University of California

Santa Barbara

Long-Context and Multimodal Language Models for Financial Forecasting

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy

in

Computer Science

by

Ross Koval

Committee in charge:

Professor Xifeng Yan, Chair

Professor William Wang

Professor Shiyu Chang

June 2026

The Dissertation of Ross Koval is approved.

Professor William Wang

Professor Shiyu Chang

Professor Xifeng Yan, Committee Chair

May 2026

Long-Context and Multimodal Language Models for Financial Forecasting

by

Ross Koval

To my family, for their endless love, support, and encouragement.
And to myself, for the curiosity to ask, the discipline to persist,
and the resolve to push the boundaries of knowledge.

Acknowledgements

I am deeply grateful to my advisor, Xifeng Yan. Without his guidance, support, and vision, this dissertation would not have been possible. His foresight and willingness to explore underexplored research directions at the intersection of language models and financial forecasting has been instrumental in shaping this dissertation and my development as a researcher. I would also like to thank my dissertation committee members William Wang and Shiyu Chang, and our collaborator Nicholas Andrews, for their insightful feedback and continued support.

I am sincerely thankful to AJO Vista, particularly Jesse Barnes and Christopher Covington, for their endless support of my academic pursuits, the freedom and flexibility to explore challenging problems, and for providing the resources necessary to conduct this research.

Finally, I would like to express my deepest gratitude to my family for their unwavering love, support, and encouragement throughout this deeply rewarding yet challenging journey.

Thesis Statement

In this thesis, I aim to develop machine learning methods for modeling financial data that is inherently domain-specific, long-context, temporally structured, and multimodal, including sequences of textual, tabular, and time series data. My goal is to design specialized AI systems that capture these structural properties and can extract, integrate, and reason over these heterogeneous information sources to generate predictive signals, enabling financial forecasting and decision-making at scale in real-world settings.

Curriculum Vitæ

Ross Koval

Education

University of California, Santa Barbara, Santa Barbara, California *Sept 2022 – June 2026*
Ph.D. in Computer Science
Advisor: Professor Xifeng Yan

Columbia University, New York, New York *Jan 2020 – Dec 2021*
M.S. in Computer Science (Machine Learning Track)

California Institute of Technology (Caltech), Pasadena, California *Sept 2013 - June 2017*
B.S. in Business Economics (Statistics Track)

Publications

[5] **Multimodal Language Models for Financial Forecasting from Interleaved Sequences of Text and Time Series**
Ross Koval, Nicholas Andrews, Xifeng Yan
IJCNLP-AACL 2025

[4] **Context-Aware Language Models for Forecasting Market Impact from Sequences of Financial News**
Ross Koval, Nicholas Andrews, Xifeng Yan
arXiv preprint, 2025

[3] **Financial Forecasting from Textual and Tabular Time Series**
Ross Koval, Nicholas Andrews, Xifeng Yan
EMNLP 2024

[2] **Learning to Compare Financial Reports for Financial Forecasting**
Ross Koval, Nicholas Andrews, Xifeng Yan
EACL 2024

[1] **Forecasting Earnings Surprises from Conference Call Transcripts**
Ross Koval, Nicholas Andrews, Xifeng Yan
ACL 2023

Teaching Experience

Teaching Assistant for CS165A: Principles of Artificial Intelligence (UCSB)

Head Teaching Assistant for ACM 157: Advanced Statistical Inference (Caltech)

Teaching Assistant for ACM 11: Mathematical Programming Methods (Caltech)

Professional Experience

Senior Quantitative Researcher – NLP, AJO Vista *2022 – 2026*

Senior Quantitative Researcher – NLP, HighVista Strategies *2019 – 2021*

Quantitative Researcher – NLP, Goldman Sachs *2017 – 2018*

Abstract

Long-Context and Multimodal Language Models for Financial Forecasting

by

Ross Koval

Financial markets produce vast amounts of multimodal data, including textual sources, such as financial news, earnings conference calls, and financial reports, coupled with structured numerical data, such as market prices and accounting statements. This combination of rich multimodal inputs and the natural availability of market-based supervision presents a compelling opportunity for the development of machine learning and natural language processing methods. However, financial data exhibits several unique challenges. Financial documents are long, complex, and domain-specific. The data has strong temporal structure and information must be interpreted in the context of prior related events. In addition, textual and numerical modalities must be jointly modeled to capture their unique patterns and interactions. While existing language models are remarkably capable, they remain limited in their ability to address these challenges, as they struggle with long-range temporal dependencies, structured numerical reasoning, and forecasting future market outcomes.

This dissertation advances financial NLP and forecasting through the development of long-context and multimodal language modeling methods tailored to the unique structural properties of financial data. To this end, we develop a sequence of methods that progressively expand the modeling scope and depth from single documents to temporally structured, multimodal sequences.

First, we begin by demonstrating that domain-adapted and long-context language models can be trained end-to-end on earnings conference calls to predict long horizon fi-

nancial outcomes. Second, we introduce methods for comparing financial documents, enabling models to capture predictive signals expressed through subtle relationships across key content. Third, we model paired and temporally aligned sequences of textual and tabular data, demonstrating that each modality contains unique structure and predictive information. Fourth, we study contextualized forecasting from sequences of financial news, proposing an efficient method for incorporating relevant historical context. Finally, we develop a multimodal architecture with modality-specific expert components for jointly modeling interleaved, mixed-frequency sequences of text and time series data.

Collectively, these contributions demonstrate the value of specialized modeling approaches designed to capture the unique structure of domain-specific language, long-range dependencies, temporal dynamics, and cross-modal interactions, leading to significant improvements in forecasting performance and economically meaningful gains in simulated investment strategies.

Contents

Chapter 1

Introduction

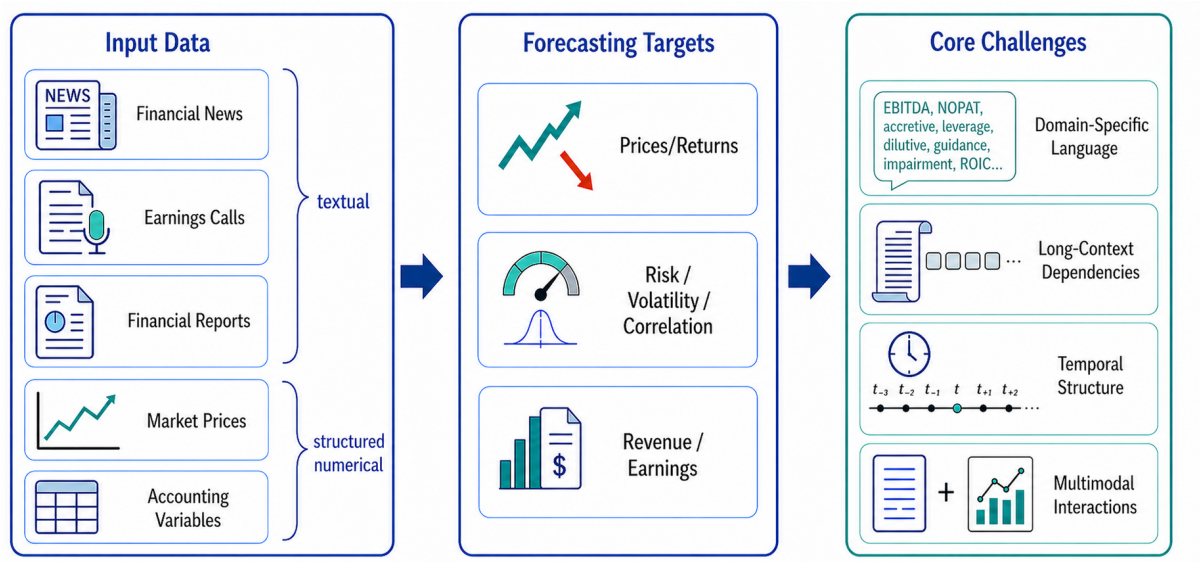

Figure 1.1: Overview of the diverse financial forecasting settings studied and core challenges addressed in this dissertation.

1.1 Introduction

Financial markets produce large volumes of textual data, including financial news, earnings conference calls, and financial reports, as well as structured numerical data, such as market prices and accounting variables. Accompanying these multimodal inputs are also many valuable market-based prediction targets available to use as supervision, including market price-based measures, such as returns and risk, and fundamental accounting-based measures, such as revenue and earnings. This combination of rich multimodal data and market-provided supervision presents a promising opportunity for the development and application of machine learning and natural language processing methods. However, financial data has unique characteristics that pose significant challenges for extracting actionable insights, including domain-specific language, long-context dependencies, temporal structure, and multimodal interactions.

1.2 Financial NLP and Forecasting

Traditionally, financial forecasting has relied on structured numerical data, such as accounting variables and market prices, to predict future firm performance and asset returns [1]. However, substantial information relevant to these market outcomes is expressed in unstructured text. For instance, financial news, earnings calls, and regulatory filings often provide qualitative context that is not directly reflected in numerical variables, including management expectations, strategic developments, risk disclosures, and interpretations of recent events. This setting motivates the use of financial NLP as a principled approach for improving financial forecasting by incorporating complementary information that is not captured by structured numerical data alone.

Early work in financial text analysis relied heavily on expert-curated dictionaries to measure sentiment and related linguistic properties [2, 3]. These studies found that generic sentiment lexicons are poorly suited for financial language, leading to the development of domain-specific resources such as the Loughran-McDonald Financial Sentiment Dictionary [3]. However, dictionary-based approaches are limited by their lack of contextual understanding, sensitivity to distribution shift, and dependence on manual curation. In fact, recent evidence suggests that corporate executives have begun to strategically reduce their usage of negative dictionary words [4], further illustrating their fragility.

More recently, pretrained language models have been shown to enable stronger contextual understanding of financial text, including sentiment analysis, summarization, question answering, and other domain-specific tasks [5–13].

Despite this progress, most existing financial language models are designed for short, text-only inputs, and are not well aligned with the unique structural demands of financial forecasting. In practice, financial prediction requires more than superficial, sentence-level language understanding. Rather, it requires models that can adapt to financial language,

reason over long and temporally structured documents, and integrate textual information with structured numerical time series.

1.3 Core Challenges in Financial Data

This dissertation addresses four central challenges in applying language models to financial forecasting.

First, financial language is highly **domain-specific** and specialized, with heavy use of jargon, implicit references, and context-dependent semantics, requiring effective domain adaptation. Models trained primarily on general-domain text often fail to capture the meaning and significance of expressions that are specific to markets, accounting, and corporate disclosures [3].

Second, financial documents are often **long-context** in nature. Earnings conference calls, annual reports, and sequences of related news articles span thousands of tokens, and the predictive signal is often distributed across multiple passages rather than localized in a single sentence. This information structure motivates the need for strong long-context understanding as a core requirement for financial NLP. While Transformer-based LMs can, in principle, process long sequences, the self-attention mechanism scales quadratically in both time and space complexity, which has historically constrained most pretrained language models to short inputs. Recent work on efficient attention mechanisms [14–17] has expanded the feasible context length, but effective utilization of long-range information remains challenging. In particular, causal decoder models have been shown to exhibit attention sinks, a phenomenon in which a disproportionate amount of attention mass concentrates on the first few tokens of the sequence, regardless of query content, as a consequence of the causal attention mask and the softmax normalization constraint [18, 19]. This positional bias contributes to the broader lost-in-the-middle effect, in which

models struggle to utilize information located in the middle of long contexts and often become distracted when provided with mixed relevance content [20].

Third, financial data is inherently **temporal**. A given financial document cannot be interpreted in isolation; its meaning depends on the history of related documents and events that precede it. For instance, a news article reporting a revenue miss carries different implications depending on whether it follows a sequence of positive or negative reports for the same company. This creates a modeling challenge that is distinct from simply having a long context: the model must reason over temporal sequences of documents, each of which may itself be long and of mixed relevance to others, and capture how information evolves and compounds over time.

Fourth, financial data is inherently **multimodal**. In realistic forecasting settings, textual documents coexist with structured numerical signals and these modalities evolve jointly over time. Effective prediction often depends on reasoning across both temporal and modality dimensions. This introduces additional challenges, since language is a compressed, high-level modality, while financial time series are raw and low-level, exhibiting high noise, irregular patterns, and lower-dimensional structure [21]. Recent work has shown that applying the same set of parameters to all modalities produces negative transfer and interference [22–25], broadly reinforcing findings in text-only language models in which sparse mixture-of-experts (MoE) architectures [26, 27] improve performance through parameter specialization, suggesting that modality-specific parameterization may be similarly beneficial in these multimodal settings. Furthermore, language models are generally poorly suited to natively represent and reason over time series data, as they are explicitly optimized for linguistic structure [28–30]. These findings motivate the need for the development of modality-specific modeling designs and cross-modal alignment strategies.

Collectively, this set of challenges highlights the need for specialized modeling ap-

proaches that are explicitly designed for the unique characteristics of financial data and forecasting.

1.4 Forecasting Tasks and Evaluation

In this dissertation, financial forecasting serves as the primary benchmark for evaluating a model's ability to understand and reason over information from textual and structured financial data. The methods developed here are evaluated across multiple forecasting tasks and under diverse settings, progressing from single-document prediction to temporally structured multimodal sequences. The prediction targets considered include both price-based measures, such as returns, risk, and correlations, and fundamental-based measures, such as earnings. These targets provide a natural source of supervision derived from future market outcomes.

Although these problems are typically formulated as supervised learning tasks, financial forecasting is fundamentally different from traditional supervision settings because the target labels are forward-looking, noisy, and only weakly predictable. Financial markets are complex adaptive systems in which many participants with heterogeneous information, beliefs, and objectives interact to determine asset prices. As a result, the relationship between input features and future outcomes is noisy, non-stationary, and highly context-dependent [31]. Predictive signals are weak individually, becoming informative only when aggregated across large datasets or long time horizons. This setting necessitates evaluation across extended time periods that span diverse economic regimes to rigorously assess whether a model captures robust predictive patterns rather than spurious artifacts of a particular market environment.

Furthermore, a critical requirement in the financial domain is the use of strictly temporally structured evaluation protocols [32, 33]. Since financial data is sequential,

models must be evaluated on held-out data that occurs strictly after the training period to prevent information leakage and reflect realistic forecasting conditions. In addition to standard predictive metrics, this thesis also evaluates model outputs through investment portfolio simulations under realistic implementation assumptions, including liquidity filters and conservative estimates of transaction costs [31, 34, 35]. This setup provides a direct measure of the economic value of model predictions and ensures that reported gains reflect net performance after implementation costs.

Taken together, this combination of diverse forecasting tasks, temporally structured evaluation, and economic simulations makes financial forecasting a real-world, challenging, and rigorous setting for evaluating long-context, temporally grounded, and multimodal language models.

1.5 Summary of Contributions

This dissertation advances financial NLP and forecasting through a sequence of contributions in the following chapters that progressively expand the modeling framework to address increasingly complex settings, from single documents to temporally structured and multimodal sequences.

First, in Chapter 2 [36], we explore domain adaptation and long-context modeling of earnings conference calls, and demonstrate that language models can be trained directly on transcript text to predict long-horizon financial outcomes, outperforming traditional models based on financial variables.

Second, in Chapter 3 [37], we study financial forecasting settings in which predictive signals are expressed through subtle relationships in key content across long financial documents. We introduce methods for aligning and comparing these documents, enabling models to learn cross-document reasoning over complex, nuanced signals. We

demonstrate that modeling fine-grained cross-document interactions yields significant improvements in forecasting company risk and correlation from the text of financial reports.

Third, in Chapter 4 [38], we extend the forecasting setting from text-only inputs to temporally aligned multimodal sequences, modeling complex interactions between textual documents and structured numerical signals across both temporal and modality dimensions. We demonstrate that each modality contains unique structure and predictive information, and the importance of targeted cross-modal fusion.

Fourth, in Chapter 5 [39], we study contextualized forecasting from sequences of financial news articles, where a given article can only be accurately interpreted in the context of prior related reporting. We propose an efficient contextualization and alignment framework for incorporating relevant historical context, improving the language model's understanding of the current news, as well as downstream forecasting and investment performance.

Finally, in Chapter 6 [40], we study multimodal forecasting from interleaved, mixed-frequency sequences of text and time series data. We propose a multimodal architecture with modality-specific expert components and a cross-modal alignment framework for jointly modeling these interleaved sequences. We demonstrate that dedicated capacity for each modality mitigates negative transfer and improves cross-modal reasoning.

Collectively, these works demonstrate that language models can be specialized to capture the unique structure of financial data, including domain-specific language, long-range dependencies, temporal dynamics, and cross-modal interactions. Across diverse forecasting settings, we demonstrate that this specialization yields significant improvements in forecasting and investment performance.

Chapter 2

Long-Context Modeling of Earnings Calls

Input:	"...we continued our positive momentum in the first quarter reporting comp sales that accelerated from our strong fourth quarter performance. during the quarter, we drove market share gains and better than expected profitability by capitalizing on the advantages of our business model with dynamic marketing, compelling brands, and providing our customers with the preferred beauty shopping experience..."
Output:	Positive Surprise, $P(y = 1) = 0.95$

Figure 2.1: Paragraph of a sample transcript from the Validation Set that resulted in a positive surprise prediction from the model.

2.1 Introduction

There is a multitude of textual data relevant to the financial markets, spanning genres such as financial news, earnings conference calls, analyst recommendation reports, social media posts, and regulatory filings. Earnings conference calls are one of the most important datasets relevant to the information flow in equity markets because they reflect a direct communication between company executives and financial analysts [41].

Many public companies in the US hold earnings conference calls quarterly, in which their executives discuss the recent performance of the firm, their prospects, and answer questions from financial analysts covering their firms. Typically, companies report results quarterly at a lag of 4-6 weeks from the end of the previous period, and hold a conference call shortly thereafter. Therefore, the company executives have substantial knowledge into the next period's results when the call is held, providing a rare opportunity to detect textual indicators, such as tone and emotion in executive and analyst language patterns. These diverse signals, which can vary from clear sentiment to more subtle signs of deception [42] or obfuscation [43], may reveal important information about the current and future prospects of the company and be used to forecast its future earnings

surprises. Earnings surprises, which measure the operating performance of a company relative to market expectations, are highly followed by equity investors and often result in high magnitude stock returns and volatility [44].

In this chapter, we introduce the problem of learning for forecast company performance from their long-form documents, specifically using the textual content of earnings call transcripts to forecast future earnings surprises. It is important to note that this is a challenging task because the forecasting horizon is long (∼3 months), producing a lot of uncertainty between the forecast and event data, and there are legal restrictions about what is allowed to be disclosed to the public during the call. As a result, it is not clear a priori if the content of call transcripts contains sufficient task signal to outperform even uninformed baselines.

In the broader literature, there has been growing interest in the relevance of textual content to financial markets that has increasingly grown in sophistication in recent years. Some initial attempts used the Harvard General Inquirer IV-4 (HGI) dictionary to measure word polarity in financial text but found there to be domain mismatch [45]. Then, Loughran and Mcdonald [46] constructed their own financial-specific sentiment dictionary (LM). Interestingly, they found that 75% of words with strong negative polarity in the HGI dictionary had a neutral sentiment in a financial context (e.g. liability, tax, excess, etc.). Further, Ke et al. [47] develop a supervised learning method to identify sentiment in WSJ news articles, and found that over 40% of the most negative words identified by their model are not present in the LM dictionary because they are not negative in the context of regulatory filings. In other words, even within the financial domain, there can be a genre mismatch between different types of financial text. This motivated us in this work to explore models that are well attuned to the language of earnings conference calls.

However, the length of these conference calls, typically ranging from 5,000 to 10,000 words per transcript, creates some challenges because many of the popular pretrained

language models only support a maximum length of 512 tokens. While there have been many advances in developing Efficient Transformers that reduce the time complexity of the self-attention mechanism, these methods are still computationally expensive to pretrain and there does not yet exist a variety of pretrained versions, such as BERT [48] and RoBERTa [49], which both contain many variants targeting specific domains, such as biology [50], medicine [51], law [52], finance [53], and many others. As far as we know, there does not exist equivalent domain specific pretrained versions of these Efficient Transformers.

In this work, we explore a variety of approaches to this novel long document classification task and make the following contributions.

1. We introduce a novel task of using the text from earnings conference calls to make long-horizon predictions of future company earnings surprises (§2.3).

2. We explore a variety of approaches for this task, including simple bag-of-words models as well as long document Transformers, such as Efficient Transformers and tailored Hierarchical Transformer models, establishing the state of the art (§2.4).

3. We demonstrate that it is possible to predict companies' future earnings surprises with reasonable accuracy from solely the textual content of their most recent conference calls (§2.5).

4. We probe the best model through different interpretability methods to reveal some intuitive explanations of the linguistic features captured, which indicates that our model is learning more powerful features than just traditional sentiment (§2.6).

2.2 Related Work

2.2.1 Long Document Classification

There are generally two different approaches to modeling long documents. First, there is a class of Efficient Transformer models that were designed for long documents, such as Transformer-XL [54], Longformer [14], BigBird [55], and Reformer [16], which modify the self-attention mechanism in the original implementation to make it more efficient to accommodate longer contexts. These models have been shown to excel at long document understanding tasks, such as classification, question-answering, and summarization.

Alternatively, a hierarchical attention approach can be used. In Hierarchical Attention Networks [56], the authors model a document in a hierarchical fashion, viewing a sentence as a sequence of words and a document as a sequence of sentences, and use self-attention and recurrent networks to produce document representations. More recent work on Hierarchical Transformers (HTs) extend this approach to use pretrained language models for segment-level embeddings and Transformers for document-level representations [57–60]. This approach naturally supports longer sequence lengths due to the product of multiple self-attention mechanisms. Although Efficient Transformers are attractive in principle, there is much less availability of them pretrained in different domains and languages, such as mBERT [48] and XLM-RoBERTa [61].

2.2.2 Financial Prediction

Since earning conference calls are highly followed by most investors, there has been a considerable amount of research performed on them. For instance, Larcker and Zakolyukina [42] use data on subsequent financial restatements to analyze conference calls for deceptive behavior. Frankel et al. [62], apply traditional ML models to extract the

sentiment of conference calls and use it to predict subsequent analyst revisions. However, as far as we know, this is the first work that uses deep learning to directly learn to predict earnings surprises from conference call transcripts in an end-to-end manner.

In Qin and Yang [63], the authors propose a deep multi-modal regression model that jointly leverages textual and audio data from a small sample of conference calls to predict short-term stock volatility. Sawhney et al. [64, 65] build on that work to leverage the network structure of stock correlations with graph networks and financial features to make joint predictions. Similarly, Sang and Bao [66] propose a multi-modal structure that jointly models the textual dialogue in the call and the company network structure. Further, Yang et al. [53, 67] propose a multi-modal model that leverages the numerical content of financial text.

In Huang et al. [6], the authors release a pretrained language model for the financial domain, termed FinBERT, and show that their model understands financial text significantly better than competing methods in many aspects, including sentiment and ESG. They constructed the model by pretraining the BERT-base architecture from scratch with a custom vocabulary on a large corpus of English text from the financial domain, consisting of conference calls, corporate filings, and analyst recommendation reports, that is commensurate in size to the BERT pretraining corpus.

2.3 Problem

2.3.1 Data

We collected manually transcribed English conference calls from the largest publicly trade companies in the US (MSCI USA Index) over January 2004 to December 2011, from FactSet Document Distributor. We also source Reported Earnings per Share (EPS)

and Analyst Consensus Estimates of EPS from FactSet Fundamentals and Consensus Estimates, respectively. To focus on the largest and most actively traded companies in the US, we filter the data such that all companies in the sample have market capitalizations above $1B USD and daily average trading volume over $50M USD. Then, we temporally partition the data into train, validation, and test sets. We use transcripts that occurred between 2005 and 2009 as the training set, those that occurred in 2010 as the validation set, and those that occurred in 2011 as the test set. It is important to note that these sets must be temporally disjoint and monotonically ascending in time to avoid look-ahead bias. We provide summary statistics in Table 2.1.

	Train	**Validation**	**Test**
Start Date	Jan 2004	Jan 2010	Jan 2011
End Date	Dec 2009	Dec 2010	Dec 2011
Sample Size	4,056	524	588
Avg # of Words	9,016	8,907	9,010
Max # of Words	26,130	19,853	17,650
Avg # of Sentences	390	409	415
Avg # of Words per Sentence	25	22	22

Table 2.1: Summary Statistics of the Earnings Call Transcript dataset on each sample split.

2.3.2 Supervised Learning Task

We propose the task to predict the direction of the next earnings surprise ES from solely the textual content of the most recent transcript as a supervised learning task. Therefore, the input is a raw, unsegmented, English transcript with maximum number of

words L_T. We set L_T to be 12,000 words (∼20,000 BERT tokens) for computational and memory constraints, and because less than 10% of the transcripts in the sample are longer than that length. We select the Standardized Unexpected Earnings [SUE; 68] as our measure of earnings surprise, which is defined to be the difference between the reported EPS of the company and the analyst consensus estimate of the EPS, scaled by the inverse of the standard deviation (dispersion) of the analyst forecasts. We measure the analyst consensus estimate as the mean of all latest valid analyst forecasts, collected 1-month following the last earnings call transcript (the one we are using to make the prediction), which serves as the closest approximation to forward-looking market expectations; this allows analysts to update their forecasts based off their perception of the conference call and recently reported company results, and yields a more challenging, but potentially more rewarding, task than if we collected analyst forecasts at an earlier time horizon. We note that there is roughly a 3-month time horizon between the earnings call and the reporting of the next earnings surprise, making this long horizon prediction task particularly challenging.

$$ES = \frac{RepEPS - Avg(EstEPS)}{Std(EstEPS)}$$

$$y = \begin{cases} 0, & ES \leq -\delta \\ 1, & ES \geq \delta \end{cases}$$

We binarize the continuous value, such that an ES above δ corresponds to a label of +1 and represents a positive surprise, while an ES below $-\delta$ corresponds to a label of 0 and represents a negative surprise. We select a value for δ of 0.10 as a balance between the sample size and the significance of the events, such that about $\frac{1}{4}$ of events are positive surprises and $\frac{1}{4}$ of events are negative surprises. We discard transcripts which

do not result in a material earnings surprise. Since this does not translate to a perfect class label balance, we randomly down-sample the majority class, such that there is an equal 50/50 split of positive surprises and negative surprises in each sample split, to more easily interpret the results. Thus, we can use accuracy score as our primary evaluation metric. We use binary cross-entropy as the loss function for this binary classification task.

While the underlying earnings surprise metric is continuous, the importance of the metric to the market is often more binary. In general, market participants are more interested in the direction of the surprise than the precise magnitude of it and typically react accordingly insofar as the surprise is a "material" event. However, there are neutral cases when the company beats or misses the forecast by a small margin (i.e. [-0.10, 0.10] in this work) in which market reaction is typically lower. These boundary cases are generally difficult for the model to learn from at this long horizon because the ex-ante true probability is approximately random and there is often some form of earnings management involved. Therefore, we choose to focus on the most important earnings surprise events and disregard the neutral class.

2.4 Methods

2.4.1 Approach

We provide a wide variety of baseline models for this task, consisting of a combination of traditional and neural models, and establish several strong baselines as well as the state-of-the-art. While the current literature suggests that domain adaptation methods, such as language model finetuning on a domain-relevant corpus, is beneficial when using generic pretrained language models in out-of-distribution tasks, [69, 70], this additional

pretraining is computationally expensive and requires large-scale datasets to be effective. Therefore, we explore the use of existing pretrained language models.

2.4.2 Bag-of-Words

For the classical ML baselines, we select bag-of-words with n-grams (BOW) with TF-IDF weighting [71], and Logistic (Logistic) and Gradient Boosted Decision Trees (GBDT) as classifiers. We also provide a simple dictionary-based model that uses the proportion of words in each category of the Loughran and McDonald (LM) financial sentiment dictionary as features to a Logistic classifier. Additionally, we consider both general (BERT-Sent) and domain-specific (FinBERT-Sent) pretrained sentiment classifiers that are applied at the sentence level and aggregated with simple majority-rule voting. BERT-Sent is BERT-base-uncased [48] finetuned on IMDB movie reviews [72] and FinBERT-Sent is FinBERT [6] finetuned on manually labeled sentences from financial analyst research reports.

Given that the positive autocorrelation of earnings surprises is documented in the financial literature [73], we provide a simple autoregressive time-series baseline AR(1) that fits a logistic classifier on the continuous value of the firm's previous earnings surprise. In general, the autocorrelation beyond the most recent quarter is much lower. While the resulting performance is far below that of the best long-document models we provided, it is important to note that the signal contained in the text is likely largely distinct from and complementary to the information contained in the lagged surprise variables.

2.4.3 Short Context Models

For the short context models that only support up to 512 tokens of text, we consider multiple variants of FinBERT, including first, last, and random 512 tokens (truncation),

as well as various forms of aggregation, including mean & max pooling over time, to aggregate segment embeddings into document representations.

2.4.4 Hierarchical Transformers

We provide various forms of Hierarchical Transformers (HTs) with different pretrained segment encoders, and train them end-to-end on our supervised learning task. HTs take a hierarchical approach by dividing each long document into shorter non-overlapping segments of maximum length L. To do so, we use greedy sentence chunking to recursively add sentences to each segment until the length of the segment exceeds L. We choose this segmentation strategy to avoid breaking up and mixing sentences into chunks to better preserve their syntax and semantics. Since it is not clear what the optimal value of L should be, we treat it as an additional hyperparameter and tune it over $\{32, 64, 128\}$.

Segment Encoder

We initialize the segment encoder with pretrained models, which typically supports a maximum sequence length of 512 tokens. We explore BERT [48] and FinBERT [6]. This model produces contextualized embeddings of all tokens in each segment and we extract the last hidden state of the first [CLS] token as our segment representation [48].

