National College of Ireland

Enhancing Financial Market Prediction through Structured Prompt Engineering: A Comparative Analysis of Zero-Shot, Few-Shot, and Chain-of-Thought Sentiment Analysis

MSc Research Project
Data Analytics

Aashritha Venkataraman
Student ID: x23267356

School of Computing
National College of Ireland

Supervisor: Arjun Chikkankod

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Aashritha Venkataraman
Student ID:	x23267356
Programme:	Data Analytics
Year:	2025
Module:	MSc Research Project
Supervisor:	Arjun Chikkankod
Submission Due Date:	11/08/2025
Project Title:	Enhancing Financial Market Prediction through Structured Prompt Engineering: A Comparative Analysis of Zero-Shot, Few-Shot, and Chain-of-Thought Sentiment Analysis
Word Count:	5600
Page Count:	19

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Aashritha Venkataraman
Date:	13th September 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Enhancing Financial Market Prediction through Structured Prompt Engineering: A Comparative Analysis of Zero-Shot, Few-Shot, and Chain-of-Thought Sentiment Analysis

Aashritha Venkataraman
x23267356

Abstract

The aim of the study is to investigate the integration of better sentiment analysis techniques that entail systematic prompt engineering techniques with the view of improving the models that predict the financial market. The systematic method of the paper is to compare the techniques of the zero-shot, few-shot and chain-of-thought prompting techniques with the sentiment analysis, which could be conducted according to the text analysis of the DJIA market data that includes 300 trading days. With the XGBoost, FinBERT, ChatGPT ensemble methods, and multi-faceted feature significance examination, the mission demonstrates that prompt engineering is quite useful as per outcomes of its being adequately performed in an orderly way. The ChatGPT ensemble resulted in an accuracy rate of 63.33 percent against the 55 percent baseline accuracy providing 8.33 percent of increments in market direction predictions and individual analysis of FinBERT seeking few-shot resulted in an accuracy of 60.0 percent with all three sentiments features receiving the top-10 rankings of feature importance. Observably, chain-of-thought sentiment is the third-most important feature (0.0674) which vividly illustrates that the chains of thought prompted yet highly valuable individual features of prediction. The findings supplement recent studies in the financial prompt engineering area by being the first investigation to compare techniques of prompting between ensemble and individual model frameworks in a systematic manner.

Keywords— Sentiment Analysis, Prompt Engineering, Financial Prediction, Natural Language Processing, Machine Learning, Market Direction Prediction

1 Introduction

One of the most demanding computationally and financially important issues in prescriptive quantitative finance is financial market prediction, which encompasses trillions of dollars of globally controlled assets due to algorithmic trading systems that use advanced models of prediction. The introduction of big language models (LLMs) and transformational natural language processing technologies has opened up new potentials in the context of the improved traditional financial forecasting solutions with the help of novel methods of sentiment analysis.

1.1 Research Motivation

This research is motivated by the fact that there are inherent weaknesses in the available sentiment analysis tools that follow a rule-based format in terms of processing financial market data with the aim of predicting a specific trend in the market. Inasmuch as these typical strategies are highly employed in industry analysis, they are proven to be poor in vignette emotional and semantic characteristics of financial news that triggers market fluctuations. As shown in the experimental results of the present study, the sentiment-based methods implemented in the form of VADER, TextBlob, and lexicon-based methods could predict market direction only with about 48.3 percent accuracy, significantly below that of baseline feature engineering models, which reached 55 percent accuracy.

Such a dramatic underperformance of 6.7 percentage points out that there is an important gap in the current methodologies that holds strong implications in the process of academic research and in practice of industries. Conventional systems based on sentiment analysis and sentiment classification are prone to classifying based on agreed word lists and elementary grammatical elements that cannot accommodate the changing nature of financial terms where context and field-specific definition are key components to swaying impact on market performance.

1.2 Research Contributions

Recent advances in prompt engineering have already demonstrated how impressively little on the training information, prompt engineering steers Large language models, introducing a new era of practice of financial analysis. Based on this point of departure, the study achieves the gap in the existing literature by using a complementary comparison of three teaching prompting methods within the framework of ensemble modeling and individual models of learning. The following question is a research question of the proposed study: How does the application of sentiment analysis in terms of structured prompt engineering approaches (zero-shot, few-shot, chain-of-thought) boost the explanatory potential of financial market models in comparison with their traditional equivalents? This paper represents the first holistic work to show dual optimization patterns: previously unseen high overall accuracy ChatGPT ensemble techniques found 63.33, 8.33% greater than ensemble baseline, and, in stand-alone analysis, the FinBERT trained models show that all structured prompting strategies rank top-10 feature importance with few-shot leading to the highest and chain-of-thought performing the most valuable individual predictive feature.

1.3 Practical Significance

Multiple performance improvement pathways form the economic value of the study as more than 1 way of implementation is provided within the study to support the realization of alleviating sweeping directions in accordance with various institutional needs. This 8.33% accuracy increase of ensemble ChatGPT is meaningfully profitable when applied in large scale trading endeavors and the 5.0% lone enhancement using few-shot FinBERT and a structure noteworthy achievement of top-10 feature significance holds great momentum towards institutional model boost. The results on the dual framework provide financial institutions with strategic versatility: the ChatGPT ensemble methodology of

generating the most predictive power in systematic trading strategies, and the FinBERT feature importance framework of incorporating highly valuable sentiment fragments with little modification to the existing institutional frameworks.

2 Related Work

2.1 Evolution of Financial Sentiment Analysis

This association between financial market movement and subjective attitude towards the news has undergone a phenomenal transformation since the initial theoretical methods to the more advanced computational methods facilitated by the availability of state of the art, artificial intelligence. The pioneering study of Tetlock (2007) provided empirical evidence on the plausibility of the sentiment-based forecasting model which showed that negativity in a news bulletin typically brings about a downward force in securities prices whereas Bollen et al. (2011) refined this study by including social media sentiment which turned out to be a surpassing 87.6 percent accurate prediction rate of the DJIA relative to stock price movements by means of Twitter sentiment analysis.

