Measuring Consistency in Text-based Financial Forecasting Models

Linyi Yang[1,2*], **Yingpeng Ma**[1,2*], **Yue Zhang**[1,2†]
[1]Institute of Advanced Technology, Westlake Institute for Advanced Study
[2]School of Engineering, Westlake University
yanglinyi,mayingpeng,yuezhang@westlake.edu.cn

Abstract

Financial forecasting has been an important and active area of machine learning research, as even the most modest advantage in predictive accuracy can be parlayed into significant financial gains. Recent advances in natural language processing (NLP) bring the opportunity to leverage textual data, such as earnings reports of publicly traded companies, to predict the return rate for an asset. However, when dealing with such a sensitive task, the consistency of models – their invariance under meaning-preserving alternations in input – is a crucial property for building user trust. Despite this, current financial forecasting methods do not consider consistency. To address this problem, we propose FinTrust, an evaluation tool that assesses logical consistency in financial text. Using FinTrust, we show that the consistency of state-of-the-art NLP models for financial forecasting is poor. Our analysis of the performance degradation caused by meaning-preserving alternations suggests that current text-based methods are not suitable for robustly predicting market information. All resources are available at https://github.com/yingpengma/FinTrust.

1 Introduction

NLP techniques have been used in various financial forecasting tasks, including stock return prediction, volatility forecasting, portfolio management, and more (Ding et al., 2014, 2015; Qin and Yang, 2019; Xing et al., 2020; Du and Tanaka-Ishii, 2020; Yang et al., 2020a; Sawhney et al., 2020). Despite the increased performance of NLP models on financial applications, there has been pushback questioning their trustworthiness, and robustness (Chen et al., 2022; Li et al., 2022). Recently, the causal explanation has been viewed as one of the promising directions for measuring the robustness and thus improving the transparency of models (Stolfo et al., 2022; Feder et al., 2022). Among them, consistency has been viewed as a crucial feature, reflecting the systematic ability to generalize in semantically equivalent contexts and receiving increasing attention in tasks such as text classification and entailment (Jin et al., 2020; Jang et al., 2022).

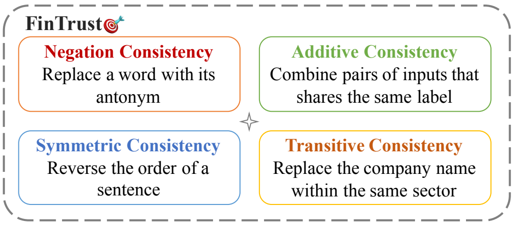
FinTrust🎯
Negation Consistency
Replace a word with its antonym
Additive Consistency
Combine pairs of inputs that shares the same label
Symmetric Consistency
Reverse the order of a sentence
Transitive Consistency
Replace the company name within the same sector

Figure 1: Examples of four consistency transformations used in FinTrust 🎯.

Previous text-based financial forecasting methods have mostly considered stock movement prediction based on various sources of data, including financial news (Xu and Cohen, 2018; Zhang et al., 2018), analyst reports (Kogan et al., 2009; Rekabsaz et al., 2017), and earnings conference calls (Qin and Yang, 2019; Keith and Stent, 2019; Li et al., 2020; Chen et al., 2021b). While most work evaluates their methods using accuracy and profit gains based on the final outcome in the market (Sawhney et al., 2021b; Yang et al., 2022), consistency evaluation remains largely unexplored. The only exception (Chuang and Yang, 2022) focuses on evaluating the implicit preferences in Pre-trained Language Models (PLMs) but not the consistency in predictive models. The lack of evaluation in behavior consistency, an important characteristic of human decisions, hinders the deployment of financial forecasting models in real-world scenarios.

The main objective of this work is to explore a wholistic measure for stock movement prediction, integrating consistency as a criterion of trustworthiness. To this end, we define *behavior consistency*

[*] Equal contribution. Yingpeng Ma did this work during his internship at Westlake University.

[†] Correspondence to: zhangyue@westlake.edu.cn.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 13751–13765
July 9-14, 2023 ©2023 Association for Computational Linguistics

of text-based models in the financial domain. Regarding the intrinsic characteristics of financial text data, we consider four types of logical consistency tests. As shown in Figure 1, these transformations include Negation Consistency, Symmetric Consistency, Additive Consistency, and Transitive Consistency. Taking negation consistency as an example, given an input *"the cost of raw materials has been greatly decreased"*, if the token *"decreased"* is changed to *"increased"*, the model prediction is expected to be flipped accordingly.

Based on the above logical transformations, we introduce FinTrust 🎯, a new evaluation tool that enables researchers to measure consistency in PLMs and text-based financial forecasting models. Using FinTrust 🎯, we design three tasks to investigate the influence of these logical transformations. First, we assess implicit preference in PLMs such as BERT (Devlin et al., 2018) and FinBERT (Yang et al., 2020b), especially for economic words. Second, we measure the accuracy of stock movement prediction on a real-world earnings call dataset after the meaning-preserving modifications. Finally, we propose a realistic trading simulation to see if simple meaning-preserving modifications can wipe out positive returns.

Experiments on several baseline models, including previous best-performing architectures (Ding et al., 2015; Qin and Yang, 2019; Yang et al., 2020a) and the machine learning classifier (Chen et al., 2015) show that all current methods exhibit a significant decline in the performance of stock movement predictions when evaluating on FinTrust compared to their original results. Notably, some models demonstrate a level of accuracy that is even lower than that of a random guess after undergoing logical consistency transformation, and most methods fail to surpass the performance of the simplest Buy-all strategy in the trading simulation. These results suggest that existing text-based financial models have robustness and trustworthiness issues, which can limit their use in practical settings.

To our knowledge, FinTrust 🎯 is the first evaluation tool for probing if the relatively accurate stock movement prediction is based on the right logical behavior. We release our tool and dataset at Github[†], which can assist future research in developing trustworthy FinNLP methods.