Document Encoder

We use the standard Transformer architecture [74] with multi-head self-attention and sinusoidal positional encodings as the document encoder. The document encoder is responsible for taking the segment embeddings and producing contextualized segment representations by allowing each segment to attend to all other segments in the document and share information. We apply max pooling over time and concatenate with the first

state to arrive at our document representation. We tune the number of layers over $\{2, 3, 4\}$ and set the number of attention heads in each layer to be 6.

2.4.5 Efficient Transformers

We select BigBird [55] as our Efficient Transformer baseline model because it has been shown to exhibit state-of-the-art performance on long document classification and question-answering tasks [55]. The model applies a combination of local (sliding window), random, and global attention to sparsely approximate the full self-attention matrix. We also experimented with Longformer [14] but found BigBird to be more efficient and effective on this task, likely due to the increased number of global attention tokens and smaller attention window sizes. This result may suggest that the number of global attention tokens can be traded-off against the attention window size to improve efficiency and maintain effectiveness in Efficient Transformer models.

Since there are no versions of BigBird pretrained on the financial domain, we use the RoBERTa-base checkpoint that was warm-started from RoBERTa-base and further pretrained on a large corpus of long documents. We continue the MLM pretraining process on our in-task dataset to adapt to financial language [70]. We tune the block size over $\{32,\ 64,\ 84\}$ and the number of random blocks over $\{3,\ 4,\ 5\}$. Please see Appendix A for more details.

We also provide two simple, heuristic-based extraction baselines in which we first extract the most salient (a priori) sentences in each transcript, defined to be those that contain forward-looking statements (FLSE) according to Li [75] or positive/negative sentiment words (LMSE) according to the LM dictionary, and pass the resulting abridged text to BigBird, thereby reducing the text length by about 75% and 67%, respectively, and potentially avoiding the truncation of the most relevant sentences. However, we do

not find these extraction steps to be effective and discuss them further below.

2.4.6 Implementation Details

We train all models for a maximum of 10 epochs and select the checkpoint with the highest validation accuracy for further evaluation. BigBird and Hierarchical Transformers both contain approximately 130M parameters. We include more details on the implementation and training process in Appendix A.

2.5 Results

2.5.1 Comparison

Model	Test Accuracy
AR(1)	58.33%
BERT-Sent	49.58%
FinBERT-Sent	51.27%
LM + Logistic	60.20%
BOW + Logistic	67.68%
BOW + GBDT	**71.43%**
FinBERT – First 512	63.44%
FinBERT – Last 512	62.24%
FinBERT – Random 512	62.07%
FinBERT + Mean Pooling	66.84%
FinBERT + Max Pooling	**73.28%**
BigBird	**75.87%**
BigBird + MLM	74.30%
BigBird + FLSE	65.65%
BigBird + LMSE	72.98%
Hierarchical BERT	70.24%
Hierarchical FinBERT	**76.56%**

Table 2.2: Comparison of classifier performance on the test set across the neural and non-neural baselines. The best model in each class is indicated in bold.

As shown in Table 2.2, the pretrained sentiment models with simple majority-rule voting, BERT-Sent and FinBERT-Sent, fail to perform much better than random chance, indicating the difficulty of the task. Surprisingly, we observe that a simple bag-of-ngrams with TF-IDF weighting performs comparatively well on this dataset when used with GBDTs, outperforming many of the simple neural baselines. This indicates that substantial signal can be captured through non-linear interaction of normalized unigram/bigram features. It also indicates the models that try to reduce the length of text through either various forms of truncation or simple aggregation, likely dilute the signal and do not possess the ability to identify and capture the most salient portions of the transcript.

Further, we observe the importance of domain alignment within the Hierarchical Transformer models, with FinBERT performing significantly better than BERT for this task. Given earnings conference calls were a large component of the pretraining corpus, the benefit of FinBERT is expected. However, we do not find that domain adaptation of BigBird via further MLM pretraining to be beneficial in this setting, likely because of the small size of the training set.

Interestingly, we find that the addition of FLSE and LMSE is detrimental to the performance of BigBird, and that FLSE performs considerably worse than the best simple baselines, suggesting that the signal contained in the task requires the additional text to contextualize statements about forward-looking performance with information about the past and present, and supports the view that a model that can simultaneously process the full transcript in an end-to-end manner is required for strong performance on this task.

# Tokens	Test Accuracy
1,000	70.09%
2,000	72.64%
5,000	74.34%
10,000	75.72%
20,000	76.56%

Table 2.3: Comparison of model performance when truncating the text in the test set at various length cutoffs in terms of number of FinBERT tokens.

2.5.2 Document Length

In Table 2.3, we apply the trained Hierarchical FinBERT model to the test set with different truncation lengths. Our results also indicate the performance degradation when truncating long documents to shorter lengths and demonstrate the need for models to support the ability to process longer documents, as there is more than a 6% gap in test set accuracy between using the first 1,000 tokens and using the first 20,000 tokens, indicating that truncation loses valuable information contained in the middle of the transcript. While the first and last portions of the transcript may be the most relevant, it is clear that the middle portions also contain predictive value.

2.5.3 Training Efficiency

Model	Time
Hierarchical FinBERT	1.00
BigBird	1.40
Longformer	1.79

Table 2.4: Comparison of model finetuning times per epoch normalized to Hierarchical FinBERT.

In Table 2.4, we observe that the hierarchical structure of our model is quite efficient for processing long documents (20K tokens). Compared to BigBird and Longformer,

which can only process the first 4,096 tokens, we observe an approximately 50% speed up in finetuning time.

2.6 Model Interpretability and Analysis

Since Hierarchical FinBERT was the best performing model, we probe its predictions through a variety of interpretability methods to better understand the linguistic features important to this task.

2.6.1 LIME

First, we conduct Local Interpretable Model-agnostic Explanations (LIME) [76] as a word-level attribution test, which constructs a sparse linear model across the perturbed neighborhood of each sample to approximate the influence of the top 10 features (words) to the model's prediction. We sort by the words that have the highest feature importance summed over 100 random samples from test set (given length of the transcripts, performing this analysis on the full test set was not computationally feasible).

Top Words for Positive Surprise:	higher, strong, great, raising, beat, congrats, outperformance, benefitted, quarter, formidable, improvement
Top Words for Negative Surprise:	bad, challenging, negatives, impacted, caused, tough, unfavorable, because, capacity, offset

Figure 2.2: Top Words according to largest LIME weights summed over a random selection of 100 samples from the test set.

As shown in Figure 2.2, many of the phrases with the highest importance have a strong sentiment attached to them. For instance, "congratulations" and "great quarter"

are important features, which are used by financial analysts to praise the performance of the company. Since it appears that the top positive words are more intuitive and have larger magnitude weights, we conjecture that positive sentiment is more easily expressed, while negative sentiment may often manifest in what is not said. We also note that words such as "because" and "caused" may be used to try to explain away poor performance.

Category	**# words**	**% words**	**% sentences**	**coeff**	**p-value**
Positive	347	1.29	21.69	53.73	0.000
Negative	2345	0.65	11.99	-38.56	0.000
Uncertain	297	0.88	15.61	16.90	0.117
Litigious	903	0.13	2.50	14.59	0.284
Constraining	184	0.10	1.30	-31.97	0.050
Strong Modal	19	0.48	9.39	-9.51	0.106
Weak Modal	27	0.40	7.56	43.04	0.017

Table 2.5: Multivariate Linear Regression of model predictions onto LM financial sentiment variables on the test set. All standard errors are clustered by firm and year-month, and covariates include year-month fixed effects to account for intra-firm residual correlation. The p-value represents a two-sided significance test.

2.6.2 LM Sensitivity Analysis

To further understand model behavior, we perform another interpretability test using the LM financial dictionary and the predictions of Hierarchical FinBERT. We provide an overview of the summary statistics of the dictionary and results in Table 2.5.

To do so, we compute the proportion of total words in each transcript that belong to each LM category as financial sentiment variables. Then, we extract the model predictions ($P(y = 1)$) on the test set and regress them onto the financial sentiment variables. We observe that the model's predictions are positively associated with positive financial sentiment, and negatively associated with negative and constraining financial sentiment.

We also see smaller associations with strong modal (negative), litigious (positive), and uncertain (positive), but these are less statistically significant. We note that the

LM dictionary was created based on word meaning in a sample of firm regulatory filings, which are distinct in style from conference calls, so there may be some domain mismatch. While some of the variables are statistically significant at the 95% level, the linear model has an adjusted R^2 of 10.3% (without the fixed effects), indicating that the trained model is capturing more than just LM sentiment. We note the negative relationship with strong modal words, such as "must," "best," "clearly," which may be used in persuasive writing to convince the audience of a particular viewpoint, and we conjecture that this may be a potential sign of executives trying to control the market narrative towards their company. In fact, Loughran and Mcdonald [46] find that firms with higher proportions of strong modal words in their regulatory filings are more likely to report material weakness in their accounting controls. Conversely, there appears to be a positive relationship with weak modal and uncertain words, such as "may," "depends," "appears," which may be a reflection of executives's honest portrayal of their expectations about the future.

2.6.3 LM Sentence Masking

We also conduct a masking-based interpretability method in which we remove all sentences in the test set that contain at least one word from any LM financial sentiment category and perform inference on the masked test set. We report the model performance in Table 2.6. We observe that while the trained model is capturing the sentiment conveyed by each category of words, it is not overly reliant on any single category and appears to be multi-faceted and balanced in its ability to utilize multiple types of signals inherent in the transcripts. This indicates that the model is relatively robust to perturbations in the input space, suggesting that it may be less susceptible to manipulation (e.g. if executives try to avoid certain words or phrases they think the market would react negatively to) than keyword based approaches.

LM Category	Accuracy
None	76.56%
Positive	70.41%
Negative	72.11%
Uncertain	69.73%
Litigious	69.39%
Constraining	69.90%
Strong Modal	69.56%
Weak Modal	70.69%

Table 2.6: Best model performance on the test set after dropping all sentences containing at least 1 word in each respective LM financial sentiment category.

2.6.4 Forward-Looking Statements

Since a typical conference call contains information about the past, present, and future performance of the company, we wish to understand the importance of forward-looking content to the predictions. In particular, we define sentences that contain certain keywords, such as "will," "expect," "believe," etc., to be forward-looking statements (FLS) according to Li [75]. We then examine the relationship between the number of forward-looking statements in the text (NFLS) and the model performance in Figure 2.3. Further, we observe that model performance generally increases for larger values of NFLS. We conjecture that higher values of NFLS provide the model with more signal about the firm's future prospects.

2.6.5 Comparison

While the Hierarchical Transformer models are able to process almost the full transcript in an efficient manner, the best model (FinBERT) only outperforms BigBird by about 0.70 absolute percentage points, even though it is able to process more than 5 times the number of tokens. This result seems to indicate that the BigBird architecture, which applies the global attention simultaneously with local attention rather than in a

Figure 2.3: Breakdown of model performance on test set by tercile of number of sentences in the text that contain forward-looking statements.

hierarchical fashion, is more effective in this setting, perhaps because of the ability of the model to inject global context into the token-level representations. Therefore, we would expect BigBird to outperform the HTs if it could be extended to support longer sequence lengths and/or adapted to the financial domain. However, given that BigBird is already 50% slower than the HTs, the training time may become intractable without adjusting to significantly smaller block sizes, and we leave it to future research to identify the best approaches to efficiently extend Efficient Transformer models, such as BigBird, to support longer sequence lengths [77].

2.7 Conclusion

In conclusion, in this chapter we propose a novel task that uses transcripts from earnings conference calls to predict future earnings surprises. We formulate the problem as a long document classification task and explore a variety of different approaches to address it. While the length and language of the calls presents challenges for generic pretrained language models, we establish several strong baselines. We demonstrate that it is possible to predict companies' future earnings surprises with reasonable accuracy

from the solely the text of their earnings conference calls. Further, we probe the model through multiple interpretability methods to uncover intuitive linguistic features that go beyond traditional sentiment analysis.

Chapter 3

Cross-Document Modeling of Financial Reports

3.1 Introduction

Investors are faced with the consumption of a myriad of textual datasets relevant to financial markets, spanning genres such as news, social media posts, and financial reports. Public companies in the US are required to publish annual reports detailing the current operations of the firm, recent financial performance, and discussing future prospects. However, these reports contain over 25,000 words in length and large amounts of financial and legal jargon. As noted in Cohen et al. [78], this length and linguistic complexity have increased significantly over time as a result of increased government regulations and business complexity, making it difficult for investors to efficiently process the salient information contained in these reports.

Despite these challenging characteristics, financial reports do contain meaningful information about future company performance. For instance, Cohen et al. [78] show that large year-over-year changes to the language of company reports indicates a significant negative signal about their future performance and can predict financial variables, such as earnings, profitability, and bankruptcy. While their methods are shown to be effective, they only use simple string similarity measures to compare reports.

In a different application, given the detailed information about company business operations contained in these reports, there is an opportunity to identify relationships between companies that can help predict their future market correlation. Public companies are related to each other in various forms and this relationship governs the co-movement of their stock prices. Therefore, the ability to predict that relationship in advance from their reports is valuable to investment managers. These relationships can take various forms, including, having similar products, sharing technologies, or being exposed to the same economic risk factors. [79–81].

In this work, we explore these applications by curating a dataset of paired financial

Annual Report - 2014	Annual Report - 2015
Our operations and facilities are subject to extensive federal, state and local laws and regulations relating to the exploration for, and the development, production and transportation of, oil and natural gas, and operating safety...	Our operations and facilities are subject to extensive federal, state and local laws and regulations relating to the exploration for, and the development, production and transportation of, oil and natural gas, and operating safety...
Results of Operations Our oil and gas sales increased $35.5 million (9%) in 2013 to $420.3 million from $384.8 million in 2012. Oil sales in 2013 increased by $50.7 million (28%) from 2012 while our natural gas sales decreased by $15.2 million (8%) from 2012. The increase in oil sales was attributable to the 29% growth in oil production offset by a 1% decrease in our realized oil prices in 2013...	Depending upon future prices and our production volumes, our cash flows from our operating activities may not be sufficient to fund our capital expenditures, and we may need additional borrowings. ...If commodity prices remain low, we may also recognize further impairments of our producing oil and gas properties if the expected future cash flows from these properties becomes insufficient to recover their carrying value, and we may recognize additional impairments. **Results of Operations**

Figure 3.1: Comparison of a sample of passages from consecutive annual reports from the validation dataset of the Risk Prediction task that highlights the salient sentences that were added that potentially indicate an increase in future company risk.

reports and introducing two novels tasks that exploit the cross-document interaction between them to make long-horizon financial predictions. We experiment with a comprehensive set of end-to-end methods to model the interaction between these long financial documents. We find that it is possible to predict stock risk and pairwise correlation solely from text and that methods that allow for a more sensitive and fine-grained interaction between them provide significant benefit. In addition, we find that these text-based models provide considerable value beyond standard financial variables.

We provide a simple yet effective method that can compare arbitrarily long documents at a fine-grained level and identify subtle similarities and differences between them. We train this model end-to-end to allow the model to learn directly from the future financial

outcomes associated with each pair of reports, so it can learn to identify subtle, task-specific similarities and differences that are most predictive.

In summary, we make the following contributions:

1. We propose two novel and challenging financial prediction tasks, including forecasting future long-horizon stock risk and pairwise correlation, that both require the ability to recognize subtle similarities and differences between long financial documents (§3.3). To the best of our knowledge, this is the first work to consider and effectively model the cross-document interactions between paired reports for financial prediction in an end-to-end manner.

2. We systematically investigate and experiment with a comprehensive set of methods for these tasks, including tailored document-level and sentence-level Transformers that achieve strong performance, establishing the state-of-the-art (§3.5).

3. We demonstrate that while the tasks are challenging and many simple methods perform poorly, it possible to predict company risk and correlation with performance well-above random chance from solely the text of their financial reports by modeling the cross-document relationship at a fine-grained level with tailored pretraining objectives (Table 3.2).

4. We probe the best performing model through quantitative and qualitative interpretability methods to reveal insight into the underlying task signal (§3.7).

Broader Impact We hope this work will inspire future research in long document similarity and cross-document modeling with two challenging tasks, particularly as the context size for LLMs continues to grow.

3.2 Related Work

In the broader NLP literature, there has been great interest recently in extending the context length of Transformer-based language models to be able to efficiently process long documents [14, 16, 54, 55, 82]. These methods attempt to approximate full self-attention with more efficient computation and have been shown to excel at long document understanding tasks.

In a related area, long document similarity involves identifying the relationship between two long documents. While semantic similarity has been of interest for a while, most work has focused on short text at the sentence or paragraph-level [83]. However, semantic similarity at the document-level is more challenging because long documents often contain content spanning multiple topics and relationships between them may exist at different levels. Despite the difficulty, the problem has varied applications, including citation recommendation, plagiarism detection, coreference resolution, and multi-document summarization. In Zhou et al. [84], the authors propose a cross-document attention component into HAN [56] to enable the comparison between documents at different levels. Further, in Caciularu et al. [85], the authors consider a similar setting and propose a novel pretraining approach for Cross-Document Language Modeling (CDLM) with a dynamic attention mechanism that allows the model to learn cross-document relationships. They demonstrate that their model has a strong understanding of the relationship between documents and delivers SOTA performance on a variety of multi-document tasks. Other methods have attempted to perform an alignment between related documents at the sentence-level for retrieval applications, but typically pretrain encoders in a self-supervised manner, without finetuning them end-to-end on the target task. [86–88].

3.2.1 Financial Prediction

In addition to Cohen et al. [78] which inspired this work, there have been other works that examine using single firm reports for financial forecasting tasks, but primarily in isolation without any comparison to other related documents. For instance, Kogan et al. [89] extract textual features from the most recent financial report to predict stock volatility, while Koval et al. [36] directly learn to predict companies' future earnings surprise from the text of their conference call transcripts. Other work in this area has combined textual reports with multimodal data, such as audio, tabular, and financial features to enhance predictions [64, 90–92].

3.3 Problem Statement

We propose two novel tasks designed to evaluate the ability to recognize subtle similarities and differences between long financial documents that are predictive of long-horizon financial outcomes. It is important to note that since the reports occur at an annual frequency, we choose target variables at the 1-year horizon, which produces a lot of uncertainty between the forecast and outcome date, and makes these long-horizon prediction tasks particularly challenging. We also believe that the long-horizon requires the ability to capture more intricate, subtle signals than similar short-horizon tasks. In addition, this choice is consistent with prior work [89–91] and the premise from Cohen et al. [78] that the text-based signal contained in these reports is related to business risk that potentially can take up to multiple quarters to materialize on company performance.

3.3.1 Risk Prediction

Risk prediction is a valuable tool for investment managers when constructing a portfolio of financial assets. While there are many measures of financial risk, Maximum Drawdown (MDD) has become an important one, which measures the most significant percentage decline in the value of an asset over a given period of time [93–96]. Therefore, we choose this as a target variable and use consecutive financial reports as task inputs to learn to identify subtle yet important signals of company risk. Given the Management Discussion and Analysis (MDA) section of the current financial report $D_{i,t}$ for firm i at year t, we wish to learn to compare and contrast it with the previous report $D_{i,t-1}$ to predict whether the company will experience an abnormal decline (MDD) over the next year. While there may signal in only considering the current report, we believe it can be considerably enhanced when contextualized with the previous report to better capture the salient risks factors facing the company. Given daily price data $P_{i,t}$ for company i at time t, we compute the MDD over the next year T as the magnitude of the largest price decline from peak to trough [97]:

$$\text{MDD}_{i,t} = |\min_{t_1 \in \{t,T\}} \frac{P_{i,t_1}}{\max_{t_0 \in \{0,t_1\}} P_{i,t_0}} - 1|$$

For each year, we label companies with the 20% largest drawdowns during each year in the sample as High Risk ($y = 1$) and those in the bottom 80% as Normal Risk ($y = 0$):

$$y_{i,t} = \begin{cases} 0, & \text{Percentile}(\text{MDD}_{i,t}) < 0.80 \\ 1, & \text{Percentile}(\text{MDD}_{i,t}) \geq 0.80 \end{cases}$$

We carefully choose this task formulation and target variable for a few different reasons. First, it is common for investment managers and the financial literature to segment portfolios into quintiles, approximating a live trading setting in which investment managers are faced with the decision to exclude certain high-risk stocks from their portfolio at each decision point. Therefore, it has the potential to help them reduce portfolio risk. Second, we carefully selected Maximum Drawdown (MDD) as our target variable because of its ability to capture the effect of extreme events that occur anytime within the time horizon, since it computes the bottom of market prices attained over the horizon, and use the stock's MDD relative to other stocks within the same time period to indicate High Risk stocks to remove the impact of broad market movements and focus on stock-specific events.

Since the dataset is imbalanced by design and High Risk is a clear positive minority class, we use the F1-score as our primary measure of performance evaluation. We believe it accurately reflects the trade-off for an investment manager who is faced with the decision to include or exclude a stock in their portfolio because misclassifying a High Risk stock as Normal Risk is more costly than misclassifying a Normal Risk stock as High Risk.

3.3.2 Correlation Prediction

In addition to risk, the correlation matrix of stock returns is an equally important measure for the risk management practices of investment managers [98, 99]. Therefore, we also propose the task to predict the future correlation between companies' stock prices by learning to identify similarities and differences between their financial reports. We introduce this task to evaluate the ability of the model to capture various forms of relationships between companies.

For computational purposes, we take a subset of the 100 largest companies in our dataset and compute pairwise relationships to generate 4,950 company-company pairs per year. Further, we remove company pairs in which both companies belong to the same industry classification to challenge the model to identify more subtle connections that extend beyond industry keywords, leaving us with 3,836 pairs per year. We measure the relationship as the correlation between their daily stock returns over the next year from their most recent reports. To do so, we compute daily stock returns $r_{i,t}$ from daily prices $P_{i,t}$ for company i at time t and the correlation between their stock returns over the next year from t to T:

$$\text{corr}(r_{i,t}, r_{j,t}) = \sum_{t=1}^{T} \frac{(r_{i,t}-\bar{r_i})(r_{j,t}-\bar{r_j})}{\sqrt{(r_{i,t}-\bar{r_i})^2(r_{j,t}-\bar{r_j})^2}}$$

Then, we normalize them to be $N(0,1)$ within each year to account for the nonstationarity of market correlations over time:

$$y_{i,j,t} = \frac{\text{corr}(r_{i,t},r_{j,t})-\mu_t}{\sigma_t}$$

We use the Spearman Rank Correlation between the model predictions and observed correlations in each year to evaluate model performance. We use these metrics at the year-level rather than aggregated Mean-Squared Error because the relative ranking of the predictions within a given year is more important than their absolute levels given the non-stationarity of market correlations.

3.4 Data

3.4.1 Data Acquisition

To curate the dataset, we download preprocessed HTML files of company filings from the Notre Dame Software Repository for Accounting and Finance [3].

We focus our analysis on Section 7A: Management Discussion and Analysis (MDA) section from annual reports of US-based public companies. According to the SEC, this section is intended to provide management's perspective on the business results of the past year and their future prospects for the upcoming year, including information about key business risks. While there are other sections, we choose to focus on the MDA because it reflects a direct communication from company management to shareholders. We use a variety of regular expressions to extract the MDA section and filter the resulting section text for quality in a refined iterative process. We source stock price data from FactSet Prices & Returns API. Please see Appendix B for further details on the data curation process.

3.4.2 Data Statistics and Task Formulation

To prevent any form of lookahead bias, we temporally partition the dataset according to the report publication date into training (Jan 2010 – Dec 2014), validation (Jan 2015 – Dec 2015), and test (Jan 2016 – Dec 2019) splits. We do not use expanding sample windows for training/validation due to lack of computational resources, but we would expect doing so would improve results across model types and we validate this hypothesis with the best performing model.

We present an overview of the dataset with summary statistics in Table 3.1, including document length and linguistic complexity, measured with Gunning FOG Index [43]. We

confirm that the MDA section of these reports is becoming longer and more complex over time, likely making it increasingly difficult for investors to process the information contained in them.

	Train	**Validation**	**Test**
Start Date	Jan-2010	Jan-2015	Jan-2016
End Date	Dec-2014	Dec-2015	Dec-2019
# Samples	8,123	1,617	7,579
# Firms	2,574	1,572	2,170
# Words	13,092	13,455	14,354
# Sents	403	417	426
Linguistic Complexity	10.76	10.98	11.38

Table 3.1: Summary Statistics of each MDA section in the Financial Report on each sample split.

3.5 Methods

We explore a comprehensive set of baselines on these novel tasks that range from simple bag-of-words based methods to well-tailored state-of-the-art document-level and sentence-level Transformer-based models, including both generic and domain-adapted versions of each.

3.5.1 Simple Baselines

First, we establish a variety of simple baselines that indicate the difficulty of the task. **BOW + Sim + Linear** is solely based on the similarity between the reports using TF-IDF weighted, bag-of-words features while **BoW + Linear** concatenates their features together and passes them to a linear classifier. We also include a pretrained financial

sentiment classifier **FinBERT-Sent + Linear** [5] applied at the sentence-level [91]:

$$\text{FinBERT-Sent} = \frac{\#\text{PositiveSentences - }\#\text{NegativeSentences}}{\#\text{TotalSentences}}$$

The results of this baseline clearly distinguish the Risk task from traditional sentiment analysis.

Additionally, given that the positive auto-correlation of risk is well documented in the financial literature [100], we provide a simple autoregressive time-series baseline **AR(1) + Linear** that fits a linear classifier on the 1-year trailing value. While the resulting performance is below that of the best text-based models, it is important to note that the signal contained in the text is largely distinct from and complementary to it (corr < 0.20). Finally, we also include a purely company financial-based linear classifier **FinVar + Linear** with 10 common accounting and stock-price based financial variables (e.g. valuation, profitability, volatility, price momentum, etc.) to serve as a traditional financial baseline [91]. Please see Appendix B for more details on the variables used.

3.5.2 Document-Level Transformers

We consider two approaches to predicting the relationship between two long documents at the document-level, including the Bi-Encoder (BE) and the Cross-Encoder (CE).

Document Encoder

First, we select our primary document encoder to be the Longformer-base because it has been shown to excel at document matching [85]. The model applies a combination of local and global attention to efficiently approximate the full attention matrix.

For the Risk Prediction task, we provide single document baselines that only make use

of the current report D_t (**Longformer-Curr**) and previous report D_{t-1} (**Longformer-Prev**), respectively, as well as one that performs a soft "diff" operation between them, only extracting those not contained in the previous report (**Longformer-Diff**), to further justify the use of more sophisticated cross-document methods. We find that several variations of the "diff"-based approach perform worse than just using the current report, which we conjecture is for two reasons. First, the changes are subtle and difficult to identify using manual heuristics. Second, the salient sentences require the surrounding context to effectively contextualize the meaning.

Cross-Encoder (CE)

We also experiment with the Cross-Encoder approach (**Longformer-CDLM-CE**) of concatenating the document text together and use the CDLM-pretrained Longformer model from Caciularu et al. [85]. This approach implicitly interacts the tokens between the documents via the local/global attention mechanism, but the granularity of the interaction may be limited because attention is limited to a local window and special global tokens.

We follow the CDLM-framework and allocate global attention to the first `[CLS]` token and special document separator tokens `<doc-s>` and `</doc-s>`. We extend the maximum length of the model to 8192 tokens by copying over the position embeddings, and then concatenating the first 4096 tokens of each document together.

$$\text{CE}(D_i, D_j) = g([D_i; D_j])$$

Bi-Encoder (BE), Document-Level

Second, we experiment with encoding each document independently and then passing them through a 1-hidden layer MLP for interaction via concatenation of the document

embeddings, known as a Bi-Encoder approach (**Longformer-BE**). Consider the document encoder g and related documents D_i and D_j that are encoded as $g(D_i) = E_i$ and $g(D_j) = E_j$, respectively:

$$\text{BE}(E_i, E_j) = \text{MLP}([E_i; E_j; |E_i - E_j|)$$

This interaction function was inspired by Reimers and Gurevych [101] for sentence-level semantic similarity and we continue to include the absolute value difference term to impose the inductive bias that encourages the model to compare and contrast documents.

3.5.3 Sentence-Level Transformers

We also experiment with methods that operate on the sentence-level. Since the related documents have a different number of sentences in varying order, we explore a simple yet effective method to perform a soft-alignment between them.

Sentence Encoder

First, we divide each document into sentences and encode each sentence $s_i \in S_i$ and $s_j \in S_j$, using a pretrained sentence encoder f to get sentence embeddings $e_i \in E_j$ and $e_j \in E_j$, in each report, respectively. This model produces contextualized embeddings of all tokens and we extract the last hidden state of the first `[CLS]` token as the sentence representation [48].

Since the task requires the detection of subtle similarities and differences between topically similar text, it is important to have a sentence encoder that is well-attuned to semantic similarity and the financial domain. Therefore, we explore both pretrained encoders, such as **SBERT** [101] and **FinBERT** [6], as well as the **DiffCSE** [102] framework to pretrain a sentence encoder on our in-domain corpus. DiffCSE improves upon

the SimCSE [103] framework, which uses stochastic dropout-based augmentations as positive pairs and in-batch negatives with contrastive learning, by incorporating an additional Replaced Token Detection (RTD) loss that conditions upon the original sentence representation to predict the location of randomly replaced tokens that were generated by a fixed masked language model. This additional objective has been shown to make the encoder more sensitive to small yet important differences in sentences.

Cross-Document Sentence Alignment (CDSA)

The IR literature suggests that methods with token-level interactions provide a more fine-grained and powerful approach for query-document similarity tasks than those that operate at the document-level [84, 104]. With this in mind, we explore a simple yet effective extension of this approach to align and compare long financial reports at the sentence-level, which we denote as Cross-Document Sentence Alignment (**CDSA**).