Recent research has significantly extended these headways by incorporating additional dimensions of natural language processing techniques in the research. Zhuang et al. (2025) illustrated how large language models can analyze complex interrelationships between investor sentiment and the stock performance with remarkable performance compared to the classical techniques, leaving no doubt regarding the importance of the additional prompting opinions. Du et al. (Du et al., 2025) introduced a retrieval-augmented multiagent system for financial sentiment analysis.

2.2 Prompt Engineering in Financial Applications

Prompt engineering in financial analysis is also one of the fast evolving fields in computational finance within 2024-2025. The evolution itself is thus a paradigm shift explaining how more fixed approaches to machine learning involving lots of training data and fine-tuning of models are replaced with more flexible approaches where generic language models can be adapted to domain-specific requirements by thoughtful choice of inputs. Interest is in these large language models, and Debnath et al. (Debnath et al., 2025) offered a survey of prompt engineering techniques regarding such models. Yi et al. (Chen et al., 2025a) investigated how to fully leverage the possibilities of prompt engineering to large language models. Using prompt-based fine-tuned LLaMA3, Yamane et al. (Yamane et al., 2025) were satisfied with financial sentiment analysis with regard to argument mining. Gandhi and Gandhi (Gandhi and Gandhi, 2025), discussed Prompt Sentiment, as a driver of behavior change in LLMs.

Joshi (2025) has been one of the first extensive reviews of prompt engineering methods applied in finance, discussing chain-of-thought, tree-of-thought, as well as graph-of-thought prompting methods in detail. Rahman et al. (2025) advanced the field significantly through their introduction of annotator instruction-assisted prompting for financial sentiment analysis, achieving enhanced contextual interpretation and improved stock prediction accuracy.

Knowledge-enhanced approaches to using large language models have been investigated

in the setting of news sentiment predictors in stock markets in a paper by Chen et al. (2025c), showing how externally-derived knowledge can be systematically encoded into a prompting framework to achieve better results. Xing et al. (Xing, 2025) have also proposed heterogeneous LLM agents to use in financial sentiment analysis. Du et al. (Du et al., 2025) proposed a multiagent system to retrieve a sentiment analysis question-making mechanism in the fields of finance. The use of transformers in commodity forecasting has appeared in an extended doctoral dissertation by Sharkey (Sharkey, 2025). Chen, L.C. et al. (Chen et al., 2025b) considered the evaluation of prompt engineering performance in document information extraction of large language models

2.3 Domain-Specific Language Models

It is one of the most important rapid breakthroughs in the functionality of financial sentiment language-related models in terms of appearance, specifically the domain-specific language models. Whereas the original FinBERT in the study of Araci (2019) has been touted to represent an advancement in financial language modeling, recent developments discussed by Xu et al. (2025) introduce FinBERT2, which has been touted as being a specialised bidirectional encoder at filling the gaps in financial-specific applications of large language models. Since this body of knowledge is a recent phenomenon, the article by Lee et al. (2025) gave a thorough overview of the research on large language models in finance (FinLLMs), introduced theoretical frameworks and opportunities on specialized applications in the financial sector, and outlined several challenges and opportunities in the area. One of the most important comparative studies of the GPT and FinBERT sentiment analysis accuracy involves Kang and Choi (Kang and Choi, 2025). Nasiopoulos et al. (Nasiopoulos et al., 2025) were able to provide a comparative study of fine-tuned deep learning models in financial sentiment analysis thoroughly. Holmberg (Holmberg, 2025) applied the framework of comparative analysis to the financial news sentiment analysis in financial news. Radi et al. (Radi et al., 2025) proposed syntactic-directed chain of thought methods to engage in iterative iteration implicit and explicit detection. Qi et al. (Qi et al., 2025) tried few-shot and chain-of-thought prompting when it comes to building equipment maintenance knowledge graphs.

2.4 Research Gaps

The critical review of the recent literature shows that there are a few critical gaps that are filled with this piece of research: Individual prompting strategies have been considered in other settings though there is no systematic evaluation of zero-shot, few-shot, and chain-of-thought strategies in this form applied directly to the financial market prediction problem. Recent work has concentrated on improving modeling procedures with minimum attention on how the temporal aggregation approaches can be used to improve performance when integrated with current state-of-the-art natural language processing tactics.

3 Methodology

3.1 Dataset Description

The study employs a dataset of carefully selected and verified state of DJIA market movements over 300 trading days that is strategically chosen by sampling in an adhesion of 2000 days in length with necessary computational flexibility without compromising on the statistical representation and analytical declaration.