2 Related Work

Text-based Financial Forecasting. A line of work has leveraged event-based neural networks based on financial news for predicting the stock movement of S&P 500 companies (Ding et al., 2014, 2015; Xu and Cohen, 2018). By taking advantage of recent advances in NLP, recent work has shown potential in predicting stock price movements using PLMs, BERT (Devlin et al., 2018), and FinBERT (Araci, 2019; Yang et al., 2020b), with rich textual information from social media and earnings conference calls (Liu and Tse, 2013; Xing et al., 2020; Chen et al., 2021a). The considerable PLMs mainly include BERT and FinBERT. While BERT is trained on corpora from fairly general domains, FinBERT is trained on financial corpora, including earnings conference calls and analyst reports, under the same architecture as BERT. Although implicit stock market preference is in the masked token predictions task, the implicit preference has been under-explored using a logical behavior test.

In addition to building pre-trained models specially trained for financial domains, researchers have recently proposed myriad neural network architectures aimed at more accurate predictions to produce profitable gains including financial risk (volatility) and return predictions. For example, researchers (Qin and Yang, 2019; Yang et al., 2020a; Sawhney et al., 2021a) have considered predicting the volatility of publicly traded companies based on multi-model earnings conference call datasets. Also, Xu and Cohen (2018); Duan et al. (2018); Yang et al. (2018); Feng et al. (2019) leverage different textual data sources for predicting the stock movement based on the daily closing price. Unfortunately, despite the alarm over the reliance of machine learning systems on spurious patterns that have been found in many classical NLP tasks, the topic of text-based financial forecasting lacks a systematical evaluation regarding the robustness analysis from either an adversarial or consistency perspective. To this end, we present the first critical investigation of popular benchmarks by using FinTrust from the consistency perspective.

Consistency Measurement. The inductive bias of machine learning systems is greatly affected by the patterns in training data due to the nature of inductive reasoning. While a flurry of research has highlighted this issue (Gururangan et al., 2018; Srivastava et al., 2020; Garg and Ramakrishnan, 2020; Kaushik et al., 2020), recent work Jang et al. (2022)

[†] https://github.com/yingpengma/FinTrust

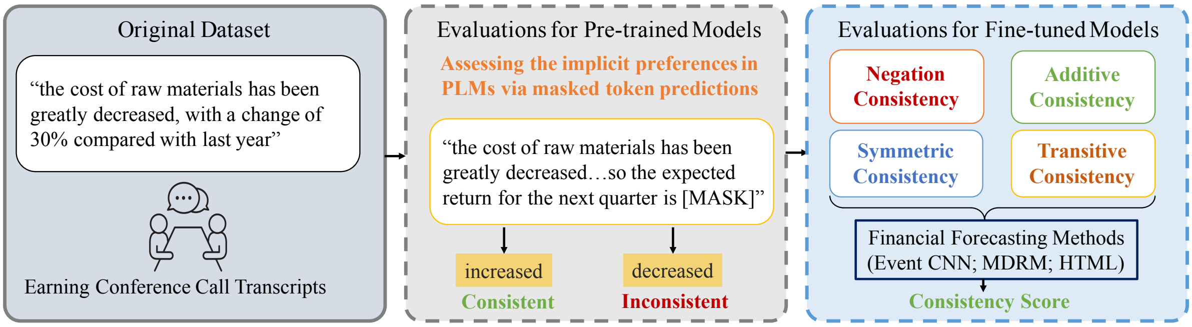
Original Dataset
“the cost of raw materials has been greatly decreased, with a change of 30% compared with last year”
Earning Conference Call Transcripts
Evaluations for Pre-trained Models
Assessing the implicit preferences in PLMs via masked token predictions
“the cost of raw materials has been greatly decreased…so the expected return for the next quarter is [MASK]”
increased
decreased
Consistent
Inconsistent
Evaluations for Fine-tuned Models
Negation Consistency
Additive Consistency
Symmetric Consistency
Transitive Consistency
Financial Forecasting Methods (Event CNN; MDRM; HTML)
Consistency Score

Figure 2: Pipelines of FinTrust 🎯 consist of evaluating pre-trained features (e.g., BERT and FinBERT) and fine-tuned text-based financial forecasting models.

shows that possible artefacts in data are more influential than the model design when leading to the problem of lacking trustworthiness. Thus, assessing the influence of data artefacts, such as consistency, becomes a crucial problem for trustworthy NLP. Elazar et al. (2021) study the consistency of PLMs (e.g., BERT, ALBERT, and RoBERTa) with regard to their knowledge extraction ability and conclude that the consistency of these models is generally low. Chuang and Yang (2022) aim to raise awareness of potential implicit stock preferences based on the finding that consistent implicit preference of the stock market exists in PLMs at the whole market.

In addition to evaluating preferences in PLMs, previous methods also attempt to evaluate the consistency of models in downstream NLP tasks, such as visual question answering (Ribeiro et al., 2018), QA (Jia and Liang, 2017; Ribeiro et al., 2019; Gan and Ng, 2019; Asai and Hajishirzi, 2020), named entity recognition (Jia et al., 2019; Wang and Henao, 2021), and natural language inference (Naik et al., 2018; Hossain et al., 2020; Camburu et al., 2020; Sinha et al., 2021). Besides, Ribeiro et al. (2020) consider using consistency for building the behavioural testing benchmark beyond accuracy. Surprisingly, these discussions have not yet been extended to text-based financial forecasting models, which require strong robustness to assist decision-making in the financial market, with the exception of our work.

3 Method

We define the pipeline of FinTrust 🎯 in Figure 2. For text-based financial models, there are two salient components, namely text representations and financial behavior. For the former, using PLMs has become a dominant approach, improving the quality of text representations in many domains. For the latter, various neural models can be built on PLMs. Correspondingly, we have two setups in the consistency evaluation, representation (Setup 1) and behavior (Setup 2), respectively.

Setup 1. In the first stage, we assess the implicit preferences in PLMs via masked token predictions. In particular, we first mask a predictable word from the original input extracted from earning conference call transcripts, such as *"the cost of raw materials has been greatly decreased...so the expected return for the next quarter is [MASK]"*. Then, we predict the masked token using PLMs and compare the probability of predicting "increased" and "decreased" for contexts from different transcripts. A higher probability of predicting "increased" would indicate that the given PLM hold logical consistency with human predictions. Conversely, it suggests that the prediction of the PLM may be influenced by spurious patterns such as favoritism towards a particular stock.