To do so, we employ a cross-attention mechanism between the sentence embeddings of both documents to perform a soft-alignment, inspired by encoder-decoder attention [74, 105], which operates at a token-level. This mechanism creates a unique and corresponding context vector for each sentence in the focal report by attention weighting all sentences in the related report, and represents the portion of information of that sentence that is contained in the other report. We apply this in both directions, for each sentence embedding $e_i \in E_i$ across sentences embeddings E_j, and for each sentence $e_j \in E_j$ across sentence embeddings E_i:

$$c_i = \sum_{e_j \in E_j} \alpha_{i,j} e_j$$

$$c_j = \sum_{e_i \in E_i} \alpha_{j,i} e_i$$

where the attention weight α is given by softmax, dot-product attention [74].

To adapt the document-level Bi-Encoder approach BE to the sentence-level, we can compare each sentence embedding e_i, e_j with the corresponding soft-aligned context vector c_i, c_j from **CDSA** using a similar interaction function:

$$\text{BES}(e_i, c_i) = \text{MLP}([e_i; c_i; |e_i - c_i|])$$

Then, we conduct simple mean pooling over all sentence-level MLP outputs:

$$m(E_i) = \frac{1}{|E_i|} \sum_{e_i \in E_i} \text{BES}(e_i, c_i);$$
$$m(E_j) = \frac{1}{|E_j|} \sum_{e_j \in E_j} \text{BES}(e_j, c_j)$$

Finally, we concatenate the pooled outputs from both reports $m(E_i)$ and $m(E_j)$ and pass them through a classifier for prediction:

$$\hat{y} = \sigma([m(E_i); m(E_j)])$$

This mechanism allows for the detection of similarities and differences across each sentence in both reports.

3.5.4 Domain Adaptive Pretraining (DAPT)

Domain adaptation is important to the success of using pretrained language models for out-of-distribution text [69, 70]. Since we believe our tasks require a nuanced understanding of financial language, we conduct domain-adaptive pretraining (**DAPT**) for all of the baseline models. To do so, we aggregate a collection of 30K paired annual reports published between 2000 and 2009, prior to the start of the training data to prevent any form of data leakage, and create an in-domain pretraining corpus for all forms of DAPT

Model	# Params	Risk Prediction					Correlation Prediction				
		$\mathbf{F1_{2016}}$	$\mathbf{F1_{2017}}$	$\mathbf{F1_{2018}}$	$\mathbf{F1_{2019}}$	**Avg**	$\boldsymbol{\rho_{2016}}$	$\boldsymbol{\rho_{2017}}$	$\boldsymbol{\rho_{2018}}$	$\boldsymbol{\rho_{2019}}$	**Avg**
Minority Class All-1	0	0.33	0.33	0.33	0.33	0.33	-	-	-	-	-
BoW + Sim + Linear	2	0.36	0.35	0.35	0.34	0.35	0.10	0.16	0.07	0.06	0.10
BoW + Linear	100K	0.41	0.39	0.38	0.37	0.38	0.14	0.20	0.19	0.19	0.18
FinBERT-Sent + Linear	2	0.38	0.38	0.38	0.38	0.38	-	-	-	-	-
AR(1) + Linear	2	0.42	0.40	0.44	0.40	0.42	0.25	0.25	0.25	0.26	0.25
FinVar + Linear	11	0.45	0.43	0.48	0.45	0.45	-	-	-	-	-
Longformer-Prev	152M	0.42	0.43	0.40	0.35	0.40	-	-	-	-	-
Longformer-Curr	152M	0.48	0.47	0.47	0.45	0.47	-	-	-	-	-
Longformer-Diff	152M	0.44	0.43	0.43	0.39	0.43	-	-	-	-	-
Longformer-BE	152M	0.48	0.47	0.47	0.44	0.47	0.11	0.24	0.19	0.08	0.15
Longformer-CDLM-CE	152M	0.51	0.48	0.50	0.44	0.48	0.12	0.26	0.20	0.13	0.18
Longformer-BE + DAPT w/ MLM	152M	0.49	0.45	0.49	0.45	0.47	0.16	0.25	0.28	0.24	0.23
Longformer-BE + DAPT w/ DiffCSE	152M	0.52	0.48	0.49	0.46	0.49	0.26	0.34	0.29	0.24	0.28**
Longformer-CE + DAPT w/ CDLM	152M	0.53	0.49	0.50	0.46	0.50	0.22	0.30	0.31	0.24	0.27
CDSA-RoBERTa	128M	0.51	0.48	0.48	0.42	0.47	0.27	0.24	0.24	0.19	0.24
CDSA-SBERT	115M	0.54	0.52	0.51	0.44	0.50	0.28	0.31	0.27	0.18	0.26
CDSA-DiffCSE	128M	0.51	0.52	0.52	0.47	0.51	0.28	0.33	0.28	0.19	0.27
CDSA-FinBERT	128M	0.53	0.51	0.53	0.48	0.51*	0.30	0.32	0.25	0.17	0.26
CDSA-FinDiffCSE	128M	0.55	0.54	0.52	0.51	0.53*	0.30	0.33	0.32	0.27	0.31**
CDSA-FinDiffCSE + AR(1)	128M	0.58	0.56	0.57	0.53	0.56	0.37	0.45	0.46	0.33	0.40
CDSA-FinDiffCSE + Expanding	128M	0.55	0.55	0.59	0.57	0.56	0.30	0.40	0.36	0.31	0.34

Table 3.2: Main Results - Model performance on the test set of the Risk and Correlation Prediction task. All performance numbers are reported in decimal and the top 2 models within each task are bolded. "-BE" indicates Bi-Encoder while "-CE" indicates Cross-Encoder document-level models as defined in §3.5. "+ Expanding" indicates that expanding training/validation sample windows was used. "+ AR(1)" indicates that a linear combination of the predictions was fit between the CDSA-FinDiffCSE and AR(1) model. *, ** indicates the performance of the best model is statistically better ($p < 0.01$) than that of the second best model according the Wilcoxon Signed-Rank Test.

in this work for fair comparison across model types.

Document-Level

For the document-level models with a Longformer backbone, we conduct DAPT across the following different pretraining objectives: long context MLM [14] denoted as **Longformer-BE + DAPT w/ MLM**, CDLM [85] with pairs of consecutive reports (**Longformer-CE + DAPT w/ CDLM**); and follow the same pretraining settings and hyperparameters as Beltagy et al. [14] and Caciularu et al. [85], respectively.

We also adapt the DiffCSE pretraining framework designed for short-context models,

to the Longformer backbone model (**Longformer-CE + DAPT w/ DiffCSE**) for more sensitive document representations by prepending and assigning global attention to the original document embedding in the RTD objective to encourage the model to use that information to predict the replaced tokens.

Sentence-Level

We also use this corpus for pretraining a more domain-adapted and sensitive sentence encoder from the RoBERTa checkpoint using the DiffCSE framework (**CDSA-FinDiffCSE**) but limit the size to 10M sentences for computational purposes, and use the same pretraining settings and hyperparameters in Chuang et al. [102]. We expect this pretraining step to be able to better differentiate topically similar yet semantically different financial language.

Finally, since the validation data (2015) and last year of the test data (2019) are 4 years apart, we experiment with an expanding window training/validation approach (**CDSA-FinDiffCSE + Expanding**) to allow the model to access more recent data and simulate a production trading environment. However, we only do this for the best performing model because it is not computationally feasible to do for all models. We also include a simple multi-modal approach (**CDSA-FinDiffCSE + AR(1)**) that fits a linear combination between the predictions of the CDSA-FinDiffCSE and AR(1) models. Please see Appendix B for further details.

Implementation Details

Finally, we train all of these baseline models on each financial prediction task with binary cross-entropy loss and mean-squared error for the Risk and Correlation prediction tasks, respectively. For fair comparison across model types, we only consider the first

4096 tokens in each report; see Appendix B for further implementation details.

3.6 Experimental Results and Analysis

The results in Table 3.2 highlight the challenging nature of both tasks, but **we find broad consistency in the relative performance results across them**, with CDSA-FinDiffCSE performing the best in both with statistical significance, and improving considerably from expanding training data. This result provides evidence that while the tasks are distinct, they both require the ability to recognize subtle similarities and differences between long documents at a fine-grained level, and this ability is directly correlated with the relative ranking of model performance.

In general, **we find that the sentence-level methods generally perform better than the document-level methods**, which we conjecture is because by they allow for a more fine-grained interaction between the document sentences before any document-level pooling. We find this effect to be more pronounced on the Correlation Prediction task, especially when the Longformer base model is not pretrained for semantic similarity. This suggests that despite the extensive, language modeling-based pretraining process of the Longformer model, it does not produce strong document embeddings without finetuning.

However, we find that our long context adaptation of the DiffCSE pretraining framework for the Longformer is well-suited for generating fine-grained document embeddings, suggesting that this is a promising direction for future work. Relatedly, **we find that pretrained models not adapted to the financial domain or pretrained with semantic similarity objectives struggle to learn the subtle task signals**. However, we observe a significant improvement across most models after DAPT, suggesting that the task requires a nuanced understanding of financial language. For both tasks, we find

that a simple multi-modal model **CDSA-FinDiffCSE + AR(1)** improves performance, particularly for the Correlation Prediction task which exhibits stronger autocorrelation. We conjecture their complementary nature is partly due to the fact that historical market patterns captures the persistence of past behavior while the text-based models identify the catalyst that causes novel behavior, suggesting the text-based methods could serve as a valuable tool to augment traditional risk management practices. However, we leave it to future work to explore more sophisticated methods to incorporate tabular data into text-based models.

3.7 Model Interpretability and Analysis

3.7.1 LM Sensitivity Analysis

To further understand model behavior on the Risk Prediction task, we perform a simple interpretability test using the LM financial dictionary [3] and the predictions of the best performing model (CDSA-FinDiffCSE). We provide an overview of the summary statistics of the dictionary and results in Table 3.3. To do so, we extract the model predicted probabilities, and regress them onto the changes in the proportion of LM dictionary words between the current and previous report to understand their linear relationship.

In Table 3.3, we observe that the model's predictions for High Risk are negatively associated with increases in positive financial sentiment, and positively associated with increases in negative, constraining, and litigious financial sentiment. While some variables are statistically significant and the results are economically intuitive, the linear model has an adjusted R^2 of just 3.4%, indicating that the trained model is capturing more powerful features than only simple changes in LM sentiment. We also note the positive correlation

Category	# words	% words	% sentences	coeff	p-value
Δ Positive	347	0.55	13.59	-4.06	0.04
Δ Negative	2345	1.32	24.92	4.12	0.04
Δ Uncertain	297	1.36	30.34	4.56	0.13
Δ Litigious	903	0.59	13.10	1.05	0.18
Δ Constraining	184	0.57	14.23	6.12	0.05
Δ Strong Modal	19	0.23	6.53	11.46	0.11
Δ Weak Modal	27	0.59	15.20	0.58	0.91

Table 3.3: Linear Regression of the model predictions onto the YoY changes in LM financial sentiment variables.

with increases in strong modal words is consistent with Loughran and Mcdonald [46], who find that firms with higher proportions of strong modal words in their quarterly reports are more likely to subsequently report material weakness in their accounting controls, which is likely a strong signal for increases in the likelihood of future High Risk behavior.

3.7.2 Case Study and Qualitative Analysis

We conduct a case study of the reports of Comstock Resources Inc, referenced (CRK) in Figure 3.1. We find that the report scores highly as High Risk by the best performing CDSA model and correctly identifies the salient risky sentences, as measured via the largest L_2 norms in the $|s_i - c_i|$ term, which we highlighted in the exhibit. As shown in Figure 3.2, the company stock price experienced a precipitous drawdown in the 6 months following the release of this report. We find that the model was able to detect subtle yet important changes in the text that predicted a large drawdown months before it occurred.

3.8 Conclusion

We curate a large-scale corpus of paired annual financial reports and introduce two novel benchmarks that require modeling complex, cross-document interactions between

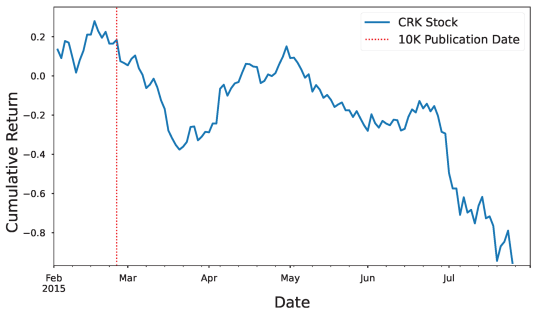

Figure 3.2: Stock Price of CRK following the publication of the 2015 Annual Report that was identified by the best performing model as High Risk based off changes from the 2014 Annual Report.

long documents. We methodically investigate a comprehensive set of methods that are well-attuned to the task, establishing the state of the art. Through analysis of the experimental results and use of interpretability methods, we reveal insights into the underlying task signals. We hope our contributions inspire further research in this important area.

Chapter 4

Multimodal Modeling of Paired Text and Time Series

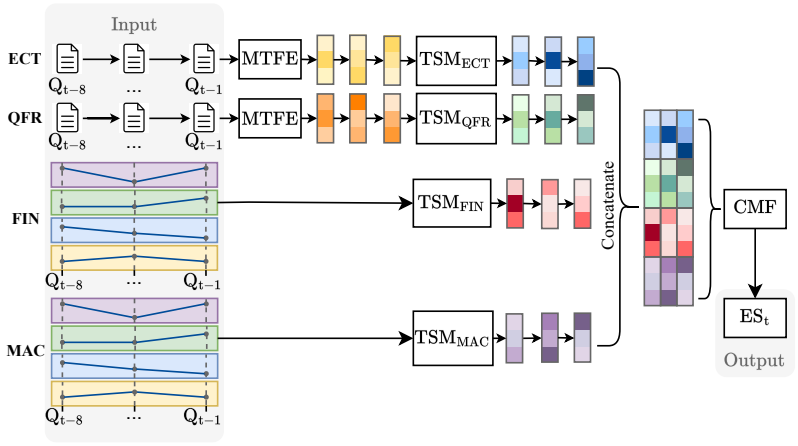

Figure 4.1: Overview of our multimodal time series prediction task and proposed multi-stage, modality-adaptive encoder fusion method. The inputs consist of the following time series: **AR** = autoregressive earnings surprise lags, **FIN** = tabular financial variables, **ECT** = earnings call transcripts, **QFR** = quarterly financial reports, and **MAC** = macroeconomic variables. These inputs are processed by the following model components: **TSM** = time series encoder, **MTFE** = multi-stage textual feature extraction, **CMF** = cross-modality fusion. The output is: **ES** = predicted Earnings Surprise (§4.3.1).

4.1 Introduction

Investors are faced with the consumption of a myriad of diverse datasets relevant to public companies, spanning modalities, genres, and sources. In general, this data is released on a quarterly basis, and includes accounting statements, financial reports, and earnings calls that comprehensively detail the current operations of the firm, recent financial performance, and discuss future business prospects and risks. While the accounting statements typically receive the most attention from investors, the textual content provided in earnings conference calls and financial reports is of equal importance as it reflects a direct communication between company executives and shareholders [41]. This textual content provides two important qualitative sources of information relative to the financial metrics. First, it provides management context about how to interpret the current and historical financial results. Second, it also allows management to express

their views about the future prospects of the company, providing the opportunity to capture the tone and sentiment from the most well-informed stakeholders. For instance, during periods of macroeconomic shocks, such as COVID-19, financial variables tend to capture backward-looking performance that may not apply in a new economic regime, while the textual commentary continues to provide forward-looking content (§4.6.2).

However, most work in financial prediction focuses on the most recently reported financial results and textual documents in isolation, without consideration of the temporal context of their historical patterns. While the most recent data point may be the most important, the historical context provides the ability to contextualize the current value and measure trends over time that help better predict future performance, as evidenced in both financial metrics and executive language patterns [78, 106, 107].

While financial analysts consume this data to make their quarterly earnings forecasts, it is difficult to quantitatively reconcile this complex, diverse information across sources and modalities to arrive at accurate estimates. While financial analyst estimates of company earnings are generally regarded as market expectations, they have been shown to exhibit predictable biases that can be exploited by machine-based methods [108, 109]. For instance, financial analysts often do not fully incorporate the subtle signals in macroeconomic shocks [110] or long financial documents [36, 62] into their earnings estimates. Therefore, we hypothesize that it is possible to use machine learning to effectively analyze this multimodal data and learn complex interactions between disparate sources of information that can be used to forecast earnings surprises months in advance of the report date.

In this work, we consider a multimodal time series forecasting problem in which we investigate the value of financial text, time series, and their interaction in predicting next quarter's company performance relative to market expectations. In doing so, we introduce a new multimodal time series dataset and challenging financial prediction task,

and propose a novel method that learns rich temporal and cross-channel features from the numerical and textual content of noisy financial time series. In summary, we make the following contributions:

1. We design a novel multimodal time series prediction task that spans different tabular and textual financial time series from diverse sources (§4.4, §4.3, Appendix C).

2. We propose a multi-stage training process to effectively extract predictive long context embeddings from long financial documents and demonstrate that these text-based features add significant value to traditional financial variables in forecasting future company performance, particularly during periods of economic shocks, which we attribute to their ability to capture forward-looking content and contextualize historical behavior (§4.5.2, Table 4.2).

3. We systematically explore the value of each modality and temporal context across financial time series and propose a simple yet effective method that allows modality-specific temporal dynamics while still capturing complex cross-modal patterns that outperforms existing methods on this challenging task (§4.5).

4. We probe our proposed method through quantitative and qualitative interpretability methods to reveal insights into the value of each modality and our proposed multimodal method.

5. We demonstrate the economic value of our model predictions in a real-world trading setting with portfolio simulations that result in economically and statistically significant gains in investment performance (§4.6.6).

Broader Impact We hope this work will inspire future research in multimodal time series modeling with textual and tabular data from different sources, particularly as the

context length and multimodal capabilities of LLMs continues to grow, as our findings suggest that small yet specialized finetuned models currently outperform LLMs on this task.

4.2 Related Work

4.2.1 Text Embeddings

In the broader NLP literature, there has been great interest recently in extending the context length of Transformer-based language models [14, 54, 55, 82, 111]. However, most of these methods are not well suited for producing semantic document embeddings that either perform well for zero-shot feature extraction or provide a strong parameter initialization for downstream finetuning. While there are many pretrained models and training methods proposed for semantic text embeddings, such as SBERT [101], SimCSE [103], and DiffCSE [102], most of them have a short context length intended for sentence or paragraph-level tasks. Recently, Wang et al. [112] demonstrate strong performance with a contrastive method for finetuning long context LLMs across diverse tasks with instructions. However, they mostly evaluate their model on short text tasks with the exception of synthetic retrieval.

4.2.2 Financial Prediction

There has been an extensive set of company financial variables derived from their accounting statements and stock market behavior that have been found to predict future firm performance [31, 34, 113–115]. Further, it has also been found that the text of company financial documents, such as earnings calls and financial reports, contains signal that is predictive of future company performance [36, 42, 46, 78, 89], but most work has

focused primarily on the most recent document in isolation without context of related documents. However, Koval et al. [37] found benefit in aligning consecutive financial reports to identify the most salient business risks. Other work has combined the textual reports with multimodal data, such as audio, graphs, videos, and tabular features to enhance predictions [63, 64, 90–92, 116–118]. Similarly, related work has found benefit in modeling cross-channel information from related time series across different sources, such as Google Trends & Weather, for predicting key financial indicators [119, 120]. However, our focus in this work is on jointly modeling the time series of paired textual financial documents and their corresponding tabular financial variables from the same firm; we do not focus on the cross-firm relationships [117].

4.3 Problem Statement

4.3.1 Task Formulation

We propose a multimodal financial time series task in which we predict a company's next quarter's earnings surprise (ES) from the time series of their financial, textual, and macroeconomic data. We believe these modalities and sources capture the information available to financial analysts when making their earnings forecasts, so our experiments present a unique exercise between the forecasting ability of human experts and machines. We select the Standardized Unexpected Earnings [68] as our measure of earnings surprise, which represent the operating performance of a company relative to market expectations and are highly followed by equity investors [44]. It is important to note that there is roughly a 3-month time horizon between prediction date and the report date, making

this long horizon prediction task particularly challenging.

$$\mathrm{ES_t} = \frac{\mathrm{RepEPS_t} - \mathrm{Avg(EstEPS_t)}}{\mathrm{Std(EstEPS_t)}}$$

Since the measure is a continuous variable with approximate normal distribution, we use Mean Squared Error (MSE) as the loss function and to evaluate performance.

Multimodal Inputs We use data from a variety of modalities and sources, described in detail below, to understand the current state of a company from different perspectives. For all variables, we use quarterly data aligned with each company's reporting period and the values of the previous 8 quarters as the input time series. In this work, we refer to data source and modality interchangeably.

4.3.2 Autoregressive Variables (AR)

The persistence of earnings surprises is well documented in the financial literature [73, 121] with streaks found to persist up to 12 quarters, so we use the historical values of the target variable **AR** as an input.

$$\mathrm{AR_t} = [\mathrm{AR}_{t-8}, ..., \mathrm{AR}_{t-1}] \in \mathbb{R}^{1\times 8}$$

4.3.3 Financial Variables (FIN)

Additionally, we include the time series of 15 well-documented firm-level financial characteristics **FIN** from the asset pricing literature [78, 91, 122, 123] derived from a company's accounting statements and stock price behavior, detailed in Appendix C, including valuation, profitability, growth, volatility, and momentum. We compute these variables on a quarterly basis and normalize them to have zero mean and unit variance.

While this set is not exhaustive, recent work [123] has showed that it largely spans the principal components of the full set of firm-level financial characteristics.

$$\mathrm{FIN_t} = [\mathrm{FIN}_{t-8}, ..., \mathrm{FIN}_{t-1}] \in \mathbb{R}^{15\times 8}$$

4.3.4 Earnings Conference Calls (ECT)

We include the text of quarterly earnings conference calls in which company executives discuss the recent performance of the firm, their prospects, and answer questions from financial analysts covering their firms. These calls provide a rare opportunity to detect diverse signals, which can vary from clear sentiment to more subtle signs of deception [42] or obfuscation [43], that may reveal important information about the current and future prospects of the company.

$$\mathrm{ECT_t} = [\mathrm{ECT}_{t-8}, ..., \mathrm{ECT}_{t-1}]$$

4.3.5 Quarterly Financial Reports (QFR)

We include the text of the Management Discussion and Analysis (MDA) section from the quarterly reports (10Q/10K) of US-based public companies. This section is intended to provide management's perspective on the business results of the past year and their future prospects for the upcoming year, including information about key business risks. [1] While there are other sections, we choose to focus on the MDA because it reflects a direct communication from company management to its shareholders.

$$\mathrm{QFR_t} = [\mathrm{QFR}_{t-8}, ..., \mathrm{QFR}_{t-1}]$$

[1] `https://www.sec.gov/files/reada10k.pdf`

4.3.6 Macroeconomic State (MAC)

Finally, we include the time series of macroeconomic data (**MAC**) that represents the broader state of the US economy at each point in time to capture exogenous demand and supply-side shocks with varying impact per industry. These monthly indices include the following variables: Industrial Production, Inflation, Consumer Sentiment, 3M Treasury Yield, 10YR Treasury Yield, Term Spread, Default Spread, and Oil Prices. Following Ball and Ghysels [110], we compute quarterly growth rates for each variable to arrive at our macroeconomic time series of 8 channels.

$$\mathrm{MAC_t} = [\mathrm{MAC}_{t-8}, ..., \mathrm{MAC}_{t-1}] \in \mathbb{R}^{6\times 8}$$

We merge this time series to align with each company's prior fiscal quarter end date.

4.4 Data

	Train	**Validation**	**Test**
Start Date	Jan-2010	Jan-2016	Jan-2017
End Date	Dec-2015	Dec-2016	Dec-2020
# Samples	7,188	1,362	5,723
# Firms	710	638	708
# Modalities	4	4	4
# Time Steps	8	8	8

Table 4.1: Summary Statistics on each sample split.

Data Acquisition and Curation We source Reported Earnings per Share (EPS) and Analyst Consensus Estimates of EPS from FactSet Fundamentals and Consensus Estimates, respectively, to compute the Earnings Surprise target variable. We collect

English conference calls from FactSet Document Distributor. We source quarterly and annual reports from Notre Dame Software Repository for Accounting and Finance. We collect the monthly macroeconomic indices from the ST. Louis FED. Please see §C.2 for further details on the extensive data curation process.

Data Statistics and Task Formulation We focus our analysis on the largest publicly trade companies in the US (MSCI USA Index) and require that each sample point possess valid data for each input variable across historical time steps. Then, we temporally partition the data into train (2010-2015), validation (2016), and test (2017-2020) sets. We provide summary statistics in Table 4.1.

4.5 Methods

We systematically explore a comprehensive set of unimodal baselines and propose a novel multimodal method that significantly outperforms the best unimodal models.

4.5.1 Multivariate Time Series

Since each of our modalities constitutes a multivariate time series with complex cross-channel interactions, we consider Time Series Mixer [124] as our backbone time series encoder because of its strong performance on multichannel time series forecasting. TSM is a simple yet effective all-MLP neural architecture that performs alternating forms of feature mixing across the time (TM) and channel dimension (CM) in each layer of the network. We denote the time series model as **TSM**, with input $X \in \mathbb{R}^{T \times C}$, T time steps, C channels, and returns features $H \in \mathbb{R}^{T \times C}$ after interaction across the temporal and channel dimensions, expressed below:

$$\mathrm{TM}_{\mathrm{l}}(X) = \mathrm{BN}(X + \sigma(W_{\mathrm{TM}} X_{*,t} + b_{\mathrm{TM}}))$$

$$\mathrm{CM}_{\mathrm{l}}(X) = \mathrm{BN}(X + \sigma(W_{\mathrm{CM}} X_{c,*} + b_{\mathrm{CM}}))$$

$$\mathrm{TSM}_{l}(X) = \mathrm{CM}_{l}(\mathrm{TM}_{\mathrm{l}}(X))$$

We perform ablations of this choice in §4.6.5. In each modality time series, we use the last $T = 8$ quarters of values as inputs. We conduct a grid search over the model architecture hyperparameters detailed in §C.6.

4.5.2 Multi-Stage Textual Feature Extraction (MTFE)

Since ECTs and QFRs are long textual documents (5K+ tokens each) and thus cannot be concatenated over the time series (50K+ tokens) and trained end-to-end with limited computational resources, we propose a novel multi-stage training method to extract predictive features from them. It is important to note that most of the text embedding literature [125] has focused on short-context texts at the sentence or paragraph-level, such as consumer reviews, which contains distinct characteristics from our problem setting.

We initialize our text encoder with the pretrained BigBird [55] checkpoint due to its long context length and strong performance in long document understanding tasks. This process consists of the following steps.

Multitask Domain Adaptation: Domain adaptation is important to the success of using pretrained language models for domain-specific text [69, 70]. Since we believe our tasks require a specialized understanding of financial language, we conduct multi-task domain-adaptive pretraining (**MT-DA**).

To do so, we adapt the BigBird-base model [55] to the financial domain by jointly performing long context masked language modeling (**LC-MLM**) and long-context Dif-

fCSE (**LC-DiffCSE**), in which we adapt the contrastive pretraining framework DiffCSE [102] designed for short-context models to long-context BigBird. We do this by prepending and assigning global attention to the original document embedding in the replaced token detection objective to encourage the model to use that information to predict the replaced tokens, resulting in more fine-grained document representations. We conduct this pretraining of over a pooled corpus of in-domain ECTs and QFRs that occur during the training date period for a total of 25K training steps.

This multi-task pretraining process serves several purposes. Firstly, **LC-MLM** adapts the model to the complex language of the financial domain. Secondly, **LC-DiffCSE** learns to produce aggregated document representations that are sensitive to topically similar but semantically different financial text. We believe both of these steps are critical towards capturing strong document representations and we demonstrate their value in Table 4.3.

Supervised Finetuning (SFT): We finetune the adapted model on the text of the last time step ECT and QFR (single document) and the corresponding earnings surprise measure to learn task-specific features.

$$\hat{Y_t} = \text{ENC}_{\text{ECT}}(\text{ECT}_t), \text{ENC}_{\text{QFR}}(\text{QFR}_t)$$

Feature Extraction: We extract features from the last layer embeddings E of the [CLS] token from the finetuned encoder ENC for each document.

$$E_{\text{ECT}} = \text{ENC}_{\text{ECT}}(\text{ECT}); E_{\text{QFR}} = \text{ENC}_{\text{QFR}}(\text{QFR})$$

These features contain richer information than solely the output predictions, allowing the time series model to capture multifaceted trends in executive language patterns over

time. We compare our approach with existing pretrained embedding models in §4.6.3.

4.5.3 Multimodal Fusion Methods

We explore three multimodal fusion methods to mix the time series across modalities with varying channel dimensionality, using TSM as the time series encoder.

There are two paradigms of approaches in multivariate time series forecasting and we explore them both here as baselines. Firstly, Channel-Mixing [124], in which all univariate channels are concatenated together and treated as a single multi-channel signal, assumes that there exists cross-channel information. In this case, since each input modality contains a different number of dimensions, we project them all into the same vector dimension and space before concatenating. We label this approach cross-modality mixer **CMM**.

Alternatively, Channel-Independence assumes that the relationship between each univariate time series is independent and should be processed separately but with shared model weights across channels [126]. This approach has found success particularly when the cross-channel signal is weak, and thus cross-channel interaction leads to overfitting. We apply this approach with the original model (**PatchTST**) and an architecture-controlled version (TSMixer), labeled modality independent, channel independent (**MICI**).

However, we believe that these two baseline approaches are not well suited to this multimodal problem for two reasons. Firstly, we hypothesize that each modality contains rich cross-channel patterns that Channel-Independence cannot capture. Secondly, we believe that differences in the temporal dynamics of each modality could lead to statistical noise and incompatibilities if relying on a single model, and that mixing modalities too early in the feature learning process is likely to lead to overfitting.

