Enhancing Financial Market Forecasting Through LLMs with Textual Alignment

by

Kai ZHAO

A Thesis Submitted to
The Hong Kong University of Science and Technology (Guangzhou)
in Partial Fulfillment of the Requirements for
the Degree of Master of Philosophy
in Artificial Intelligence Thrust

May 2024, Guangzhou

AUTHORIZATION

I hereby declare that I am the sole author of the thesis.

I authorize the Hong Kong University of Science and Technology (Guangzhou) to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I further authorize the Hong Kong University of Science and Technology (Guangzhou) to reproduce the thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

_____

Kai ZHAO

30 May 2024

Enhancing Financial Market Forecasting Through LLMs with Textual Alignment

by

Kai ZHAO

This is to certify that I have examined the above MPhil thesis
and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by
the thesis examination committee have been made.

Prof. Hui XIONG, Thesis Supervisor

Prof. Kani CHEN, Thesis Supervisor

Prof. Lei CHEN, Head of Thrust

Artificial Intelligence
30 May 2024

ACKNOWLEDGMENTS

I am profoundly grateful to the faculty of The Hong Kong University of Science and Technology (HKUST) at the Clear Water Bay and Guangzhou campuses. A special note of gratitude goes to Prof. Hui XIONG, whose mentoring has been transformative. He not only introduced me to the world of artificial intelligence but also nurtured my initial curiosity into a strong academic passion. I also extend my sincere thanks to Prof. Kani CHEN, whose steadfast support and unwavering belief in my potential have been critical as I navigate my future career in the industry. I owe profound gratitude to Prof. Lin WANG, who was a beacon of support during my most uncertain times. When I was struggling even to comprehend academic papers, Prof. WANG continually encouraged me, instilling the belief that I was capable and could succeed. I am equally thankful to Prof. Hao LIU, whose wisdom has been instrumental throughout my academic journey. Each session with him was a profound learning experience; he not only answered my questions but also taught me that the strongest motivation comes from within oneself. I am grateful to Dr. Tongge Huang for being a constant companion during my growth these past two years. His optimistic outlook and wisdom in accepting the unchangeable circumstances deeply resonated with me, especially during the job search process when his encouragement enabled my perseverance. I am also thankful to Dr. Li Chen, who motivated me to embrace challenges and persevere with my projects right from the start of my master's program, extending a generous helping hand during difficult times.

At the University of Hong Kong(HKU) Business School, in the Master of Science in Business Analytics (MScBA) program, Prof. Zhengli WANG played a pivotal role in my journey into machine learning, igniting my passion for data science. Prof. Wei ZHANG's mantra that it is "never too late to learn" has inspired me to pursue a second MPhil degree. I am thankful for the enthusiastic teaching and patient guidance of all the MScBA program professors at HKU, which have sustained my interest and excitement in this evolving field.

Special acknowledgment goes to Zihan DONG from North Carolina State University,

Raleigh, North Carolina, USA, for providing the essential datasets that enabled the completion of this work.

Lastly, the unwavering support and encouragement from my family, friends, and everyone who has motivated me throughout this journey cannot be overstated. Their belief in me has been my constant source of strength.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Enhancing Financial Market Forecasting Through LLMs with Textual Alignment

by

Kai ZHAO

Artificial Intelligence

The Hong Kong University of Science and Technology (Guangzhou)

ABSTRACT

Large language models (LLMs) have demonstrated remarkable capabilities in various natural language processing tasks, including text generation, question answering, and language understanding. Beyond the realm of language, an increasing number of researchers are applying LLMs to multimodal tasks. However, their application in time series analysis, particularly in the domain of financial market prediction, remains largely unexplored. This research aims to investigate the potential of large language models in forecasting real-world time series data, with a specific focus on stock market prices and non-fungible token (NFT) valuations.

Traditional time series analysis techniques often rely on machine learning algorithms or deep learning models that operate solely on one modality, that is numerical data. In contrast, this study proposes a novel approach that transfers the linguistic capabilities of pretrained LLMs to capture the intrinsic patterns and dynamics of time series data. By aligning the language model's understanding with the inherent structure of financial time series, this research seeks to bridge the gap between the linguistic domain and the numerical representation of time series data.

The proposed methodology involves integrating large language models' textual information with time series data using attention mechanisms and contrastive learning. Additionally, this research demonstrates the zero-shot capability of LLMs through a unified training process designed to imbue the model with an understanding of the underlying dataset patterns. Subsequently, the trained language model will be applied to predict future time series values based on its acquired knowledge and understanding.

CHAPTER 1

INTRODUCTION

Time series analysis and forecasting play a crucial role in various domains, including finance, economics, and decision-making processes. Accurate predictions of future trends and values can provide significant advantages in investment strategies, risk management, and resource allocation. However, traditional time series analysis techniques often face challenges in capturing the complex dynamics and external factors that influence real-world time series data, such as stock market prices and non-fungible token(NFT) valuations.

Most research works focusing on financial market forecasting are primarily based on typical machine learning methods and involve extensive feature engineering to utilize financial data. However, excessive feature engineering often leads to information loss during the data processing stage. Especially for financial news, which has been proven to be essential and strongly correlated with financial market prices [1], it is crucial to utilize this information effectively. Although recent studies have attempted to employ more advanced deep learning techniques to fit datasets, the large number of parameters required by deep learning models necessitates complex data. Even though some transformer models have achieved relatively good results, they tend to face overfitting issues. This is also the reason why a unified foundation model that can adapt to different kinds of downstream tasks has not been established yet in time series domain.

When CLIP[2] was proposed by OpenAI, it broke the boundary between language and vision tasks. Multimodal models and domain transfernig were becoming a new trend, and many research work have proved the multimodel's ability in transfer from domain to domain crossly. After that, large language models(LLMs) such as GPT-4[3] had become a hit worldwide in 2022. These models, trained on massive corpora of textual data, have shown promise in capturing intricate patterns and relationships within language. Even out of language tasks[4], LLMs showed it graet abilities, which lead human to achieve the final purpose of artificial intelligence - Artificial General Intelligence(AGI).

Based on LLMs' ability, some research works have attempted to transfer the capabilities of large language or visual models to time series forecasting due to the lack of data in the time series domain. According to calculations from [5, 6], the largest datasets available for time series analysis are less than 10GB. The scarcity of high-quality and sufficient data has presented a significant challenge for time series researchers in developing a foundation model that can be effectively adapted to different downstream tasks. Consequently, [5] explored the potential of LLMs in time series forecasting and achieved state-of-the-art(SOTA) results. Following this breakthrough, many large language model for time series forecasting (LLM4TS) work boosting, and it has become a new and promising topic in this area. However, latest LLM4TS' research work still remains largely unexplored, primarily due to the disconnect between the linguistic domain. There is no clear direction on how to effectively incorporate the numerical representation of time series data into a language pre-trained model. In addition, there is also a lack of a comprehensive explainability framework that can elucidate the reasons behind the potential success of LLMs in this domain.

Thus, This research aims to bridge this gap by proposing a novel approach that leverages the linguistic capabilities of large language models to understand and predict real-world time series data. The core premise of this study is that by aligning the language model's understanding with the inherent structure and dynamics of time series, it can capture the underlying patterns and external factors that influence market behavior.

One of the biggest obstacles preventing explainability and combination is that most time series data do not have corresponding textual descriptions. Although some research works [7, 8] have applied prompt descriptions, they are only used for describe a whole dataset abut not specific for a series fluctuation period, which has hindered the connection between the linguistic and numerical domains. Considering the need for many high-quality datasets, we decided to begin with financial market, as they possess a wealth of high-quality news articles and real-world price fluctuations. Besides, this markets are largely impacted by public opinion and social factors such as news articles and analyst reports, are closely aligned with market movements.

The proposed methodology involves training large language models on textual descrip-

tions and narratives related to financial markets, such as stock prices and non-fungible token (NFT) valuations. This training process is designed to imbue the language model with an understanding of the underlying financial concepts, market dynamics, and external factors that influence time series patterns. By exposing the language model to a diverse corpus of financial-related news, reports, and social media content, it can learn to associate linguistic cues with numerical patterns in time series data. Our main contributions are as followed:

- This is the first LLM based state-of-the-art(SOTA) work trained on the FNSPID dataset, the current most comprehensive and authoritative stock market dataset. We use unified-training process solve the unpracticality of the existing dataset and benchmark.
- We "open" the "black box" of LLM4TS and increase its explainability: We align the LLM's corpus and time series data based on contrastive learning and double attention mechanism, and we attempt to prove that the LLM's ability for time series forecasting (LLM4TS) stems from knowledge (pattern transferring) rather than data overfitting.
- We use zero-shot prediction trained from the stock market on non-fungible tokens (NFTs) to demonstrate the possibility of building a unified and zero-shot large time series model based on large language models, provided that we have sufficient high-quality data from different domains.