Setup 2. We evaluate text-based financial forecasting models after fine-tuning PLMs on a popularly used earnings conference call dataset (Qin and Yang, 2019). However, the consistency measurement faces significant challenges in defining the relationship between two texts, particularly when the text is a long transcript with complex logical connections, such as earnings conference call transcripts. Incorrectly defining this relationship can render consistency judgments meaningless. In line with prior research (Jang et al., 2022), we develop four logical consistency transformations customized for financial text in this work. By meaning-preserving altering the original text, we ensure generated samples have a logical relationship to the original text, thus ensuring the consistency judgment is meaningful. Below we define our text level consistency transformation first (Sec 3.1), before introducing the financial tasks for behavior study (Sec 3.2) and a wholistic metric (Sec 3.3) to integrate performance and trustworthiness.

3.1 Logical Consistency Transformations on Text Data

In FinTrust, four logical consistency transformation approaches are defined to evaluate if the model maintains the same logical behavior as humans, representing the consistency in text-based financial forecasting models.

Negation consistency refers to the ability of a model to generate converse predictions for texts with opposite meanings, i.e. $f(x) = positive \Leftrightarrow f(\neg x) = negative$, where x is the input transcript, $f(x)$ represents the output of the model, a "positive" outcome means the stock price will increase, and a "negative" outcome means the stock price will decrease. $\neg x$ is a negation consistency transformed test example flipped through predetermined rules based on the bi-grams of the most frequent words and their antonyms. We achieve this by splitting the dataset at the sentence level and flipping the meanings of sentences. Given an input *"the cost of raw materials has been greatly decreased, with a change of 30% compared with last year"*, its counterpart can be *"the cost of raw materials has been greatly increased, with a change of 30% compared with last year"*. In the financial market, a significant cost reduction may lead to optimism about the company's future prospects and an increase in stock price. Only when the model can give the correct predictions for both pairs of testing data we consider that the model is consistent with non-contradictory predictions. Otherwise, it is considered to lack negation consistency.

Symmetric consistency is the property of a model where the order of the inputs does not affect the output. It is defined as $f(S_{p1}, S_{p2}) = f(S_{p2}, S_{p1})$, where S is a sentence in the transcript, S_{pi} represents the part i of the sentence. This can be tested by reordering the segments of each sentence in the transcript and comparing the predictions before and after the reordering. For example, given the sentence *"the cost of raw materials has been greatly decreased, with a change of 30% compared with last year"*, if the prediction is reversed after reordering it to *"with a change of 30% compared with last year, the cost of raw materials has been greatly decreased"*, then the model is regarded as lacking symmetric consistency.

Additive consistency refers to the property of a model to predict the stock movement based on the combination of two inputs, x and y that share the same label. The model is expected to hold the same prediction for x, y, and the concatenation of those inputs $x + y$. If the model produces different predictions for the above three kinds of inputs, it can be regarded as lacking additive consistency. For example, if a model gives a positive prediction for the sentence *"the cost of raw materials has been greatly decreased, with a change of 30% compared with last year"*, and also gives a positive prediction for the sentence *"we believe that our products can bring convenience to everyone's life"*, then it should also make a positive prediction for the combined sentences after the concatenation.

Transitive consistency refers to the ability of a model where the perceived sentiment of a company should be reflected in the performance of the top-valued company in the same industry. It can be expressed as $f(x) = f(x')$, where x' represents transitive consistency transformed text. Specifically, for transcripts of a particular company, the top-valued company in the same industry is identified and its name is denoted as "company_name". Then occurrences of words such as "we" and "our" are replaced with "company_name" and "company_name's" respectively. For example, if the corresponding sector of the company is "Information Technology" and the top-valued company in the S&P 500 is Apple Inc., a sentence such as "*we believe that our products can bring convenience to everyone's life*" will be transformed to "*Apple Inc. believe that Apple Inc.'s products can bring convenience to everyone's life*" after transitive consistency transformation. Again, we calculate the consistency of models by considering the non-contradictory predictions over transitive instances.

3.2 Prediction Tasks in FinTrust

Consistency Measurement in PLMs. To better assess the implicit preference in PLMs, we extend the previous cloze-style prompts used in assessing stock market preference (Chuang and Yang, 2022) by considering logical changes rather than simply predicting the masked token in the input. This is crucial as if PLMs are biased, the fine-tuned model's predictions based on features learned by PLMs could be further influenced by spurious preference tendencies, which would negatively impact the effect of financial forecasting.

Stock Prediction Task. Following previous studies (Ding et al., 2015; Duan et al., 2018; Sawhney et al., 2020), we treat the stock movement prediction as a binary classification problem, where the model predicts whether the daily closing price of a given asset will increase or decrease over the next *n* days (*n*=3, 7, 15, 30) based on the content of earnings call transcripts. The output is either "increase" (positive) or "decrease" (negative).

Trading Simulation Task. We use the predictions to determine whether to buy or sell a stock after n days. For example, if the model predicts that the stock price would increase from day d to day d+30, we would buy the stock on day d and sell it on day $d + 30$. Otherwise, we execute a short sell. The previous work Sawhney et al. (2021a) simulates the trade of one hand for each stock, which allows for the potential offset of multiple forecast failures if one stock is more valuable. However, this approach is unfair under specific situations since each prediction and trade are treated equally and thus will lose the balance between trades. Therefore, we invest the same amount of money in each stock and calculate the profit ratio instead of the cumulative profit. This method does not affect the calculation of the Sharpe Ratio and allows us to explore the impact of financial forecasting consistency on performance and profitability. Notably, we do not consider the transaction cost in accordance with previous work (Sawhney et al., 2021a).

3.3 Wholistic Evaluation Metrics

We introduce the predictive evaluation metrics and the novel consistency evaluation metrics as elaborated below.