Therefore, we propose an different approach that imposes this inductive bias and

treats each modality (**AR+FIN**, **ECT**, **QFR**, **MAC**) as a separate multi-channel time series. In our modality-independent, channel mixing method (**MICM**), we first process each modality multi-channel time series independently using modality-adaptive time series models and then linearly project them into the same space and dimension.

$$X_{\text{AR+FIN}} = W_{\text{AR+FIN}}\text{TSM}_{\text{AR+FIN}}([\text{AR}; \text{FIN}])$$

$$X_{\text{ECT}} = W_{\text{ECT}}\text{TSM}_{\text{ECT}}(\text{E}_{\text{ECT}})$$

$$X_{\text{QFR}} = W_{\text{QFR}}\text{TSM}_{\text{QFR}}(\text{E}_{\text{QFR}})$$

$$X_{\text{MAC}} = W_{\text{MAC}}\text{TSM}_{\text{MAC}}(\text{MAC})$$

This approach allows the the modality-specific models to learn cross-time and cross-channel patterns that are adaptive to each modality and produce features that can be fused across modalities at a later stage. We concatenate the resulting mixed features together and introduce a cross-modality fusion (**CMF**) module to interact cross-modality channels based upon temporal similarity. This module consists of a cross-channel multihead attention mechanism (MHA) and layer normalization (LN) with residual connections.

$$X_{MM} = [X_{\text{AR+FIN}}; X_{\text{ECT}}; X_{\text{QFR}}; X_{\text{MAC}}]$$

$$X_O = \text{LN}(X_{MM} + \text{MHA}_{\text{c},:}(X_{MM}, X_{MM}, X_{MM}))$$

Again, this design imposes the inductive bias that while cross-modality channel patterns exist, they should be carefully interacted late in the network base upon temporal similarity in a learned projection space. We flatten the output and include a 2-layer MLP head for prediction:

$$\hat{Y}_t = \text{MLP}(\text{Flatten}(X_O))$$

Input	Modality	Time Steps	MSE_{2017}	MSE_{2018}	MSE_{2019}	MSE_{2020}	MSE
AR	TABULAR	1	1.33	1.52	1.60	2.50	1.77
AR	TABULAR	8	1.30	1.50	1.58	2.45	1.71
FIN	TABULAR	1	1.33	1.51	1.60	2.47	1.77
FIN	TABULAR	8	1.29	1.48	1.56	2.44	1.72
ECT	TEXT	1	1.19	1.46	1.47	2.35	1.62
ECT	TEXT	8	1.18	1.46	1.47	2.29	1.60
QFR	TEXT	1	1.27	1.46	1.64	2.29	1.68
QFR	TEXT	8	1.27	1.44	1.52	2.18	1.62
AR + FIN	TABULAR	8	1.22	1.38	1.42	2.29	1.60
AR + FIN + ECT*	MULTIMODAL	8	1.17	1.34	1.35	2.04	1.48*
AR + FIN + ECT + QFR**	MULTIMODAL	8	1.14	1.30	1.31	1.96	1.44**
AR + FIN + ECT + QFR + MAC	MULTIMODAL	8	**1.13**	**1.28**	**1.30**	**1.92**	**1.42**

Table 4.2: Main Results (smaller is better, **best in bold**): Model performance on the test set of our multimodal Earnings Surprise Prediction task. All results use our proposed multimodal method, including modality-specific **TSMixer** encoders, **MTFE** to extract textual features, and cross-modality fusion **CMF** to mix modalities. "Time Steps" indicate how many quarters of data are used in the time series model. **AR** = autoregressive earnings surprise lags, **FIN** = tabular financial variables, **ECT** = text features from earnings call transcripts, **QFR** = text features from quarterly financial reports, and **MAC** = Macroeconomic indicator variables. *, ** indicates the performance of the specified model is statistically better ($p < 0.05$) than that of the next best performing model on the test set according to the Wilcoxon Signed-Rank Test.

We conduct a grid search over the TSM model architecture over each modality in isolation (unimodal), detailed in Appendix C, allowing each modality-specific time series model to have a different representational capacity to adapt to varying complexities of temporal and cross-channel patterns.

4.6 Experimental Results and Analysis

4.6.1 Value of Multiple Modalities

The results in Table 4.2 highlight the challenging nature of the task, but **we find broad consistency in the relative performance of each method across modalities and time periods**.

We find that **all input features benefit from including the time series context**

in addition to the most recent time step albeit modestly. While the benefit is modest in magnitude, we note the consistency in positive improvements across modalities and time periods. We find that the QFRs tend to benefit the most from this temporal context. This is result is consistent with recent work [78] because there is considerable boilerplate content in financial reports that does not vary much year to year, making it difficult to identify new information, and therefore requiring the historical context to identify and contextualize the differences over time.

We also find that **both sources of textual data provide considerable value beyond the financial time series variables**. However, we find that the marginal benefit of ECTs is greater than that of the QFRs. We suspect this is because the ECTs contain richer information in the form of both less scripted language by the company executives and the inclusion of analyst question-answer exchanges that provide an additional perspective of analyst context.

We conjecture that the complementary nature of the textual and tabular data is partly due to the fact that tabular data captures the past performance while the text-based models contextualize the persistence of that performance with qualitative information and augment it with forward-looking content, which we investigate further in the next section.

4.6.2 Case Study and Qualitative Analysis

To further understand the value of the textual modalities, we analyze model behavior during the 2020 time period, which was characterized by a significant economic shock caused by the COVID-19 pandemic. While we observe larger errors during the period, we also observe the largest improvement from the incorporation of text-based time series. We believe the value of Text to be most significant in the 2020 (COVID-19) regime

shift likely because the executive commentary contained in the text contains forward-looking content about the future prospects of the company in a new economic regime while **AR+FIN** largely captures historical behavior under an old economic regime. For example, it is well known that Internet Technology companies performed well during this period as people were confined to their homes and more active on the internet, while Hotels & Restaurants struggled. This regime shift would not be captured in the historical time series of company financial performance but would be reflected in the executive discussions contained in their forward-looking disclosures. We confirm this by comparing the differences between model predictions for Q2-2020. We find that the models with text-based inputs (**AR+FIN+ECT+QFR**) have 0.47 higher average prediction values for companies in the Internet Technology industry than those that only rely on financial variables (**AR+FIN**). Conversely, we find that the models with text-based inputs have 0.38 lower average prediction values for companies in the Hotels & Restaurants industry than those that only rely on financial variables.

4.6.3 Text Encoding Methods

In Table 4.3, we compare our **MTFE** approach with strong pretrained baselines to demonstrate the value of our approach. Firstly, we compare with **Mistral-Embedding**, which is a state-of-the-art finetuned embedding checkpoint [112] of the Mistral-7B model [111], a decoder-only language model that supports long contexts. The model has been finetuned with contrastive learning on a collection of document pairs across diverse tasks. We also include a version **E5-LongEmbed** of the weakly supervised embedding model E5 [127] that was optimized for long context lengths [128]. Secondly, we also compare with short-context models, including **SBERT** [101], which was contrastively pretrained on a massive corpus of 1B weakly supervised text pairs, as well as with domain-specific

language model **FinBERT** [6]. We apply the short context models at the sentence level and average the resulting embeddings over all sentences in each document. Finally, we include a pretrained financial sentiment classifier **FinBERT-Sent** [5] applied at the sentence-level [91], detailed in Appendix C.

We find that BigBird with **MTFE** provides considerable improvement over these SOTA pretrained models, suggesting that the task requires a specialized understanding of financial language, task-specific predictive features, and document-level context. This improvement over **Mistral-Embedding** suggests that smaller encoder-only models with specialized training can outperform much larger, general-purpose decoder-only LLMs on domain-specific tasks, consistent with the recent findings of Shah and Chava [129]. We conjecture that the benefits of **MTFE** arises from a combination of specialized domain and task adaptation to capture subtle signals, and a bidirectional encoding that provides better relative importance estimation within a long context [20].

Text Encoder	**Params**	**ECT**	**QFR**
MTFE	150M	**1.60***	**1.62***
w/o LC-DiffCSE	150M	1.64	1.65
w/o LC-MLM	150M	1.66	1.68
w/o SFT	150M	1.75	1.77
FinBERT-Sent	130M	1.91	1.94
SBERT	120M	1.72	1.75
FinBERT	130M	1.75	1.77
Mistral-Embedding	7B	1.70	1.72
E5-LongEmbed	110M	1.70	1.73

Table 4.3: Results indicate MSE on the test set using different textual encoding methods for feature extraction from the specified financial documents and demonstrate the value of each step of our MTFE process. * indicates the performance of the specified model is statistically better ($p < 0.05$) than that of the next best performing model on the test set according to the Wilcoxon Signed-Rank Test.

4.6.4 Multimodal Fusion Methods

In Table 4.4, we our compare modality-independent, channel mixing with late stage cross-modality fusion method **MICM** with the cross-modality channel concatenation **CMM** (typically used in multivariate time series forecasting), and channel independence across modalities **MICI**. While all multimodal methods outperform the unimodal baselines, we find that our proposed method **MICM** performs significantly better than **CMM** and **MICI**. We believe this result is due to: (1) **differential temporal patterns across modalities and the resulting incongruence that results from treating them equally (CMM)**; (2) **the rich cross-channel interactions that exists within each modality time series that independence ignores (MICI)**.

4.6.5 Time Series Encoder

In Table 4.4, we ablate our choice of TSMixer as our time series encoder with generic neural models (MLP, Transformer) and channel-independent, time series model PatchTST to further justify our use of a specialized temporal model with cross-channel interaction in the proposed method. We find that TSMixer demonstrates strong performance on this task due to its ability to both capture intra-channel temporal patterns as well as cross-channel interactions.

4.6.6 Portfolio Simulations

In Table 4.5, we demonstrate the economic value of our model predictions using portfolio simulations. We form monthly long-short (*market-neutral*) quintile portfolios [34] by sorting stocks based on (test set) model predictions, detailed in §C.1. We follow Cong et al. [130] in reporting net portfolio performance that includes conservative estimates of the impact of transaction costs on portfolio implementation. The resulting performance

Method	MSE
TSMixer-MICM	**1.42***
TSMixer-CMM	1.49
TSMixer-MICI	1.47
MLP	1.57
Transformer	1.50
PatchTST	1.48

Table 4.4: Comparison of the performance of using different time series encoders and modality fusion methods. * indicates the performance of the specified model is statistically better ($p < 0.05$) than that of the next best performing model according to the Wilcoxon Signed-Rank Test.

of our proposed method generates strong investment performance that is that is economically and statistically better than the tabular financial time series baseline. These results demonstrate that the model predictions contain significant signal that is not priced into the market with substantial value in a real-world trading setting.

Statistic	AR	+FIN+MAC	+ECT+QFR
Net Return	4.51	7.08	9.74
Volatility	9.64	9.68	9.47
Net Sharpe Ratio	0.47	0.73	1.03

Table 4.5: Annualized portfolio statistics of simulated investment performance, expressed in percentage units. "Net" performance includes an estimate of the impact of transaction costs, detailed in §C.1.

4.7 Conclusion

In conclusion, we introduce a new multimodal financial time series prediction task. We propose a novel textual feature extraction process and multimodal fusion method that demonstrates state-of-the-art performance on this challenging task. Our extensive experimental results and interpretability analysis reveal insights into the value of each modality and our proposed method. Notably, we find that the greatest gains in perfor-

mance are achieved when jointly considering both the temporal and cross-modal context that are not possible with either context alone.

Chapter 5

Context-Aware Modeling of Financial News

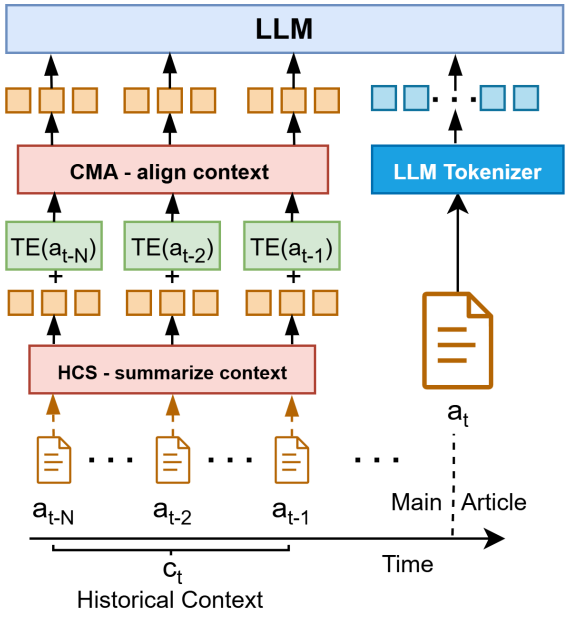

Figure 5.1: Overview of our contextualized news forecasting task and proposed Prefix Summary Context (PSC), which efficiently and effectively summarizes historical context with the Historical Context Summarizer (HCS) and aligns it in the latent space of a large LM with Cross-Model Alignment (CMA) and Context-Aware Language Modeling (CALM). The resulting PSC embeddings are prepended to the main article tokens, allowing the LLM to contextualize the main article with historical background.

5.1 Introduction

Financial news plays a critical role in the information diffusion process in financial markets and is a well-known driver of stock prices. These news articles cover a variety of events about companies, including their earnings reports, product launches, legal investigations, and changes in corporate structure. These events can have a substantial impact on company business operations and stock prices. However, it is difficult for investors to be able to identify which articles will materially impact markets because it not only requires being able to understand the sentiment of the articles, but also the more challenging task of recognizing the novelty of the information reported and gathering sufficient historical background to interpret this information in the appropriate context. We hypothesize that the most rigorous test of understanding financial news is being able

to predict the market reaction. A complete understanding requires both a local understanding of the information contained in the article but also the broader context in which that news occurs, including prior coverage of related events, as news events often develop over long periods of time.

However, it is not clear *a priori* the best way to retrieve the most relevant context for a given news article and how to optimally incorporate that context into the processing of that article. While the context length of language models has increased significantly, simply concatenating all of the articles together is not necessarily the most efficient or effective method. Firstly, the near-quadratic time and space complexity of the computation makes long contexts computationally challenging. Secondly, language models have struggled to effectively understand long documents, particularly if irrelevant context is included in their context windows, as they have been shown to exhibit positional biases and ignore content lost in the middle [20, 131, 132].

In this work, we investigate how to efficiently and effectively incorporate historical context into language models to better interpret financial news, and demonstrate that gains from our approach lead to significant improvements in financial forecasting and trading performance. In summary, we make the following key contributions:

1. We systematically evaluate how historical context improves language model understanding of financial news, demonstrating consistent gains across contextualization methods and time horizons (Table 5.2, Table 5.3).

2. We propose an efficient and effective contextualization method that demonstrates state-of-the-art performance on this challenging task compared to a comprehensive set of baselines (§5.4, Table 5.2).

3. We introduce a hybrid retrieval method that captures both the relevance and timeliness of historical context, enhancing the effectiveness of contextualization (§5.6, Ta-

ble 5.4).

4. We probe the model through multiple interpretability methods to reveal insights into the value of historical context and its significant impact on investment applications (§5.7).

Broader Impact We hope this work will inspire future research in modeling the temporal sequence of textual documents and using small LMs to efficiently provide large LMs with more context, particularly as the context length and in-context learning capabilities of LLMs continues to grow, as our findings suggest that retrieval-augmented long contexts underperform forms of context compression and summarization.

5.2 Related Work

5.2.1 Long Context Modeling

With the advances in retrieval augmented generation and in-context learning [133, 134], there is great interest in extending the context lengths of language models [14, 54, 55, 82, 111]. However, given the computational complexity of long contexts, recent work has focused on compressing them into more efficient representations [135–137] that can be leveraged by language models. These methods reduce the time complexity of self-attention at inference time, but often require expensive pretraining procedures to train the base model as a summary compressor. Other approaches use small or frozen LMs to more efficiently encode the context [138–140], but lack any form of length compression, and tend to rely on multiple stages of pretraining to learn separate cross-attention weights.

5.2.2 Financial Prediction

The impact of financial news on stock prices has been extensively studied in the academic literature. Tetlock [2, 141], Fedyk and Hodson [142], Brière et al. [143] find that investors respond to news sentiment in the short-term but the magnitude and duration of the reaction depends upon the novelty of the news content. Chen et al. [144], Xie et al. [145], Lopez-Lira and Tang [146] study the ability of language models to predict stock prices from financial news, finding that performance generally improves with model size. Ang and Lim [117], Hu et al. [147], Xu and Cohen [148], Sawhney et al. [149, 150], Agarwal et al. [151] propose methods to use multimodal streaming news and social media data to make high-frequency predictions of stock prices.

5.3 Problem Statement

Task Formulation Following Chen et al. [144], Xu and Cohen [148], we adopt the task of predicting the direction of the stock price P_t change over the course of short-term and long-term horizons (days) $h \in \{7\text{D}, 30\text{D}\}$ following the publication of the news article a_t at time t.

$$Y_t = \text{Sign}(P_{t+h} - P_t)$$

We refer to a_t as the main article and primary input, which we wish to contextualize. To enhance the understanding of the main article, we include N relevant historical articles about the same company for context:

$$c_t = \{a_{t-N}, ..., a_{t-1}\}$$

For computational efficiency, we select the $N = 5$ most recent articles about the same company as the historical context in our main experiments (Table 5.2), but investigate different methods to select the most relevant historical context in §5.6 and increase the number of historical contexts in Table 5.3.

We evaluate the performance on this binary classification task with AUC because the continuous scores are more informative than their discrete predicted classes for investment management applications.

Data Acquisition and Curation We collect *stock-specific* financial news articles in English from FactSet StreetAccount for US-based publicly-traded companies. These articles cover a variety of events and news sources. We follow the filtering criterion in Chen et al. [144] for data quality, detailed in §D.3. Our sample contains more than 3,000 public companies in the US, encompassing a diverse range of firm sizes and industries.

Data Statistics and Task Formulation We temporally partition the data into training (2010-2014), validation (2015-2016), and test (2017-2023) sets. We provide summary statistics in Table 5.1.

	Train	**Validation**	**Test**
Start Date	Jan-2010	Jan-2015	Jan-2017
End Date	Dec-2014	Dec-2016	Dec-2023
# Samples	129,146	46,931	149,409
# Companies	2,997	2,944	3,618
TE_1	15	14	12
TE_5	115	112	110

Table 5.1: Summary Statistics of the characteristics of news articles based on the average values in each sample split. TE_i indicates the average time elapsed in number of days between the publication dates of the main article and the ith most recent historical context article.

5.4 Proposed Method

Prefix Summary Context (PSC) We propose an efficient and effective method to incorporate historical context into a language model's understanding of the main article a_t, which we denote as Prefix Summary Context (PSC). Our method adapts a small language model encoder, the Historical Context Summarizer (HCS), to encode and summarize the historical context articles c_t. The main article a_t is processed with a large language model (LLM), and the summarized historical context is incorporated via learned prefix embeddings prepended to the input embeddings of the main article.

More specifically, we adapt a small language model encoder to serve as a historical context summarizer (HCS). We train HCS to learn M summary embeddings per historical news article of length L by inserting learnable summary tokens after every $L//M$ tokens, intended to aggregate information in multiple dimensions:

$$a_{t-i} = [\, s_1,\, x_1, \ldots, x_k,\, s_2,\, \ldots,\, x_L,\, s_M \,]$$

where $s_1, \ldots, s_M$ are learnable summary tokens interleaved within the input token sequence $x_1, \ldots, x_L$ of historical context a_{t-i} at regular intervals. We extract the contextualized hidden states of these summary tokens as dense summary embeddings SE of the input:

$$\mathrm{SE}(\mathrm{a_{t-i}}) = [\mathrm{HCS}(\mathrm{s_1}), .., \mathrm{HCS}(\mathrm{s_M})] \in \mathbb{R}^{\mathrm{M} \times \mathrm{d_{CE}}}$$

We expect these summary embeddings to provide multiple perspectives of the same article [152] that can be selectively attended to depending upon the content of the main article. Dense embeddings have been found to convey almost as much information as the original text itself [153], so we expect that learning them end-to-end to provide compact yet expressive representations of the original text.

To help the model distinguish the temporal relevance of each context article, we add time-based embeddings TE, based on the distance between article publication dates. We concatenate the time-enhanced summary embeddings together to form the Summary Context (SC) of $P_L = N \times M$ summary embeddings.

$$\mathrm{SC_t} = [\mathrm{SE}(\mathrm{a_{t-N}}) + \mathrm{TE}(\mathrm{a_{t-N}}), \dots] \in \mathbb{R}^{\mathrm{P_L} \times \mathrm{d_{CE}}}$$

Prior to concatenation, we set the position index of all the prefix summary embeddings to be 0 and then the first token index of the main article to be 1. This is because most large LMs use rotary position embeddings [154] in the self-attention computation, and the prefix summary embeddings, containing summarized embeddings from multiple articles, should not be treated equivalently to main article tokens. We allow the time-based embeddings to distinguish the timeliness of articles from each other. This design choice allows context length extrapolation in which more historical context articles can be provided at inference time than during training [155, 156].

Cross-Model Alignment We introduce a Cross-Model Alignment (CMA) module to learn to align the summary context (SC) representations from the historical context summarizer into the token embedding space of the large LM using multiheaded attention (MHA).

Let $E_{\text{vocab}} \in \mathbb{R}^{|V| \times d_{\text{LLM}}}$ be the token embedding matrix of the large LM, where $|V|$ is the vocabulary size and d_{LLM} is the embedding dimension. We compute the Prefix Summary Context (PSC) using multihead attention with query $Q = \mathrm{SC} \cdot W^Q$, and keys and values $K = V = E_{\text{vocab}}$, where $W^Q \in \mathbb{R}^{d_{\text{CE}} \times d_{\text{LLM}}}$ projects the context summarizer's

summary embeddings to the dimensionality of the LM's token embeddings:

$$\mathrm{PSC} = \mathrm{MHA}(Q, K, V) \in \mathbb{R}^{P_L \times d_{\mathrm{LLM}}}$$

This cross-attention mechanism allows each summary embedding in SC to attend over the entire vocabulary of the large LM, producing aligned representations that lie in the span of the large model's token embeddings. This step can be viewed as an alignment between the semantic spaces of the two language models, a concept inspired by alignment methods in other domains, such as vision [157, 158] and time series [159–162]. This design ensures that the Summary Context is compatible in the input embedding space of the large LM, and in doing so, imposes the inductive bias that PSC can be explained by a learned weighted combination of the token embeddings of the large LM.

We prepend PSC to the token embeddings of the most recent article, which are passed to the large LM's transformer layers for contextualization:

$$\mathrm{CE_t} = \mathrm{LLM}([\mathrm{PSC_t}; \mathrm{E_{vocab}}[a_t]])$$

where $\mathrm{E_{vocab}}[a_t]$ represents the token embeddings for the main article a_t.

Overall, our design provides two key advantages:

1. It exploits the power of a large LM on the main article, which is most important to the prediction.

2. It enables efficient incorporation of more historical articles as context with a small LM that summarizes the key information.

Further, we note that our method introduces a minimal number of trainable parameters and allows the historical context to interact with each other through the LLM's

transformer layers, contrary to interleaved cross-attention layers [138, 139, 163] which introduces significantly more parameters. We conduct ablations of our architecture design in §D.6, which demonstrate their value.

Context-Aligned Language Modeling (CALM) To enable the historical context summarizer to distill and encode relevant information that enhances the LLM's comprehension of the main article, we pretrain the context summarizer (HCS) and alignment module (AM) with context-aligned language modeling (CALM), a pretraining objective inspired by retrieval-augmented language modeling [133]. In this step, we learn to use PSC to predict the tokens in the main article. Given the encoded historical context $\mathrm{PSC_t}$, main article a_t, and token index i, the loss function is:

$$\mathrm{L_{CALM}} = -\sum_{\mathrm{i}} \log \mathrm{p}(\mathrm{a_{i,t}}|\mathrm{PSC_t}; \mathrm{a_{1,t}}, ..., \mathrm{a_{i-1,t}})$$

This objective encourages the context summarizer and alignment module to learn to encode contextual information that is predictive of future news text, thereby improving the LLM's ability to interpret the market impact of the main article.

Implementation Details

We initialize **Mistral-7B** [111] as the large LM and **DeBERTa-base** [164] as the historical context summarizer. During pretraining, we freeze the large LM and update only the context summarizer and alignment module. During supervised finetuning, we use LoRA [165] to finetune the large LM, while fully training all newly introduced parameters. Please see §D.12 for more details.

Model Class	Method	# Params	# Contexts	7D	30D
Zero-Shot LLMs	LMD-Sent [3]	0	0	50.71	51.32
	FinBERT-Sent [5]	110M	0	51.19	51.38
	FinMA-7B, Prompting [145]	7B	0	51.37	51.64
	FinMA-7B, Prompting [145]	7B	5	51.49	52.18
	Llama3-8B, Prompting [146]	8B	0	51.22	51.47
	Llama3-8B, Prompting [146]	8B	5	51.68	51.94
	Llama3-70B, Prompting [146]	70B	0	52.76	53.35
	Llama3-70B, Prompting [146]	70B	5	53.31	53.81
Financial Forecasting	HAN [147]	20M	5	53.17	54.32
	FAST [149]	110M	5	53.62	54.75
	HYPHEN [151]	110M	5	55.38	56.37
Single-Article Baseline	SINGLE	7B	0	55.73 [0.07]	56.50 [0.06]
Long-Context Baselines	CONCAT-FULL	7B	5	56.43 [0.09]	57.52 [0.08]
	CONCAT-PREFIX	7B	5	56.90 [0.09]	57.74** [0.09]
	MDS [166] - SUM + CONCAT	7B	5	56.62 [0.09]	57.61 [0.08]
	MDS [166] - CONCAT + SUM	7B	5	56.47 [0.09]	57.22 [0.07]
	HIERARCHICAL [167]	7B	5	56.95* [0.13]	57.66 [0.12]
Proposed	**PSC**	7B	5	**58.24*** [0.09]	**59.12**** [0.08]

Table 5.2: Main Results (higher is better): Model performance on the test set of our contextualized news prediction task at different forecasting horizons. All models after and including "SINGLE" use Mistral-7B as the base model. All results indicate the AUC of the model's predicted probabilities reported in percentage units. "[]" indicate the standard deviation of the results for that method over 3 training runs with different random seeds. *, ** indicate that the performance of our proposed model is statistically better ($p < 0.01$) than that of the highest-performing baseline according to the Wilcoxon Signed-Rank Test. The last row (bolded) indicates our proposed method PSC.

5.5 Baselines

We explore a comprehensive set of baseline methods for modeling long contexts, including single-article and multi-article prediction. For all experiments in §5.4 and Table 5.2, single-article predictions are made solely based on the most recent article, while multi-article prediction incorporates historical context based on the previous N articles about the same company. We explore more sophisticated methods to retrieve the most relevant context in §5.6.

5.5.1 Zero-Shot Baselines

First, we establish some zero-shot, pretrained baselines to highlight the difficulty of the task. We begin with the Loughran McDonald financial sentiment dictionary (**LMD-Sent**) and the pretrained financial sentiment classifier **FinBERT-Sent** [5]. We also include leading open-source LLMs: Llama3-Instruct-8B, Llama3-Instruct-70B, **Llama3, Prompting** [168], as well a financial instruction-tuned model **FinMA** [145]. For each model, we prompt it to predict the direction of the market reaction, according to the prompt developed in Lopez-Lira and Tang [146]. We explore multiple variations of this prompt, detailed in §D.8, and report the one with the best validation performance. The consistently poor performance of these baselines and their inability to effectively leverage historical context highlights the challenges LLMs face in interpreting the impact of complex factors driving market reactions.

5.5.2 Financial Forecasting Baselines

We include financial domain-specialized baseline methods, including HAN [147], FAST [149], and HYPHEN [151], which were designed to make stock movement predictions from sequences of financial text, and currently represent the state-of-the-art on this type of financial forecasting task.

5.5.3 Long-Context Baselines

We evaluate several strong long-context modeling baselines to assess the value of incorporating historical context. For all experiments, we use Mistral-7B [111], a 7B-parameter model with an 8K context window, to ensure fair comparison against our proposed method. As a primary baseline, we finetune Mistral-7B on single news articles (**SINGLE**) to isolate the effect of adding historical context.

Concatenation The most direct contextualization method is to simply concatenate the previous articles and the current article together, denoted as **CONCAT-FULL**, allowing the self-attention mechanism to interact them. We also include a version **CONCAT-PREFIX** that concatenates the first paragraph of each historical context article. We sort them in ascending order of time and delimit each article with special tokens. Since the context length of most LLMs exceeds 8K tokens, this approach can easily accommodate more than 10 articles. In both cases, we use Mistral-7B to encode this long context and finetune on the task. This represents the most direct approach of leveraging long-context capabilities without specialized contextualization. We extract the representation of the main article tokens with mean pooling, now contextualized with historical articles.

Multi-Document-Summarization MDS involves the summarization of a set of related documents, which is a natural approach to extract the most relevant information from the historical context. We use QAMDEN [166], a state-of-the-art MDS model shown to outperform GPT-4 [166]. We consider two strategies. First, we summarize the historical context and prepend this summary to the main article, denoted as **MDS-SUM+CONCAT**. Alternatively, we concatenate the historical context and the main article together, and summarize them jointly, denoted as **MDS-CONCAT+SUM**. We pass the resulting output to **Mistral-7B**, which is then finetuned on the task as in §5.5.1.

Hierarchical Encoding Alternatively, a hierarchical encoding approach can be used [58, 59, 167, 169, 170]. In this approach, adapted from Ivgi et al. [167] and denoted as **HIERARCHICAL**, each article is encoded independently, and the resulting token representations are concatenated and globally contextualized across articles by a transformer encoder, further detailed in §D.10.