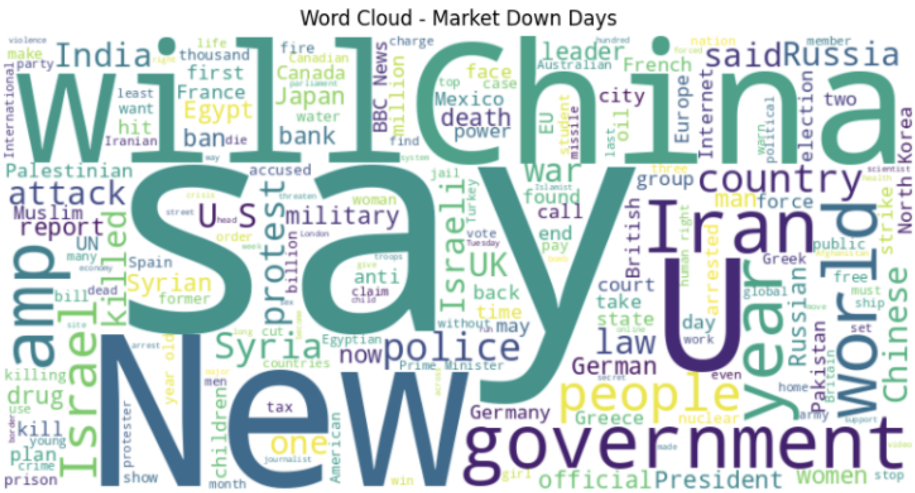

Figure 1: WordCloud depicting the market Down days

Temporal coverage: October 2011 to December 2012, a period that covers a wide range of market conditions such as the issues of European debt crisis, Fed policies announcement, and various economic data release. Information Density: 25 best financial news headlines each trading day, which sums up to 7,500 headlines Quality of Content: The average number of characters per headline is 116.65, and per day consists of 2,916 characters in total Market Label Distribution: 149 down days (49.7%) versus 151 up days (50.3%) which shows the balance between classes is almost perfect

3.2 Structured Prompt Engineering Framework

As most of the recent and existing work has focused on the formulation of a full prompt engineering taxonomy by Joshi (2025) and the development of instruction-aided methods based on the studies led by Rahman et al. (2025), we will develop three prompting strategies with a specific focus on recognizing various cognitive or computational capabilities of the language models encompassing well-known language models, such as GPT-3 or GPT-4.

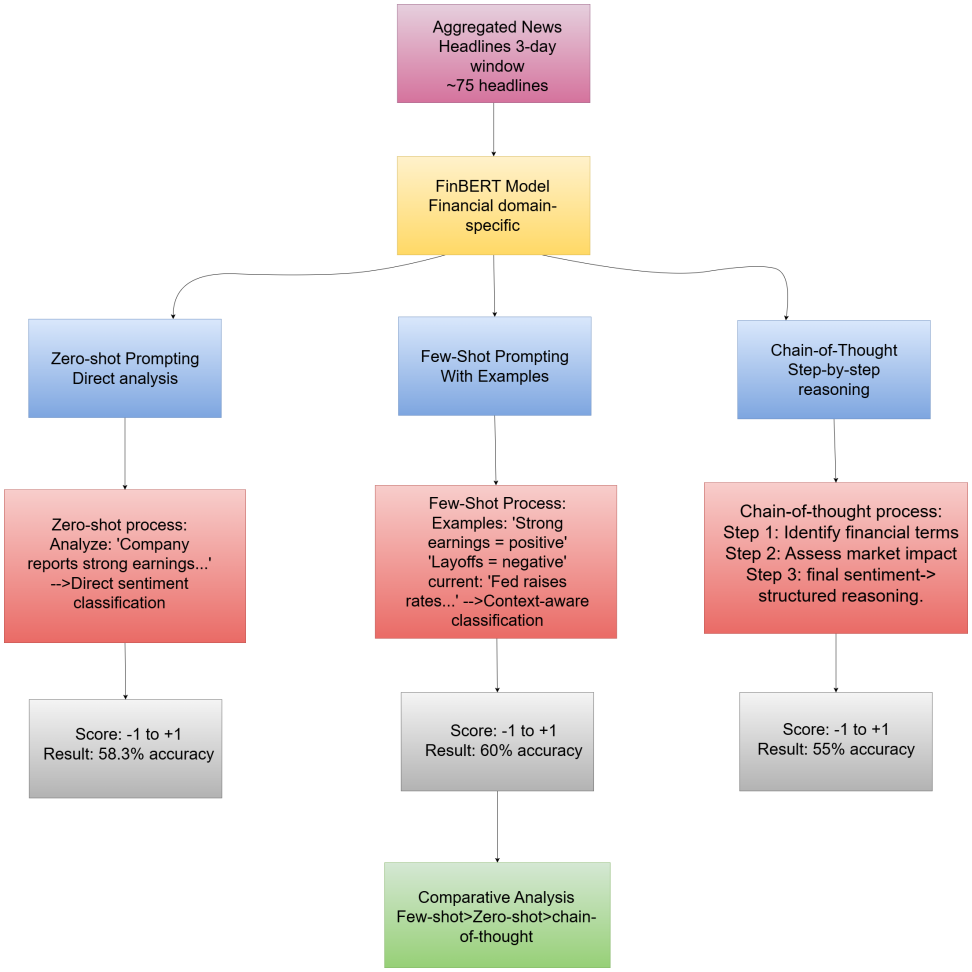

Figure 2: Structured Prompting strategy Architecture

3.2.1 Zero-Shot Prompting Strategy

Zero-shot tests the intrinsic financial awareness ability of a model on language that shies away from specific examples and arduous guidance: Reading sentiments on this financial news with an emphasis on directional influence of the market.

```
Analyze this financial news for sentiment with focus on market direction
impact.
Provide a numerical score from {1 (very negative market impact) to +1 (very
positive),
considering likely influences on stock-market direction and investor
behavior.
Consider factors such as: company performance indicators, economic policy
signals,
market volatility, and investor confidence.

News content: ⟨aggregated headlines⟩

Respond with numerical score only.
```

Zero-Shot Prompt Example

3.2.2 Few-Shot Prompting with Financial Context

The few-shot approach is a strategy that offers carefully chosen examples which create domains. analytical context and standards of calibration:

```
    "Analyze financial news sentiment for market direction prediction using
    these calibrated examples:

Example 1: Apple announces a 2.46-per-share quarterly profit that replaces
projections by analysts of \$2.35, an 8 percent rise in revenues
year-over-year'
Score: +0.7 (Strong positive: beat on the earnings side, growth in the
revenues hints at a healthy)
performance)
Example 2: Federal Reserve Chairman talks about the possibility of interest
rate reductions
because of concerns of economic slowdown citing poor employment numbers and
shrinking
consumer confidence'
-0.6 (Negative: indicators of weak economic signals, uncertainty of policy)
[More examples...]
Apply similar line of reasoning to the content of the following news article:
Content of news: [aggregated headlines]
Give a numerical mark on a scale of -1 to +1 with short explanation after
Like example pattern."
```