Predictive Evaluations. For stock prediction, we use three metrics to measure performance: Accuracy, F1 score, and Matthews correlation coefficient (MCC). These metrics are calculated as follows:

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \tag{1}$$

For a given confusion matrix:

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \tag{2}$$

$$MCC = \frac{tp \times tn - fp \times fn}{\sqrt{(tp + fp)(tp + fn)(tn + fp)(tn + fn)}} \tag{3}$$

We use both Profit Ratio and Sharpe Ratio for the trading simulation task as performance indicators. Return R, and investment I is involved in calculating the Profit Ratio.

$$ProfitRatio = \frac{R}{I} \tag{4}$$

The Sharpe Ratio measures the performance of an investment by considering the average return R_x, risk-free return R_f, and standard deviation of the investment $\sigma(R_x)$.

$$SharpeRatio = \frac{R_x - R_f}{\sigma(R_x)} \tag{5}$$

Consistence Evaluations. Based on logical transformations, we propose the consistency evaluation metrics of consistency, aiming to measure text-based financial forecasting models from a consistency perspective as a complementary metric to accuracy. Assuming that C is a set of four logical consistencies. To begin with, we define the consistency score ($Consis$), elaborated as follows:

$$Consis = \frac{\sum_{i=1}^{|C|} C_i}{|C|} \tag{6}$$

where the C set contains Negation consistency $Consis^N$, Symmetric consistency $Consis^S$, Additive consistency $Consis^A$, Transitive consistency $Consis^T$. We give the formal definition of those four metrics, respectively. The consistency of $Consis^N$ is calculated as:

$$Consis^N = \frac{\sum_{i=1}^{|D|} \begin{cases} 0 & (f(x_i) = f(x_i^N)) \\ 1 & (f(x_i) \neq f(x_i^N)) \end{cases}}{|D|} \tag{7}$$

where D is the original test set, x_i is the test sample in the original test set, i.e. $x_i \in D$. x_i^N is the new test sample obtained by negation consistency transformation on x_i, and $f(x)$ is the prediction of the model (positive or negative) for the input x. In terms of the symmetric, additive, and transitive transformations, the value equals 0 when $f(x_i) \neq f(x_i^N)$ while equals 1 when $f(x_i) = f(x_i^N)$.

4 Experiments

We first evaluate the explicit preferences in PLMs. Then we assess the ability of text-based models to make consistent predictions on the stock movement and finally test the profitability of these predictions using a trading simulation.

4.1 Dataset

Earnings Call Data. We use the publicly available Earning Conference Calls dataset by (Qin and Yang, 2019), which includes transcripts of 576 earnings calls from S&P 500 companies listed on the American Stock Exchange, obtained from the Seeking Alpha website. It also includes the meta-information on the company affiliations and publication dates.

Financial Market information. We also collect historical price data (closing price) for the traded companies listed in S&P 500 from Yahoo Finance for the period from January 1, 2017, to January 31, 2018. This data was used to calculate the label of stock price movement and profitability.

Data Processing. Following (Qin and Yang, 2019; Yang et al., 2020a), we split the dataset into mutually exclusive train/validation/test sets in a 7:1:2 ratio in chronological order to ensure that future information is not used to predict past price movements. We also construct logical consistency datasets based on the original test set using the above-mentioned four logical consistency transformations. The size of our evaluation dataset is four times the size of the original one since we ensure that each sample in the original test set corresponds to four logical consistency test samples. To facilitate future research, we release our dataset and the evaluation toolkit in **FinTrust**.

4.2 Models

Representation Models. We conduct experiments on popular PLMs, including BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), DistilBERT (Sanh et al., 2019), and FinBERT (Yang et al., 2020b). The vocabulary of FinBERT is different from the others as it contains domain-specific terms in the financial market, including company names.

Predictive Models. Regarding the forecasting models, we evaluate several baselines, including the traditional machine learning and state-of-the-art transformer-based methods, detailed as follows.

- **HTML:** Yang et al. (2020a) propose a hierarchical transformer-based framework to address the problem of processing long texts in earnings call data. It utilizes a pre-trained WWM-BERT-Large model to generate sentence representations as inputs for the model.

- **MRDM:** Qin and Yang (2019) propose the first method to treat volatility prediction as a multi-modal deep regression problem, building benchmark results and introducing the earnings conference call dataset.

PLM	Params	Neg	Pos	Consistency
BERT-base	110M	+	+	71.33%
BERT-base	110M	+	-	55.87%
BERT-base	110M	-	+	86.79%
BERT-large	340M	+	+	75.67%
BERT-large	340M	+	-	67.60%
BERT-large	340M	-	+	83.74%
RoBERTa-base	125M	+	+	77.79%
RoBERTa-base	125M	+	-	69.17%
RoBERTa-base	125M	-	+	86.40%
RoBERTa-large	355M	+	+	**82.70%**
RoBERTa-large	355M	+	-	**76.67%**
RoBERTa-large	355M	-	+	**88.72%**
FinBERT	110M	+	+	72.40%
FinBERT	110M	+	-	56.27%
FinBERT	110M	-	+	88.53%
DistilBERT	66M	+	+	70.13%
DistilBERT	66M	+	-	57.92%
DistilBERT	66M	-	+	82.33%

Table 1: The results of the consistency measurement in PLMs via masked token predictions, splitting by negative and positive token predictions. '+' denotes that the attitude of the word with the specific polarity will be predicted while '-' means that we do not consider tokens with a specific polarity.

- **Event:** Ding et al. (2015) adapt Open IE for event-based stock price movement prediction, extracting structured events from large-scale public news without manual efforts.

- **XGBoost:** Chen et al. (2015) propose a gradient-boosting decision tree known as the classical machine learning baseline.

5 Results and Discussion

We report the results of three tasks defined in Section 3.2 and the consistency score calculated by the consistency evaluation metrics in this section. Furthermore, we present extensive ablation studies and discussions to support in-depth analyses of each component in FinTrust.