5.6 Experimental Results and Analysis

For all experimental results, we would like to highlight the large size of the test set (150K samples), allowing us to detect most small differences in model performance with statistical significance. **We demonstrate the substantial practical value of significant but seemingly modest absolute gains in classifier performance in §5.7**.

Historical Context The main results in Table 5.2 highlight the challenging nature of the task, but **we find broad consistency in the relative performance of contextualization methods across time horizons**.

Overall, we find that all forecasting horizons benefit consistently and significantly from the inclusion of historical context. Contrary to Chen et al. [144], we do not find a significant degradation in model performance for longer time horizons, which we hypothesize is due to our contextualization method. From a practical perspective, we highlight that longer-term financial prediction is typically more useful for investors since it reduces transaction costs, so these gains are likely to be even more valuable in real-world applications, which we confirm in §5.7 with portfolio simulations.

Context Length In Table 5.3, we compute the forecasting performance and language model (cross-entropy) loss as a function of the number of historical context articles with the PSC contextualization method. In this experiment, we only train the model with 5 context articles but test with up to 20 since our method allows for context length extrapolation.

In this result, we view the LM loss as an approximation of LM's understanding of the main article [171, 172]. The results provide evidence towards our hypothesis that model performance on our task demonstrates language understanding as we find that both

forecasting performance *and* LM loss improves with the number of contexts incorporated.

Moreover, we find that the method performance continues to improve up to $N = 15$ context articles, but the gains diminish after 10. We note that the method's ability to generalize to more articles at inference time is particularly valuable as the number of news articles over a given period of time varies considerably with the size and industry of the company. Furthermore, as shown in Table 5.1, the average age of context articles $c_{t-N}|N \geq 5$ is more than 3 months, which supports our hypothesis that, despite the age, the context is helpful in interpreting the recent article, as the market is clearly no longer reacting to the old context article.

Contextualization Method While we find that all contextualization methods provide a material benefit over their corresponding single-article baselines, we find that approaches that summarize the relevant context into a more concise and informative form, either in the language or the latent space, tend provide better performance than simple concatenation, which is the most computationally expensive.

We conjecture that the observed trend is due to several factors. Firstly, concatenation does not impose the inductive bias that the most recent article is the most critical to the prediction. Secondly, concatenation is likely the most sensitive contextualization method to positional bias and irrelevant context. Recent work has shown that even SOTA commercial LLMs exhibit positional bias in the information that they pay attention to, irrespective of its relevance, and can become confused by irrelevant context [20, 131, 132, 173]. Since the most relevant historical context may not be optimally positioned within the context window, or may only be partially relevant to the main article, we suspect that these challenges are exacerbated on a task with such a low signal-to-noise ratio.

# Contexts	LM Loss	7D	30D
0	2.10	55.73	56.50
1	1.95	56.60	57.47
2	1.84	57.48	58.39
5	1.80	58.24	59.12
10	1.75	58.59	59.44
15	**1.72**	**58.66**	59.50
20	1.73	58.64	**59.52**

Table 5.3: Results indicate the forecasting performance and LM loss of the PSC contextualized model while varying the number of the most recent articles to include as context.

Retrieval Method In the experiments so far, we assume that the $N = 5$ most recent historical articles about the same company provide the best context for the main article (**Time**). However, the most recent articles may not always be the most relevant context as news stories can develop over longer time periods.

We test this hypothesis by instead retrieving the $N = 5$ most relevant articles about the same company according to semantic similarity. To do this, we experiment with three strong but generic retrievers: **SBERT** [101], **Contriever** [174], **InstructOR** [175]. Additionally, we create our own domain-specific, text-encoder (**FinSim**) pretrained on the training data with the DiffCSE framework [102]. We further incorporate time into the measure by applying an exponential decay with a half-life of 6 months to the similarity score to capture both context similarity and timeliness (**TimeFinSim**). Given main article a, historical context c, time elapsed t, and half-life H:

$$\mathrm{TimeFinSim(a, c, t)} = \mathrm{FinSim(a, c)} \cdot \mathrm{e}^{-\frac{\ln(2)}{\mathrm{H}}\mathrm{t}}$$

For each target article, we retrieve the $N = 5$ closest articles from the same company within the last year and pass them to PSC for contextualization in temporal order. We include more details about the retrievers in §D.11.

As shown in Table 5.4, we find that semantic similarity-based retrieval is only better than time-based when the retriever is specialized to the financial domain (**FinSim**) and that a hybrid approach (**TimeFinSim**) performs significantly better. We conjecture that since the retrieval set is all financial articles about the same company, generic retrievers are not sufficiently discriminative to connect related events together.

Retriever	# Contexts	7D	30D
Time	5	58.24	59.12
SBERT	5	58.02	58.80
Contriever	5	58.10	58.96
InstructOR	5	58.17	59.11
FinSim	5	58.81	59.74
Proposed-TimeFinSim	5	**59.12**	**60.15**

Table 5.4: Results indicate the AUC performance of different text retriever models used to retrieve the $N = 5$ most relevant articles to contextualize the main article with PSC contextualization method.

5.7 Model Analysis and Interpretability

In this section, we explore the behavior of the model from different perspectives and reveal insights into the value of historical context and potential impact on investment applications.

Staleness of News Financial news articles do not always report novel information. Stale news may cover information that has already been priced into the market, in which case it can have a different market impact. To better understand the impact of stale news on our results and the value of contextualization, we must first introduce a measure of news staleness. For this purpose, we define article staleness to be the average similarity of the article to the preceding $N = 5$ articles about the same company, using SBERT

Historical Context, N=1	Main Article
Harte Hanks announces sale of Trillium business to Syncsort for $112m in cash...Harte Hanks will use substantially all of the net proceeds from the sale to retire its outstanding credit facility ...additional details will be provided at a later date..	Harte-Hanks to delay 10-K filing...complications required to account for the sale of the company's trillium business (and calculate the tax therefor),...delays and complications in performing its annual goodwill impairment analysis, and additional time-consuming review, testing and evaluation.

Figure 5.2: Sample news that highlights the value of historical context in understanding the recent news article. Without knowing the terms of the sale, filing delays would typically be interpreted negatively. However, the sale allows the company to pay off its outstanding debt. We vary the number of historical context articles (N) and show that more context leads to consistently improved language understanding and forecasting performance.

[101] for document embeddings E:

$$\text{staleness}(a_t) = \frac{1}{5}\sum_{i=1}^{N=5} \text{sim}(E(a_t), E(a_{t-i}))$$

This metric will assign higher staleness scores to articles which overlap substantially with previous articles about the same company, and estimate how much novel information is reported. We categorize the test data into 3 levels of staleness based upon this measure.

As shown in Table 5.5, we find that the historical contextualization provides consistent benefit across all articles but has the largest relative improvement for the stalest articles relative to their histories, as the model can learn to adjust its predictions based upon the novelty of the information and its effect within the related context. It is also important to note that high article novelty (low staleness) is likely associated with less relevant contexts, which could be viewed as distracting information and ultimately confuse the model from the main article [20]. However, our method still provides value in these

instances, suggesting that it is robust to less relevant contexts, which we attribute to its ability to distill relevant information.

Staleness Level	Single, N=0	Contextualized, N=5	Δ
1	56.94	58.31	1.37
2	56.51	59.37	2.86
3	56.39	59.72	3.33

Table 5.5: Results indicate the 30D AUC performance of the single article model baseline and the PSC contextualized model, as well as the difference between them (Δ), across articles with different levels of staleness relative to their histories, reported in percentage points.

Case Study and Qualitative Analysis We conduct a qualitative analysis of model predictions to further understand the effects of historical contextualization and its impact on the model behavior. To this end, we sort the difference in model predictions between the single article and historical contextualized model, and examine the sample with the largest difference. In this example, displayed in Figure 5.2, the single-article model predicts that the filing delay reported in the most recent article will lead to a negative return, but the additional context of an article from a few months prior reveals the reason is due to the sale of a business unit that will enable the company to pay down its debt and reduce its interest burden. As a result, investors viewed the news favorably and the stock rallied considerably over the following month.

Portfolio Simulations In Table 5.6, we demonstrate the economic value of our methodology with portfolio simulations. We form monthly long-short (*market-neutral*) portfolios [34] by sorting stocks based on the average 30D-horizon model predictions from the past month, detailed in §D.1. For comparison, we simulate two common trading strategies: Price Momentum [176] and the Fama-French 6-Factor pricing model [177], described further in §D.2. Following Cong et al. [130], we include net performance that includes

conservative estimates of the impact of transaction costs on portfolio implementation.

The resulting performance of our contextualization method generates investment performance that is economically and statistically better than the single and multi-article baselines. Relative to the single article baseline, we note that a seemingly modest absolute gain in AUC (∼2.5%) from our approach results in a more than 50% relative improvement in investment performance. These results demonstrate that the contextualized model provides significant predictive value beyond the single-article baseline in a real-world trading setting.

Method	Net Return	Volatility	Net Sharpe Ratio
Price Momentum	5.30	18.33	0.29
Fama-French 6-F	6.99	14.51	0.48
SINGLE, $N = 0$	9.04	13.40	0.67
CONCAT, $N = 5$	11.35	13.30	0.85
HIERARCH, $N = 5$	12.04	13.55	0.89
Proposed-PSC, $N = 5$	**14.13**	**13.33**	**1.06**
+ TimeFinSim	**15.02**	**13.25**	**1.14**

Table 5.6: Annualized portfolio statistics of simulated investment performance, expressed in percentage units. N denotes the number of historical context articles used. "Net" performance includes an estimate of the impact of transaction costs, detailed in §D.1.

5.8 Conclusion

We explore the value of historical context in the ability of language models to understand the market impact of financial news, and find that it provides a consistent and significant improvement in performance across methods and time horizons. To this end, we propose an efficient and effective contextualization method that allows the use of a large LM on the main article while using a small LM to summarize the historical context. Through multiple interpretability tests, we reveal insights into the value of historical context and demonstrate that it directly translates to improvements in simulated investment

performance.

Chapter 6

Multimodal Modeling of Interleaved Text and Time Series

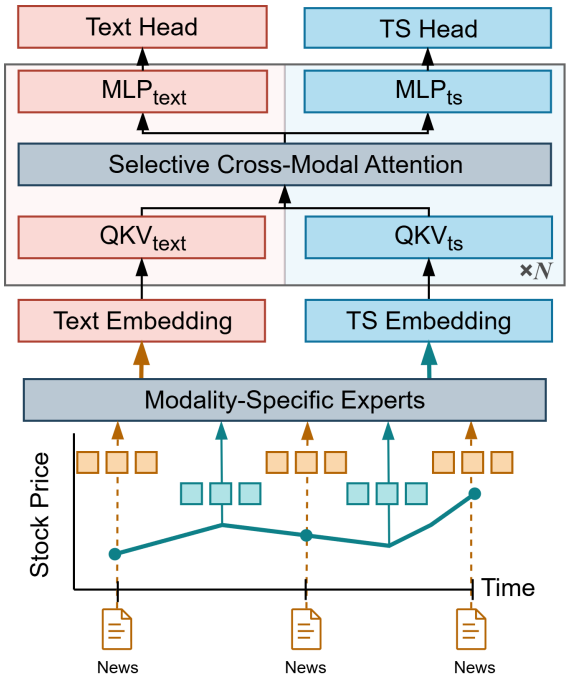

Figure 6.1: Overview of our multimodal forecasting task and proposed model (MSE-ITT), which processes interleaved sequences of tokens of news articles (Text) and discretized stock returns (TS). MSE-ITT incorporates modality-specific experts to capture distinct patterns in text and time series, while enabling joint reasoning across modalities through selective cross-modal attention.

6.1 Introduction

Text and time series provide complementary perspectives on financial markets. News articles describe company events, such as earnings announcements, product launches, and mergers, while stock prices reflect how markets react to these events over time. When these modalities are combined, it presents a promising yet challenging opportunity for large language models (LLMs) to enhance financial forecasting by reasoning over temporally aligned but semantically distinct inputs.

To better understand their complementary nature, consider how time series and text contribute distinct but related signals. The time series data offers global context into both short and long-term price behavior, allowing the model to learn patterns driven

by investor biases [176, 178, 179]. Moreover, the interleaved stock prices act as implicit supervision, revealing how markets have previously responded to similar news, and exposing the model to repeated cause-and-effect dynamics. Conversely, news articles provide both retrospective context about historical market behavior and forward-looking signals that may reinforce or contradict current price trends. By jointly modeling these modalities, LLMs have the potential to learn cross-modal interactions and context-aware representations that would not be possible from either modality alone.

However, effectively integrating long sequences of text and time series in a unified model remains challenging. Simple strategies, such as converting numerical data into strings of digits, fail to capture their distinct structures. Language is discrete, syntax-rich, and compositional, while time series are continuous, stochastic, and governed by temporal dependencies. Furthermore, news arrives at irregular intervals, while stock prices are observed daily. These structural differences make it difficult for pretrained LLMs, which are optimized for language, to extract meaningful signals from time series inputs. Instead, we argue that modeling these multimodal sequences requires modality-specific components that can respect and exploit the unique structure of each input type, while enabling joint reasoning across modalities.

To address this, we propose a unified multimodal architecture with modality-specific experts and design a pretraining objective to effectively learn cross-modal interactions. In summary, we make the following key contributions.

1. We design a multimodal architecture (**MSE-ITT**) that can effectively model interleaved sequences of text and time series, demonstrating state-of-the-art performance on a challenging financial forecasting task compared to a comprehensive set of baselines (§6.4, Table 6.2).

2. We introduce a unified cross-modal alignment framework (**SALMON**) with a dy-

namic, salient token weighting mechanism (**STW**) to effectively learn time-series-specific features that enhance language understanding (§6.4.2, Table 6.3).

3. We develop an interpretability method that reveals insights into the value of time series context and demonstrate that it translates to significant economic gains in investment applications (§6.7.2, Table 6.5, Table 6.6).

Broader Impact We hope this work encourages future research on LLMs that reason over interleaved sequences of text and time series, particularly in domains where structured and unstructured data interact over time, such as finance, healthcare, and climate. Our findings highlight the limitations of treating time series equivalently to language and underscore the importance of dedicated mechanisms for structured time series inputs.

6.2 Related Work

6.2.1 Multimodal Time Series Forecasting

LLMs have greatly improved their capabilities in understanding multimodal inputs, such as text, images, audio, and video [180–183]. However, they have demonstrated challenges in understanding time series data. While some work has found benefit in using LLMs to perform time series forecasting with zero-shot learning [184] or finetuning [159], other work has found that their pretrained weights do not provide positive transfer [28] and struggle to reason about them effectively [29] without specialized encoding [30]. Moreover, recent work on contextualized forecasting has explored integrating text and time-series data using two main strategies. One paradigm converts time series into strings or embeddings, allowing text to condition LLM predictions [159, 185, 186]. However, this assumes LLMs can natively interpret time series and limits the model's ability

to learn modality-specific patterns. Another approach uses frozen language models to extract fixed text features for multivariate time series models [187, 188]. While this allows modality-specific patterns, it prevents deep cross-modal interaction and inhibits the reasoning capabilities of LLMs.

6.2.2 Financial Prediction

Recent work has shown that language models can predict stock returns from news articles [144–146]. Separately, historical price patterns have been shown to be predictive [176, 178, 179]. These findings have motivated recent work exploring methods to integrate textual and time series data to improve forecasting [38, 117, 148, 189–191]. However, most of these models process each modality independently and generally rely on late-fusion methods, limiting cross-modal interaction during representation learning.

6.3 Problem Statement

6.3.1 Multimodal Inputs

In our main problem formulation, we consider a multimodal sequence of time-stamped inputs, aperiodically arriving news articles (text), and daily stock returns (time series), described below.

Time Series Inputs As time series inputs X_t, we consider the daily stock return r_t of the company over the last 1-year (252 trading days).

$$X_t = \{r_{t-252}, ..., r_{t-1}\}$$

While we demonstrate that our method generalizes to multivariate time series in §E.9, we focus on the univariate case in our main experiments to more precisely study the effects of the cross-modal interaction method in a controlled setting.

Textual Inputs As textual inputs Y_t, we consider the N most recent news articles a_t about the company prior to time t:

$$Y_t = \{a_{t-N}, ..., a_{t-1}\}$$

For computational efficiency, we select the $N = 10$ most recent articles about the same company that occurred within the last 1-year.

6.3.2 Task Formulation

Following Chen et al. [144], Xu and Cohen [148], we adopt the task of predicting the direction of the stock price P_t change over the course of short-term and long-term horizons (days) $h \in \{7\text{D}, 30\text{D}\}$ at prediction time t.

$$D_t = \text{Sign}(P_{t+h} - P_t)$$

We evaluate the performance on this binary classification task with AUC because the continuous scores are more informative than discrete classes for investment management applications.

Data Acquisition and Curation We curate financial news articles in English from the FNSPID Dataset [192], for US-based public companies, which covers a variety of company events and news sources. We perform a curation process, following the filtering criterion in Chen et al. [144] for data quality, detailed in §E.4. Our sample contains more

than 3,000 public companies in the US, encompassing a diverse range of firm sizes and industries.

Data Statistics and Task Formulation We temporally partition the data into training (2010-2017), validation (2018-2019), and test (2020-2024) sets. We provide summary statistics in Table 6.1.

	Train	**Validation**	**Test**
Start Date	Jan-2010	Jan-2018	Jan-2020
End Date	Dec-2017	Dec-2019	Dec-2024
# Samples	155,146	36,931	115,611
# Companies	2,591	2,249	3,564

Table 6.1: Summary statistics of the characteristics of financial news articles and stock return time series in each sample split.

6.4 Proposed Method

6.4.1 Model Design

In this section, we introduce our proposed modality-specific experts multimodal model for interleaved sequences of text and time series (**MSE-ITT**), illustrated in Figure 6.1. The use of mixture-of-experts (MOE) layers [26, 27, 193] in LLMs have become pervasive because of their improved representational capacity and compute efficiency, allowing subnetworks to specialize in different regions of the input. Inspired by recent advances in multimodal MOE-based LMs for text and images [22–25], we design a modality-specific experts architecture that builds on Llama3-8B [194], an autoregressive LM with strong language capabilities, and extend it with components tailored for temporal and cross-modal understanding.

Modality-Specific Experts To process both modalities effectively, we introduce dedicated modality-specific (TS) parameters into each layer of the pretrained LM. These modality-specific components are responsible for processing time series inputs, enabling the model to capture time-series patterns without disrupting the pretrained language capabilities of the base LM. This separation reduces cross-modal interference and imposes an inductive bias that respects the structural differences between text and time series. At the same time, it allows for joint modeling of interleaved sequences. We maintain causal masking in the self-attention layers to preserve temporal causality.

These added parameters include layer normalization (LN), multi-headed attention projections (QKV), and feedforward MLPs (MLP):

$$\mathbf{h}^q_{\text{text}},\, \mathbf{h}^k_{\text{text}},\, \mathbf{h}^v_{\text{text}} = \text{QKV}_{\text{text}}\big(\text{LN}_{\text{text}}(\mathbf{h}_{\text{text}})\big)$$

$$\mathbf{h}^q_{\text{ts}},\, \mathbf{h}^k_{\text{ts}},\, \mathbf{h}^v_{\text{ts}} = \text{QKV}_{\text{ts}}\big(\text{LN}_{\text{ts}}(\mathbf{h}_{\text{ts}})\big)$$

$$\mathbf{h}_{\text{text}} = \text{MLP}_{\text{text}}\big(\text{LN}_{\text{text}}(\mathbf{h}_{\text{text}})\big)$$

$$\mathbf{h}_{\text{ts}} = \text{MLP}_{\text{ts}}\big(\text{LN}_{\text{ts}}(\mathbf{h}_{\text{ts}})\big)$$

In this design, the text ($\mathbf{h}_{\text{text}}$) and time series ($\mathbf{h}_{\text{ts}}$) hidden states are routed to separate modules, before performing joint self-attention over the multimodal sequence:

$$\mathbf{h} = \text{softmax}\Big([\mathbf{h}^q_{\text{text}}; \mathbf{h}^q_{\text{ts}}][\mathbf{h}^k_{\text{text}}; \mathbf{h}^k_{\text{ts}}]^\top\Big)\,[\mathbf{h}^v_{\text{text}}; \mathbf{h}^v_{\text{ts}}]$$

We initialize these dedicated TS-parameters from the pretrained values in the corresponding LM layer. We omit the residual connections and scaling for brevity.

Selective Cross-Modal Attention Prior work has shown that early layers of pre-trained LMs capture low-level, syntactic patterns [195–197], while deeper layers encode higher-level, global semantics. In addition, attention weights have been found to exhibit noise due to positional biases [20, 198–200]. Based on these findings, our hypothesis is that deeper layers are more likely to benefit from time-series context and that introducing cross-modality attention at early layers may be harmful, as it risks disrupting low-level, modality-specific representations with noisy time-series signals. To address this, we restrict cross-modality attention to the top half of the network layers (16-32), allowing the lower layers (1-16) to focus on learning more localized, modality-specific features. This design choice results in efficiency gains, and, as shown in our ablation studies (Table 6.4), improves language understanding and task performance.

Interleaved Multimodal Sequence We design the input as a temporally ordered sequence of multimodal tokens, where a_t is a sequence of tokens for news article a and r_t is the stock return with timestamp t.

$$x_{1:L} = \{r_{t-252}, ..., a_{t-N}, ..., r_{t-1}, ..., a_{t-1}\}$$

Since the news articles arrive at irregular intervals, we employ pointwise embedding tokenization [201] to accommodate a variable number of time steps between news articles, rather than using fixed-length patches. This allows flexible modeling of sequences with variable temporal resolution and supports application to other domains with irregular event streams.

Because financial data is known to be statistically noisy, we discretize the continuous time series values into a fixed number of bins B, enabling more robust representations and reducing sensitivity to outliers. We learn embeddings (TSEmb) for each bin [202,

203] with quantile binning to ensure a well-calibrated distribution. We tune B over the validation set and ablate this design choice in §E.7. This choice allows us to transform the inputs into a unified sequence of embeddings z_i:

$$\mathbf{z}_i = \begin{cases} \text{TextEmb}\big(x_i\big), & \text{if mod}(x_i) = \text{text}, \\ \text{TSEmb}\big(x_i\big), & \text{if mod}(x_i) = \text{ts}, \end{cases}$$

We pass them through MSE-ITT to obtain contextualized hidden states $\mathbf{h}$:

$$\mathbf{h}_{1:L} = \text{MSE-ITT}\big(\mathbf{z}_{1:L}\big) = [\,\mathbf{h}_1, \ldots, \mathbf{h}_L\,]$$

This interleaved input design enables the model to leverage the pretrained LLM's relative positional encodings [154], to reflect the natural temporal order of events, rather than relying on learnable time embeddings [204] only at the input layer. Because rotary encodings impose a strong locality bias throughout the network, preserving temporal order in the input allows the model to better exploit this inductive bias in its attention patterns. It also ensures that the relative distance between article tokens reflects their actual separation in time, spaced by time series embeddings corresponding to the number of days between events.

6.4.2 SALMON

To further improve multimodal understanding, we introduce Salience-Aware Language Modeling over Interleaved Modalities (**SALMON**), a unified pretraining objective, that jointly predicts the next text token and the next time series token in an interleaved multimodal sequence, with a dynamic salience-based weighting mechanism. Since we discretize the time series into discrete tokens, both objectives can be trained using cross-

entropy loss, with separate projection heads for text tokens (U_{text}) and time series tokens (U_{ts}):

$$P_\theta(x_i) = \begin{cases} \text{softmax}(U_{\text{text}}\mathbf{h}_{i-1}), & \text{if mod}(x_i) = \text{text}, \\ \text{softmax}(U_{\text{ts}}\mathbf{h}_{i-1}), & \text{if mod}(x_i) = \text{ts}, \end{cases}$$

$$\mathcal{L}(x;\theta) = -\sum_{i\in\mathcal{I}_{\text{text}}} \log P_\theta(x_{i,\text{text}}) \; - \sum_{i\in\mathcal{I}_{\text{ts}}} \log P_\theta(x_{i,\text{ts}})$$

To enable cross-modal learning without disrupting pretrained language capabilities, we freeze the pretrained text-branch parameters and train only the newly introduced TS-specific parameters. This joint objective encourages the model to learn to identify features in the time series that are predictive of future news text, and learn aligned representations that capture how text and time series data co-evolve together.

Salient Token Weighting (STW) While SALMON enables joint language modeling across modalities, standard cross-entropy loss places equal weight on all next token predictions. However, we hypothesize that not all textual tokens should theoretically or equally benefit from time-series context. In typical news text, many tokens, such as function or filler words, are easily predicted using neighboring tokens. While these tokens may not benefit from TS-context, we suspect that specific salient tokens, such as those sentiment-charged or related to market behavior, could benefit substantially. For example, consider the news headline in Figure 6.2.

We hypothesize that the easily predicted words dominate the loss function and that a selective weighting mechanism can help focus the training signal on tokens where TS-context provides significant mutual information, thereby improving the alignment process by reducing weight on noisy, irrelevant tokens.

To systematically identify such salient tokens without relying on external sentiment

Model Class	Model	Base LM	Input	7D	30D
Zero-Shot LLM	GPT-4o, Direct [146]	GPT-4o	text	52.45	53.31
	GPT-4o, CoT	GPT-4o	text	53.15	54.33
	GPT-4o, Direct	GPT-4o	ts	50.92	49.77
	GPT-4o, CoT	GPT-4o	ts	48.70	47.90
	GPT-4o, Direct [185]	GPT-4o	text, ts	52.09	52.71
	GPT-4o, CoT [205]	GPT-4o	text, ts	50.56	53.05
Unimodal	TS-Only [126]	None	ts	52.93	53.88
	Text-Only [144]	Llama3-8B	text	53.76	54.13
MMTSF	TimeLLM [159]	Llama2-7B	text, ts	53.79	55.10
	TaTs [188]	Llama2-7B	text, ts	54.48	54.81
	TTSR [30]	Mistral-7B	text, ts	55.93	56.17**
	TimeMDD [187]	Llama3-8B	text, ts	55.15	55.25
	Hybrid-MMF [186]	Llama3-8B	text, ts	55.96*	55.84
SFF	FinMA [145]	Llama2-7B	text, ts	51.11	52.15
	MTFE-MICM [38]	BigBird-125M	text, ts	55.44	54.49
	FININ [189]	RoBERTa-125M	text, ts	52.47	53.13
	StockTime [206]	Llama3-8B	text, ts	55.36	55.85
	MAT [207]	FinBERT-110M	text, ts	54.43	53.81
	MM-iTransformer [191]	FinBERT-110M	text, ts	54.16	53.57
Proposed	**MSE-ITT**	Llama3-8B	text, ts	**57.94*** [0.08]	**58.48**** [0.07]

Table 6.2: Main Results (higher is better): Model performance on the test set of our multimodal prediction task at different forecasting horizons. All results indicate the AUC of the model's predicted probabilities reported in percentage units. "[]" indicate the sample standard deviation of the results over 3 training runs with different random seeds. *, ** indicate that the performance of our proposed model is statistically better ($p < 0.01$) than the next best performing model according to DeLong's test. The last row indicates our proposed method MSE-ITT.

dictionaries, we design a contrastive estimation approach inspired by recent advances in contrastive decoding [208, 209] and long-context training [199, 210]. To this end, we leverage the LM with text-only inputs as a contrastive baseline. Specifically, we compute two versions of token-level predicted probabilities: one using text-only inputs and another using the full multimodal input. Since we freeze the text-only parameters of the model, we can compute the baseline forward pass efficiently with no gradients attached. The ratio of these probabilities (i.e. the likelihood ratio) reflects how much the TS context improves the prediction of each token, and serves as a proxy for token salience.

Mathematically, assume that $P_\theta(x_{i,\text{text}} \mid x_{j<i,\text{text}})$ is the probability of the i-th text

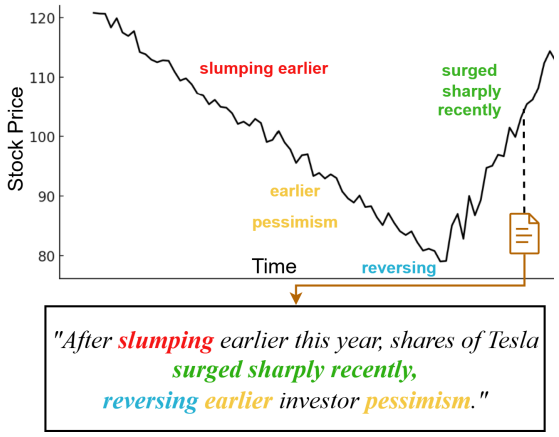

Figure 6.2: Illustration of our cross-modal alignment framework, SALMON, which learns to align historical stock price behavior and news articles with a unified objective. The Salient Token Weighting (STW) mechanism dynamically assigns higher weight to tokens that benefit most from time-series context, improving cross-modal alignment.

token given preceding text tokens, and $P_\theta(x_{i,\text{text}} \mid x_{j<i,\text{text}}, x_{j<i,\text{ts}})$ is the probability when preceding time series context is also provided. We define the token-level weight as:

$$W(x_{i,\text{text}}) = \frac{P_\theta(x_{i,\text{text}} \mid x_{j<i,\text{text}}, x_{j<i,\text{ts}})}{P_\theta(x_{i,\text{text}} \mid x_{j<i,\text{text}})}$$

such that W directly measures how much the prediction of each token improves when time series context is available. Values greater than 1 indicate tokens that benefit from time series context, while values less than 1 indicate tokens where time series context is less helpful or potentially distracting. Finally, the weights are normalized to have mean one per sequence $\tilde{W}$ and then applied to each textual token in the cross-entropy loss. Mathematically, let $x_{i,\text{text}}$ and $x_{i,\text{ts}}$ denote the i-th text and time series tokens, respectively, and $c_{j<i} = (x_{j<i,\text{text}}, x_{j<i,\text{ts}})$ denote the previous context, then the STW loss is formed on textual inputs x_{text} by:

$$\mathcal{L}_{STW}(x_{\text{text}}; \theta) = -\sum_{i \in \mathcal{I}_{\text{text}}} \tilde{W}(x_{i,\text{text}}) \cdot \log P_\theta(x_{i,\text{text}} \mid c_{j<i})$$

This weighting mechanism amplifies the learning signal for textual tokens that benefit most from time series context. Since, at the beginning of training, the estimated token-level weights are not meaningful as the model has not yet learned to effectively leverage the time series context, we warm-start $W = 1$ for the first 20% of training steps, and then relax this constraint to $W \in [0.1, 10.0]$ as the models learn to better utilize this information. We perform this cross-modal pretraining step on the input training data, and ablate the value of it and the token weighting mechanism in Table 6.3.