Our experimental results illustrate that we have integrated the strengths of large language models with the numerical representation of time series data by combining linguistic pattern transferring with time series analysis. We unlock new possibilities in financial forecasting and decision-making processes.

CHAPTER 2

SCIENTIFIC BACKGROUND AND LITERATURE REVIEW

2.1 Literature Review for Financial Market

Financial markets are deeply influenced by the volume and sentiment of world-of-mouth, such as news articles and social media posts[9, 10] . This textual abundance means that perceived "hotness" or trending interest often drives market fluctuations significantly. Our analysis primarily focuses on the stock market and the NFT (Non-Fungible Token) market.

Stock Market In the early stages, stock market price prediction relied on statistical and machine learning methods to forecast prices by combining and analyzing complex indices and numerical features. From 2010 onwards, researchers began exploring the application of advanced techniques in this domain.

The journey into using machine learning for stock price prediction can be traced back to foundational works like that by Yuhong Li and Weihua Ma in 2010[11]. Their study emphasized the versatility of artificial neural networks (ANNs) in financial economics, particularly highlighting their capability to model nonlinear relationships inherent in stock market data. This work set the stage for the integration of more sophisticated machine learning models in financial predictions, establishing a critical foundation for subsequent developments in the field.

By 2013, the field began incorporating elements of natural language processing (NLP), as demonstrated by Boyi Xie et al.[12], who explored how semantic frames could predict stock price movements. This approach utilized linguistic analysis to infer market sentiments, showcasing an early attempt to blend textual analysis with quantitative data. The following year, Xiaodong Li and colleagues [13] further advanced this approach by analyzing the

impact of news sentiment on stock prices, marking a significant step towards the holistic incorporation of unstructured data sources in predictive models.

The adaptation of more complex data patterns and the introduction of deep learning significantly advanced stock price predictions. In 2017, Liheng Zhang et al.[14], work on discovering multi-frequency trading patterns illustrated how deep learning could identify intricate patterns across different time scales, enhancing prediction accuracy. By 2018, deep reinforcement learning (DRL) began to take center stage, with researchers like Zhuoran Xiong et al.[15] demonstrating its practical applications in stock trading, optimizing decisions based on dynamic market conditions.

The trend towards more sophisticated AI techniques became more pronounced by 2019. Derek Yang et al.'s[16] introduction of fully parameterized quantile functions for distributional reinforcement learning highlighted the shift towards models capable of assessing a range of potential outcomes rather than single-point predictions. In 2021, the development of FinRL-Podracer by Zechu Li et al.[17] exemplified high-performance, scalable deep reinforcement learning frameworks specifically tailored for quantitative finance, pushing the boundaries of speed and efficiency in trading algorithms.

By 2022, the adoption of transformer-based models, as explored by Qiuyue Zhang et al.[18], marked a significant innovation in stock market prediction. Their use of attention mechanisms to better understand and predict stock movements showcased the integration of cutting-edge AI technologies, which are particularly adept at handling sequential data and capturing temporal dependencies.

The synthesis of these advancements provides a comprehensive view of how stock price prediction has evolved from relatively simple machine learning models to complex systems integrating deep learning, NLP, and AI-driven techniques. The current landscape, as detailed in the survey from 2023, points towards a future where multi-modal data integration, advanced neural architectures, and a deeper understanding of market dynamics through AI will drive further innovations. The emphasis on transparency, interpretability, and ethical considerations in AI applications remains paramount as these technologies become more intertwined with financial decision-making processes.

Non-Fungible Tokens Market Analogous to the financial market, news articles and social media discussions surrounding Non-Fungible Tokens (NFTs) serve as vital and high-quality data sources for developing accurate price prediction models. As an emerging and rapidly evolving market, NFT price prediction has garnered significant attention from researchers, investors, and enthusiasts alike. However, in contrast to traditional financial markets, NFT prices are even more heavily influenced by social media and public sentiment. This heightened impact of social media on NFT valuations underscores the importance of leveraging these data sources for effective price forecasting in this domain.

A groundbreaking study by [19, 20] serves as a testament to the inextricable link between social media sentiment and NFT prices. By bridging the gap between Twitter data and the NFTs, the authors demonstrate the remarkable effectiveness of social media features in forecasting the value of NFT assets. However, it is essential to acknowledge the limitations of this work, such as the incomplete integration of NFT market data and the primary focus on profile-related data rather than the semantic content of tweets.

Given the overwhelming evidence[21, 22, 23] of social media's influence on NFT prices, it is imperative to incorporate this data when developing robust price prediction models. The dominant role of social media in shaping NFT prices cannot be overstated.

In conclusion, almost most of financial related products are widely regarded as sentiment-driven or popularity-based collectibles, with their values heavily influenced by public opinion and market hype. We could simply take this time series have a high relativity with those "text". Given this premise, incorporating social media data and news articles into financial market price prediction models emerges as a crucial research direction, which could fill the gap between text and time series alignment in LLM.

From that motivation, the objective of this research is to establish an time series prediction framework that integrates rich text information into large language models. By gaining a profound understanding of the key factors influencing the real world series and developing highly interpretable models, we aim to contribute to this rapidly evolving field and provide reliable and transparent predictions for researchers and decision-makers.

2.2 Literature Review for Large Language Models for Time Series Forecasting

Research efforts exploring the application of LLMs for time series forecasting can be broadly categorized into two approaches: prompt-based 2.1 and model-based2.2. Prompt-based studies primarily focus on enhancing the performance of closed-sourced LLMs through prompt engineering and time series description. Prompt engineering is to constrain the model in a special situation. Descriprion is to transfer time series data into more natural language guided way, in which LLM could understand. They all attempt to fine-tune the models in a manner that aligns with human reasoning and knowledge. Conversely, model-based researches involve modifying the open-sourced architectural components of LLMs and fine-tuning their parameters.

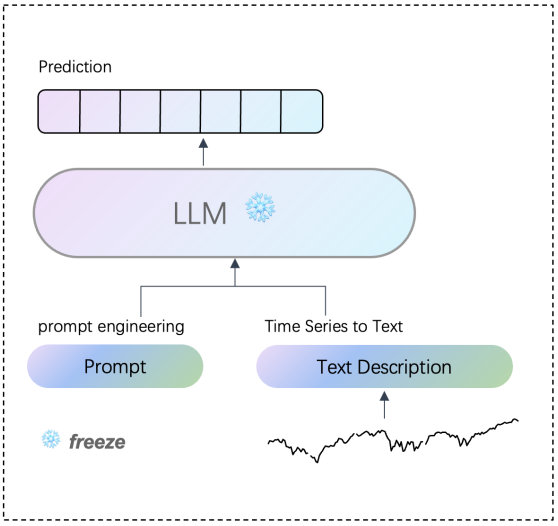

Figure 2.1: Prompt-Based LLM4TS

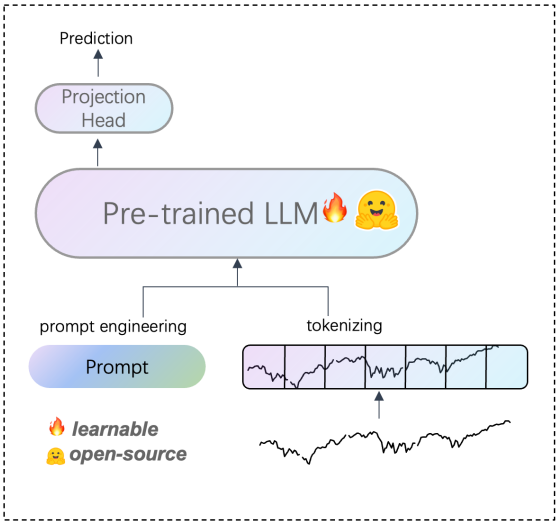

Figure 2.2: Structure-Based LLM4TS

Prompt-Based Approaches In the domain of financial time series forecasting, a recent investigation [24] employed Large Language Models (LLMs) to tackle pivotal challenges - explainability by integrating diverse signals such as financial news and LLM-generated company profiles, and augmenting model capabilities. Concentrating on NASDAQ-100 stocks, the researchers adopted methods to categorize price fluctuations into bins, and executed experiments utilizing zero-shot/few-shot inference with GPT-4, alongside fine-tuning the Open LLaMA model. The study illustrated that the Chain of Thoughts (CoT) mechanism

could enhance the performance of LLMs by enabling effective reasoning across textual and numerical data. Significantly, LLMs demonstrated adeptness in merging information from multiple modalities, whereas the refined models like Open-LLaMA provided explainable and interpretable forecasts, vital for financial applications. This seminal work signifies a step forward in crafting more advanced and dependable financial forecasting models by leveraging sophisticated language models, with a pronounced focus on explainability.

