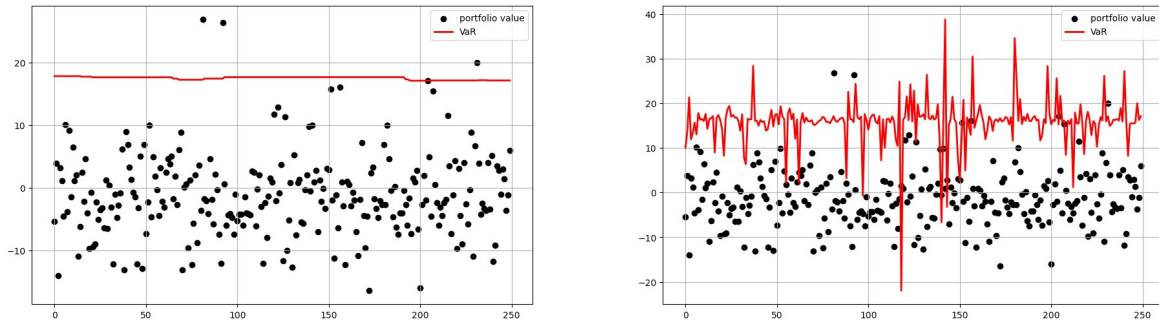
**Figure 6: This figure provides a visual representation of the Rolling Window methodology in action. By establishing a fixed window length, the approach systematically progresses the training set forward, day by day, in alignment with the passage of time.**

- Advances in neural information processing systems* 33 (2020), 12449–12460.
- [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] Pierre Cizeau, Yanhui Liu, Martin Meyer, C-K Peng, and H Eugene Stanley. 1997. Volatility distribution in the S&P500 stock index. *Physica A: Statistical Mechanics and its Applications* 245, 3–4 (1997), 441–445.
  - [6] 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.
  - [7] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821* (2021).
  - [8] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. 2000. Learning to forget: Continual prediction with LSTM. *Neural computation* 12, 10 (2000), 2451–2471.
  - [9] Shihao Gu, Bryan Kelly, and Dacheng Xiu. 2020. Empirical asset pricing via machine learning. *The Review of Financial Studies* 33, 5 (2020), 2223–2273.
  - [10] Luckyson Khaideem, Snehanishu Saha, and Sudeepa Roy Dey. 2016. Predicting the direction of stock market prices using random forest. *arXiv preprint arXiv:1605.00003* (2016).
  - [11] 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.
  - [12] John Kruschke. 2014. Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan. (2014).
  - [13] Young-Chan Lee. 2007. Application of support vector machines to corporate credit rating prediction. *Expert Systems with Applications* 33, 1 (2007), 67–74.
  - [14] Charalampos M Liapis, Aikaterini Karanikola, and Sotiris Kotsiantis. 2023. Investigating deep stock market forecasting with sentiment analysis. *Entropy* 25, 2 (2023), 219.
  - [15] 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).
  - [16] Saloni Mohan, Sahitya Mullanpudi, Sudheer Sammeta, Parag Vijayvergia, and David C Anastasiu. 2019. Stock price prediction using news sentiment analysis. In *2019 IEEE fifth international conference on big data computing service and applications (BigDataService)*. IEEE, 205–208.
  - [17] 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.
  - [18] Sima Siami-Namini, Neda Tavakoli, and Akbar Siami Namin. 2019. The performance of LSTM and BiLSTM in forecasting time series. In *2019 IEEE International conference on big data (Big Data)*. IEEE, 3285–3292.
  - [19] 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)*. Ieee, 1310–1315.
  - [20] Xingyou Song, Oscar Li, Chansoo Lee, Daiyi Peng, Sagi Perel, Yutian Chen, et al. 2024. Omnipred: Language models as universal regressors. *arXiv preprint arXiv:2402.14547* (2024).
  - [21] Wataru Souma, Irena Vodenska, and Hideaki Aoyama. 2019. Enhanced news sentiment analysis using deep learning methods. *Journal of Computational Social Science* 2, 1 (2019), 33–46.
  - [22] Aki Vehtari, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Brkner. 2021. Rank-normalization, folding, and localization: An improved R for assessing convergence of MCMC (with discussion). *Bayesian analysis* 16, 2 (2021), 667–718.
  - [23] Dan Wang, Zhi Chen, Ionuț Florescu, and Bingyang Wen. 2023. A sparsity algorithm for finding optimal counterfactual explanations: Application to corporate credit rating. *Research in International Business and Finance* 64 (2023), 101869.
  - [24] 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 (2022), 24824–24837.
  - [25] 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: A holistic financial benchmark for large language models. *arXiv preprint arXiv:2402.12659* (2024).
  - [26] Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. Fingpt: Open-source financial large language models. *arXiv preprint arXiv:2306.06031* (2023).
  - [27] Linyi Yang, Tin Lok James Ng, Barry Smyth, and Rihui Dong. 2020. Htl: Hierarchical transformer-based multi-task learning for volatility prediction. In *Proceedings of The Web Conference 2020*. 441–451.
  - [28] 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).
  - [29] Beichen Zhang. 2020. *Financial Risk Disclosure Return Premium: A Topic Modeling Approach*. Master’s thesis. Stevens Institute of Technology.
  - [30] 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*. 349–356.
  - [31] Chong Zhang, Xinyi Liu, Mingyu Jin, Zhongmou Zhang, Lingyao Li, Zhengting Wang, Wenyue Hua, Dong Shu, Suiyuan Zhu, Xiaobo Jin, et al. 2024. When AI Meets Finance (StockAgent): Large Language Model-based Stock Trading in Simulated Real-world Environments. *arXiv preprint arXiv:2407.18957* (2024).
  - [32] 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 (2024), 39–57.
  - [33] Wentao Zhang, Lingxuan Zhao, Haochong Xia, Shuo Sun, Jiaze Sun, Molei Qin, Xinyi Li, Yuqing Zhao, Yilei Zhao, Xinyu Cai, et al. 2024. FinAgent: A Multi-modal Foundation Agent for Financial Trading: Tool-Augmented, Diversified, and Generalist. *arXiv preprint arXiv:2402.18485* (2024).