Implementation Details Our method is parameter and compute efficient as we freeze the pretrained text-branch parameters, and only finetune the newly added TS-branch parameters efficiently with LoRA [165], during both cross-modal alignment and task finetuning.

6.5 Baselines

We provide a comprehensive set of strong baselines to evaluate the benefits of our approach, spanning commercial LLMs, state-of-the-art multimodal time series forecasting models, and specialized financial models for stock movement prediction.

Unimodal First, we provide simple unimodal baselines that indicate the independent forecasting ability of each input modality. **Text-Only** indicates training a classifier on frozen Llama3-8B embeddings of the text-only inputs [47]. **TS-Only** indicates training PatchTST [126] on only the time series inputs.

Zero-Shot LLMs We include the zero-shot performance of GPT-4o [211] with a variety of different prompting configurations, following [185], establishing state-of-the-art commercial LLM baselines and highlighting the difficulty of the task. We provide dif-

ferent prompting methods, including direct prediction (**Direct**) and Chain-of-Thought (**CoT**) [205, 212]. The time series inputs are converted to text strings and we tune their formatting with validation set performance [184, 185]. We also provide text-only and time-series only baselines to indicate the relative ability of LLMs to reason over each modality, detailed in §E.5.

Multimodal Time Series Baselines We implement a variety of baselines specialized for multimodal time series forecasting (**MMTSF**). These include TaTs [188], TimeMDD [187], TimeLLM [159], Hybrid-MMF [186], and TTSR [30]. All of these models are finetuned on the training data according to their original implementations, described further in §E.2.

Financial Forecasting Baselines We include a variety of baselines specialized for multimodal stock movement prediction (**SFF**). These include FinMA [145], MTFE-MICM [38], FININ [189], StockTime [206], MAT [207], and MM-iTransformer [191]. All of these models are finetuned on the training data, detailed in §E.2.

6.6 Experimental Results and Analysis

6.6.1 Main Results

Overall, we find that our proposed model delivers meaningful gains in forecasting performance across time horizons compared to a strong set of diverse baselines, and that these gains translate to significant improvements in investment simulations (Table 6.6). As we further demonstrate with ablation experiments (Table 6.3, Table 6.4), these improvements stem from key architectural design choices grounded in inductive biases. First, the model captures modality-specific structure by introducing dedicated

parameters for time series data, while maintaining a unified architecture that enables joint reasoning across modalities. Second, our SALMON objective aligns time series features in the latent space of the LLM for enhanced cross-modal understanding, while our dynamic salient weighting mechanism enhances cross-modal alignment by focusing on key tokens that benefit most from time series context.

Time Series Context The main results in Table 6.2 highlight the challenging nature of the task and that the value of multimodal context highly depends upon the method of integration. While some methods benefit significantly from multimodal context, others underperform unimodal baselines.

We highlight that across both general and specialized baselines, there is a clear trend that methods that initially encode the time series with separate parameters from the LLM perform better than those that exclusively rely on LLM's native ability to extract meaningful time series features in either the text or latent space. Similarly, we find that models that simply treat text features as additional channels within a multivariate time series model underperform those that harness the powerful reasoning capabilities of LMs. These findings reinforce our design decision for modality-specific experts within a unified architecture.

Zero-Shot LLMs We highlight two interesting findings for GPT-4o. Firstly, GPT-4o performs better with text-only inputs than multimodal inputs, failing to benefit from the time series context. Secondly, prompting with chain-of-thought improves the ability to GPT-4o to analyze the textual inputs while failing to improve its ability to reason about the time series inputs. Overall, these results clearly highlight the challenges that LLMs exhibit in reasoning over time series inputs [29], particularly in the financial domain, which are often noisy and lack consistent patterns.

Method	7D	30D
MSE-ITT w/o SALMON	56.93	57.14
+SALMON w/o STW	57.56*	57.89*
+SALMON w/ STW	57.94**	58.48**

Table 6.3: Results demonstrate the value of our proposed Cross-Modal alignment objective (SALMON) with Salient-Token Weighting (STW) compared to the baseline MSE-ITT model finetuned without any pretraining. *, ** indicate that the performance of the model is statistically better ($p < 0.05$) than the previous model according to DeLong's test.

6.6.2 Ablation Studies

We conduct extensive ablations of our multimodal architecture and cross-modal alignment process, highlighting the contribution of each design choice to overall model performance.

Cross-Modal Alignment In Table 6.3, we find that SALMON pretraining significantly improves both language understanding capabilities and task performance by aligning representations across modalities. Additionally, our STW mechanism provides additional benefits over standard equal token weighting by selectively identifying and emphasizing tokens that benefit most from TS context, leading to more effective cross-modal representation learning.

Modality-Specific Experts In Table 6.4, we provide a direct comparison of our method against baseline models that share weights between the text and time-series inputs. These results demonstrate that our multimodal model effectively leverages time series context to improve its language understanding capabilities with unified pretraining (20% reduction in LM loss). Compared to sharing parameters, our modality-specific experts architecture improves both language and time series understanding capability, mitigating cross-modal interference and allowing unique modality patterns.

Model	LM Loss	30D
Text-Only	2.20	54.13
Shared Parameters	2.00	55.80
Separate QKV	1.85	56.78
Separate MLP	1.91	56.31
Cross-Modal Attention, Layers 1-16	1.96	56.00
Cross-Modal Attention, Layers 1-32	1.81	56.51
MSE-ITT	1.78	57.14

Table 6.4: Results demonstrate the value of our proposed MSE-ITT architecture in both language understanding (LM Loss) and financial forecasting (30D) performance, compared to two sets of baselines that (1) share parameters across text and TS inputs and (2) perform cross-modal attention in different layers of the network. Note that the LM Loss results include cross-modal alignment training (SALMON) but the 30D results do NOT.

Selective Cross-Modal Attention We include baselines in which cross-modal attention is performed in early layers of the network, rather than just the second half. As shown in Table 6.4, cross-modal attention in the early layers of the network is detrimental to performance, supporting our hypothesis that early LM layers focus on low-level features that can become disrupted by premature cross-modal interactions. We believe these findings could generalize to other multimodal contexts (e.g. healthcare, climate, etc.) in which modalities possess different properties but shared temporal alignment.

6.7 Model Analysis and Interpretability

In this section, we explore the behavior of the model and reveal insights into the value of time series context and its potential impact on investment applications.

6.7.1 Real-World Case Study

To further illustrate how our model benefits from multimodal reasoning, we examine a real-world event sequence involving the proposed Pfizer, Allergan merger, shown in Figure 6.3. This event unfolded over several months in 2015 and 2016, with headlines describing early merger rumors, confirmation of deal terms, political opposition, and ultimately, the termination of the transaction.

While a headline such as "Merger terminated by mutual consent" may appear negative in isolation, it follows a series of prior developments in which investors reacted negatively to news suggesting the deal would proceed. The accompanying stock return history reveals this clearly: mild declines on initial merger rumors, followed by a significant sell-off once the deal was announced, and a strong rebound when the US Treasury issued rules to block the deal. On April 6, 2016, when Pfizer and Allergan officially terminated the merger, the stock surged.

This price behavior indicates that investors viewed the termination as favorable, even though the text of the headline appears quite negative. A text-only model, seeing only the phrase "merger terminated", predicts a low probability (0.32) of a positive return. The time series-only model, reacting only to the long-term price trend of the stock price, driven negative by the merger rumors, predicts a score of (0.21).

However, our multimodal model (MSE-ITT), which jointly reasons over the headline, prior news, and accompanying stock price movements, assigns a high probability (0.79) to a positive return following the announcement. This prediction reflects the model's ability to perform a form of counterfactual reasoning: although the headline ("merger terminated") appears negative in isolation, the model recognizes that it cancels a deal investors had previously responded to unfavorably. By conditioning on prior market reactions to similar events, MSE-ITT infers that the termination is likely to be viewed positively. This

Date	News Headline	Stock Return	Interpretation
29 Oct 2015	Pfizer approaches Allergan about record merger	-1.9%, ... , -0.00%	Rumor; investors mildly optimistic
23 Nov 2015	$160 B deal announced; Pfizer to relocate to Ireland	–2.6%, ... , +2.3%	Fear of dilution and tax-inversion risk
5 Apr 2016	U.S. Treasury issues rules likely to scuttle tax inversion	+2.1%	Hope deal will be abandoned
6 Apr 2016	Merger **terminated** by mutual consent	+5.0%	Relief rally

Figure 6.3: Example sequence of news events and market reactions about the potential Pfizer-Allergan merger. The news articles contain persuasive language, yet the accompanying stock returns reveal the market's true perception: optimism on early rumors, a sharp sell-off once the deal terms are set, and a relief rally when the takeover is cancelled. Jointly modeling these inputs can more effectively interpret the outcome of such events.

deep integration of modalities enables the model to capture temporally grounded event semantics, where the sentiment of a headline depends critically on historical context.

Furthermore, this case study illustrates how unimodal models, lacking either narrative or market information, misinterpret semantically ambiguous events. In contrast, MSE-ITT aligns text and time series through token-level weighting that links financial language to observed outcomes, allowing it to resolve sentiment ambiguity and deliver more accurate, context-aware forecasts.

6.7.2 Value of Time Series Context

To further investigate where and when time-series context is most beneficial, we use the LM Financial Dictionary [3] to classify the financial sentiment of words. Following §6.4.2, we compute the benefit of TS-context (likelihood ratio) for each group of words across the test set with our MSE-ITT model after SALMON pretraining.

In Table 6.5, we report median statistics on the likelihood ratio for sentiment words, as defined by the LM Financial Dictionary [3], which indicates the marginal benefit that time series context provides for predicting the identity of the tokens. These results demonstrate

Word Category	Likelihood Ratio
Stop Words	0.71
Non-Sentiment	1.45
All-Sentiment	1.83
Positive	2.17
Negative	1.74
Litigious	1.76
Uncertainty	1.63
Constraining	1.23
Weak Modal	2.95
Strong Modal	1.28

Table 6.5: Median likelihood ratios across word categories from the LM Financial Dictionary [3], computed on the test set. These ratios quantify the marginal benefit of time series context for predicting words in each category, revealing the strongest gains for sentiment-charged words.

that sentiment-charged words benefit from the time series context significantly more than non-sentiment words, and dramatically more than stop words. These findings highlight the complementary nature of time series and textual data and reinforce the design of our **STW** loss function §6.4.2. The time series context provides valuable information for improving language understanding and interpreting the impact of news events.

6.7.3 Portfolio Simulations

In Table 6.6, we demonstrate the economic value of our methodology with portfolio simulations. We form monthly long-short (*market-neutral*) portfolios [34] by sorting stocks based on the 30D model predictions from the past month, detailed in §E.3. For comparison, we simulate unimodal baselines and the best performing multimodal baselines, described further in §E.3. Following Cong et al. [130], we include net performance that includes conservative estimates of the impact of transaction costs on portfolio implementation.

The test period (Jan 2020 - Dec 2024) spans 5-years of highly diverse market regimes,

including the COVID-19 crash and recovery, stimulus-driven expansion, and inflation and rate hikes, ensuring robustness across economic conditions. The resulting performance of our unified multimodal model generates investment performance that is economically and statistically better than the best performing multimodal baselines. These results demonstrate that our model provides significant predictive value in a real-world trading setting.

Method	Net Return	Volatility	Net Sharpe Ratio
TS-Only [126]	5.99	13.11	0.46
Text-Only [144]	8.60	10.47	0.82
TTSR [30]	12.37	11.28	1.10
Hybrid-MMF [186]	11.60	10.19	1.13
MTFE-MICM [38]	10.23	10.39	0.99
StockTime [206]	13.91	12.65	1.10
Proposed, MSE-ITT	**17.01**	**11.26**	**1.51**

Table 6.6: Annualized portfolio statistics of simulated investment performance, expressed in percentage units. "Net" performance includes an estimate of the impact of transaction costs, detailed in §E.3.

6.8 Conclusion

We propose a unified multimodal architecture for modeling interleaved sequences of text and time series data, and introduce a cross-modal alignment framework with a salient token weighting mechanism that learns to align representations across modalities with a focus on the most informative tokens. Our approach demonstrates state-of-the-art performance on a challenging financial forecasting task, and our ablation experiments confirm the contribution of our design decisions. These forecasting improvements translate to economically meaningful gains in portfolio simulations, underscoring the real-world value of our approach. These findings highlight the need for modality-specific structures and joint reasoning in multimodal LMs, with broader implications for domains in which text and time series co-evolve.

Chapter 7

Conclusion and Future Work

7.1 Summary of Contributions

In summary, financial data exhibits distinctive structure that requires specialized modeling approaches. Financial text is long, complex, and domain-specific, often expressing predictive signals through subtle linguistic cues, comparisons across documents, and relationships among prior events. Financial data is also inherently temporal and multimodal, comprising timestamped sequences of textual and structured numerical data. Information in one modality at a given time contextualizes contemporaneous signals in other modalities, while both must often be interpreted in the context of historical information. Therefore, effective financial forecasting requires models that can reason jointly over domain-specific language, long-range dependencies, temporal dynamics, and cross-modal interactions.

To address these challenges, this dissertation advances financial NLP and forecasting through a sequence of contributions that progressively extend the modeling framework from single documents to long-context, temporally structured, and multimodal sequences. Overall, we demonstrate that models designed and trained to capture the unique structure of financial data deliver significant improvements in forecasting and investment performance.

First, in Chapter 2 [36], we study forecasting from individual long documents. We demonstrate that language models can be adapted to the financial domain, extended to long-contexts, and trained directly on full-length transcripts to predict long-horizon financial outcomes, and that such models can outperform traditional approaches based on manually engineered financial variables. This work highlights the importance of domain adaptation and long-context modeling for capturing subtle signals embedded in financial disclosures.

Second, in Chapter 3 [37], we study forecasting settings in which predictive signals are

expressed through subtle relationships across long, complex documents. We formulate two novel document comparison tasks, predicting long-horizon company risk and correlation, that require cross-document reasoning over complex, nuanced signals. We introduce methods for aligning and comparing long documents, enabling models to capture subtle similarities and differences in content that are informative for forecasting. This work emphasizes the importance of comparative reasoning in financial analysis.

Third, in Chapter 4 [38], we extend the forecasting setting from text-only inputs to paired, temporally aligned sequences of structured and unstructured financial data. We develop models that jointly process textual and numerical inputs over time, capturing dependencies across both modality and temporal dimensions. We demonstrate that each modality contains distinct structure and predictive information, the importance of targeted multi-stage fusion for effective multimodal integration, and that the resulting predictions yield economic value in simulated trading settings. This work demonstrates the importance of modality-specific modeling and interactions between sequences of text and tabular data in dynamic settings.

Fourth, in Chapter 5 [39], we study contextualized forecasting from sequences of financial news articles. In this setting, the market impact of a given article depends not only on its own content, but also on prior related events. We propose methods for efficiently encoding and aligning historical context within a language model, addressing key challenges in long-context modeling such as positional bias, the lost-in-the-middle effect, and context overload. This work shows that targeted historical contextualization improves language modeling of current news, generalizes to longer article histories at inference time, and leads to more accurate forecasting and stronger simulated investment performance.

Finally, in Chapter 6 [40], we study multimodal forecasting from mixed-frequency, interleaved sequences of financial news and stock return time series. We propose a uni-

fied, multimodal architecture and alignment framework with modality-specific expert components that better capture the distinct statistical properties of text and time series data, while enabling joint reasoning over temporally ordered multimodal inputs. We demonstrate that parameter specialization for each modality mitigates negative transfer, while the unified architecture and dynamic cross-modal alignment framework enables more effective cross-modal learning. These gains translate to improvements in language understanding, forecasting, and investment performance.

Collectively, these works represent significant progress towards developing effective long-context and multimodal language models specialized for the distinctive structure of financial data and forecasting tasks. Across diverse settings, we demonstrate that the methods developed in this dissertation consistently enable language models to effectively extract predictive signals from both textual and structured data, and that these modeling improvements lead to economically meaningful gains in forecasting and investment performance.

7.2 Future Directions

My long-term research objective is to develop advanced AI systems that can perform complex financial reasoning, forecasting, and decision-making tasks at or exceeding expert-level performance. The work in this dissertation represents significant progress toward this goal by developing long-context and multimodal language models that can effectively extract predictive signals from financial data. However, several important challenges remain. In particular, future systems must move beyond predictive modeling alone toward models that can reason explicitly across modalities and optimize decision-making directly. We highlight three promising directions for future research below.

Chain-of-Thought Financial Reasoning First, the models developed in this dissertation are not explicitly trained to produce structured reasoning processes. Although they demonstrate strong predictive performance, their reasoning remains implicit in hidden representations rather than expressed through a sequence of interpretable intermediate steps. Given the promise of Reinforcement Learning (RL) for training language models to align human preferences [213] and solve complex math and coding problems [214], including those that cannot be solved without the generation of thousands of intermediate "thinking" tokens, this is a promising direction. More specifically, language models could be trained to generate explicit reasoning traces that mirror financial analysis: decomposing evidence from inputs, forming intermediate judgments about key drivers, and justifying how these judgments support downstream forecasts. While supervised finetuning from expert demonstrations can teach models to imitate human financial reasoning, such approaches are limited by the quality, scale, and biases of the available demonstrations, effectively imposing a ceiling on model performance. In contrast, learning from verifiable outcomes through reinforcement learning offers an opportunity to surpass expert-level performance by directly optimizing for task objectives. In particular, Reinforcement Learning from Verifiable Rewards (RLVR) [214] provides a promising framework for financial forecasting, where models can learn to generate and refine financial rationales through market feedback, enabling the emergence of reasoning processes that are both interpretable and directly optimized for forecasting performance.

Multimodal Latent Reasoning Second, while recent LLMs trained with RL have shown strong reasoning capabilities in text-only settings, they often struggle to reason over multimodal inputs, such as images, videos, and time series [215–217]. Recent work, has shown that VLMs often fail to benefit from chain-of-thought reasoning and attribute this limitation to the fact that the reasoning computation is performed in the textual

token space rather than within modality-specific representation space. In particular, they show that reasoning in the latent visual space yields improved performance [216, 217].

Our findings in [40] further highlight these limitations in the context of financial reasoning. We observe that GPT-4o [211] performs better with text-only inputs than with multimodal inputs that include time series data, suggesting that current LLMs struggle to utilize structured numerical signals effectively. Moreover, while chain-of-thought prompting improves the model's ability to reason over textual inputs, it does not lead to improvements in reasoning over time series data. These results suggest that future financial reasoning models should not rely solely on textual reasoning traces. Instead, they may need mechanisms for reasoning within modality-specific latent spaces, where models can manipulate representations of time series, tabular data, and text before integrating them into a unified forecast. This direction is particularly important for the financial domain, where meaningful signals often emerge from interactions between narrative events and noisy, temporally structured numerical data.

Decision-Aware Learning Third, the work in this dissertation primarily focuses on forecasting key financial targets, such as asset returns, risk, and related quantities. In modern investment management, these predictions are typically used as the primary inputs to a downstream portfolio construction process, where trading decisions are obtained by solving an optimization problem, most commonly through mean-variance optimization over expected returns, risk estimates, and transaction costs. Although this modular paradigm is effective and widely adopted in practice, it separates prediction from decision-making, and therefore may fail to fully capture the interactions between these components and deliver optimal outcomes.

In particular, forecasts of returns, risk, and transaction costs are inherently interdependent, and modeling them independently may result in suboptimal decisions when inte-

grated in a downstream optimization step. This motivates the development of decision-aware learning approaches in which models are trained directly to optimize decision-relevant objectives rather than intermediate prediction targets. In this paradigm, the learning objective is aligned with the ultimate goal of the system, such as maximizing risk-adjusted returns or portfolio utility.

This limitation has motivated some recent work that has begun to explore this direction. For example, [130] propose a deep learning framework that directly learns trading strategies from stock-level features, optimizing model parameters through reinforcement learning with rewards based on simulated portfolio returns. This class of approaches provides a promising direction for aligning model training with real-world objectives, and has the potential to better capture complex interactions across assets and over time.

However, decision-aware learning in financial settings remains challenging. The portfolio construction problem is inherently high-dimensional, involving a large number of assets, constraints, and transaction costs, leading to a complex and likely non-convex optimization landscape. In addition, the feedback signal is typically sparse, delayed, and noisy, as portfolio performance is only realized over time and influenced by many external factors. These properties make it difficult to train stable and generalizable models with end-to-end approaches. Future work may therefore require reinforcement learning frameworks tailored to financial environments, including methods for robust reward modeling under noise and more precise credit assignment over long horizons with delayed and stochastic feedback.

Ultimately, the development of decision-aware models that can integrate forecasting, reasoning, and optimization into a unified learning framework would represent an important step towards AI systems capable of making effective financial decisions at scale in real-world environments.

Appendix A

Long-Context Modeling of Earnings Calls

A.1 Implementation Details and Training Process

We use Scikit-learn and XGBoost to develop the non-neural baseline models. We develop all Transformer-based models in PyTorch and source all pretrained checkpoints from HuggingFace.

For the BOW models, we remove stop words, create both unigrams and bigrams from the resulting 50,000 most frequent phrases vocabulary, and apply Term Frequency-Inverse Document Frequency weighting [TF-IDF; 71, 218] to create features. For the CNN model, we initialize the word embeddings with pretrained weights from GLOVE [100D; 219], select the 50,000 most frequent words as our vocabulary, and truncate all transcripts after the first 12,000 words.

We perform all experiments on a single Tesla A100 GPU with 40GB in memory. We use AdamW to optimize all parameters. We tune the hypeparamters of each neural model by conducting a limited grid search over learning rates $\in \{5e-6, 1e-5, 5e-5\}$, weight

decay $\in \{1e-4, 1e-3, 1e-2\}$ and batch size $\in \{32, 64, 128\}$, based off validation set accuracy score. For computational constraints, we train all models using FP16 precision training, and apply gradient checkpointing to satisfy GPU memory constraints, and clip gradient norms. It takes approximately 10 minutes per epoch of supervised finetuning for the Hierarchical Transformer models and 15 minutes per epoch of training for BigBird with block size of 64.

We conduct the MLM pretraining process for BigBird on the training set for a maximum of 10 epochs or until the MLM loss on the validation set increases. This pretraining process takes multiple days of run time and indicates the difficulty of pretraining these Efficient Transformers models on domain relevant text. We tune the block size over $\{32,\ 64,\ 84\}$ and the number of random blocks over $\{3,\ 4,\ 5\}$.

Appendix B

Cross-Document Modeling of Financial Reports

B.1 Data Curation

To extract the MDA section from the HTML files, we begin by searching for strings that begin with "Item 7: Management Discussion and Analysis" and conclude with "Item 7A: Quantitative and Qualitative Disclosures", as well as other variations of these patterns in a refined and iterative process to achieve the best coverage. This process required an extensive amount of text processing that was required to extract the relevant sections required many different regular expressions, extensive trial-and-error, and a significant amount of manual quality filtering. We next pair reports for companies based on their fiscal calendar and reporting dates, allowing for delays and differences in publication dates. Finally, we filter the section text for validity and quality, such as ensuring each text has at least 500 words. We also choose focus on annual rather than quarterly reports because their formatting is more standardized and consistent.

B.2 Text-Based Baseline Models

We use Scikit-learn develop the BoW models. We apply the following text preprocessing steps to create input features: remove stop words and rare words; create both unigrams and bigrams; and apply Term Frequency-Inverse Document Frequency weighting [TF-IDF; 71, 218].

We develop the neural models in PyTorch and source pretrained checkpoints from HuggingFace. We perform several variations of the Longformer-Diff model over different ways to measure sentence-sentence similarity, only reporting the configuration with the best result on the validation set in Main Results for brevity, including Jaccard Similarity and Cosine Similarity between SBERT pretrained sentence embeddings. We also vary the cutoff threshold over $\{0.10, 0.25, 0.50, 0.75, 0.90\}$ to define a sentence in the current report that is sufficiently different from those in the previous report.

We use an Expanding training/validation window for the best performing model (CDSA-FinDiffCSE + Expanding) to simulate a live trading setting in which we do walk-forward prediction by expanding the training and validation set by 1-year as we predict on the next year of the test set. For instance, when we make predictions on the test set for 2018, we use training data from 2010-2016 and 2017 as validation data. We only due this for the best performing model to provide a proof-of-concept because it is too computationally expensive to do for all models.

B.2.1 Financial-Based Baseline Model

We select 10 commonly used market price and accounting-based financial variables available at the time of the report from the literature [91], including dividend yield, valuation, growth, profitability, medium-term price momentum, short-term price reversal, volatility, leverage, liquidity, and size. This baseline is not intended to be comprehensive

in including all possible relevant financial variables to the prediction task but rather to serve as a reasonable baseline approximating common risk factor models employed in the financial industry against which to reference and compare the value of text-based models. There may be other relevant financial variables such as those source from the options or corporate credit market to which we do not have access and is out of the scope of this text-based focused work.

B.3 Training Details and Hyperparameter Tuning

All neural network-based experiments are performned on a single Tesla A100 GPU with 40GB in memory and use AdamW to optimize all parameters. We tune the hypeparamters with a grid search over learning rates $\in \{3e-5, 5e-5, 7e-5\}$, weight decay $\in \{1e-3, 1e-2\}$ and batch size $\in \{32, 64\}$, based off validation set performance. We train all models for 10 epochs and select the best checkpoint based off validation set performance for test evaluation. For computational constraints, we use mixed precision training and gradient checkpointing to satisfy GPU memory constraints. It takes approximately 30 minutes per epoch of supervised finetuning for the sentence-level models and 60 minutes per epoch for the document-encoder models.

B.4 DAPT Pretraining Details

We conduct the DAPT process for the document-level, Longformer backbone models for a maximum of 25K training steps or until the loss on the validation set increases, using the same hyperparameter configuration and settings as Caciularu et al. [85]. This pretraining process takes multiple days of run time for each framework and indicates the difficulty of pretraining these Efficient Transformers models on domain relevant text.

We conduct the DAPT process for the sentence encoder with the DiffCSE framework for a maximum of 100K training steps or until validation loss increases. For the Longformer DAPT w/ Diffcse model, we use Longformer base as the fixed generator (masked language model) model because there are no widely accepted distilled or smaller versions. For both sentence and document encoders, we tune the RTD loss weight in the DiffCSE objective over $\{0.01, 0.05, 0.10, 0.50\}$ according to validation set performance. Please see Chuang et al. [102] for more details on the framework.

Appendix C

Multimodal Modeling of Paired Text and Time Series

C.1 Portfolio Simulations

In Table 4.5, we demonstrate the economic value of our model predictions using portfolio simulations. We form monthly long-short (market-neutral) quintile portfolios [34] by sorting stocks based on (test set) model predictions from the most recently reported quarter, and buying those in the top 20% and shorting those in the bottom 20% on a monthly basis in equal proportions.

In Table 4.5, we report net portfolio performance that includes conservative estimates of the impact of transactions costs on portfolio implementation. We follow the turnover-based method used in Cong et al. [130], which conservatively estimates the annual transaction cost as 0.01 times the annual 1-way portfolio turnover, to compute net returns. Since the data is updated quarterly and the investment universe consists of large, liquid US stocks, the proposed trading strategy is likely to have relatively low turnover and incur very modest transaction costs.

C.2 Data Curation

We merge all data across modalities to align with the quarterly fiscal period end dates for each company. Therefore, for each Earnings Surprise forecast quarter date (e.g. Q2 2020), we select data that was available within 5 business days of the end of the previous quarter fiscal end date, including all financial variables, quarterly financial reports, earnings conference calls, and macroeconomic data. Although some of this data is available at a higher frequency than once a quarter, such as stock price-based financial variables or monthly economic indicators, we only select the values of such data that were available as of 5 business days after the last fiscal period end date. Thus, our predictions are always made 3 months in advance of the earnings report date. We leave it to future work to explore the value in updating the predictions at a higher frequency.

C.3 Pretrained Language Models

We develop all Transformer-based models in PyTorch and source all pretrained checkpoints from HuggingFace.

C.4 Textual Feature Extraction

We include pretrained financial sentiment classifier **FinBERT-Sent + Linear** [5] applied at the sentence-level [91]:

$$\text{FinBERT-Sent} = \frac{\#\text{PositiveSentences - }\#\text{NegativeSentences}}{\#\text{TotalSentences}}$$

C.5 Firm Financial Variables

We select 15 commonly used market price and accounting-based financial variables available at the time of the report date from the definitions and cluster classifications in Swade et al. [123]. This set includes dividend yield (Value), earnings-to-price (Value), sales-to-price (Value), book value-to-price (Value), sales growth (Growth), earnings growth (Growth), gross profit to assets (Profitability), net income to equity (Profitability), net income to assets (Profitability), medium-term price momentum (Momentum), short-term price reversal (Reversal), price volatility (Low Risk), market leverage (Debt Issuance), share turnover (Low Risk), and market capitalization (Size). This set of variables is not exhaustive but has been shown to span the set of principle components of firm characteristics.