In opposite, LLMTIME[25] focused on the tokenization of time series, an innovative approach is presented where time series inputs are adapted to the input format of large language models without the necessity for training. The study highlights that different models employ various tokenization techniques, and these methods significantly influence the model's performance. The proposed tokenization strategy involves separating each number with a space, ensuring individual tokenization of each digit, and using commas to delineate each timestep within the time series. Employing artificially generated time series, the authors empirically demonstrate that the ability of LLMs to predict time series is attributable to their capacity to identify low-complexity explanations within the data, enabling them to extrapolate numerical sequences in a zero-shot manner.

In the innovative study "Promptcast" [26], time series forecasting is reimagined as a dialog task through the novel approach of transforming numerical inputs into natural language sentences using specific prompt templates. This groundbreaking method marks the first instance of addressing general time series prediction issues through natural language generation. Among the models, the input and output are all natural language rather than numbers. The authors advocate that prompt engineering enables language models to effectively utilize auxiliary information, such as time of day and contextual semantics, thereby deepening the models' comprehension of the interplay between this additional data and the time series. Extensive experiments validate the efficacy and generalizability of language models in time series forecasting, demonstrating that they can achieve results on par with conventional numerical approaches, thereby positioning prompt-based forecasting as a promising avenue in the field.

Further enhancing the research landscape, the investigators constructed the first-ever

prompt-based time series forecasting task dataset, offering a valuable resource for future studies. A deep dive into why language models excel in this domain revealed the pivotal role of prompt engineering. By employing prompts, language models can fully leverage contextual information, significantly enhancing their ability to understand and predict time series data.

The researchers also aim to inspire their peers with new investigative directions, such as generating textual prompts suitable for numerical data and circumventing biases introduced by fixed templates. They propose a visionary idea: leveraging language models to autonomously generate descriptions for time series data, which could pave new pathways for forthcoming research endeavors.

Model-Based Approaches One of the most representative and pioneering research works in editing LLMs' structrure in this domain is GPT4TS [5], which marked a significant milestone in the exploration of LLMs for time series forecasting (LLM4TS). GPT4TS proposed a unified, frozen state-of-the-art(SOTA) model by segmenting time series data into patches and incorporating them into a large language model, such as GPT-2. By frozen GPT's self-attention and FFN layers, this groundbreaking work demonstrated the remarkable performance of LLMs in kinds of time series downstream tasks such as forecasting, imputation, detection, classfication and few shot learning. GPT4TS paving the way for an outpouring of subsequent investigations in the field and building a unified foundation model in time series.

Based on prompt-based LLMs development, recent strcture-based model try to combine prompt to enhance model's ability. Building upon GPT4TS's seminal work, Time-LLM [7] stands as a notable advancement, leveraging the foundations laid by GPT4TS. Time-LLM aimed to unlock the untapped potential of LLMs by employing prompts and textual information. Notably, it was the first study to attempt aligning textual data with time series and reducing the gap between the two domains while integrating them into an LLM framework. This work utilized text prototypes and attention mechanisms, enabling each time series patch to be aligned with a corresponding text prototype. Additionally, Time-LLM incorporated prompts containing information about seasonality, trends, and lagged values to further enhance the LLM's capabilities in capturing temporal dependencies.

Similar to Time-LLM, both studies aimed to establish an alignment between time series data and language corpus. However, TEST[8] argued that the direct alignment of text prototypes and time series segments might not yield human-interpretable correspondences by visualizing experimental results. Except for compulsorily aligning text and time series, TEST employed soft prompts and contrastive learning across different time series window segments in an attempt to enhance the model's robustness. Remarkably, TEST achieved SOTA performance even without the need for fine-tuning. Nevertheless, this approach potentially reintroduced opacity, rendering the model's inner workings opaque and hindering its explainability.

By challenging the direct text-time series alignment approach and proposing alternative techniques such as soft prompts and contrastive learning, TEST[8] contributed to advancing the field while also highlighting the persistent challenge of maintaining explainability in LLM-based time series forecasting models.

Rather than fine-tuning the last or several projection layers, TEMPO[27] takes a distinct approach by decomposing the time series model into trends, seasonality, and residual patterns. Through experiments, the study demonstrates that the seasonal component significantly influences the model's predictive performance, implying that the model heavily relies on the overall repetitive patterns within the data. This analysis offers insightful perspectives into how the model interprets and utilizes the decomposition preprocessing step, providing a solid foundation for model optimization and enhancement. Notably, TEMPO's experimental results substantially outperform state-of-the-art models such as TIME-LLM and GPT4TS.

Despite the advancements made by structure-based LLM4TS models, researchers are still faced with two critical challenges. Firstly, there lacks a clear understanding and empirical evidence to elucidate the underlying reasons behind the transferability of large language models to the time series domain. Secondly, no dedicated effort has been made to explicitly focus on aligning textual data with time series information, a crucial step towards enhancing the gap between these two modalities. It is these two gaps that motivate the present work, aiming to provide insights into the successful application of LLMs for time series forecasting and to develop strategies for improving the interpretability of such models by bridging the

linguistic and series representations.

By addressing these two challenges, this study aims to contribute to the theoretical foundations of LLM4TS while also proposing practical solutions to enhance the transparency and trustworthiness of these models, ultimately enabling their wider adoption in real-world decision-making scenarios involving time series data.

CHAPTER 3

METHODS

Problem Definition A time series is defined as a sequence of data points arranged in chronological order, intended to represent the status or behavior of a process or activity at different time points. Formally, given a time index set $\mathcal{T} = \{1, 2, \ldots, T\}$, a time series can be represented as $\{x_t\}_{t \in \mathcal{T}}$, where x_t denotes the observation at time t. The objective of time series prediction is to forecast future values $\{x_{T+1}, x_{T+2}, \ldots\}$ based on historical observations $\{x_1, x_2, \ldots, x_T\}$.

Evaluation Metrics To assess the performance of time series prediction models, we commonly employ the Mean Squared Error (MSE) and the Mean Absolute Error (MAE) as evaluation metrics. Given a series of true values $\{y_t\}_{t \in \mathcal{T}}$ and predicted values $\{\hat{y}_t\}_{t \in \mathcal{T}}$, these metrics are defined as follows:

- **Mean Squared Error (MSE)**: This metric calculates the average of the squares of the errors, that is, the average squared difference between the estimated values and the actual value. Mathematically, it is expressed as:

$$\text{MSE} = \frac{1}{T} \sum_{t=1}^{T} (y_t - \hat{y}_t)^2 \tag{3.1}$$

- **Mean Absolute Error (MAE)**: This metric measures the average magnitude of the errors in a set of predictions, without considering their direction. It is the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight. It is defined as:

$$\text{MAE} = \frac{1}{T} \sum_{t=1}^{T} |y_t - \hat{y}_t| \tag{3.2}$$

The smaller the values of MSE and MAE, the smaller the error between predictions and true values, indicating better performance of the predictive model. The loss function used for training the model is the Mean Squared Error (MSE) - $Loss_{MSE}$, which measures the average of the squares of the errors between the estimated values and the actual values.

3.1 Model Overview

Model Structure The overall model structure is illustrated in Figure 3.1. First, we select abundant textual information as input. We then employ a straightforward projection technique, such as a linear probe or PCA, to reduce the dimensionality of this information into concise language embeddings (text prototypes). The goal is to merge these text prototypes with time series patches. We leverage attention mechanisms and contrastive learning to integrate and align the time series data with the textual information, striving to make these representations comprehensible to the large language model (LLM). The merged tokens are then input into the LLM along with a simple prompt. Finally, a regression number will be output from a projection head.