5.1 Predictive Results

Consistency Measurement in PLMs. The results of explicit preferences in PLMs are presented in Table 1. In general, we find that all PLMs exhibited relatively low consistency, ranging from 70.13% to 82.7%, which falls significantly short of the level of robustness expected in the financial market. Also, we observe that PLMs typically demonstrated lower consistency when tested on

Metrics	**ACC**					**F1**					**MCC**				
Period	Avg	3	7	15	30	Avg	3	7	15	30	Avg	3	7	15	30
HTML	**0.546**	0.442	0.531	0.566	0.646	**0.671**	0.571	0.619	0.713	0.780	**0.078**	0.052	0.056	0.032	0.175
+FinTrust	**0.521**↓	0.465↑	0.527↓	0.529↓	0.564↓	**0.647**↓	0.608↑	0.629↑	0.648↓	0.703↓	**0.040**↓	0.019↓	0.058↑	0.019↓	0.063↓
MRDM	**0.555**	0.504	0.513	0.584	0.619	**0.670**	0.541	0.663	0.722	0.754	**0.059**	0.079	0.007	0.107	0.044
+FinTrust	**0.504**↓	0.465↓	0.511↓	0.507↓	0.535↓	**0.622**↓	0.569↑	0.667↑	0.578↓	0.674↓	**0.017**↓	-0.024↓	0.038↑	0.032↓	0.023↓
Event	**0.542**	0.416	0.522	0.593	0.637	**0.694**	0.582	0.682	0.736	0.776	**0.122**	0.078	0.097	0.189	0.123
+FinTrust	**0.512**↓	0.447↑	0.504↓	0.529↓	0.569↓	**0.656**↓	0.598↑	0.663↓	0.658↓	0.705↓	**0.006**↓	-0.032↓	-0.023↓	0.013↓	0.068↓
XGB	**0.515**	0.434	0.487	0.584	0.558	**0.561**	0.448	0.500	0.641	0.653	**0.018**	-0.093	-0.027	0.147	0.043
+FinTrust	**0.507**↓	0.462↑	0.502↑	0.531↓	0.531↓	**0.545**↓	0.456↑	0.518↑	0.584↓	0.622↓	**-0.004**↓	-0.076↑	-0.002↑	0.045↓	0.014↓

Table 2: Performance and robustness evaluation of stock movement prediction for multiple baselines using FinTrust. Significant performance decay has been observed on all methods using the Student T-test over 10 times run, p<0.05.

negative tokens than positive tokens (on average 63.91% – negative vs. 86.09% – positive). This suggests that popular PLMs tend to exhibit stereotypes when predicting negative tokens.

From a model-level perspective, our results indicate that FinBERT, which utilizes a domain-specific training corpus during the pre-training phase, can slightly improve consistency compared to BERT-base. Besides, we show that the increase in parameter size brings significant benefits for improving consistency, given that BERT-large and RoBERTa-large both outperform their base-sized versions (75.67% vs. 71.33% – BERT; 82.70% vs.77.79% –RoBERTa). In particular, RoBERTa-achieves the highest consistency across three settings, indicating its high robustness. In contrast, DistilBERT achieves the lowest consistency.

Strategy	**Profit Ratio**	**Sharpe Ratio**
HTML	3.752	0.266
+ FinTrust	3.359↓	0.229↓
Δ↓	-10%	-14%
Event	3.720	0.263
+ FinTrust	3.535↓	0.245↓
Δ↓	-5%	-7%
MRDM	3.495	0.241
+ FinTrust	2.384↓	0.138↓
Δ↓	-32%	-43%
XGB	-0.515	-0.126
+ FinTrust	0.296↑	0.032↑
Δ↑	158%	75%
Buy-all	3.681	0.259
Random	-0.271	-0.105
Short-sell-all	-3.681	-0.259

Table 3: Performance on the trading simulation. '+FinTrust' represents the performance using the input after the transformation.

Stock Movement Prediction. The results of stock movement prediction over text-based financial forecasting models are shown in Table 2. We evaluate multiple baselines by comparing the results of models on the original test set to the results tested on transformed datasets (shown as +FinTrust). It is noteworthy that the accuracy of some predictions is even lower than that of random guess, especially for the short-time prediction (n=3). Furthermore, we demonstrate that the effect of logical consistency transformations on traditional performance indicators varies depending on the time period, but the average performance of all models decreased significantly over three metrics. In particular, models show extraordinary vulnerability when it comes to predicting the long-term stock return (n=15 and 30), as transformations in all settings decrease accuracy when the time period is 15 and 30 days.

From the model perspective, regarding the ratio of performance decay, XGBoost is the least impacted, and MRDM is the most affected. This can be because traditional machine learning models, such as XGBoost, have fewer parameters than deep learning models and are therefore less affected by artefacts. Despite this, the accuracy on FinTrust achieved by models is only slightly more accurate than the random guess (e.g., **0.504** on MRDM, **0.507** on XGBoost). The vulnerability of these models, including state-of-the-art methods, hinders the deployment of NLP systems in the real financial market and should be taken more seriously.

Trading Simulation. We compare three simple trading strategies (Buy-all, Short-sell-all, and Random) with four baselines. The results are shown in Table 3. It can be seen that HTML and Event have higher yields and can exceed simple trading strategies. However, after conducting consistency transformations, positive returns of these two methods are much reduced, even lower than the simple Buy-all strategy. Methods such as MRDM and XGBoost gain lower returns than Buy-all, with MRDM experiencing the highest drop of about **32%-43%**. Even though the returns of XGBoost improved significantly after the transformations, it still remained much lower than the Buy-all strategy and the other three baselines. Hence, we contend that the increase in XGBoost's returns does not have a strong reference value. We conclude that most methods

Period		3	7	15	30
Event	AVG	**0.730**	**0.739**	0.644	**0.692**
	Add	0.903	0.947	0.664	0.805
	Neg	0.106	0.035	0.044	0.018
	Sym	0.947	0.982	0.929	0.973
	Tra	0.965	0.991	0.938	0.973
HTML	AVG	0.699	0.628	**0.688**	0.684
	Add	0.894	0.655	0.876	0.743
	Neg	0.115	0.212	0.177	0.009
	Sym	0.894	0.796	0.841	0.991
	Tra	0.894	0.850	0.858	0.991
MRDM	AVG	0.597	0.706	0.524	0.650
	Add	0.664	0.894	0.301	0.735
	Neg	0.248	0.062	0.053	0.053
	Sym	0.655	0.894	0.805	0.885
	Tra	0.823	0.973	0.938	0.929
XGB	AVG	0.566	0.595	0.593	0.653
	Add	0.522	0.487	0.496	0.504
	Neg	0.071	0.133	0.124	0.354
	Sym	1.000	0.973	0.973	0.991
	Tra	0.673	0.788	0.779	0.761

Table 4: The consistency score calculated by $Consis$.

show unacceptably poor performance caused by lacking consistent logical behavior.