C.6 Training Details and Hyperparameter Tuning

We perform all experiments on a single Tesla A100 GPU with 40GB in memory. We use AdamW to optimize all parameters. We conduct an extensive grid search over the neural architecture of the time series encoder for each modality in isolation, including number of layers $\{1, 2, 4, 6, 8, 10\}$ and hidden sizes $\{16, 32, 64, 128, 256, 512\}$, as well as learning rates $\{1e-5, 5e-4, 1e-4, 5e-3, 1e-3, 5e-3\}$ and batch sizes $\{32, 64, 128, 256\}$. We train all models for 20 epochs and select the best checkpoint based off validation set performance for test evaluation. For computational constraints, we train all models using mixed precision training, and apply gradient checkpointing to satisfy GPU memory constraints, and clip gradient norms.

C.7 MT-DA Pretraining Details

We conduct the MT-DA pretraining process for the document-level, BigBird backbone models for a maximum of 25K training steps or until the loss on the validation set increases, using the same hyperparameter configuration and settings as Chuang et al. [102]. This pretraining process takes multiple days of run time for each framework and indicates the difficulty of pretraining these Efficient Transformers models on domain relevant text. We adapt the DiffCSE objective to the long context BigBird model by prepending and assigning global attention to the original document embedding in the RTD objective to encourage the model to use that information to predict the replaced tokens, resulting in more fine-grained document representations. We conduct this pretraining process over a pooled corpus of in-domain ECTs and QFRs that occur during the training date period for a total of 25K training steps. We use pretrained checkpoint of BigBird-base as the fixed generator (masked language model) model because there are no widely accepted distilled or smaller versions. We tune the tradeoff between the LC-MLM and LC-DiffCSE loss weight in the multitask DA objective over $\{0.1, 0.25, 0.5, 0.75, 0.90\}$ according to validation set performance. Please see Chuang et al. [102] for more details on the DiffCSE framework.

Appendix D

Context-Aware Modeling of Financial News

D.1 Portfolio Simulations

In Table 5.6, we demonstrate the economic value of our model predictions using portfolio simulations. In these results, all contextualization methods use *Time* as the retrieval method (i.e. we retrieve the 5 most recent articles about the same company as the historical context), except for the last row which uses our proposed hybrid retrieval method to perform the context article selection.

First, we filter the investment universe of stocks according to sufficient liquidity requirements to ensure feasibility of portfolio implementation, including a minimum market capitalization of $250M and daily average value of shares traded of $1M.

Then, we form monthly long-short (market-neutral) quintile portfolios according to Fama and French [34]. To do so, we sort stocks based on the average 30D model predictions from news articles about companies in the past 1 month. We demean score values by industry averages to remove any potential effect of industry bets or risks. Then our

portfolios are formed by buying those in the top 20% of average scores and shorting those in the bottom 20% of average scores on a monthly basis in equal proportions. Please note that these portfolios are market-neutral and therefore have essentially no correlation with broad market indices.

In Table 5.6, we include conservative estimates of the impact of transactions costs on portfolio implementation. We follow the turnover-based method used in Cong et al. [130], which conservatively estimates the annual transaction cost as 0.01 times the annual 1-way portfolio turnover.

D.2 Fama-French 6-Factor Model (FF-6)

We use the Fama-French 6-Factor model [177], which consists of 6 financial factors, including size, beta, value, momentum, profitability, and investment, to make predictions of stock returns by estimating the linear regression model over the 10 years prior to the start of the simulation (Jan 2007 - Dec 2016) using monthly stock returns. We use the estimated model to make forward stock return predictions using these 6 financial factors on a monthly basis. Then, we form long-short quintile portfolios based on the monthly predictions [34].

D.3 Data Curation

Since FactSet StreetAccount uses human curation to only provide relevant, market-moving news in the newsfeed, we do not perform any additional steps to filter the data for impactful news.

Following Chen et al. [144], if the news article was published after market open (9:30 am), then we consider the news article to be published the next trading day. We only

include articles originally written in English according to the following criteria [144]: they are tagged relevant for only one company; they are longer than 100 characters or shorter than 10,000 characters; they contain less than 10% of numerical characters; they have less than 90% Jaccard similarity to a previous article (to remove potential duplicate articles). We require that each main article possess at least 5 historical articles about the same company within the past year as well as available stock returns.

There is a slight class imbalance so we randomly downsample the majority class to ensure balanced class ratios. All models are optimized using Binary Cross-Entropy as the loss function.

D.4 Pretrained Language Models

We develop all Transformer-based models in PyTorch and source all pretrained checkpoints from HuggingFace.

D.5 Prefix Summary Context

In the Prefix Summary Context (PSC) method, we tune the number of attention heads in the alignment module (AM) over {2, 4, 8, 16} and the number of summary embeddings per historical context article over {1, 2, 5, 10} according to validation set performance.

D.6 Architecture Ablations

In Table D.1, we ablate the choice of the alignment module between the representation space of the context summarizer and large LM, including both a **Linear** projection

and also a fully connected **MLP** network. We find that the use of the cross-attention mechanism in the alignment module to be important for convergence, particularly for direct finetuning without CALM pretraining. We conjecture that this effectively forces the resulting attention output to be within the convex hull spanned by the large model's token embeddings, and suspect that this makes them more likely to be in-distribution and thus receive attention weight. We expect this improvement to be more pronounced when using common parameter efficient finetuning methods that do not update the LMs token embeddings, not allowing them to adapt distributions to the new prefix. The challenges of direct finetuning from continuous prompts [220] and different modality spaces [136, 138] have been confirmed in the literature, which consequently use a staged training approach.

Alignment Method	**30D AUC**
Linear	57.50
MLP	57.37
Proposed-CMA	58.33
Linear + CALM	58.29
MLP + CALM	58.67
Proposed-CMA + CALM	59.12

Table D.1: Results indicate the 30D AUC performance of different alignment methods between the output of the context summarizer and the input embeddings of the large LM with and without CALM pretraining.

D.7 Pretrained Baselines

We include pretrained financial sentiment classifier **FinBERT-Sent + Linear** [5] applied at the sentence-level [91]:

$$\text{FinBERT-Sent} = \frac{\#\text{PositiveSentences - }\#\text{NegativeSentences}}{\#\text{TotalSentences}}$$

To be consistent across all text embedding models, we apply mean pooling at the

token-level to aggregate contextualized token representations to the sequence level.

For some baseline models in which additional inputs or modalities are incorporated, such as proprietary sentiment scores or stock price time series, we do not include these additional inputs (and we do not have access to them) in the model in order to present a fair comparison of model types and isolate the effects of historical contextualization.

D.8 Prompting

For the zero-shot Llama-3-8B and Llama-3-70B baselines, we have explored a variety of prompting strategies, listed below. We do not find much variation in the performance of each, but find (1) developed by Lopez-Lira and Tang [146], to perform best according to the validation set and report those results in Table 5.2.

(1) `Forget all your previous instructions. Pretend you are a financial expert. You are a financial expert with stock recommendation experience. Answer "YES" if good news, "NO" if bad news in the first line. Then elaborate with one short and concise sentence on the next line. Is this news article good or bad for the stock price of {company name} in the next {time horizon} days? Article: {news article}`

(2) `Classify the following financial news article(s) as 'Positive' if it will lead to a positive stock return in the next {time horizon} days and 'Negative' if it will lead to a negative stock return in the following {time horizon} days. Please only respond with 'Positive' or 'Negative': {news article}.`

(3) `Predict if the following financial news article is 'Positive' if it will cause a positive stock return in the next {time horizon}`

days and 'Negative' if it will cause a negative stock return in the following {time horizon} days. Please only respond with 'Positive' or 'Negative': {news article}.

(4) You are a financial analyst with expertise in market reactions. Based on the following news article(s), predict the likelihood of a positive stock return for {company name} in the next {time horizon} days. Respond with a probability between 0% and 100%. Article: {news article}.

For each prompt response (except for #4), we extract the softmax probabilities for the two corresponding positive and negative tokens and re-normalize to arrive at a continuous score. If historical context is provided, then the historical news articles are prepended to the main article.

D.9 Statistical Significance

In Table 5.2, we report the standard deviation of results across 3 different training runs with different random seeds. The variability of results from the random seeds results from the randomness in the training process caused by random initialization of the classification layer weights and the random order of training samples during stochastic gradient descent optimization. We also report the results from the Wilcoxon Signed-Rank test, which is a pairwise test of statistical significance conducted with bootstrapped samples from the test set.

D.10 Hierarchical Encoding

For the hierarchical approach, we select a randomly-initialized, causal Transformer to globally contextualize the local (article-level) token representations. We apply mean pooling across the token embeddings of the most recent article a_t. We tune the number of layers $\in \{2, 4, 8\}$ and the number of attentions head $\in \{2, 4, 8\}$ according to validation set performance.

D.11 Retrieval Models

For Instructor embedding models [175] for context retrieval, we use the following prompt: "Represent the financial news article for retrieval: {news article}". We use the instructor-base to ensure similar parameter counts across retrieval models.

For FinSim, we conduct the DiffCSE-Domain pretraining process over the training data for a maximum of 25K training steps or until the loss on the validation set increases, using the same hyperparameter configuration and settings as Chuang et al. [102]. We initialize the model from the pretrained checkpoint of DeBERTa-base and additionally use that as the fixed generator (masked language model) model. We tune the tradeoff between the contrastive and replaced token detection (RTD) objective over $\{0.10, 0.25, 0.50, 0.75, 0.90\}$ according to validation set performance. Please see Chuang et al. [102] for more details on the framework.

For TimeFinSim, we use the time distance from the historical news article date to the main article date to apply an exponential decay (half-life = 180 days) to the FinSim similarity score to capture both the similarity and timeliness of the historical context. Given main article a, historical context c, time elapsed t, and half-life H:

$$\mathrm{TimeFinSim(a, c, t)} = \mathrm{FinSim(a, c)} \cdot \mathrm{e}^{-\frac{\ln(2)}{\mathrm{H}}\mathrm{t}}$$

D.12 Training Details and Hyperparameter Tuning

We perform all experiments on a single Tesla A100 GPU with 80G in memory. We use AdamW to optimize all parameters. For all finetuned models, we use an effective batch size of 32 with gradient accumulation. We train all models for up to 5 epochs with early stopping based off validation set performance for test evaluation.

For long-context models based on Mistral-7B, we finetune with low-rank adapters (LoRA) [165]. We tune the learning rate over {1e-6, 3e-6, 5e-6, 7e-6, 1e-5} and LoRA rank parameter over $r \in \{16, 32, 64\}$ according to validation set performance. We apply LoRA adapters to all linear layers of the base model.

In all main baselines, based on Mistral-7B, we apply mean pooling across the hidden states of the main article tokens, which is then passed to a classification layer to make a binary prediction.

For computational constraints, we train all models using mixed precision training and gradient checkpointing to satisfy GPU memory constraints, and clip gradient norms. For DeBERTa-based models, we finetune in FP16 precision, while we use BF16 for Mistral-based models. For LoRA-based finetuning, we always set the value of the alpha parameter to be equal to double the value of rank parameter.

Appendix E

Multimodal Modeling of Interleaved Text and Time Series

E.1 Pretrained Language Models

We implement all models in PyTorch and source all pretrained checkpoints from HuggingFace.

E.2 Baseline Models

For our proposed model (MSE-ITT), for supervised classification, we train a classification head on top of the end-of-sequence token's last hidden representations to make binary predictions.

For some baseline models in which additional inputs or modalities are incorporated, such as proprietary sentiment scores [189], we do not include these additional inputs (and we do not have access to them) in the model in order to present a fair comparison of model types and isolate the effects of multimodal fusion between sequences of text

and time series. Some baseline models [187–189] explore the use of various LLMs for textual encoding. In such cases, we select the LM with the best reported performance for evaluation. For models that patch the time series input into non-overlapping consecutive chunks [30, 206], if the patch length used is not provided in the implementation details, then we tune the value over $\{1, 5, 10\}$ based on validation set performance. Further, for some multivariate time series baseline models that require a one-to-one mapping between text and time series inputs across time steps [38, 188], if there are no news articles on a given time step, then we simply carry forward the previous news article (embeddings) until the next news article (embeddings) are available. Additionally, in the MAT baseline [207], the authors do not report the number of topics or the base pretrained LM used in BERTTopic [221], so we resort to using the hidden representations produced from their sentiment model FinBERT [5] as text features. Further, for the FinMA [145] baseline, we provide the text and time series sequences in the same format as for GPT-4o §E.5, such that the time series is converted to raw text and interleaved between news articles in temporal order.

For our proposed model, we use learnable special tokens to denote the beginning and end of each news article and modality [85], and include the article timestamp in the article text.

E.3 Portfolio Simulations

In Table 6.6, we demonstrate the economic value of our model predictions using portfolio simulations.

To perform these simulations in a realistic setting, we first filter the investment universe of stocks according to sufficient liquidity requirements to ensure feasibility of portfolio implementation, including a minimum market capitalization of $250M and daily

average value of shares traded of $1M.

Then, we form monthly long-short (market-neutral) quintile portfolios according to Fama and French [34]. To this end, we sort stocks based on the average 30D model predictions from news articles about companies in the past 1 month. Then our portfolios are formed by buying those in the top 20% of average scores and shorting those in the bottom 20% of average scores on a monthly basis in equal proportions. Please note that these portfolios are market-neutral and therefore have essentially no correlation with broad market indices.

In Table 6.6, we include conservative estimates of the impact of transactions costs on portfolio implementation. We follow the turnover-based method used in Cong et al. [130], which conservatively estimates the annual transaction cost as 0.01 times the annual 1-way portfolio turnover. Therefore, the net return of the portfolio is the gross return minus the estimated transaction costs.

E.4 Data Curation

The FNSPID dataset spans multiple news sources, including Nasdaq, Reuters, CNBC, Benzinga, and cover a variety of company events, including product launches, earnings reports, and mergers. We truncate each article after the first 128 words during all experiments for computational efficiency. We only include articles originally written in English according to the following criteria [144]: they are tagged relevant for only one company; they are longer than 100 characters or shorter than 10,000 characters; they contain less than 10% of numerical characters; they have less than 90% Jaccard similarity to a previous article (to remove potential duplicate articles). We require that each company possess at least 5 news articles within the past year as well as available stock returns to be in our sample. In addition, we apply further quality filtering to ensure that

the dataset contains high-quality, event-driven news articles that are specific to individual stocks. We systematically remove broad market summaries, sector-level reports, and generic financial commentary by filtering out headlines containing specific keywords and patterns. This process allows us to isolate stock-specific events. There is a slight class imbalance so we randomly downsample the majority class to ensure balanced class ratios.

E.5 Zero-Shot LLMs

For the zero-shot GPT-4o baselines, we have explored a variety of prompting strategies, listed below. For this set of baselines, the time series inputs are converted to text strings and we tune their form {decimal, percentage}, digits of precision {2, 3, 4}, and whether to include a space in between digits, with validation set performance [184, 185]. We do not find much variation in the performance of each, but tune them according to the validation set and report those results in Table 6.2.

To this end, we construct a natural language prompt that interleaves company-specific financial news with daily stock return sequences in chronological order. The model is tasked with predicting future price movement on a bounded scale.

Each prompt follows the format:

```
Given the company's daily stock return over the last year
interleaved in temporal order with recent news about the company,
sorted chronologically:

{interleaved_input}

Predict the company's future stock price movement over the next
{horizon} days, on a scale from 0 to 100, where 100 indicates strong
positive movement and 0 indicates strong negative movement.
{prompt_style}
```

```
Prediction:
```

The field {`interleaved_input`} contains alternating sequences of news articles and their corresponding stock returns (e.g., `News Article: ... Stock Returns: ...`). The field {`prompt_style`} controls the reasoning strategy used by the model and takes one of two forms:

- **Direct**: The model immediately outputs a prediction.
- **Chain-of-Thought (CoT)**: The model is instructed to "think step-by-step" before producing a prediction, following recent findings that CoT improves LLM reasoning over time series data [205].

This design enables a uniform comparison of reasoning strategies across interleaved multimodal inputs, without any task-specific fine-tuning.

E.6 Statistical Significance

In Table 6.2, we report the sample standard deviation of results for our proposed method across 3 different training runs with different random seeds. The variability of results across random seeds stems from the randomness in the training process caused by random initialization of the classification layer weights and the random batch order of training samples during stochastic gradient descent optimization. We also report the results from DeLong's pairwise test of statistical significance between model AUC scores [222].

E.7 Time Series Discretization

For time-series discretization, we tune the number of discrete time-series bins $B \in \{4, 8, 16, 32, 64\}$ and their embedding dimension $d_{ts} \in \{32, 64, 128, 256, 512\}$. To map the time-series embeddings into the token embedding space of the LM, we learn a simple linear map to increase dimensionality from d_{ts} to d_{text}.

In Table E.1, we compare our discretization approach to alternative continuous embedding methods. While the benefits of discretization are indeed modest and clearly not the primary source of our performance gains, they dramatically simplify and unify the interleaved multimodal (SALMON) pretraining objective. By converting all tokens, regardless of modality, to a common discrete space, we can apply a consistent cross-entropy loss to all discrete tokens, without inducing a complex multitask setup that requires careful calibration of continuous and discrete loss terms. In the continuous setting, we tune the weight on the TS (Mean Squared Error) loss over $w_{ts} \in \{0, 10, 100, 1000\}$ according to validation set performance.

Method	**LM Loss**	**7D**	**30D**
Discrete	1.78	57.94	58.48
Linear	1.83	57.54	58.10
MLP	1.83	57.62	58.13

Table E.1: Results demonstrate the value of time series discretization in our multimodal architecture in both language understanding (LM Loss) and financial forecasting (7D, 30D) performance, compared to baselines that use continuous embeddings via linear and nonlinear transformations. These results include SALMON pretraining.

E.8 Training Details and Hyperparameter Tuning

We perform all experiments on a single NVIDIA H100 GPU with 80G in memory. We use AdamW to optimize all parameters. For all finetuned models, we use an effective

batch size of 64 with gradient accumulation. We train all models for up to 5 epochs based on validation set performance for test evaluation. All supervised models are optimized using Binary Cross-Entropy as the loss function.

We tune the learning rate over {1e-6, 3e-6, 5e-6, 7e-6, 1e-5} and LoRA [165] rank parameter over $r \in \{16, 32, 64\}$ according to validation set performance. We apply LoRA adapters to all linear layers. We use a simple 2-layer MLP classification layer to project the token hidden states to make a binary prediction for all language models used.

For computational constraints, we train all models using mixed precision training and gradient checkpointing to satisfy GPU memory constraints, and clip gradient norms. For Llama-based models, we finetune in BF16 precision. For LoRA-based finetuning, we always set the value of the alpha parameter to be equal to double the value of rank parameter.

E.9 Multivariate Time Series

In this section, we demonstrate that our MSE-ITT model architecture can support numerical time series with multiple channels.

To this end, we select 15 commonly used market price and accounting-based financial variables available at the time of the report date from the definitions and cluster classifications in Swade et al. [123]. This set includes dividend yield (Value), earnings-to-price (Value), sales-to-price (Value), book value-to-price (Value), sales growth (Growth), earnings growth (Growth), gross profit to assets (Profitability), net income to equity (Profitability), net income to assets (Profitability), medium-term price momentum (Momentum), short-term price reversal (Reversal), price volatility (Low Risk), market leverage (Debt Issuance), share turnover (Low Risk), and market capitalization (Size).

Since some variables have different frequencies, ranging from daily to quarterly, we

up-sample them all to daily frequency by forward-filling previous values.

We fit different discretized embeddings for each channel following our approach in §6.4. After this, we simply concatenate their embeddings together vertically at a daily frequency and similarly interleave with the news articles according to timestamp. To extend our SALMON pretraining objective to the multivariate setting, we learn a separate output projection head for each channel and compute the discretized token for each channel independently and average the loss across channels.

Method	**LM Loss**	**30D**
Univariate	–	57.14
Univariate w/ SALMON	1.78	58.48
Multivariate	–	58.45
Multivariate w/ SALMON	1.74	59.89

Table E.2: Results indicate the performance of our MSE-ITT multimodal architecture when extended to the multivariate setting in both language understanding (LM Loss) and financial forecasting (30D) performance.

Bibliography

[1] Andrew Y Chen and Tom Zimmermann. Open source cross-sectional asset pricing. Critical Finance Review, Forthcoming, 2021.

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

[3] Tim Loughran and Bill McDonald. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. The Journal of finance, 66(1):35–65, 2011.

[4] Sean Cao, Wei Jiang, Baozhong Yang, and Alan L. Zhang. How to talk when a machine is listening: Corporate disclosure in the age of AI. The Review of Financial Studies, 36(9):3603–3642, 2023. doi: 10.1093/rfs/hhad021.

[5] Dogu Araci. Finbert: Financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:1908.10063, 2019.

[6] Allen H Huang, Hui Wang, and Yi Yang. Finbert: A large language model for extracting information from financial text. Contemporary Accounting Research, 2022.

[7] Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. FinGPT: Open-source financial large language models, 2023. URL `https://arxiv.org/abs/2306.06031`.

[8] Xiao-Yang Liu, Guoxuan Wang, Hongyang Yang, and Daochen Zha. FinGPT: Democratizing internet-scale data for financial large language models, 2023. URL `https://arxiv.org/abs/2307.10485`.

[9] Raj Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. When FLUE meets FLANG: Benchmarks and large pretrained language model for financial domain. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 2322–2335, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.148. URL `https://aclanthology.org/2022.emnlp-main.148/`.

[10] Rajdeep Mukherjee, Abhinav Bohra, Akash Banerjee, Soumya Sharma, Manjunath Hegde, Afreen Shaikh, Shivani Shrivastava, Koustuv Dasgupta, Niloy Ganguly, Saptarshi Ghosh, and Pawan Goyal. ECTSum: A new benchmark dataset for bullet point summarization of long earnings call transcripts. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 10893–10906, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.748. URL https://aclanthology.org/2022.emnlp-main.748/.

[11] Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. FinQA: A dataset of numerical reasoning over financial data. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 3697–3711, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.300. URL https://aclanthology.org/2021.emnlp-main.300/.

[12] Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. ConvFinQA: Exploring the chain of numerical reasoning in conversational finance question answering. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 6279–6292, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.421. URL https://aclanthology.org/2022.emnlp-main.421/.

[13] Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. PIXIU: A large language model, instruction data and evaluation benchmark for finance, 2023. URL https://arxiv.org/abs/2306.05443.

[14] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer, 2020.

[15] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big bird: Transformers for longer sequences. In Advances in Neural Information Processing Systems, volume 33, 2020.

[16] Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In International Conference on Learning Representations, 2020.

[17] Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity, 2020.

[18] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In International Conference on Learning Representations, 2024.

[19] Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view, 2024.

[20] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. arXiv preprint arXiv:2307.03172, 2023.

[21] Annan Yu, Danielle C Maddix, Boran Han, Xiyuan Zhang, Abdul Fatir Ansari, Oleksandr Shchur, Christos Faloutsos, Andrew Gordon Wilson, Michael W Mahoney, and Yuyang Wang. Understanding transformers for time series: Rank structure, flow-of-ranks, and compressibility. arXiv preprint arXiv:2510.03358, 2025.

[22] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the Next Token and Diffuse Images with One Multi-Modal Model, 2024. URL https://arxiv.org/abs/2408.11039.

[23] Weijia Shi, Xiaochuang Han, Chunting Zhou, Weixin Liang, Xi Victoria Lin, Luke Zettlemoyer, and Lili Yu. LMFusion: Adapting Pretrained Language Models for Multimodal Generation, 2024. URL https://arxiv.org/abs/2412.15188.

[24] Xi Victoria Lin, Akshat Shrivastava, Liang Luo, Srinivasan Iyer, Mike Lewis, Gargi Ghosh, Luke Zettlemoyer, and Armen Aghajanyan. MoMa: Efficient Early-Fusion Pre-training with Mixture of Modality-Aware Experts, 2024. URL https://arxiv.org/abs/2407.21770.

[25] Weixin Liang, Junhong Shen, Genghan Zhang, Ning Dong, Luke Zettlemoyer, and Lili Yu. Mixture-of-Mamba: Enhancing Multi-Modal State-Space Models with Modality-Aware Sparsity. arXiv preprint arXiv:2501.16295, 2025.

[26] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. arXiv preprint arXiv:1701.06538, 2017.

[27] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. Journal of Machine Learning Research, 23(120):1–39, 2022.

[28] Mingtian Tan, Mike A. Merrill, Vinayak Gupta, Tim Althoff, and Tom Hartvigsen. Are Language Models Actually Useful for Time Series Forecasting? In Amir Globerson, Lihua Mackey, Danielle Belgrave, Alice Fan, Ulrich Paquet, Jakub Tomczak, and Cheng Zhang, editors, Advances in Neural Information Processing Systems 37, pages 60162–60191. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/6ed5bf446f59e2c6646d23058c86424b-Abstract-Conference.html.

[29] Mike A. Merrill, Mingtian Tan, Vinayak Gupta, Tom Hartvigsen, and Tim Althoff. Language Models Still Struggle to Zero-shot Reason about Time Series. In Findings of the Association for Computational Linguistics: EMNLP 2024, pages 3512–3533, Miami, Florida, USA, 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-emnlp.201.

[30] Winnie Chow, Lauren Gardiner, Haraldur T. Hallgrímsson, Maxwell A. Xu, and Shirley You Ren. Towards Time Series Reasoning with LLMs. arXiv preprint arXiv:2409.11376, 2024. Presented at the NeurIPS 2024 Workshop on Time Series in the Age of Large Models.

[31] Shihao Gu, Bryan Kelly, and Dacheng Xiu. Empirical asset pricing via machine learning. The Review of Financial Studies, 33(5):2223–2273, 2020.

[32] Rob Arnott, Campbell R. Harvey, and Harry Markowitz. A backtesting protocol in the era of machine learning. The Journal of Financial Data Science, 1(1):64–74, 2019. doi: 10.3905/jfds.2019.1.064.

[33] Paul Glasserman and Caden Lin. Assessing look-ahead bias in stock return predictions generated by gpt sentiment analysis. The Journal of Financial Data Science, 6(1):25–42, 2024. doi: 10.3905/jfds.2023.1.143.

[34] Eugene F Fama and Kenneth R French. A five-factor asset pricing model. Journal of financial economics, 116(1):1–22, 2015.

[35] Yining Cong, Yichen Tang, Zhen Wang, and Jiebo Luo. Alphaportfolio: Direct construction through reinforcement learning and interpretable ai. Proceedings of the AAAI Conference on Artificial Intelligence, 35(1):417–424, 2021.

[36] Ross Koval, Nicholas Andrews, and Xifeng Yan. Forecasting earnings surprises from conference call transcripts. In Findings of the Association for Computational Linguistics: ACL 2023, pages 8197–8209, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.520. URL https://aclanthology.org/2023.findings-acl.520.

[37] Ross Koval, Nicholas Andrews, and Xifeng Yan. Learning to compare financial reports for financial forecasting. In Yvette Graham and Matthew Purver, editors,

Findings of the Association for Computational Linguistics: EACL 2024, pages 500–512, St. Julian's, Malta, March 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-eacl.34.

[38] Ross Koval, Nicholas Andrews, and Xifeng Yan. Financial forecasting from textual and tabular time series. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, Findings of the Association for Computational Linguistics: EMNLP 2024, pages 8289–8300, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.486. URL https://aclanthology.org/2024.findings-emnlp.486/.

[39] Ross Koval, Nicholas Andrews, and Xifeng Yan. Context-aware language models for forecasting market impact from sequences of financial news. arXiv preprint arXiv:2509.12519, 2025.

[40] Ross Koval, Nicholas Andrews, and Xifeng Yan. Multimodal language models for financial forecasting from interleaved sequences of text and time series. In Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 987–1001, 2025.

[41] Stephen Brown, Stephen A Hillegeist, and Kin Lo. Conference calls and information asymmetry. Journal of Accounting and Economics, 37(3):343–366, 2004.

[42] David F Larcker and Anastasia A Zakolyukina. Detecting deceptive discussions in conference calls. Journal of Accounting Research, 50(2):495–540, 2012.

[43] Brian J Bushee, Ian D Gow, and Daniel J Taylor. Linguistic complexity in firm disclosures: Obfuscation or information? Journal of Accounting Research, 56(1): 85–121, 2018.

[44] Jeffrey T Doyle, Russell J Lundholm, and Mark T Soliman. The extreme future stock returns following i/b/e/s earnings surprises. Journal of Accounting Research, 44(5):849–887, 2006.

[45] S McKay Price, James S Doran, David R Peterson, and Barbara A Bliss. Earnings conference calls and stock returns: The incremental informativeness of textual tone. Journal of Banking & Finance, 36(4):992–1011, 2012.

[46] Tim Loughran and Bill Mcdonald. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. Journal of Finance, 66, 2011. ISSN 00221082. doi: 10.1111/j.1540-6261.2010.01625.x.

[47] Zheng Tracy Ke, Bryan T Kelly, and Dacheng Xiu. Predicting returns with text data. Technical report, National Bureau of Economic Research, 2019.

[48] Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT, pages 4171–4186, 2019.

[49] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692, 2019.

[50] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4):1234–1240, 2020.

[51] Jiuxiang Gu, Jason Kuen, Vlad I. Morariu, Handong Zhao, Nikolaos Barmpalios, Rajiv Jain, Ani Nenkova, and Tong Sun. Unified pretraining framework for document understanding. volume 1, 2021.

[52] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. Legal-bert: The muppets straight out of law school. arXiv preprint arXiv:2010.02559, 2020.