Soft Alignment For soft alignment, we utilize the corpus from the language model and directly project the embedding space into a more compact spatial representation, referred to as text prototypes. In this context, the entire corpus is denoted as V, and the text prototypes are represented as V', where $V' \ll V$. This reduction in dimensionality significantly decreases computational costs.

Hard Alignment For hard alignment, we summarize the daily news articles by utilizing the GPT-3.5 API with a predefined prompt from[24]. The prompt and reponse exmaple are as showed in Table 3.2. For longer texts that exceed the maximum token limit of GPT, we chunk the entire article and summarize each chunk individually. Through this process, we obtain accurate descriptions of the daily news. However, time series prediction typically involves sequence lengths exceeding fifty days, and incorporating excessive news content may lead to computationally intensive calculations.

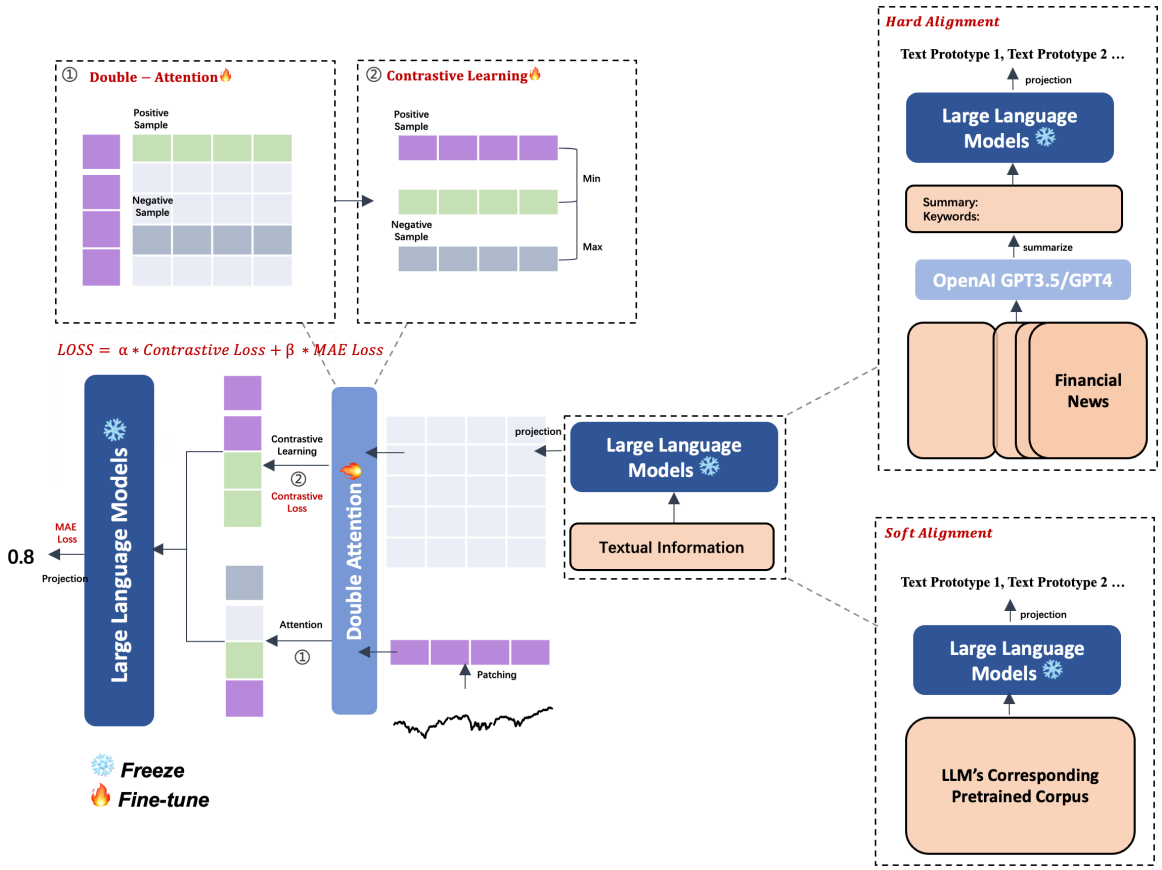

Figure 3.1: LLM4FN Model Structure

3.2 Time Series and Text Alignment

We believe that the key to improving LLM's performance in time series prediction lies in enabling the model to understand the time series input effectively. From an overarching perspective, a robust alignment and appropriate input representation will lead the LLM to comprehend our input and output data better. We first employ an attention mechanism to identify similar words from the time series data, and then leverage positive and negative pairs through contrastive learning to further constrain the representational space. The details are as follows:

3.3 Learnable Double Attention Mechanism and Contrastive Learning

Attention mechanisms have profoundly impacted our understanding of sequence modeling, primarily through the Transformer architecture, which utilizes attention to dynamically

model:
gpt-3.5-turbo-1106
role-system:
template = f"""Please summarize the following noisy but possible news data extracted from web page HTML, and extract keywords of the news. The news text can be very noisy due to it is HTML extraction. Give formatted answer such as Summary: ..., Keywords: ... The news is supposed to be for company symbol stock. You may put 'N/A' if the noisy text does not have relevant information to extract."""
role-user:
{news}
response:
Summary: The news discusses the stock market futures, Amazon's stock split, and Apple's introduction of Apple Pay Later service. It also covers Kohl's negotiations for a potential sale and the recent rise in oil prices due to easing COVID lockdowns in China and Saudi Arabia's increased selling prices.
Keywords: stock market, futures, Amazon, stock split, Apple, Apple Pay Later, Affirm, Goldman Sachs, Mastercard, Kohl's, sale, oil prices, China, Saudi Arabia

Figure 3.2: Summary of News

weigh the importance of different elements in the input data. However, traditional attention mechanisms, as defined in Equation 3.3, rely solely on non-parametric interactions between queries and keys to compute attention weights. These models are limited because they do not allow for parameterized transformations within the attention operation itself, which can be crucial for capturing more complex dependencies.

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3.3}$$

Inspired by the way Graph Convolutional Networks (GCNs)[28] incorporate Laplace operation into their node feature aggregation mechanism, we propose a novel double attention mechanism that not only calculates attention weights but also allows these weights to interact through a parameterized function. This method enhances the model's ability to integrate and learn from the interaction between different types of data, such as time series and textual embeddings.

Double Attention Formulation: Our approach, referred to as *Double Attention*, extends the standard attention mechanism by introducing an additional learnable matrix W, which

interacts with the outputs of two parallel attention computations involving different ‘value‘ representations. The mechanism is defined as follows:

$$\text{DOUBLE-ATTENTION}(Q, K) = LeakyRelu(W \cdot Attention_T || W \cdot Attention_S)$$

$$Attention_T = Att(Q, K, V_1) = softmax\left(\frac{LeakyReLU(QW_1K^TW_2)}{\sqrt{d_k}}\right)V_1$$

$$Attention_S = Att(Q, K, V_2) = softmax\left(\frac{LeakyReLU(QW_1K^TW_2)}{\sqrt{d_k}}\right)V_2$$

$Q = V_1 =$ Text, $\quad K = V_2 =$ Series

$Attention_T =$ Attention for Text, $Attention_S =$ Attention for Series, $||$ means multiple or concat

(3.4)

In this formulation, Q (Queries) and K (Keys) represent word embeddings and patched time series, respectively. Two parallel computations are performed, in which V_1 and V_2 alternately represent the ’value’ data, specifically using word embeddings for one attention computation and time series patches for the other, hence the name "double attention." This approach allows the model to dynamically focus on and integrate information from both textual and time-series data, depending on their relevance to the task at hand. The concatenated output of two distinct attention computations is then transformed by the learnable matrix W and a non-linear activation function, allowing the model to learn how to best integrate information from both textual and sequential data sources effectively. This novel double attention mechanism facilitates a more flexible and powerful way to combine different data modalities, enhancing the model’s ability to perform complex reasoning over multimodal inputs. It is particularly suitable for applications such as multimodal sentiment analysis, contextual forecasting, and integrated language-time series tasks, where the interplay between text and sequential data is critical.