## A Experiment Details

### A.1 Dataset

For dataset usage, we strategically partitioned it into a training set and a test set, adhering to an 8:2 ratio. It is crucial to highlight that the division of the data was conducted on a temporal basis,



**Figure 7:** This figure presents two plots that compare daily Value at Risk predictions (red curve) with the actual returns of the asset (black dot). It visualizes the percentage of actual returns exceeding the predicted VaR. For instance, with a predefined VaR of 0.05, observing approximately 5% of the actual returns surpassing the predicted VaR curve indicates a high degree of prediction accuracy. The left plot showcases the VaR forecast using the historical method, illustrating how this traditional technique estimates risk in relation to the asset’s actual performance. The right plot, on the other hand, using a fully connected neural network for VaR prediction, offering a modern computational approach to risk assessment.

ensuring that the dates of the data in the training set always precede those in the test set. This temporal segregation is imperative for maintaining the integrity of our predictive model. By structuring the dataset in this manner, we ensure that the training process is consistently oriented towards predicting future risks based on past data, a fundamental principle for the accuracy and reliability of our forecasting methodology.

## A.2 Baseline Setup

We compare our approach to volatility prediction to several important baselines as described below.

- **Classical Methods:** The analysis incorporates the GARCH model, a classical auto-regressive model for predicting volatility, as well as its various derivatives [6, 11]. These models are widely recognized and commonly employed in the realm of volatility prediction. Primarily, they are tailored for short-term volatility forecasting and may not perform as efficiently in predicting average volatility over an extended period (e.g., n-day volatility).
- **LSTM [8]:** Long Short-Term Memory Networks (LSTMs) are a popular choice for financial time series prediction due to their efficacy in handling sequential data. In the context of volatility prediction, we select a straightforward LSTM model as a benchmark.
- **MT-LSTM+ATT [15]:** merges the prediction of average n-day volatility with the forecasting of single-day volatility, using attention-enhanced LSTM as the foundational learning models.
- **HAN (Glove):** This baseline implements a Hierarchical Attention Network that incorporates dual-layered attention mechanisms at both the word and sentence levels. Initially, every word within a sentence is transformed into a word embedding through the pre-trained Glove 300-dimensional embeddings. Subsequently, these embedded sentences are processed by a Bi-GRU encoder [4], and in parallel, another Bi-GRU encoder is utilized to formulate a representation of each document as a series of sentences. This representation

of the document is then fed into the concluding regression layer to generate predictions.

- **MRDM [17]:** The MRDM model first introduced a multi-modal deep regression approach for volatility prediction tasks. It utilizes pre-trained GloVe embeddings and bespoke acoustic features, which are processed through individual BiLSTMs to generate uni-modal contextual embeddings. These embeddings are subsequently merged and input into a two-layer dense network for further processing.
- **HTML [27]:** 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-3.5-Turbo:** We evaluated the efficacy of using GPT-3.5-Turbo for direct financial risk prediction. The input for this test was earnings conference calls, and we instructed GPT-3.5-Turbo to generate numerical risk forecasts based on the provided EC, utilizing a specified prompt setting. In the experiment, we set the temperature as zero.

## B Detailed Experiment Results

### B.1 Detailed analysis of VaR values

From the analysis presented in the figure, a notable observation emerges: the plot on the left, which uses the historical method for VaR prediction, appears relatively flat, indicating a consistent, albeit less responsive, forecast over time. In contrast, the plot on the right, using a fully connected neural network, exhibits a more zigzag pattern, reflecting greater responsiveness to daily information changes. This contrast suggests that AI techniques, such as neural networks or LLMs, offer a dynamic advantage by more effectively incorporating daily updates into the model, as opposed to relying solely on historical scenarios. The ability of AI-driven models to adapt to new information underscores their potential for providing more accurate and timely risk assessments.