[53] Linyi Yang, Tin Lok James Ng, Barry Smyth, and Riuhai Dong. Html: Hierarchical transformer-based multi-task learning for volatility prediction. In Proceedings of The Web Conference 2020, pages 441–451, 2020. doi: 10.1145/3366423.3380128.

[54] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 2978–2988, 2019. doi: 10.18653/v1/p19-1285.

[55] Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big bird: Transformers for longer sequences. volume 2020-December, 2020.

[56] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1480–1489, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1174. URL https://aclanthology.org/N16-1174.

[57] Raghavendra Pappagari, Piotr Zelasko, Jesus Villalba, Yishay Carmiel, and Najim Dehak. Hierarchical transformers for long document classification. In 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pages 838–844. IEEE, 2019. doi: 10.1109/ASRU46091.2019.9003958.

[58] Liu Yang, Mingyang Zhang, Cheng Li, Michael Bendersky, and Marc Najork. Beyond 512 tokens: Siamese multi-depth transformer-based hierarchical encoder for long-form document matching. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management, pages 1725–1734, 2020.

[59] Xingxing Zhang, Furu Wei, and Ming Zhou. Hibert: Document level pre-training of hierarchical bidirectional transformers for document summarization. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 5059–5069, 2019.

[60] Andriy Mulyar, Elliot Schumacher, Masoud Rouhizadeh, and Mark Dredze. Phenotyping of clinical notes with improved document classification models using contextualized neural language models. arXiv preprint arXiv:1910.13664, 2019.

[61] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8440–8451, 2020.

[62] Richard M. Frankel, Jared N. Jennings, and Joshua A. Lee. Using natural language processing to assess text usefulness to readers: The case of conference calls and earnings prediction. SSRN Electronic Journal, 2018. doi: 10.2139/ssrn.3095754.

[63] Yu Qin and Yi Yang. What you say and how you say it matters: Predicting stock volatility using verbal and vocal cues. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 390–401, 2019. doi: 10.18653/v1/p19-1038.

[64] Ramit Sawhney, Piyush Khanna, Arshiya Aggarwal, Taru Jain, Puneet Mathur, and Rajiv Ratn Shah. Voltage: Volatility forecasting via text audio fusion with graph convolution networks for earnings calls. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 8001–8013, 2020. doi: 10.18653/v1/2020.emnlp-main.643.

[65] Ramit Sawhney, Puneet Mathur, Ayush Mangal, Piyush Khanna, Rajiv Ratn Shah, and Roger Zimmermann. Multimodal multi-task financial risk forecasting. In Proceedings of the 28th ACM international conference on multimedia, pages 456–465, 2020.

[66] Yunxin Sang and Yang Bao. Predicting corporate risk by jointly modeling company networks and dialogues in earnings conference calls. arXiv preprint arXiv:2206.06174, 2022.

[67] Linyi Yang, Jiazheng Li, Ruihai Dong, Yue Zhang, and Barry Smyth. Numhtml: Numeric-oriented hierarchical transformer model for multi-task financial forecasting. arXiv preprint arXiv:2201.01770, 2022.

[68] Henry A Latane and Charles P Jones. Standardized unexpected earnings–1971-77. The journal of Finance, 34(3):717–724, 1979.

[69] Xiaochuang Han and Jacob Eisenstein. Unsupervised domain adaptation of contextualized embeddings for sequence labeling. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 4238–4248, 2019.

[70] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don't stop pretraining: Adapt language models to domains and tasks. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8342–8360, 2020.

[71] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. Information processing & management, 24(5):513–523, 1988.

[72] Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies, pages 142–150, 2011.

[73] Itay Kama. On the market reaction to revenue and earnings surprises. Journal of Business Finance & Accounting, 36(1-2):31–50, 2009.

[74] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.

[75] Feng Li. The information content of forward- looking statements in corporate filings-a naïve bayesian machine learning approach. Journal of Accounting Research, 48, 2010. ISSN 00218456. doi: 10.1111/j.1475-679X.2010.00382.x.

[76] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?" explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pages 1135–1144, 2016. doi: 10.1145/2939672.2939778.

[77] Jason Phang, Yao Zhao, and Peter J Liu. Investigating efficiently extending transformers for long input summarization. arXiv preprint arXiv:2208.04347, 2022.

[78] Lauren Cohen, Christopher Malloy, and Quoc Nguyen. Lazy prices. The Journal of Finance, 75(3):1371–1415, 2020.

[79] Lauren Cohen and Andrea Frazzini. Economic links and predictable returns. The Journal of Finance, 63(4):1977–2011, 2008.

[80] Gerard Hoberg and Gordon Phillips. Text-based network industries and endogenous product differentiation. Journal of Political Economy, 124(5):1423–1465, 2016.

[81] Charles MC Lee, Stephen Teng Sun, Rongfei Wang, and Ran Zhang. Technological links and predictable returns. Journal of Financial Economics, 132(3):76–96, 2019.

[82] Mandy Guo, Joshua Ainslie, David C Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. Longt5: Efficient text-to-text transformer for long sequences. In Findings of the Association for Computational Linguistics: NAACL 2022, pages 724–736, 2022.

[83] Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. arXiv preprint arXiv:1708.00055, 2017.

[84] Xuhui Zhou, Nikolaos Pappas, and Noah A. Smith. Multilevel text alignment with cross-document attention. 2020. doi: 10.18653/v1/2020.emnlp-main.407.

[85] Avi Caciularu, Arman Cohan, Iz Beltagy, Matthew Peters, Arie Cattan, and Ido Dagan. CDLM: Cross-document language modeling. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2648–2662, 2021.

[86] Dvir Ginzburg, Itzik Malkiel, Oren Barkan, Avi Caciularu, and Noam Koenigstein. Self-supervised document similarity ranking via contextualized language models and hierarchical inference. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 3088–3098, 2021.

[87] Luca Di Liello, Siddhant Garg, Luca Soldaini, and Alessandro Moschitti. Paragraph-based transformer pre-training for multi-sentence inference. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2521–2531, 2022.

[88] Luca Di Liello, Siddhant Garg, Luca Soldaini, and Alessandro Moschitti. Pre-training transformer models with sentence-level objectives for answer sentence selection. arXiv preprint arXiv:2205.10455, 2022.

[89] Shimon Kogan, Dimitry Levin, Bryan R. Routledge, Jacob S. Sagi, and Noah A. Smith. Predicting risk from financial reports with regression. 2009. doi: 10.3115/1620754.1620794.

[90] Qi Feng, Han Chen, and Ruohan Jiang. Analysis of early warning of corporate financial risk via deep learning artificial neural network. Microprocessors and Microsystems, 87, 2021. ISSN 01419331. doi: 10.1016/j.micpro.2021.104387.

[91] Emmanuel Alanis, Sudheer Chava, and Agam Shah. Benchmarking machine learning models to predict corporate bankruptcy. arXiv preprint arXiv:2212.12051, 2022.

[92] Puneet Mathur, Mihir Goyal, Ramit Sawhney, Ritik Mathur, Jochen L Leidner, Franck Dernoncourt, and Dinesh Manocha. Docfin: Multimodal financial prediction and bias mitigation using semi-structured documents. In Findings of the Association for Computational Linguistics: EMNLP 2022, pages 1933–1940, 2022.

[93] Malik Magdon-Ismail and Amir F Atiya. Maximum drawdown. Risk Magazine, 17 (10):99–102, 2004.

[94] Alexei Chekhlov, Stanislav Uryasev, and Michael Zabarankin. Portfolio optimization with drawdown constraints. In Supply chain and finance, pages 209–228. World Scientific, 2004.

[95] Wesley R Gray and Jack Vogel. Using maximum drawdowns to capture tail risk. Available at SSRN 2226689, 2013.

[96] Peter Nystrup, Stephen Boyd, Erik Lindström, and Henrik Madsen. Multi-period portfolio selection with drawdown control. Annals of Operations Research, 282 (1-2):245–271, 2019.

[97] Mikica Drenovak, Vladimir Ranković, Branko Urošević, and Ranko Jelic. Mean-maximum drawdown optimization of buy-and-hold portfolios using a multi-objective evolutionary algorithm. Finance Research Letters, 46:102328, 2022.

[98] Paul Embrechts, Alexander McNeil, and Daniel Straumann. Correlation and dependence in risk management: properties and pitfalls. Risk management: value at risk and beyond, 1:176–223, 2002.

[99] Torben G Andersen, Tim Bollerslev, Peter Christoffersen, and Francis X Diebold. Practical volatility and correlation modeling for financial market risk management. In The risks of financial institutions, pages 513–548. University of Chicago Press, 2007.

[100] Dimos S Kambouroudis, David G McMillan, and Katerina Tsakou. Forecasting stock return volatility: A comparison of garch, implied volatility, and realized volatility models. Journal of Futures Markets, 36(12):1127–1163, 2016.

[101] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084, 2019.

[102] Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljačić, Shang-Wen Li, Wen-tau Yih, Yoon Kim, and James Glass. Diffcse: Difference-based contrastive learning for sentence embeddings. arXiv preprint arXiv:2204.10298, 2022.

[103] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6894–6910, 2021.

[104] Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, pages 39–48, 2020.

[105] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014.

[106] Ferhat Akbas, Chao Jiang, and Paul D Koch. The trend in firm profitability and the cross-section of stock returns. The Accounting Review, 92(5):1–32, 2017.

[107] Dashan Huang, Huacheng Zhang, and Guofu Zhou. Twin momentum: Fundamental trends matter. Available at SSRN 2894068, 2019.

[108] Tim De Silva and David Thesmar. Noise in expectations: Evidence from analyst forecasts. Technical report, National Bureau of Economic Research, 2021.

[109] Jules H Van Binsbergen, Xiao Han, and Alejandro Lopez-Lira. Man versus machine learning: The term structure of earnings expectations and conditional biases. The Review of Financial Studies, 36(6):2361–2396, 2023.

[110] Ryan T Ball and Eric Ghysels. Automated earnings forecasts: Beat analysts or combine and conquer? Management Science, 64(10):4936–4952, 2018.

[111] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. arXiv preprint arXiv:2310.06825, 2023.

[112] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Improving text embeddings with large language models. arXiv preprint arXiv:2401.00368, 2023.

[113] Robert Novy-Marx. The other side of value: The gross profitability premium. Journal of financial economics, 108(1):1–28, 2013.

[114] Tarun Chordia and Lakshmanan Shivakumar. Earnings and price momentum. Journal of financial economics, 80(3):627–656, 2006.

[115] Liping Chen, Yan Deng, Xi Wang, Frank K. Soong, and Lei He. Speech bert embedding for improving prosody in neural tts. volume 2021-June, 2021. doi: 10.1109/ICASSP39728.2021.9413864.

[116] Yunxin Sang and Yang Bao. DialogueGAT: A graph attention network for financial risk prediction by modeling the dialogues in earnings conference calls. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, Findings of the Association for Computational Linguistics: EMNLP 2022, pages 1623–1633, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.117. URL https://aclanthology.org/2022.findings-emnlp.117.

[117] Gary Ang and Ee-Peng Lim. Guided attention multimodal multitask financial forecasting with inter-company relationships and global and local news. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6313–6326, 2022.

[118] Puneet Mathur, Atula Neerkaje, Malika Chhibber, Ramit Sawhney, Fuming Guo, Franck Dernoncourt, Sanghamitra Dutta, and Dinesh Manocha. Monopoly: Financial prediction from monetary policy conference videos using multimodal cues. In Proceedings of the 30th ACM International Conference on Multimedia, pages 2276–2285, 2022.

[119] Dawei Zhou, Lecheng Zheng, Yada Zhu, Jianbo Li, and Jingrui He. Domain adaptive multi-modality neural attention network for financial forecasting. In Proceedings of The Web Conference 2020, pages 2230–2240, 2020.

[120] Defu Cao, Yixiang Zheng, Parisa Hassanzadeh, Simran Lamba, Xiaomo Liu, and Yan Liu. Large scale financial time series forecasting with multi-faceted model. In Proceedings of the Fourth ACM International Conference on AI in Finance, pages 472–480, 2023.

[121] Roger K Loh and Mitch Warachka. Streaks in earnings surprises and the cross-section of stock returns. Management Science, 58(7):1305–1321, 2012.

[122] Eugene F. Fama and Kenneth R. French. A five-factor asset pricing model. Journal of Financial Economics, 116(1):1–22, 2015. ISSN 0304-405X. doi: https://doi.org/10.1016/j.jfineco.2014.10.010. URL https://www.sciencedirect.com/science/article/pii/S0304405X14002323.

[123] Alexander Swade, Matthias X Hanauer, Harald Lohre, and David Blitz. Factor zoo (. zip). Available at SSRN, 2023.

[124] Si-An Chen, Chun-Liang Li, Nate Yoder, Sercan O Arik, and Tomas Pfister. Tsmixer: An all-mlp architecture for time series forecasting. arXiv preprint arXiv:2303.06053, 2023.

[125] Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. Mteb: Massive text embedding benchmark. arXiv preprint arXiv:2210.07316, 2022.

[126] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. arXiv preprint arXiv:2211.14730, 2022.

[127] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. arXiv preprint arXiv:2212.03533, 2022.

[128] Dawei Zhu, Liang Wang, Nan Yang, Yifan Song, Wenhao Wu, Furu Wei, and Sujian Li. Longembed: Extending embedding models for long context retrieval. arXiv preprint arXiv:2404.12096, 2024.

[129] Agam Shah and Sudheer Chava. Zero is not hero yet: Benchmarking zero-shot performance of llms for financial tasks. arXiv preprint arXiv:2305.16633, 2023.

[130] Lin William Cong, Ke Tang, Jingyuan Wang, and Yang Zhang. Alphaportfolio: Direct construction through deep reinforcement learning and interpretable ai. Available at SSRN 3554486, 2021.

[131] Amirkeivan Mohtashami and Martin Jaggi. Random-access infinite context length for transformers. Advances in Neural Information Processing Systems, 36, 2024.

[132] Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. Long-context llms struggle with long in-context learning. arXiv preprint arXiv:2404.02060, 2024.

[133] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In International conference on machine learning, pages 3929–3938. PMLR, 2020.

[134] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. A survey on in-context learning. arXiv preprint arXiv:2301.00234, 2022.

[135] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 3045–3059, 2021.

[136] Tao Ge, Jing Hu, Xun Wang, Si-Qing Chen, and Furu Wei. In-context autoencoder for context compression in a large language model. arXiv preprint arXiv:2307.06945, 2023.

[137] Alexis Chevalier, Alexander Wettig, Anirudh Ajith, and Danqi Chen. Adapting language models to compress contexts. arXiv preprint arXiv:2305.14788, 2023.

[138] Howard Yen, Tianyu Gao, and Danqi Chen. Long-context language modeling with parallel context encoding. arXiv preprint arXiv:2402.16617, 2024.

[139] João Monteiro, Étienne Marcotte, Pierre-André Noël, Valentina Zantedeschi, David Vázquez, Nicolas Chapados, Christopher Pal, and Perouz Taslakian. Xc-cache: Cross-attending to cached context for efficient llm inference. arXiv preprint arXiv:2404.15420, 2024.

[140] Sijun Tan, Xiuyu Li, Shishir Patil, Ziyang Wu, Tianjun Zhang, Kurt Keutzer, Joseph E Gonzalez, and Raluca Ada Popa. Lloco: Learning long contexts offline. arXiv preprint arXiv:2404.07979, 2024.

[141] Paul C Tetlock. All the news that's fit to reprint: Do investors react to stale information? The Review of Financial Studies, 24(5):1481–1512, 2011.

[142] Anastassia Fedyk and James Hodson. When can the market identify old news? Journal of Financial Economics, 149(1):92–113, 2023.

[143] Marie Brière, Karen Huynh, Olav Laudy, and Sébastien Pouget. Stock market reaction to news: Do tense and horizon matter? Finance Research Letters, 58: 104630, 2023.

[144] Yifei Chen, Bryan T Kelly, and Dacheng Xiu. Expected returns and large language models. Available at SSRN 4416687, 2022.

[145] Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. Pixiu: A large language model, instruction data and evaluation benchmark for finance. arXiv preprint arXiv:2306.05443, 2023.

[146] Alejandro Lopez-Lira and Yuehua Tang. Can chatgpt forecast stock price movements? return predictability and large language models. arXiv preprint arXiv:2304.07619, 2023.

[147] Ziniu Hu, Weiqing Liu, Jiang Bian, Xuanzhe Liu, and Tie-Yan Liu. Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction. In Proceedings of the eleventh ACM international conference on web search and data mining, pages 261–269, 2018.

[148] Yumo Xu and Shay B Cohen. Stock movement prediction from tweets and historical prices. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1970–1979, 2018.

[149] Ramit Sawhney, Arnav Wadhwa, Shivam Agarwal, and Rajiv Shah. Fast: Financial news and tweet based time aware network for stock trading. In Proceedings of the 16th conference of the european chapter of the association for computational linguistics: main volume, pages 2164–2175, 2021.

[150] Ramit Sawhney, Shivam Agarwal, Megh Thakkar, Arnav Wadhwa, and Rajiv Ratn Shah. Hyperbolic online time stream modeling. In Proceedings of the 44th International ACM SIGIR conference on research and development in information retrieval, pages 1682–1686, 2021.

[151] Shivam Agarwal, Ramit Sawhney, Sanchit Ahuja, Ritesh Soun, and Sudheer Chava. Hyphen: Hyperbolic hawkes attention for text streams. In Proceedings of the 60th Annual Meeting of The Association of Computational Linguistics, Dublin, May 2022. Association for Computational Linguistics.

[152] Haw-Shiuan Chang, Ruei-Yao Sun, Kathryn Ricci, and Andrew McCallum. Multi-cls bert: An efficient alternative to traditional ensembling. arXiv preprint arXiv:2210.05043, 2022.

[153] John X Morris, Volodymyr Kuleshov, Vitaly Shmatikov, and Alexander M Rush. Text embeddings reveal (almost) as much as text. arXiv preprint arXiv:2310.06816, 2023.

[154] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. Neurocomputing, 568:127063, 2024.

[155] Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. arXiv preprint arXiv:2306.15595, 2023.

[156] Hongye Jin, Xiaotian Han, Jingfeng Yang, Zhimeng Jiang, Zirui Liu, Chia-Yuan Chang, Huiyuan Chen, and Xia Hu. Llm maybe longlm: Self-extend llm context window without tuning. arXiv preprint arXiv:2401.01325, 2024.

[157] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. Advances in neural information processing systems, 36, 2024.

[158] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In International conference on machine learning, pages 19730–19742. PMLR, 2023.

[159] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, et al. Time-llm: Time series forecasting by reprogramming large language models. arXiv preprint arXiv:2310.01728, 2023.

[160] Tian Zhou, Peisong Niu, Liang Sun, Rong Jin, et al. One fits all: Power general time series analysis by pretrained lm. Advances in neural information processing systems, 36:43322–43355, 2023.

[161] Chenxi Sun, Yaliang Li, Hongyan Li, and Shenda Hong. Test: Text prototype aligned embedding to activate llm's ability for time series. arXiv preprint arXiv:2308.08241, 2023.

[162] Peiyuan Liu, Hang Guo, Tao Dai, Naiqi Li, Jigang Bao, Xudong Ren, Yong Jiang, and Shu-Tao Xia. Taming pre-trained llms for generalised time series forecasting via cross-modal knowledge distillation. arXiv preprint arXiv:2403.07300, 2024.

[163] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems, 35:23716–23736, 2022.

[164] Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. arXiv preprint arXiv:2111.09543, 2021.

[165] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685, 2021.

[166] Avi Caciularu, Matthew E Peters, Jacob Goldberger, Ido Dagan, and Arman Cohan. Peek across: Improving multi-document modeling via cross-document question-answering. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1970–1989, 2023.

[167] Maor Ivgi, Uri Shaham, and Jonathan Berant. Efficient long-text understanding with short-text models. Transactions of the Association for Computational Linguistics, 11:284–299, 2023.

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

[169] Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. arXiv preprint arXiv:2007.01282, 2020.

[170] Amanda Bertsch, Uri Alon, Graham Neubig, and Matthew Gormley. Unlimiformer: Long-range transformers with unlimited length input. Advances in Neural Information Processing Systems, 36, 2024.

[171] Julian Salazar, Davis Liang, Toan Q Nguyen, and Katrin Kirchhoff. Masked language model scoring. arXiv preprint arXiv:1910.14659, 2019.

[172] Zhengxiao Du, Aohan Zeng, Yuxiao Dong, and Jie Tang. Understanding emergent abilities of language models from the loss perspective. Advances in neural information processing systems, 37:53138–53167, 2024.

[173] Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. Ruler: What's the real context size of your long-context language models? arXiv preprint arXiv:2404.06654, 2024.

[174] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning. arXiv preprint arXiv:2112.09118, 2021.

[175] Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A Smith, Luke Zettlemoyer, and Tao Yu. One embedder, any task: Instruction-finetuned text embeddings. arXiv preprint arXiv:2212.09741, 2022.

[176] Narasimhan Jegadeesh and Sheridan Titman. Returns to buying winners and selling losers: Implications for stock market efficiency. The Journal of Finance, 48(1):65–91, 1993. doi: 10.1111/j.1540-6261.1993.tb04702.x.

[177] Eugene F Fama and Kenneth R French. Choosing factors. Journal of financial economics, 128(2):234–252, 2018.

[178] Narasimhan Jegadeesh. Evidence of predictable behavior of security returns. The Journal of Finance, 45(3):881–898, 1990. doi: 10.1111/j.1540-6261.1990.tb05110.x.

[179] Bryan T. Kelly, Tobias J. Moskowitz, and Seth Pruitt. Understanding momentum and reversal. Journal of Financial Economics, 140(3):726–743, 2021. doi: 10.1016/j.jfineco.2020.06.024.

[180] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. Advances in neural information processing systems, 36:34892–34916, 2023.

[181] Dongting Li, Chenchong Tang, and Han Liu. Audio-llm: Activating the capabilities of large language models to comprehend audio data. In International Symposium on Neural Networks, pages 133–142. Springer, 2024.

[182] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. arXiv preprint arXiv:2405.09818, 2024.

[183] Guo Chen, Yin-Dong Zheng, Jiahao Wang, Jilan Xu, Yifei Huang, Junting Pan, Yi Wang, Yali Wang, Yu Qiao, Tong Lu, et al. Videollm: Modeling video sequence with large language models. arXiv preprint arXiv:2305.13292, 2023.

[184] Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew G Wilson. Large language models are zero-shot time series forecasters. Advances in Neural Information Processing Systems, 36:19622–19635, 2023.

[185] Andrew Robert Williams, Arjun Ashok, Étienne Marcotte, Valentina Zantedeschi, Jithendaraa Subramanian, Roland Riachi, James Requeima, Alexandre Lacoste, Irina Rish, Nicolas Chapados, and Alexandre Drouin. Context is Key: A Benchmark for Forecasting with Essential Textual Information. arXiv preprint arXiv:2410.18959, 2024.

[186] Kai Kim, Howard Tsai, Rajat Sen, Abhimanyu Das, Zihao Zhou, Abhishek Tanpure, Mathew Luo, and Rose Yu. Multi-Modal Forecaster: Jointly Predicting Time Series and Textual Data. arXiv preprint arXiv:2411.06735, 2024.

[187] Haoxin Liu, Shangqing Xu, Zhiyuan Zhao, Lingkai Kong, Harshavardhan Kamarthi, Aditya B. Sasanur, Megha Sharma, Jiaming Cui, Qingsong Wen, Chao Zhang, and B. Aditya Prakash. Time-MMD: Multi-Domain Multimodal Dataset for Time Series Analysis. In Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, December 10-16, 2024, Vancouver, BC, Canada, Track on Datasets and Benchmarks, 2024. URL https://openreview.net/forum?id=1kP0Nn0xDE.

[188] Zihao Li, Xuanye Lin, Zeyu Liu, Jiaru Zou, Zhenyu Wu, Ling Zheng, Dawei Fu, Yan Zhu, Hendrik Hamann, Hanghang Tong, et al. Language in the Flow of Time: Time-Series-Paired Texts Weaved into a Unified Temporal Narrative. arXiv preprint arXiv:2502.08942, 2025.

[189] Mengyu Wang, Shay B. Cohen, and Tiejun Ma. Modeling News Interactions and Influence for Financial Market Prediction. In Findings of the Association for Computational Linguistics: EMNLP 2024, pages 3302–3314, Miami, Florida, USA, 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-emnlp.189.

[190] Chang Zong, Jian Shao, Weiming Lu, and Yueting Zhuang. Stock movement prediction with multimodal stable fusion via gated cross-attention mechanism. arXiv preprint arXiv:2406.06594, 2024.

[191] Shangyang Mou, Qiang Xue, Jinhui Chen, Tetsuya Takiguchi, and Yasuo Ariki. Mm-itransformer: A multimodal approach to economic time series forecasting with textual data. Applied Sciences, 15(3):1241, 2025.

[192] Zihan Dong, Xinyu Fan, and Zhiyuan Peng. FNSPID: A Comprehensive Financial News Dataset in Time Series. arXiv preprint arXiv:2402.06698, 2024.

[193] Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. St-moe: Designing stable and transferable sparse expert models. arXiv preprint arXiv:2202.08906, 2022.

[194] Aaron Grattafiori and et al. The Llama 3 herd of models. arXiv preprint arXiv:2407.21783, July 2024.

[195] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of BERT's attention. In Tal Linzen, Grzegorz Chrupała, Yonatan Belinkov, and Dieuwke Hupkes, editors, Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 276–286, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4828. URL https://aclanthology.org/W19-4828/.

[196] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In NeurIPS, 2021. URL https://arxiv.org/abs/2107.00135.

[197] Cheng Zhang, Jinxin Lv, Jingxu Cao, Jiachuan Sheng, Dawei Song, and Tiancheng Zhang. Unravelling the semantic mysteries of transformers layer by layer. The Computer Journal, page bxaf034, 2025.

[198] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. arXiv preprint arXiv:2309.17453, 2023.

[199] Tianzhu Ye, Li Dong, Yuqing Xia, Yutao Sun, Yi Zhu, Gao Huang, and Furu Wei. Differential transformer. arXiv preprint arXiv:2410.05258, 2024.

[200] Olga Golovneva, Tianlu Wang, Jason Weston, and Sainbayar Sukhbaatar. Multi-token attention. arXiv preprint arXiv:2504.00927, 2025.

[201] Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. Time-moe: Billion-scale time series foundation models with mixture of experts. arXiv preprint arXiv:2409.16040, 2024.

[202] Stephan Rabanser, Tim Januschowski, Valentin Flunkert, David Salinas, and Jan Gasthaus. The effectiveness of discretization in forecasting: An empirical study on neural time series models. arXiv preprint arXiv:2005.10111, 2020.

[203] Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. arXiv preprint arXiv:2403.07815, 2024.

[204] Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. 2024.

[205] Mingtian Tan, Mike A. Merrill, Zachary Gottesman, Tim Althoff, David Evans, and Tom Hartvigsen. Inferring Event Descriptions from Time Series with Language Models. arXiv preprint arXiv:2503.14190, 2025.

[206] Shengkun Wang, Taoran Ji, Linhan Wang, Yanshen Sun, Shang-Ching Liu, Amit Kumar, and Chang-Tien Lu. Stocktime: A time series specialized large language model architecture for stock price prediction. arXiv preprint arXiv:2409.08281, 2024.

[207] Hajar Emami Gohari, Xuan-Hong Dang, Syed Yousaf Shah, and Petros Zerfos. Modality-aware transformer for financial time series forecasting. In Proceedings of the 5th ACM International Conference on AI in Finance, pages 677–685, 2024.

[208] Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 12286–12312, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.687. URL https://aclanthology.org/2023.acl-long.687/.

[209] Hongyi Yuan, Keming Lu, Fei Huang, Zheng Yuan, and Chang Zhou. Speculative contrastive decoding. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 56–64, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-short.5. URL https://aclanthology.org/2024.acl-short.5/.

[210] Lizhe Fang, Yifei Wang, Zhaoyang Liu, Chenheng Zhang, Stefanie Jegelka, Jinyang Gao, Bolin Ding, and Yisen Wang. What is wrong with perplexity for long-context language modeling? arXiv preprint arXiv:2410.23771, 2024.

[211] OpenAI. Gpt-4o: Openai's new omnimodal model. https://openai.com/index/gpt-4o, 2024. Accessed: 2025-04-23.

[212] Alex Kim, Maximilian Muhn, and Valeri Nikolaev. Financial statement analysis with large language models. arXiv preprint arXiv:2407.17866, 2024.

[213] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. Advances in neural information processing systems, 30, 2017.

[214] DeepSeek-AI. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024.

[215] Mike A Merrill, Mingtian Tan, Vinayak Gupta, Thomas Hartvigsen, and Tim Althoff. Language models still struggle to zero-shot reason about time series. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, Findings of the Association for Computational Linguistics: EMNLP 2024, pages 3512–3533, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.201. URL https://aclanthology.org/2024.findings-emnlp.201/.

[216] Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Hao Chen, Emad Barsoum, Muhao Chen, and Zicheng Liu. Latent visual reasoning. arXiv preprint arXiv:2509.24251, 2025.

[217] Guohao Sun, Hang Hua, Jian Wang, Jiebo Luo, Sohail Dianat, Majid Rabbani, Raghuveer Rao, and Zhiqiang Tao. Latent chain-of-thought for visual reasoning. arXiv preprint arXiv:2510.23925, 2025.

[218] Ho Chung Wu, Robert Wing Pong Luk, Kam Fai Wong, and Kui Lam Kwok. Interpreting tf-idf term weights as making relevance decisions. ACM Transactions on Information Systems, 26, 2008. ISSN 10468188. doi: 10.1145/1361684.1361686.

[219] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pages 1532–1543, 2014.

[220] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. arXiv preprint arXiv:2101.00190, 2021.

[221] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. arXiv preprint arXiv:2203.05794, 2022.

[222] Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics, pages 837–845, 1988.