5.2 Consistency Score

Results. We show the results of the consistency score (defined in Section 3.4) in Table 4. It can be seen that Event has the highest consistency score (*Consis*) and XGBoost has the lowest *Consis*. Regarding the average consistency over four transformations, Event achieves three of the four highest consistency scores. XGBoost tends to make contradictory predictions in terms of the lowest scores in three settings. Additionally, all methods perform poorly on negation consistency, consistent with findings in the PLMs evaluation (Table 1).

Correlation Analysis. We examine the correlation between the indicators of consistency and accuracy. Importantly, we find that our consistency score does not align with traditional performance indicators such as accuracy, evidenced by the fact that the most consistent model (Event) is not necessarily the highest in accuracy (HTML). The overall Pearson correlation coefficient between the consistency score and accuracy is only 0.314, indicating a low-level correlation. This suggests that the proposed consistency score can be used as a complementary evaluation metric for accuracy in future research on text-based financial forecasting.

5.3 Discussion

Human Evaluation. To assess the effectiveness of our consistency transformation method in preserving the original meaning, we conduct a human annotation study. Two annotators are employed from the author list and be required to label each sample and its four consistency transformations. Both of them received an advanced degree in computer science. The Inter-Annotator Agreement score is calculated to be 0.98, based on an evaluation of 40 samples and their 160 transformed samples. The average consistency score for human annotators is 0.975, indicating that our method successfully preserves the original meaning in most cases.

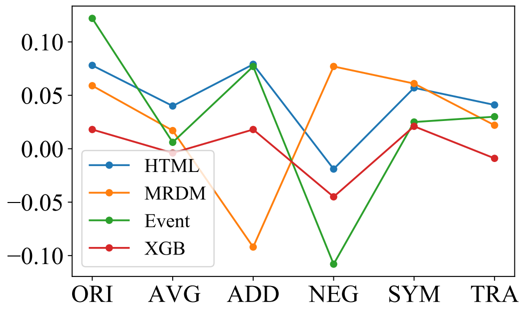

Figure 3: The ablation study of FinTrust in stock movement prediction on the average MCC over periods.

Ablation Study. We show the ablation results in stock movement prediction of four transformations in Figure 3. We find that evaluations on the FinTrust lead to significant performance decay for most settings compared to the original performance, which illustrates the individual influence of transformations. In particular, we show that models usually underperform when evaluating the *negation transformation*, with the exception of MRDM. It suggests that current models lack the ability to provide non-contradictory predictions.

6 Conclusion

We proposed FinTrust, an evaluation tool that assesses the trustworthiness of financial forecasting models in addition to their accuracy. Results on FinTrust show that (1) the consistency of state-of-the-art models falls significantly short of expectations when applied to stock movement prediction; (2) predictions with such a low logical consistency can lead to severe consequences, as evidenced by poor performance in a trading simulation test. Our empirical results highlight the importance of perceiving such concerns when developing and evaluating text-based financial models, and we release our dataset for facilitating future research. Despite this, how to evaluate the consistency of large-scale language models (LLMs) is still an open question towards the financial forecasting task.

Limitation

While our pipeline is designed to be applicable to any financial text dataset, the evaluation dataset is transformed solely on earnings conference calls. We will expand the scope of experiments to include other financial text sources such as news articles and social media posts. Finally, the current trading simulation does not take transaction costs into account. Going forward it will be necessary to consider more sophisticated trading policies.

Ethics Statement

This paper honors the ACL Code of Ethics. The dataset used in the paper does not contain any private information. All annotators have received enough labor fees corresponding to their amount of annotated instances. The code and data are open-sourced under the CC-BY-NC-SA license.

Acknowledgements

We would like to thank anonymous reviewers for their insightful comments and suggestions to help improve the paper. This publication has emanated from research conducted with the financial support of the Pioneer and “Leading Goose" R&D Program of Zhejiang under Grant Number 2022SDXHDX0003, the 72nd round of the Chinese Post-doctoral Science Foundation project 2022M722836, and the financial support from the rxhui.com company. Yue Zhang is the corresponding author.

References

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Akari Asai and Hannaneh Hajishirzi. 2020. Logic-guided data augmentation and regularization for consistent question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5642–5650.

Oana-Maria Camburu, Brendan Shillingford, Pasquale Minervini, Thomas Lukasiewicz, and Phil Blunsom. 2020. Make up your mind! adversarial generation of inconsistent natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4157–4165.

Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2021a. Evaluating the rationales of amateur investors. In *The World Wide Web Conference*.

Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2021b. From opinion mining to financial argument mining. *Springer Briefs in Computer Science*, pages 1–95.

Chung-Chi Chen, Hiroya Takamura, and Hsin-Hsi Chen. 2022. Fintech for social good: A research agenda from nlp perspective. *arXiv preprint arXiv:2211.06431*.

Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, et al. 2015. Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4.

Chengyu Chuang and Yi Yang. 2022. Buy tesla, sell ford: Assessing implicit stock market preference in pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 100–105, Dublin, Ireland. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2014. Using structured events to predict stock price movement: An empirical investigation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1415–1425.

Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2015. Deep learning for event-driven stock prediction. In *Proceedings of the 24th International Conference on Artificial Intelligence*, page 23272333, Buenos Aires, Argentina.

Xin Du and Kumiko Tanaka-Ishii. 2020. Stock embeddings acquired from news articles and price history, and an application to portfolio optimization. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 3353–3363.

Junwen Duan, Yue Zhang, Xiao Ding, Ching Yun Chang, and Ting Liu. 2018. Learning target-specific representations of financial news documents for cumulative abnormal return prediction. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING-18)*, pages 2823–2833.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Amir Feder, Katherine A Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E Roberts, et al. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158.

Fuli Feng, Huimin Chen, Xiangnan He, Ji Ding, Maosong Sun, and Tat-Seng Chua. 2019. Enhancing stock movement prediction with adversarial training. *arXiv preprint arXiv:1810.09936*.