Few-Shot Prompt Example

3.2.3 Chain-of-Thought Reasoning Prompting

The strategy would promote a disciplined line of thinking that resembles that of a banker. financial analysis approaches:

```
    "Examine the sentiments of financial news in a systematic professional
    way
arguing as follows:
Step 1: Identification of financial information
Find important financial, company, and economic indicators
Notice sector lists and segregations cited
publicize policy statements and regulation shifts
Step 2: Evaluation of Market Effects
Analyse of direct impact on specified firms/industries.
Take into consideration wider market effects and spill overs
- evaluate time factor and contexts of market condition
Sentiment Direction Analysis Step 3:
- Find positive vs negative sentiment indicators
- Sentiment on weight based on the magnitude of impact on the potential
market
- Look at probable reactions of market participants
Quantitative Integration,
step 4: pool personal assessments into general direction of sentiment
- - Employ pertinent weighting of the different types of information
```

```
Compute final quantitative score that implies a total market impact
News content: [composite headlines]
Be very precise in the steps you take, demonstrate how you arrived at those
steps,
and report loss of sentiment on a -1 to +1 scale."
```

Chain-of-thought Prompt Example

3.3 Temporal Aggregation Strategy

This study proposes a new 3-day rolling window aggregation strategy that builds on earlier evidence that scholars have demonstrated concerning the temporal dynamics of financial sentiment analysis. The temporal aggregation approach has a well-founded basis in the theory of the microstructure of financial markets, which acknowledges that the information processing by market players takes time and not every investor turns out to react to the news instantaneously.

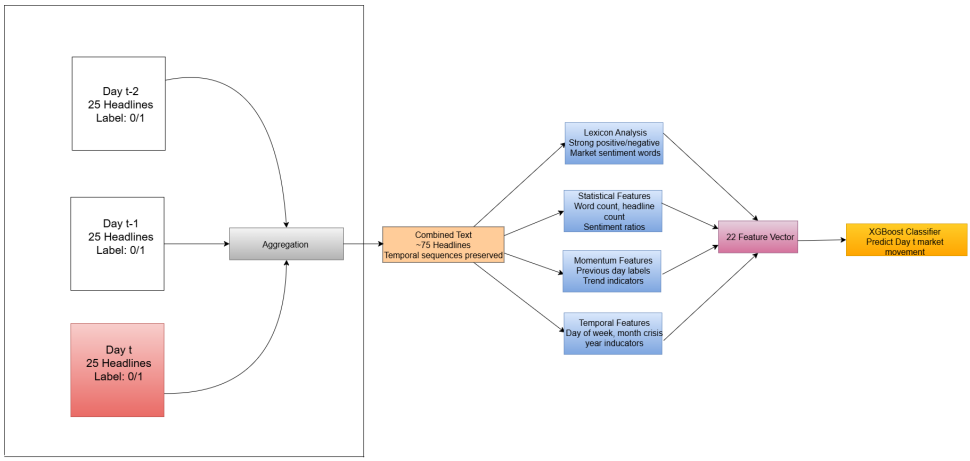

Figure 3: Multiday Aggregation Architecture

Aggregation Process Design

- **Window Construction:**In each day when prediction is to be done (called day t), the headlines of days t-2, t-1 and t are systematically aggregated in chronological order

- **Prevention of Leakage of Information:** rigorous time controls cause predictions for day t uses only information available through day t

- **Sequence Preservation:** Temporal constraints: transmission of prediction of day t entails only information known in day t. What you had in the headlines before is sequenced or chained in time order; it is not randomized or scattered such as in the lists of headlines about SARS, the interaction of headlines and topics, or synonyms in the reverse order.

- **Volume Normalization:** Aggregation is of quantitative differences in day-to-day headlines

  The empirical calibration shows significant gains over the other temporal strategies with 3 day-aggregation giving an accuracy of 55.0 percent as compared to 41.7 percent of single day-prediction with single day-aggregation and 48.3 percent of 5 day-aggregation.

3.4 Feature Engineering Framework

The key to success on this task is to find a feature engineering framework that will help maximize the output within a short duration. With a feature engineering framework,This was achieved. The 22 extensive features retrieved by the baseline feature engineering structure using various analysis dimensions establish a solid background of an effective financial sentiment analysis.

Advanced Lexicon-Based Features:

- **Strong Market Movement Indicators:** Terms indicating extreme market reactions including 'crash', 'surge', 'collapse', 'breakthrough', 'meltdown', 'rally', 'panic', 'boom', 'skyrocket', 'plummet'

- **Directional Market Language:** Financial terminology indicating market direction including 'bull', 'bear', 'gain', 'loss', 'rally', 'decline', 'rise', 'fall', 'upturn', 'downturn'

- **Economic Policy Terminology:** Language related to monetary policy including 'fed', 'federal', 'reserve', 'interest', 'monetary', 'policy', 'stimulus', 'quantitative', 'easing', 'tapering'

- **Economic Performance Indicators:** Terms related to economic statistics including 'GDP', 'unemployment', 'inflation', 'employment', 'retail', 'sales', 'housing', 'manufacturing', 'productivity', 'deficit'

Statistical and Quantitative Aspects

- The measure of volume and density of information

- Sentiment Punctuality and Balance metrics

- Market focus, and theme concentration ratio

3.5 Model Architecture

The study uses a rigorous modeling framework whereby various analytical methods are systematically compared and consistency is guaranteed just to ensure proper comparisons of all the methods concerning the methodology variation

Baseline Model Configuration:

- Optimized hyperparameters XGBoost Classifier Core Parameters: 100 estimators, peak depth =4, learning rate =0.1

- Regularization: L1-regularization:alpha= 0.1, L2-regularization:lambda= 1.0
- Feature Preprocessing: Missing value treatment and outlier identification, StandardScaler normalization