Contrastive Learning

From the double attention mechanism, we could obtain weighted scores, which could be understood as relevance between text and time series data3.5. To further enhance the fusion

of multimodal sources, we employ the concept of contrastive learning. Specifically, we select the highest score from the text corpus as the positive sample ($test_p$) and the lowest score as the negative sample ($test_n$). An additional loss function is introduced to optimize this contrastive learning process3.7.

$$\text{ATTENTION SCORE} = softmax\left(\frac{LeakyReLU(QW_1K^TW_2)}{\sqrt{d_k}}\right) \tag{3.5}$$

In detail, we could identify the indices of the maximum and minimum scores from ATTENTION SCORE, and then use these indices to select the positive and negative samples:

$$\begin{aligned} &test_p = T_{\arg\max(S)} = \text{positive sample} \quad test_n = T_{\arg\min(S)} = \text{negative sample} \\ &S = \text{ATTENTION SCORE} \end{aligned} \tag{3.6}$$

Normally, by maximizing the similarity between time series data and positive words embeddings, and minimizing the similarity with negative ones, we effectively narrow the representational space for these two modalities. The loss function can be formulated as 3.7. Constraining the latent space of time series and language embeddings, the positively paired data is then concatenated followed by a projection layer, enhancing the integration of the modalities. Subsequently, a single encoder is used to project the outputs from both the attention mechanism and the contrastive learning process, yielding a well-integrated and learned representation of the time series and text data. Finally, this representation will be input into a pretrained large language model, facilitating a deeper understanding of the underlying multimodal interactions.

$$Loss_{contrast} = \sum_{i=1}^{N} [\max(0, m + \text{sim}(P, \text{test}_n) - \text{sim}(P, \text{test}_p))] \tag{3.7}$$

Where:

- m is a margin, a hyper-parameter.
- P are the patched time series.

- $\text{sim}(a, b)$ represents any similarity function, we use cosine similarity in this paper.

Learning objective

As previously mentioned, our primary objective is to minimize the Mean Squared Error (MSE), which focuses on reducing the error magnitude within our model predictions. However, considering the integration of time series with linguistic features, it becomes necessary to incorporate additional loss components that cater to the directionality of these elements. Consequently, our composite loss function is formulated as follows:

$$Loss = (1 - \alpha) \cdot Loss_{MSE} + \alpha \cdot Loss_{contrast}$$

Here, α is a weighting factor that balances the impact between the MSE loss and the Contrastive Loss. This balance is crucial as it allows the model to not only focus on minimizing prediction error but also on enhancing the differentiation capabilities between distinct sequences. Empirical evidence suggests that α values below 0.5 are generally effective. Through extensive experimentation, an optimal range for α has been identified between 0.15 and 0.30. This range provides a robust trade-off, significantly enhancing model performance by integrating the predictive accuracy with the ability to distinguish between nuanced differences in data.

CHAPTER 4

EXPERIMENT

4.1 Datasets

FNSPID In our study on financial stock market forecasting, we utilize the FNSPID[29] dataset (Financial News and Stock Price Integration Dataset). This comprehensive dataset is specifically engineered to augment the accuracy of market predictions by merging multimodal features that include both quantitative data, such as stock prices, and qualitative data from financial news. Spanning from 1999 to 2023, it encompasses 29.7 million stock price entries and 15.7 million financial news records across 4,775 companies in the S&P 500, obtained from four leading stock market news outlets. The inclusion of diverse data types, particularly authoritative news articles, makes FNSPID exceptionally suitable for our analysis. However, the stock market dataset is incomplete and contains some noise. The collected data are sometimes discontinuous and exhibit gaps. For example, AAPL has only two years of data, despite being a public company for over 30 years. More severely, in some companies, we found gaps exceeding 20 days. These interruptions in daily data can disrupt patterns, making them harder for models to recognize.

ETT-small To compare the effectiveness of our framework, we utilize the ETT-small[30] datasets with other models. Specifically the ETT-small-h1 and ETT-small-m1. These datasets are meticulously curated to enhance the accuracy of predictive models by integrating multivariate features. They contain a wealth of data points—approximately 70,080 per variant—capturing intricate patterns across eight features, including the crucial 'oil temperature' of electrical transformers. Spanning two years, these datasets encompass a comprehensive array of data entries from two regions in a Chinese province, capturing the intricacies of electrical consumption and transformer conditions.

NFT In our analysis of Non-Fungible Tokens (NFTs), we utilize the NFT dataset[31], comprising transaction and Twitter activity data from 19 prominent NFT collections and their respective copycat collections, ranked by Opensea. This comprehensive dataset encompasses a total of 2,515,400 assets, 1,712,883 tweets, and 6,237,735 transactions. It furnishes intricate details about each collection, including the collection name, asset count, contract addresses, associated Twitter accounts, originality status, and tallies of filtered tweets and transactions. The dataset serves as a crucial resource for scrutinizing the impact of social media engagement on NFT prices, offering invaluable insights into market dynamics, community participation, and the potential for forecasting price trends based on social media interactions.

4.2 Main Results

All experimental results of LLM4FN (our models) are based on soft alignment. We will discuss the reasons and details in the ablation study. We apply most recent SOTA methods in time series forecasting research to the FNSPID stock market datasets. We follow the author's benchmarks [29]. We train and test individual models for each company, and we use MSE (Mean Squared Error) and MAE (Mean Absolute Error) to measure the performance of different models. To test real-world efficiency, we also follow the dataset's guide of using the 50 most famous companies. The experimental results Table 4.1 are as follows:

Table 4.1: Experiment for FNSPID datasets

FNSPID Dataset with different time length	**# 50**		**# 50 w/ full**	
	MSE	**MAE**	**MSE**	**MAE**
Transformer	6.5270	1.2331	11.8409	1.8499
DLinear	**0.0166**	**0.0718**	**0.0003**	**0.0064**
PatchTST	0.0170	0.0727	0.0004	0.0070
GPT4TS	0.0169	0.0716	0.0015	0.0288
TimeLLM	0.0170	0.0727	0.0010	0.0136
LLM4FN	**0.0166**	**0.0703**	**0.0003**	**0.0054**

Red: the best, **Black**: the second best. This table presents the initial results for the FNSPID datasets. '# 50' means we use the benchmark from FNSPID, '# 50 w/ full' means we use complete data for those 50 companies stock price prediction.

Among the models evaluated, GPT4TS[5], TimeLLM[7], and LLM4FN (ours) are Large Language Models for Time Series (LLM4TS), all based on transformer architecture. Our model, LLM4FN, achieved the best performance, followed closely by DLinear[32], a fully linear model. Interestingly, despite the dominance of our model, DLinear outperformed all other transformer-based state-of-the-art models in last two years. This highlights an intriguing phenomenon: a simple linear model can surpass most advanced transformer models in this context. This result aligns with our expectations due to the nature of benchmark approach—training: we train a separate model for each individual company resulting in a limited data volume per model. This limited data size is insufficient for Transformers, particularly those based on large language models, to effectively learn and generalize. With their substantial number of parameters, these models are especially prone to over-fitting when faced with restricted data volumes, leading to suboptimal performance. Moreover, Transformer-based methods excel at capturing semantic information in long sequence. Considering the incompleteness of the FNSPID dataset, as discussed in section 4.1, the time lengths, ranging from 400 to 2000, and the time windows of 50 timesteps are too short for effective learning. Consequently, the suboptimal performance of the models on the FNSPID dataset, which requires predictions over just a three-day window, is not surprising. In this context, simpler models perform better.

To validate our hypothesis, we use a more complete set of stock prices from the time these companies were listed. After filling the gaps, the number of timesteps increases from 2000 to 8000. Although these datasets are still not long enough for transformer-based models to fully learn the patterns, we observe a significant improvement when using complete data compared to incomplete data. Notably, the performance of transformer-based models shows a remarkable surge, demonstrating their potential when provided with more comprehensive datasets.

4.3 Ablation Study

4.3.1 Hard Alignment and Soft Alignment

The performance metrics for hard alignment indicate a Mean Squared Error (MSE) of 0.0377 and a Mean Absolute Error (MAE) of 0.01132. These values are more than twice as poor as those for soft alignment, which achieves an MSE of 0.0166 and an MAE of 0.0703. When compared to other state-of-the-art models in recent years, hard alignment is significantly inferior either. The reason for this becomes evident when we visualize the corresponding textual and time series pairs, as shown in Figure 4.1.