Wee Chung Gan and Hwee Tou Ng. 2019. Improving the robustness of question answering systems to question paraphrasing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6065–6075.

Siddhant Garg and Goutham Ramakrishnan. 2020. Bae: Bert-based adversarial examples for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6174–6181.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*.

Md Mosharaf Hossain, Venelin Kovatchev, Pranoy Dutta, Tiffany Kao, Elizabeth Wei, and Eduardo Blanco. 2020. An analysis of natural language inference benchmarks through the lens of negation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9106–9118, Online. Association for Computational Linguistics.

Myeongjun Jang, Deuk Sin Kwon, and Thomas Lukasiewicz. 2022. Becel: Benchmark for consistency evaluation of language models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3680–3696.

Chen Jia, Xiaobo Liang, and Yue Zhang. 2019. Cross-domain NER using cross-domain language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2464–2474, Florence, Italy. Association for Computational Linguistics.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8018–8025.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.

Katherine Keith and Amanda Stent. 2019. Modeling financial analysts' decision making via the pragmatics and semantics of earnings calls. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, ACL 19, pages 493–503, Florence, Italy.

Shimon Kogan, Dimitry Levin, Bryan R. Routledge, Jacob S. Sagi, and Noah A. Smith. 2009. Predicting risk from financial reports with regression. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 272–280.

Hao Li, Jie Shao, Kewen Liao, and Mingjian Tang. 2022. Do simpler statistical methods perform better in multivariate long sequence time-series forecasting? In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 4168–4172.

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

Shouwei Liu and Yiu Kuen Tse. 2013. Estimation of monthly volatility: An empirical comparison of realized volatility, garch and acd-icv methods. *Finance Research Letters*.

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

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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

Navid Rekabsaz, Mihai Lupu, Artem Baklanov, Alexander Dür, Linda Andersson, and Allan Hanbury. 2017. Volatility prediction using financial disclosures sentiments with word embedding-based ir models. In

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1712–1721.

Marco Tulio Ribeiro, Carlos Guestrin, and Sameer Singh. 2019. Are red roses red? evaluating consistency of question-answering models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6174–6184.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging nlp models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Ramit Sawhney, Arshiya Aggarwal, and Rajiv Shah. 2021a. An empirical investigation of bias in the multimodal analysis of financial earnings calls. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3751–3757.

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

Ramit Sawhney, Puneet Mathur, Ayush Mangal, Piyush Khanna, Rajiv Ratn Shah, and Roger Zimmermann. 2020. Multimodal multi-task financial risk forecasting. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 456465. Association for Computing Machinery.

Koustuv Sinha, Prasanna Parthasarathi, Joelle Pineau, and Adina Williams. 2021. UnNatural Language Inference. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7329–7346, Online. Association for Computational Linguistics.

Megha Srivastava, Tatsunori Hashimoto, and Percy Liang. 2020. Robustness to spurious correlations via human annotations. In *International Conference on Machine Learning*, pages 9109–9119. PMLR.

Alessandro Stolfo, Zhijing Jin, Kumar Shridhar, Bernhard Schölkopf, and Mrinmaya Sachan. 2022. A causal framework to quantify the robustness of mathematical reasoning with language models. *arXiv preprint arXiv:2210.12023*.

Rui Wang and Ricardo Henao. 2021. Unsupervised paraphrasing consistency training for low resource named entity recognition. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5303–5308, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Frank Xing, Lorenzo Malandri, Yue Zhang, and Erik Cambria. 2020. Financial sentiment analysis: an investigation into common mistakes and silver bullets. In *Proceedings of the 28th international conference on computational linguistics*, pages 978–987.

Yumo Xu and Shay B Cohen. 2018. 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.

Linyi Yang, Jiazheng Li, Ruihai Dong, Yue Zhang, and Barry Smyth. 2022. Numhtml: Numeric-oriented hierarchical transformer model for multi-task financial forecasting. In *AAAI*.

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

Linyi Yang, Zheng Zhang, Su Xiong, Lirui Wei, James Ng, Lina Xu, and Ruihai Dong. 2018. Explainable text-driven neural network for stock prediction. In *2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS)*, pages 441–445. IEEE.

Yi Yang, Mark Christopher Siy UY, and Allen Huang. 2020b. Finbert: A pretrained language model for financial communications. *arXiv preprint arXiv:2006.08097*.

Xi Zhang, Yunjia Zhang, Senzhang Wang, Yuntao Yao, Binxing Fang, and S Yu Philip. 2018. Improving stock market prediction via heterogeneous information fusion. *Knowledge-Based Systems*, 143:236–247.

A Transitive Consistency

Example. We show an example to understand better the motivation for using Transitive Consistency when measuring the consistency of FinNLP models. Given *"Nektar Therapeutics gave investors strong confidence after Earnings Conference Call on March 1, 2017, and its stock price soared 79.43% in the following month."*. As a leading company in the same Sector (Health Care), Johnson & Johnson (JNJ) was also affected by this and increased by 1.91% over the same period, which confirmed the rationality of selecting transitive consistency as one of the measurement methods.

B Full Ablation Results

We report the ablation study results of four different types of logical transformation based on the fine-tuned forecasting models in Table 5. We use *italics* to indicate the performance before consistency transformation, use **bold** to express the performance that has been reduced after consistency transformation, and do not deal with other parts that have not decreased, for the convenience of readers. All detailed return changes in trading simulation based on text-based fine-tuned forecasting models are also shown in Table 6. "+FinTrust " means the average impact of the four transformations.

C Additional experimental details

The model settings involved in the paper are all aligned with the parameters and training details described in the corresponding article Yang et al. (2020a); Qin and Yang (2019); Ding et al. (2015); Chen et al. (2015). The total computational budget is about 50 GPU hours, using a GeForce RTX 3090. All models use the highest performance among ten repeated experiments using different seeds and ensure reproducibility.