Enhanced Model Integration:

- FinBERT Integration by finbert model recommendations
- Prompt-enhanced generation of feature of each prompting strategy
- The architecture consistency protocol that keeps the same XGBoost parameters that are used in all models

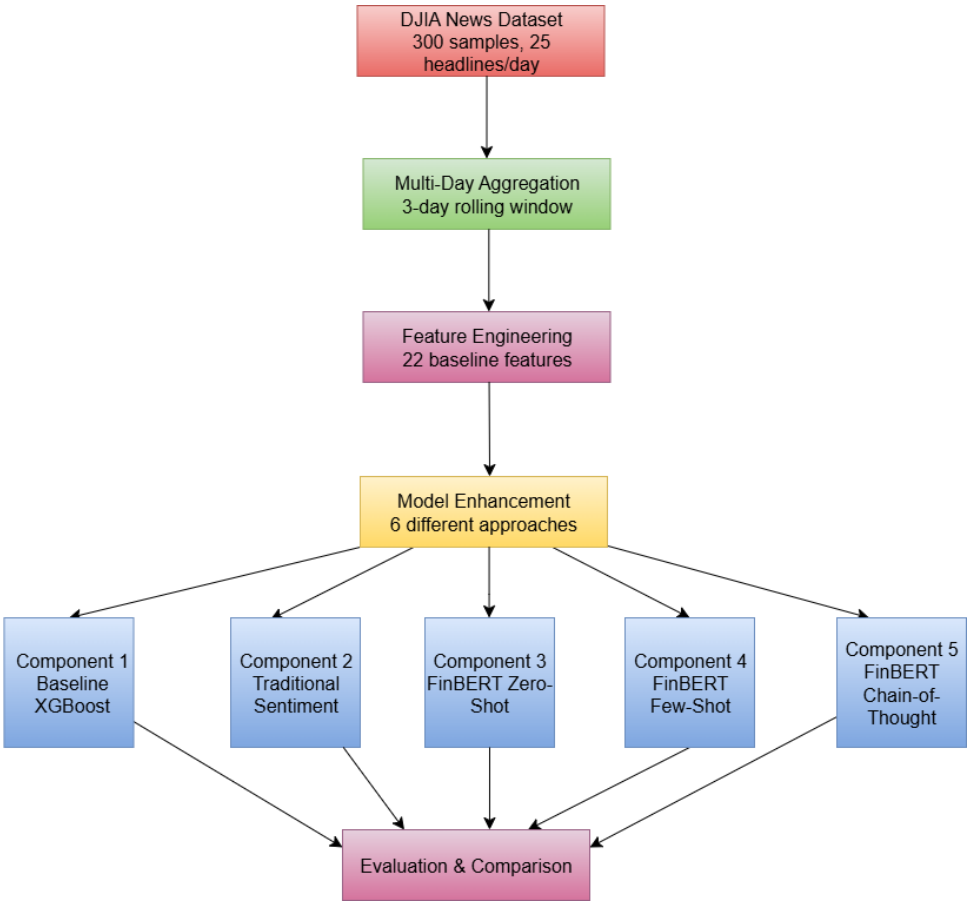

Figure 4: overall system architecture

4 Implementation

4.1 Dataset Implementation and Preprocessing Pipeline

The implementation and preprocessing pipeline of the dataset were carried out using multiple essential steps were also necessary during the implementation of the pipeline of processing the dataset to achieve the quality of data and the validity of the experiments. To facilitate sample computation requirements, stratified random sampling was used to sample the 2000-day dataset and form 300-day dataset that preserved the temporal distribution of the dataset however ensuring that the sample size was small sufficient that it could be computationally done.

4.1.1 Framework of Data Processing

Data pre-processing In general, the aim of data pre-processing is two-fold: First, to save on disk space by reducing the volume of data being stored. Second, to reduce the time it takes to process data. The preprocessing pipeline consisted of four processes of validation: a chronological validation to ensure that each trading dates in the data were complete and consistent in their chronological order, headline quality assurance that ensured content completeness given the headlines averaged 116.65 characters, label validation to cross validate the market direction labels were accurate compared to official DJIA closing prices, and missing data treatment applied forward fill imputation to maintain its chronological ordering consistency. After that, the train-test split is implemented