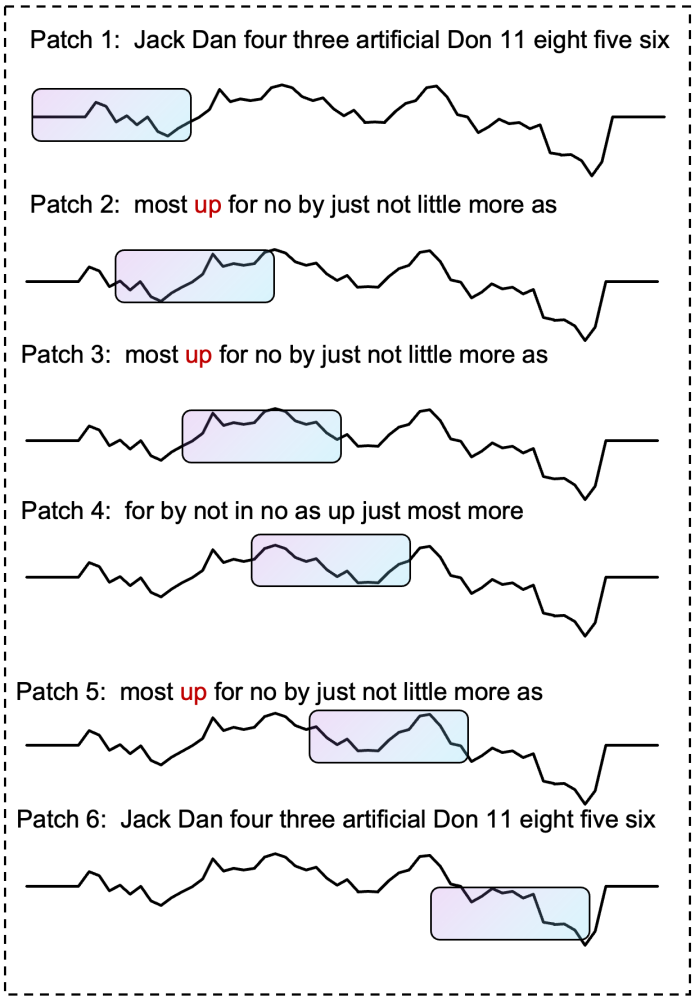

Figure 4.1: Positive Samples and Corresponding Time Patches

In this figure, six time series patches are presented, each aligned with its corresponding textual information. We selected the top 10 nearest neighbors from encoded embeddings for positive samples, resulting in the depicted alignments. Notably, patches 2, 3, and 5 share

the same positive text. Both patches 2, 3 and 5 exhibit an increasing trend in the series. Patches 1 and 6, which share the same positive samples, illustrate a downward trend in textual information.

Mirchandani's research [33] has demonstrated that large language models (LLMs) are capable of pattern learning. Soft alignment leverages this ability, using language to guide time series forecasting and learning, even if the language used is logically and semantically incomprehensible.

In contrast, hard alignment relies heavily on the semantics of the news summaries (Figure 3.2). Without the expertise of a financial analyst, relying solely on the model's computational capabilities proves ineffective.

4.3.2 Attention Mechanism and Contrastive Learning

We did ablation study to explore the contribution and effectiveness of contrastive learning and attention work in a more authoritative dataset (ETDataset [30]) and reliable existing LLM4TS model (Time-LLM[7]). Because our aim is not simply push the performance boundary further by using a larger pre-trained model, such as LLama, we use GPT-2 in this ablation study to prove our workable framework. We use ETT-h1 and ETT-m1 for study. For both of these dataset, we predict long-term time series by 96 timesteps.

From Table 4.2, we could see for ETTh1-96 and ETTm1-96 dataset, GPT-2 with contrastive learning and double attention reached the best result. Our frame work only using 6 layers pre-trained GPT-2 surpass pre-trained Llama with 8 layers and GPT-2 with 12 layers. Followed by GPT-2 with contrastive learning because it constrain latent space with a loss function, which is proved to be valuable. No matter which improvement, they all exceed Time-LLM for GPT-2 with 6 layers.

Additionally, it is challenging to determine which component of our LLM4FN framework contributes more significantly to the overall performance. The functionalities of contrastive learning and double attention are tightly integrated, as the attention scores generated by the double attention network are crucial for the subsequent comparison and contrasting processes in contrastive learning.

Thus, from experiment study Table 4.2, we could say that double attention and contrastive learning could imrpove pre-trained LLM's performance.

Table 4.2: Ablation Study for Contrastive Learning and Double Attention

default GPT-2(6)	ETTh1-96 MSE	ETTh1-96 MAE	ETTm1-96 MSE	ETTm1-96 MAE
TIME-LLM	0.3946	0.4185	0.3110	0.3612
LLM4FN w/o DA	0.3853	0.4116	**0.2921**	**0.3495**
LLM4FN w/o CL	**0.3796**	**0.4090**	0.2932	0.3495
LLM4FN	**0.3718**	**0.4023**	**0.2887**	**0.3442**
TIME-LLM (LLama (8))	0.3890	0.4151	0.2970	0.3815
TIME-LLM (GPT-2 (12))	0.3850	0.4115	0.3060	0.3843

Red: the best, **Black**: the second best. Default methods are based on a 6-layer GPT-2 as the backbone, except for models specifically marked. The notation '(#)' indicates a #-layer model. DA represents "Double Attention" and CL represents "Contrastive Learning".

4.4 Unified Model and Zero Shot Transferring

In our prior experiments, we focused on training individual models tailored to each company's unique dataset. This approach, while effective in its specific context, is unrealistic and over-fitting, consumes lots of time, and deviates significantly from the concept of large, foundational models that are renowned for their versatility and exceptional performance across a range of downstream tasks on complex, cross-domain or unseen data.

This realization prompted a pivotal shift in our research objectives. Inspired by the capabilities of large language models and large vision models, which excel across diverse applications without task-specific training, we sought to develop a universal model. Our aim was to transcend the limitations of individualized training and create a model robust enough to generalize across various financial domains without the need for retraining.

To bridge this gap, we embarked on an experiment employing unified training and zero-shot learning techniques. For unified traininig, showed as Figure 4.2, we conform all fifty companies data by chunking them into equal-sized batches, and test on the same part as same as individual models. For Zero-shot learning, our strategy initially involved training on a dataset from a subset of companies in FNSPID's benchmark and then testing the model's

predictive accuracy on stock prices for an entirely different set of 50 companies. However, this approach not only challenges but also aims to overcome the inherent limitations posed by the varying data characteristics of different companies, such as dataset size and time sequence variations.