	ACC					F1					MCC				
	3	7	15	30	Avg	3	7	15	30	Avg	3	7	15	30	Avg
HTML	0.442	0.531	0.566	0.646	*0.546*	0.571	0.619	0.713	0.780	*0.671*	0.052	0.056	0.032	0.175	*0.078*
HTML-Add	0.407	0.540	0.584	0.619	**0.538**	0.579	0.662	0.715	0.726	0.671	0.000	0.085	0.104	0.127	0.079
HTML-Neg	0.602	0.522	0.416	0.363	**0.476**	0.737	0.625	0.522	0.532	**0.604**	0.089	0.073	-0.113	-0.123	**-0.019**
HTML-Sym	0.425	0.522	0.566	0.637	**0.538**	0.564	0.620	0.684	0.776	**0.661**	0.004	0.036	0.066	0.123	**0.057**
HTML-Tra	0.425	0.522	0.549	0.637	**0.533**	0.552	0.609	0.671	0.776	**0.652**	-0.016	0.037	0.021	0.123	**0.041**
HTML-Avg	0.465	0.527	0.529	0.564	**0.521**	0.608	0.629	0.648	0.703	**0.647**	0.019	0.058	0.019	0.063	**0.040**
MRDM	0.504	0.513	0.584	0.619	*0.555*	0.541	0.663	0.722	0.754	*0.670*	0.079	0.007	0.107	0.044	*0.059*
MRDM-Add	0.416	0.496	0.434	0.496	**0.460**	0.507	0.655	0.289	0.627	**0.520**	-0.079	-0.073	-0.073	-0.145	**-0.092**
MRDM-Neg	0.619	0.513	0.434	0.381	**0.487**	0.746	0.667	0.600	0.539	**0.638**	0.153	0.161	-0.018	0.013	0.077
MRDM-Sym	0.425	0.531	0.584	0.628	**0.542**	0.504	0.683	0.697	0.753	**0.659**	-0.067	0.101	0.111	0.100	0.061
MRDM-Tra	0.398	0.504	0.575	0.637	**0.529**	0.521	0.663	0.727	0.776	0.672	-0.105	-0.037	0.108	0.123	**0.022**
MRDM-Avg	0.465	0.511	0.507	0.535	**0.504**	0.569	0.667	0.578	0.674	**0.622**	-0.024	0.038	0.032	0.023	**0.017**
Event	0.416	0.522	0.593	0.637	*0.542*	0.582	0.682	0.736	0.776	*0.694*	0.078	0.097	0.189	0.123	*0.122*
Event-Add	0.425	0.522	0.575	0.637	**0.540**	0.558	0.671	0.652	0.745	**0.657**	-0.007	0.044	0.116	0.157	**0.077**
Event-Neg	0.531	0.478	0.381	0.381	**0.442**	0.686	0.638	0.545	0.539	**0.602**	-0.151	-0.049	-0.246	0.013	**-0.108**
Event-Sym	0.416	0.504	0.593	0.628	**0.535**	0.571	0.671	0.726	0.767	**0.684**	0.003	-0.092	0.138	0.051	**0.025**
Event-Tra	0.416	0.513	0.566	0.628	**0.531**	0.577	0.675	0.707	0.767	**0.681**	0.025	0.004	0.042	0.051	**0.030**
Event-Avg	0.447	0.504	0.529	0.569	**0.512**	0.598	0.663	0.658	0.705	**0.656**	-0.032	-0.023	0.013	0.068	**0.006**
XGB	0.434	0.487	0.584	0.558	*0.515*	0.448	0.500	0.641	0.653	*0.561*	-0.093	-0.027	0.147	0.043	*0.018*
XGB-Add	0.398	0.504	0.593	0.575	0.518	0.433	0.533	0.657	0.676	0.575	-0.156	0.006	0.160	0.064	0.018
XGB-Neg	0.549	0.469	0.398	0.451	**0.467**	0.622	0.444	0.404	0.492	**0.490**	0.062	-0.064	-0.187	0.011	**-0.045**
XGB-Sym	0.434	0.496	0.575	0.566	0.518	0.448	0.513	0.636	0.662	0.565	-0.093	-0.010	0.127	0.058	0.021
XGB-Tra	0.469	0.540	0.558	0.531	0.524	0.318	0.581	0.638	0.658	**0.549**	-0.115	0.076	0.078	-0.075	**-0.009**
XGB-Avg	0.462	0.502	0.531	0.531	**0.507**	0.456	0.518	0.584	0.622	**0.545**	-0.076	0.002	0.045	0.014	**-0.004**

Table 5: Ablation study results of four different types of logical transformation based on the fine-tuned forecasting models. Compared to the original results, the decreased performance is presented in **bold**.

Strategy	Profit Ratio	Sharpe Ratio	Transformations	Profit Ratio	Sharpe Ratio	Transformations	Profit Ratio	Sharpe Ratio
HTML-Original	3.752	0.266	HTML-ADD	2.282↓	0.125↓	HTML-NEG	3.720↓	0.263↓
HTML+FinTrust	3.359↓	0.229↓	HTML-SYM	3.720↓	0.263↓	HTML-TRA	3.713↓	0.263↓
Event-Original	3.720	0.263	Event-ADD	3.646↓	0.256↓	Event-NEG	3.347↓	0.226↓
Event+FinTrust	3.535↓	0.245↓	Event-SYM	3.494↓	0.241↓	Event-TRA	3.652↓	0.256↓
MRDM-Original	3.495	0.241	MRDM-ADD	0.605↓	-0.026↓	MRDM-NEG	3.512↑	0.243↑
MRDM+FinTrust	2.384↓	0.138↓	MRDM-SYM	1.674↓	0.070↓	MRDM-TRA	3.743↑	0.266↑
XGB-Original	-0.515	-0.126	XGB-ADD	0.972↑	0.006↑	XGB-NEG	-0.833↓	-0.067↑
XGB+FinTrust	0.296↑	-0.032↑	XGB-SYM	-0.072↑	-0.087↑	XGB-TRA	1.118↑	0.020↑

Table 6: The ablation study of the trading simulation based on text-based fine-tuned forecasting models.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitation Section.

☑ A2. Did you discuss any potential risks of your work?
Ethics Statement.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 4.

☑ B1. Did you cite the creators of artifacts you used?
Section 4.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section 4.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 4.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 4.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 4.

C ☑ Did you run computational experiments?

Section 5.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 5.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 4.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.