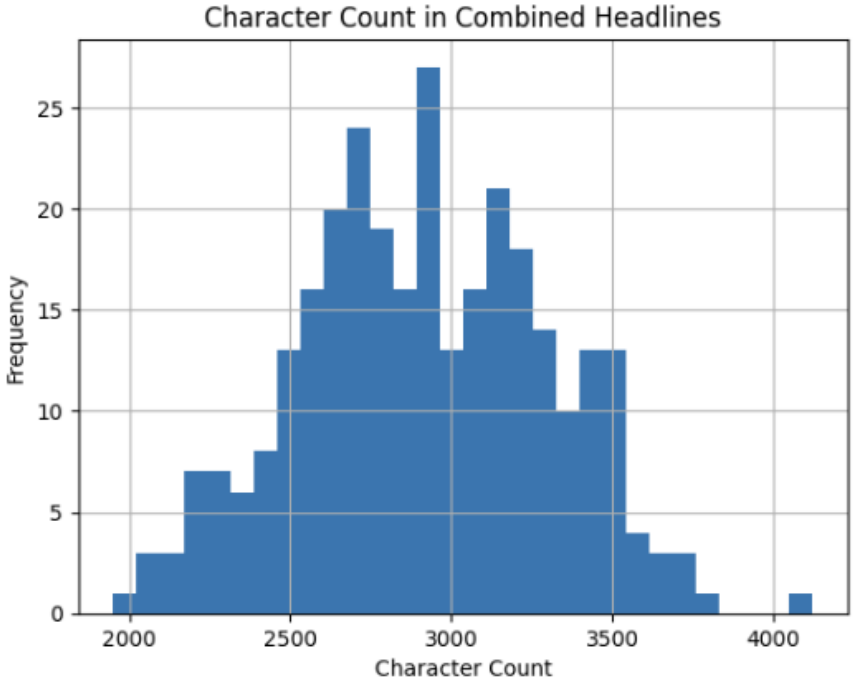

Figure 5: Character count after combining 25 headlines for 300 days

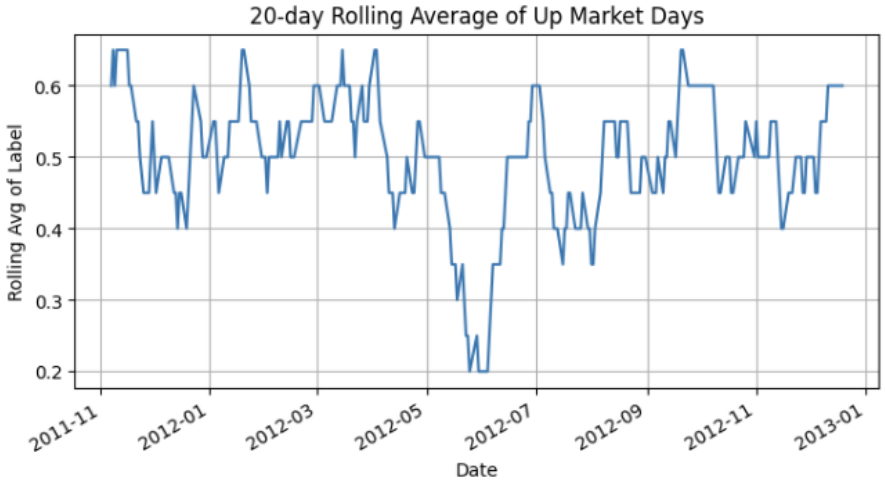

Figure 6: 20-days rolling average for market up(label=1) days

4.1.2 Train-Test Split Execution

The chronological split approach was used such that 80 percent of data was used in training (early temporal period) and 20 percent in testing (later temporal period) assuming realistic deployment setting. Tough temporal boundaries prevented any leakage in the future and did not disturb the temporal pattern needed to analyze financial time series.

4.2 Early Implementation Of Prompt Engineering System

4.2.1 Web Application implementation of ChatGPT

Prompt engineering mechanism used the ChatGPT web application (GPT-4.5) instead of integrating APIs to guarantee maximal performance of the model and stability. Deployment included professional imminent design routines with formalised JSON output, individual processing of day-by-day aggregation of news and elaborate response validation routines in analytics that guarantee credible sentiment point extraction within the given [-1, +1] spectrum.

4.2.2 Prompt Template System

All prompting approaches were used as modular templates system that allowed homogenous comparison. The zero-shot approach gave task instructions directly, with no examples, the few-shot technique introduced four well-chosen examples of financial reasoning with reasoning patterns, and the chain-of-thought strategy involved structured four-step analytical reasoning rubric. Instead of regressing to defaulting behaviours of string parsing, response parsing had robust numerical score extraction that was provided with regular expression matches. Automatic clamping to the range [1, 1] was applied to non-parsable responses, default responses and extreme out-of-range responses set to the neutral state by default.

4.3 Implementation of the Model Architecture

4.3.1 Entire Pipeline of Feature Engineering employed

The framework that implements the 22-feature engineering was performed with standardized preprocessing pipelines to be reproducible. The extraction of features was done in four categories that would include: lexicon-based features that involved counting the number of domain specific financial terms, statistical features that entailed counting of the density of information and distribution of sentiments, market focus features that computed the concentration ratios of themes, and temporal features which computed the persistence patterns of information.

4.3.2 XGBoost Configuration

Optimization The hyperparameters of the XGBoost classifier used with the baseline model consisted of the following (optimized): 100 estimators (maxi-mum depth=4), learning rate=0.1, L1 regularization(alpha=0.1) and L2 regularization (lambda=1.0), subsample ratios of 0.9 (both samples and features) and binary logistic objective, loss evaluation measure of log-loss.