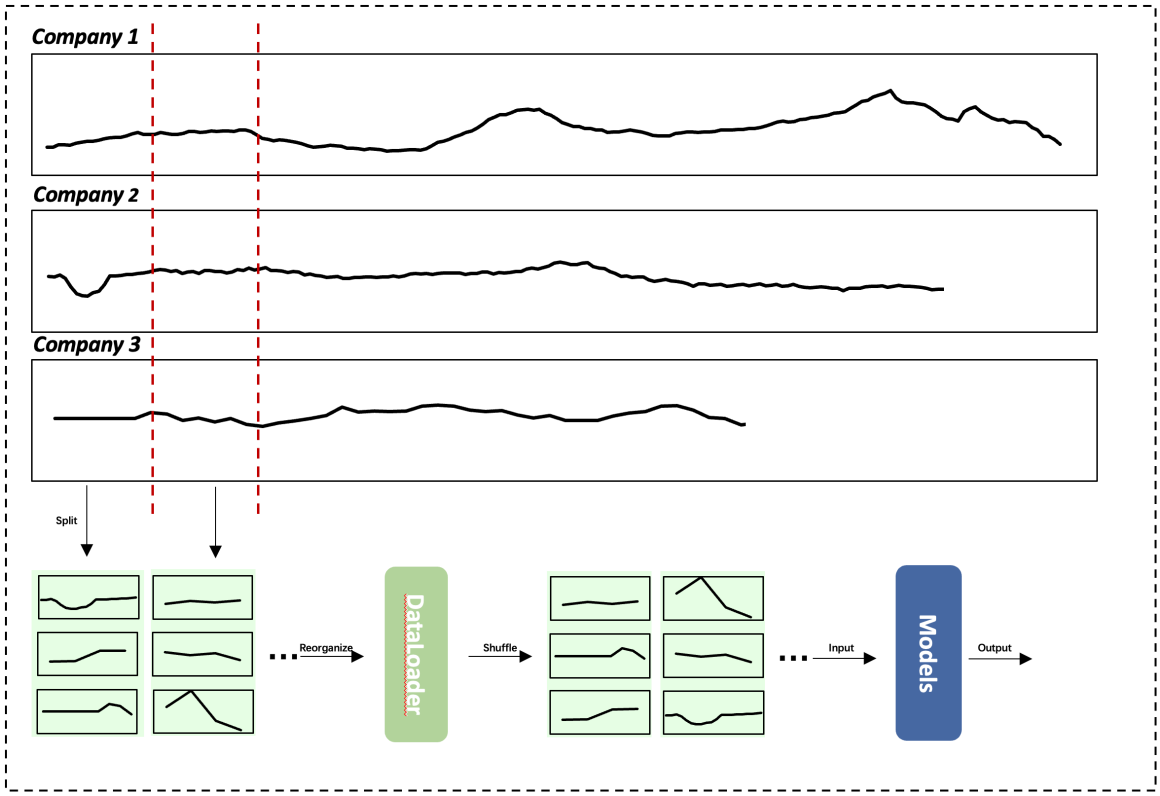

Figure 4.2: Unified Training

Note: Illustrated with examples from three companies

We chose not to adopt an sequential learning method from one company to the next due to the risk of catastrophic forgetting—where earlier learned information is overwritten as new data is introduced. This risk is exacerbated as more companies are added to the training sequence, causing earlier patterns to be lost. Additionally, treating the data from multiple companies as a series of separate objectives complicates the learning process; balancing the weights of different losses can lead to either slow updates or diluted model objectives.

To address these challenges, we developed a method called "batch extraction and reloading strategy" – involving the extraction and shuffling of company batches. By organizing the data into manageable batches and shuffling them each epoch, we significantly mitigate the risk of catastrophic forgetting. This method, however, disrupts the long-term dependencies

within the datasets within individual companies or specific industries. While these dependencies are crucial for capturing domain-specific nuances, our overarching goal was to create a model that prioritizes general applicability and robustness over domain confinement.

This innovative approach aims to lay the groundwork for a foundational model capable of universally predicting all financial-related time series without prior exposure, thereby revolutionizing the application of predictive modeling in financial analysis.

Table 4.3: **UNIFIED-TRAINING** Results for the FNSPID Dataset

Unified Training	FNSPID-50 MSE	FNSPID-50 MAE
DLinear	0.0208	0.0883
PatchTST	0.0187	0.0806
GPT4TS	0.0159	0.0661
TimeLLM	0.0160	0.0674
LLM4FN	**0.0152**	**0.0608**

Red: the best, **Black**: the second best. This table summarizes the preliminary metrics (MSE, MAE) for different models on the FNSPID-50 dataset based on unified training. All LLM4TS models are based on 6-layer GPT-2.

Table 4.4: **ZERO-SHOT** Transferring Results for the FNSPID Dataset

Zero-Shot	FNSPID-50 MSE	FNSPID-50 MAE
DLinear	0.5452	0.5777
PatchTST	0.0373	0.1130
GPT4TS	0.0192	0.0827
LLM4FN (freeze)	**0.0185**	**0.0616**
LLM4FN (unfreeze)	**0.0184**	**0.0611**

Red: the best, **Black**: the second best. This table summarizes the preliminary metrics (MSE, MAE) for different models on the FNSPID-50 dataset. 'Freeze' refers to freezing layers during training, while 'unfreeze' indicates that the freed forward layers are left trainable. All methods are based on a 6-layer GPT-2 as the backbone.

Unified-Training The results of our experiments with the unified FNSPID dataset are presented in Table 4.3. Consistent with our theoretical predictions in Section 4.2, transformer-based methods typically underperform with sparse training data. Yet, our analysis demonstrates that these methods excel within a more extensive, complex and unified dataset, particularly when data volume exceeds three gigabytes. This extensive data integration significantly enhances the model's ability to generalize across diverse financial scenarios, which contrasts sharply with the performance of DLinear. Unlike in individual dataset contexts where DLinear was effective, it could not maintain its efficacy in the unified setting. In stark contrast, all models under the LLM4TS umbrella, especially our LLM4FN model, showcased superior performance, outstripping benchmarks set by conventional models. The

exceptional efficacy of LLM4FN is largely due to its sophisticated integration of textual and time-series data, providing robust predictive capabilities across varying financial domains. Also, when we go back to compare it's performance with individual training dataset 4.1, all LLM based models' abilities are improved.

Zero-Shot Transferring Our zero-shot learning experiment results are detailed in Table 4.4. This innovative approach tested the LLM4FN model's adaptability without prior direct exposure to the test data's specific financial domain. Notably, LLM4FN, both in its 'frozen' and 'unfrozen' parameter states, achieved the lowest mean squared error (MSE) and mean absolute error (MAE) scores among the tested models. This indicates a robust model capability to generalize from training on a subset of financial data to accurately predicting outcomes on an entirely different set. The superior performance of the 'unfrozen' version, where the forward layers were adaptable during training, underscores the benefit of allowing the model dynamic adjustments when faced with new data types. This flexibility is crucial for applications in dynamic environments such as financial markets, where market conditions and data patterns can change unpredictably. The use of a unified dataset enriched with sufficient news content and time series data has significantly propelled our advancements towards a truly adaptable and effective unified model. Employing a pre-trained GPT 6-layer model as the backbone further highlights the potential of scalable, transferable models in achieving high-accuracy predictions in zero-shot settings.

Table 4.5: Experimental Results for NFT Dataset

	OthersideMeta		**Moonbirds**		**Doodles**		**CryotoKitties**		**BoredApeYC**	
	MSE	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**
PatchTST	0.0004	0.0049	0.0060	0.0446	0.0196	0.0810	0.0184	0.0858	0.0393	0.1314
GPT4TS	0.0004	0.0046	0.0067	0.0501	0.0196	0.0813	0.0179	0.0870	0.0395	0.1314
zero-shot LLMFN	**0.0003**	**0.0045**	**0.0053**	**0.0395**	**0.0177**	**0.0683**	**0.0162**	**0.0768**	**0.0011**	**0.1111**

Red: the best.

NFT Evaluation We employed a unified training model, training on FNSPID-50 datasets and directly applying it to the NFT dataset for prediction. This approach aimed to validate the model's powerful zero-shot transfer capability. Unlike our model, all other models were trained using supervised learning on the NFT dataset itself. In other words, the competing

models were trained and tested on the same NFT dataset, while our approach involved testing a pretrained model, without any additional training, directly on the NFT dataset. Remarkably, our model achieved leading positions across all metrics, demonstrating its superior zero-shot transfer performance. The experimental results from table 4.5 showcase the effectiveness of our approach. Across the various NFT datasets, including OthersideMeta, Moonbirds, Doodles, CryotoKitties, and BoredApeYC, our LLMFN Zero-Shot model consistently outperformed the PatchTST and GPT4TS baselines, achieving the lowest Mean Squared Error (MSE) and Mean Absolute Error (MAE) scores. This highlights the model's ability to generalize well to unseen NFT data without requiring explicit training on the target domain. Overall, these findings underscore the significant potential of our unified training approach and zero-shot transfer capabilities for real-world applications in the NFT space and beyond. The strong performance achieved without fine-tuning on the target dataset is particularly promising, as it reduces the need for extensive data collection and model retraining for each new domain.

Figure 4.3: Text-prototypes for LLM4TS

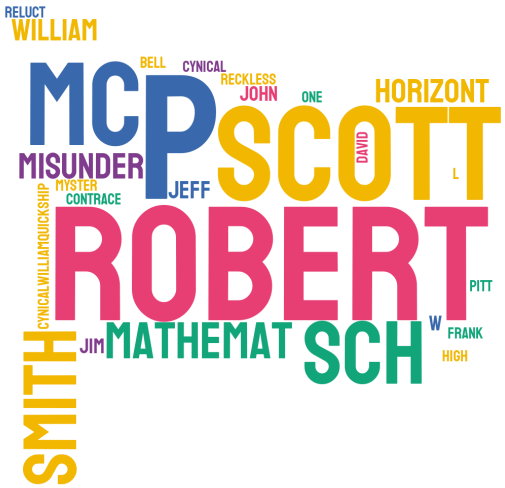

Figure 4.4: Text-prototypes of Other LLM4TS Models

Supplementary Experiment We also visualize the text prototypes after utilizing unified training on the FNSPID dataset through the top 50 companies. As shown in Figure 4.3, our models' text prototypes are all related to the financial market, including terms like "investors," "growers," and "leasing." In contrast, the text alignment methods used in LLM4TS often

produce prototypes involving human names, showed in Figure 4.4, such as "Robert" and "Scott." Additionally, we removed all stop words (e.g., "a," "an," "the," "to") to improve visualization clarity. For our model, stop words account for only 10%, whereas in other methods, they exceed 90%. From this comparison, we can confidently conclude that our alignment method enables the model to learn the context of the dataset, thereby improving time series forecasting.

CHAPTER 5

CONCLUSION AND FUTURE WORK

Through numerous experiments and ablation studies, we have demonstrated that integrating textual data with time series significantly enhances the performance of large language models in time series forecasting. This integration is particularly effective in unified training with a large amount of high-quality data, and zero-shot learning within similar fields. We are confident that with access to larger, high-quality datasets and more extensive pre-training, we can develop a unified model for financial forecasting.

Our results show that even relatively small pre-trained models can deliver excellent performance. Moving forward, we plan to enrich our framework by incorporating a broader range of high-quality data and by utilizing larger, more advanced pre-trained models. We will also consider unfreezing the models' parameters and tuning Parameter-Efficient Fine-Tuning (PEFT). This approach will not only help in refining the predictive accuracy of our models but also in extending their applicability across different domains within financial forecasting.

Ultimately, our goal is to push the boundaries of what is currently possible in predictive analytics for finance, making strides towards models that can accurately predict market movements under a variety of conditions and from minimal initial data. This effort will involve continued research into model scalability, data integration techniques, and the exploration of new methodologies for pre-training models specifically tailored for the complexities of financial data.

BIBLIOGRAPHY

[1] G. Gidofalvi and C. Elkan, "Using news articles to predict stock price movements. 2001," University of California, San Diego: Department of Computer Science and Engineering.

[2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark et al., "Learning transferable visual models from natural language supervision," in International conference on machine learning. PMLR, 2021, pp. 8748–8763.

[3] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat et al., "Gpt-4 technical report," arXiv preprint arXiv:2303.08774, 2023.

[4] T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, E. Hambro, L. Zettlemoyer, N. Cancedda, and T. Scialom, "Toolformer: Language models can teach themselves to use tools," Advances in Neural Information Processing Systems, vol. 36, 2024.

[5] T. Zhou, P. Niu, L. Sun, R. Jin et al., "One fits all: Power general time series analysis by pretrained lm," Advances in neural information processing systems, vol. 36, 2024.

[6] R. Godahewa, C. Bergmeir, G. I. Webb, R. J. Hyndman, and P. Montero-Manso, "Monash time series forecasting archive," arXiv preprint arXiv:2105.06643, 2021.

[7] M. Jin, S. Wang, L. Ma, Z. Chu, J. Y. Zhang, X. Shi, P.-Y. Chen, Y. Liang, Y.-F. Li, S. Pan et al., "Time-llm: Time series forecasting by reprogramming large language models," arXiv preprint arXiv:2310.01728, 2023.

[8] C. Sun, Y. Li, H. Li, and S. Hong, "Test: Text prototype aligned embedding to activate llm's ability for time series," arXiv preprint arXiv:2308.08241, 2023.

[9] Y. Wang, "Volatility spillovers across nfts news attention and financial markets," International review of financial analysis, vol. 83, p. 102313, 2022.

[10] W. Khan, M. A. Ghazanfar, M. A. Azam, A. Karami, K. H. Alyoubi, and A. S. Alfakeeh, "Stock market prediction using machine learning classifiers and social media, news," Journal of Ambient Intelligence and Humanized Computing, pp. 1–24, 2022.

[11] Y. Li and W. Ma, "Applications of artificial neural networks in financial economics: a survey," in 2010 International symposium on computational intelligence and design, vol. 1. IEEE, 2010, pp. 211–214.

[12] B. Xie, R. Passonneau, L. Wu, and G. G. Creamer, "Semantic frames to predict stock price movement," in Proceedings of the 51st annual meeting of the association for computational linguistics, 2013, pp. 873–883.

[13] X. Li, H. Xie, L. Chen, J. Wang, and X. Deng, "News impact on stock price return via sentiment analysis," Knowledge-Based Systems, vol. 69, pp. 14–23, 2014.

[14] L. Zhang, C. Aggarwal, and G.-J. Qi, "Stock price prediction via discovering multi-frequency trading patterns," in Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, 2017, pp. 2141–2149.

[15] X.-Y. Liu, Z. Xiong, S. Zhong, H. Yang, and A. Walid, "Practical deep reinforcement learning approach for stock trading," arXiv preprint arXiv:1811.07522, 2018.

[16] D. Yang, L. Zhao, Z. Lin, T. Qin, J. Bian, and T.-Y. Liu, "Fully parameterized quantile function for distributional reinforcement learning," Advances in neural information processing systems, vol. 32, 2019.

[17] Z. Li, X.-Y. Liu, J. Zheng, Z. Wang, A. Walid, and J. Guo, "Finrl-podracer: high performance and scalable deep reinforcement learning for quantitative finance," in Proceedings of the second ACM international conference on AI in finance, 2021, pp. 1–9.

[18] Q. Zhang, C. Qin, Y. Zhang, F. Bao, C. Zhang, and P. Liu, "Transformer-based attention network for stock movement prediction," Expert Systems with Applications, vol. 202, p. 117239, 2022.

[19] A. Kapoor, D. Guhathakurta, M. Mathur, R. Yadav, M. Gupta, and P. Kumaraguru, "Tweetboost: Influence of social media on nft valuation," in Companion Proceedings of the Web Conference 2022, 2022, pp. 621–629.

[20] J. Luo, Y. Jia, and X. Liu, "Understanding nft price moves through social media keywords analysis," arXiv preprint arXiv:2209.07706, 2022.

[21] J. Christopher Westland, "Periodicity, elliott waves, and fractals in the nft market," Scientific Reports, vol. 14, no. 1, p. 4480, 2024.

[22] M. Choi, H. J. Lee, S. H. Park, S. W. Jeon, and S. Cho, "Stock price momentum modeling using social media data," Expert Systems with Applications, vol. 237, p. 121589, 2024.

[23] D. Costa, L. La Cava, and A. Tagarelli, "Show me your nft and i tell you how it will perform: Multimodal representation learning for nft selling price prediction," arXiv preprint arXiv:2302.01676, 2023.

[24] X. Yu, Z. Chen, Y. Ling, S. Dong, Z. Liu, and Y. Lu, "Temporal data meets llm–explainable financial time series forecasting," arXiv preprint arXiv:2306.11025, 2023.

[25] N. Gruver, M. Finzi, S. Qiu, and A. G. Wilson, "Large language models are zero-shot time series forecasters," Advances in Neural Information Processing Systems, vol. 36, 2024.

[26] H. Xue and F. D. Salim, "Promptcast: A new prompt-based learning paradigm for time series forecasting," IEEE Transactions on Knowledge and Data Engineering, 2023.

[27] D. Cao, F. Jia, S. O. Arik, T. Pfister, Y. Zheng, W. Ye, and Y. Liu, "Tempo: Prompt-based generative pre-trained transformer for time series forecasting," arXiv preprint arXiv:2310.04948, 2023.

[28] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," arXiv preprint arXiv:1609.02907, 2016.

[29] Z. Dong, X. Fan, and Z. Peng, "Fnspid: A comprehensive financial news dataset in time series," arXiv preprint arXiv:2402.06698, 2024.

[30] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting," in The Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Virtual Conference, vol. 35, no. 12. AAAI Press, 2021, pp. 11 106–11 115.

[31] J. Luo, Y. Jia, and X. Liu, "Understanding nft price moves through tweets keywords analysis," in Proceedings of the 2023 ACM Conference on Information Technology for Social Good, 2023, pp. 410–418.

[32] A. Zeng, M. Chen, L. Zhang, and Q. Xu, "Are transformers effective for time series forecasting?" in Proceedings of the AAAI conference on artificial intelligence, vol. 37, no. 9, 2023, pp. 11 121–11 128.

[33] S. Mirchandani, F. Xia, P. Florence, B. Ichter, D. Driess, M. G. Arenas, K. Rao, D. Sadigh, and A. Zeng, "Large language models as general pattern machines," arXiv preprint arXiv:2307.04721, 2023.