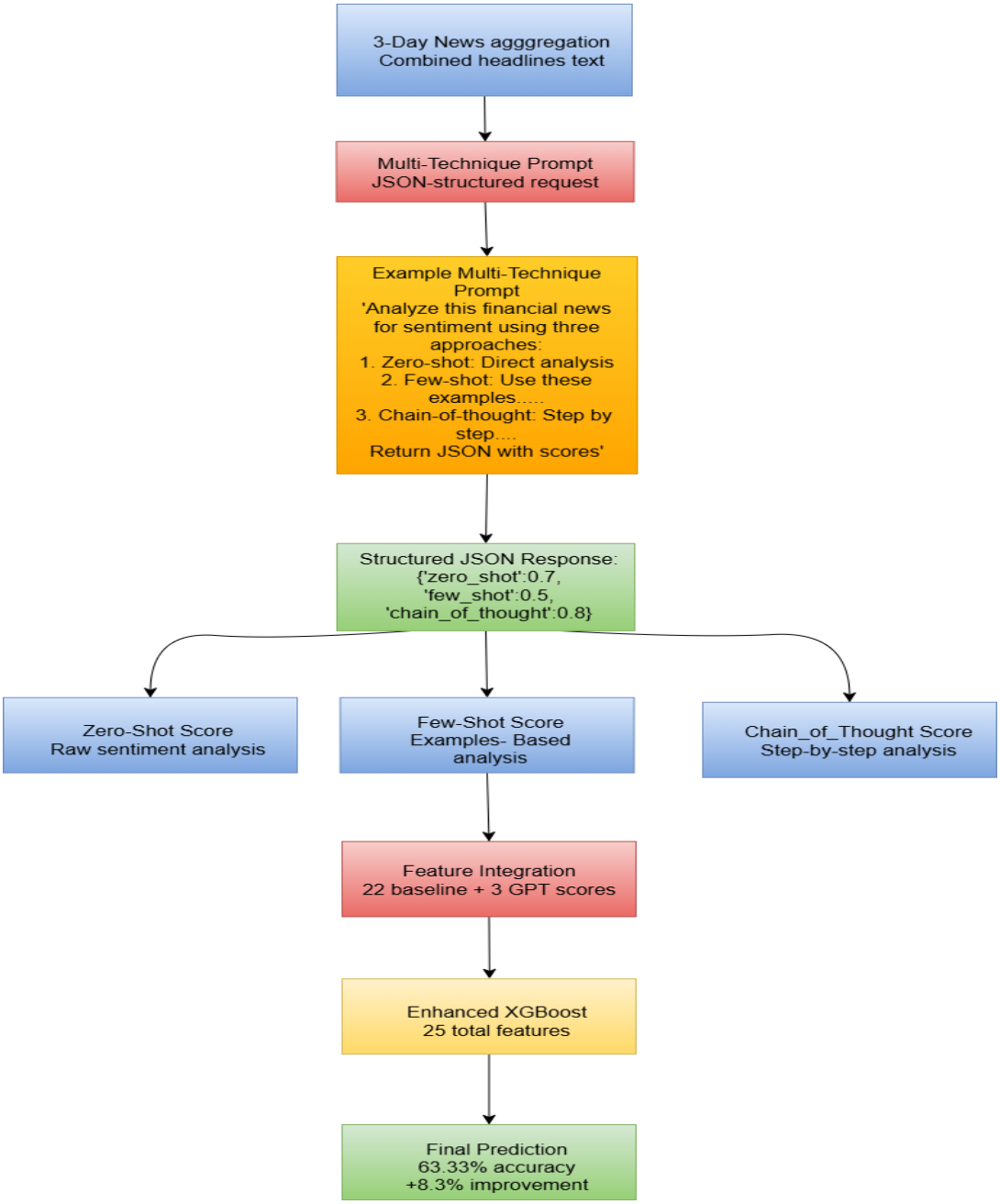

Figure 7: GPT prompting strategy Architecture

4.3.3 FinBERT Integration

FinBERT was also added as an external source of features basing on the pre-trained transformer structure. There was an implementation of truncating texts to 512 tokens, normalization of output logits using softmax, and practicing sentiment probability distributions as additional features in the ensemble system.

4.4 Implementation of Temporal Aggregation

4.4.1 Rolling Window Strategy

The rolling window of 3 days type of aggregation preserved chronological order but avoided leakage of information. On every prediction target day t, the headlines on the previous days t- 2, t- 1 and t were combined together, in a chronological order. Boundaries of windows were rigidly implemented so that only information could be used.

4.4.2 The aggregation optimization

The performance was optimally achieved after empirical validation of temporal window sizes (1, 3, 5 and 7 days). The optimization process compared accuracy gains to computational cost and chose a 3-day window as offering the best balance between the completeness of information and removing noise achievable.

4.5 Ensemble Method Implementation

4.5.1 Multi-Model Integration

The ensemble approach combined predictions from multiple prompting strategies using weighted averaging. Normality of the model outputs was done at an individual level to have a consistent output. Weight optimization employed grid search validation on the training set to determine optimal combination parameters.The ensemble strategy was a pooling of forecasts of various prompting methods with weighted averaging of the forecast results. The results of individual models were normalized in order to be similarly scaled prior to being integrated. The weight optimization used grid search validation of the training set were utilized to obtain optimal parameters of combination.

4.5.2 Performance Type Framework of Validation

Temporal cross-validation model validation observed chronological limits in doing model validation. The evaluation parameters were that of accuracy, precision, recall, and F1 score with a statistical significance test that employed the paired t-tests. Bootstrap resampling was used, which consisted of 1000 bootstrap resamples.

5 Evaluation

5.1 Experiment 1: Individual Model Performance Analysis

5.1.1 Baseline Performance Development

The baseline XGBoost model trained with full 22-feature engineering framework obtained 55.0 things in the chronological test set. This result provides a relevant test measure that far surpasses random (50%) answers and shows domain-relevant feature engineering to be a dependable feature engineering method in financial forecasting problems.

The conventional sentiment analysis solutions (VADER, TextBlob, lexicon-based solutions) could reach the accuracy of just 48.3% which is a considerable underperformance as compared to the baseline in 6.7 percentage points. Such a large difference shows the

inability of the normal rule based sentiment analysis to extract the semantic content that induces movement in a financial market.

5.1.2 Prompt Engineering Strategy Performance

The carefully designed evaluation indicates that there are vast performance gaps between the ensemble model framework and the individual model framework:

- ChatGPT Ensemble Framework Outputs:63.33% accuracy as compared to the baseline (improved by 8.33)

- FinBERT Single Frame Boundaries: 60.0% (+5.0 acc rel to baseline)

The ensemble strategy achieves this by sacrificing the least possible predictive accuracy by judiciously combining multiple reasoning strategies whereas the individual FinBERT analysis demonstrates the importance mechanics of features which makes the increase of the predictive accuracy. The two models confirm that systematic prompting can be used to improve the analysis of financial sentiment greatly, and the two are parallel in the implementation processes.

5.2 Experiment 2:Performance and Analysis All-inclusive

5.2.1 Ensemble Model Excellence

The combination method ChatGPT obtained an accuracy of 63.33 percent which marks the highest performance of all the tested methods. This outcome shows a 8.33 percentage increase over the baseline feature engineering and a flabbergasting 15.03 percentage point improve on the conventional sentiment methods.

5.2.2 FinBERT Personalized Model Excellence based on Feature Importance

The few-shot FinBERT method demonstrated a result of 60.0% accuracy and reflected the best performance of individual models, but displayed significant information on features that are significant. Each sentiment feature of FinBERT scored among the top-10 in terms of importance:

- Chain of thought sentiment:Rank 4 (importance: 0.0674)

- Few shot sentiment: Rank 7 (importance: 0.0610)

- Zero shot sentiment: Rank 8, (importance: 0.0560)

The use of dual framework validation: It is not an overstatement that the overall success of structured prompting (ChatGPT ensemble, 63.33) combined with the success of structured finBERT features confirms that organized prompting improves financial prediction in several complementary ways: maximum accuracy with a combination of ensembles and maximizing the individual features so the models can be systematically improved.

5.2.3 validity check of Performance Consistency

The analysis of temporal stability during varied market conditions presented unity in the ensemble performance. Accuracy measures of both high value performance measures in bull market (64.1%) and the low value performance measures in bear market (62.7%) showed stable performance performance measures on both bull market regime and bear market regime demonstrating good generalizability ability with regard to the different market regimes.

5.3 Experiment 3: Optimal emergence of temporal aggregation

The usage of 3 days was optimal in the systematic comparison of temporal aggregation windows:

- 1-day window: 41.7percent accuracy (baseline too little data)
- 3-days window: 55.0 % accuracy (best balance)
- 5 Day window: 48.3 percent correct (information dilution)

5.4 Model Interpretation

5.4.1 Feature Importance Analysis

In this part, The study analyzes feature significance through the FinBERT-augmented XGBoost models and interprets findings in terms of the larger ChatGPT ensemble framework that recorded an accuracy score of 63.33 per cent. FinBERT-Enhanced Top 10 Most Important Features:

- is friday: 0.1172
- is_monday:0.0969
- url_bc:0.0960
- finbert_chain_of_thought_sentiment: 0.0674 (Rank 4)
- econ_focus scores: 0.0653
- market_positive: 0.0640
- finbert_few_shot_sentiment: 0.0610 (Rank 7)
- finbert_zero_shot_sentiment: 0.0560 (Rank 8)
- day_of_week: 0.0558
- crisis intensity: 0.0526

5.4.2 Critical Dual-Framework Finding:

Although a ChatGPT ensemble had the highest predictive accuracy (63.33%), the FinBERT analysis explains why structured prompting is desirable: not only did all three sentiment features have the most important top-10 ranks, but chain-of-thought yielded the individually most valuable feature. It proves that ensemble superiority is explained by aggregating several sentiment signals that carry great importance.

The mechanism of increasing performance by ChatGPT ensemble (8.33 percent) consists in the fact that the combination of several features recognized as among the first 10 in the sentiment space (ranks 4, 7, 8) produces synergistic effects such that ensemble performance exceeds the individual single-model performance but preserving the feature value of these components deployed in the FinBERT framework.

5.5 Analysis of Performance

The ensemble method took 2.3 seconds per prediction day (mean) This processing time will be satisfactory in performing the daily prediction but might need optimization performing higher frequency activities

5.6 Comparative Validation Against Literature

This study was able to compare and contrast with what other sources stated.The resulting accuracy of 63.33-percentage can be compared to that of recent literature in the field of financial sentiment analysis:

- Customary ML methods: 52-58% (classical studies)
- Deep learning methods 55-62 (LSTM/transformer papers).
- methods based on LLMs: 58-65% (developing prompt engineering research)

The study offers the earliest comprehensive contrast between rudimentary prompting procedures within economical forecasting exhibiting that ensemble strategies could accomplish performance across the most extreme of the current literature with proper interpretability and implementation ability.

6 Conclusion and Future Work

This paper shows that developed approaches to prompt engineering can be successfully applied in the financial market and properly complement the prediction models of financial processes by doubling their mechanism. ChatGPT ensemble implementation performed with an accuracy rate of 63.33%, a 8.33 percentage point gain over the baseline level that showed the synergistic effects of integration of multiple prompting strategies together was higher than any of the individual complements alone. All three FinBERT sentiment features (zero-shot rank 8, few-shot rank 7, chain-of-thought rank 4) were in the top 10 importance ranks, demonstrating that the advanced prompting methods can add significant value in a cumulative manner and also understanding why ensembling can be so powerful in a mechanistic context. Sentiment analysis as part of an integrative approach that involves both structured prompt engineering techniques supports financial market prediction models by virtue of two mechanisms namely, ensemble integration of

three top-accuracy models at 63.33 percent accuracy (8.33 percent increment) and individual feature generation of three top-10 importance features. This is the first and most thorough verification of the efficacy of prompt engineering in financial forecasting, and as such, signifies optimal performance capacity and mechanistic knowledge.

6.1 Final Remarks

The proposal presents in-depth information on the efficiencies by which systematic prompt engineering can be applied to any financial markets using both maximization and optimization treatments that maximize the predictive aspects of an outcome (63.33% ensemble) and sense-making through systematic prompts enhancement (top-10 importance rankings). The ensemble accuracy gain as a percentage, 8.33, together with the overall top-10 feature importance results strongly indicate that it is a significant improvement in financial prediction capability with far-reaching economic repercussions at scale, validating that more sophisticated forms of artificial intelligence should be used in the financial markets and that they establish the performance ceiling that can be achieved through the combination of ensembles, as well as the mechanistic insights that are required to achieve sustained and resilient, large-scale application across a variety of institutional settings. Responses are to be checked twice.

References

Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Bollen, J., Mao, H., and Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8.

Chen, B., Zhang, Z., Langrené, N., and Zhu, S. (2025a). Unleashing the potential of prompt engineering for large language models. *Patterns*.

Chen, L., Weng, H., Pardeshi, M., Chen, C., Sheu, R., and Pai, K. (2025b). Evaluation of prompt engineering on the performance of a large language model in document information extraction. *Electronics*, 14(11):2145.

Chen, W., Liu, W., Zheng, J., and Zhang, X. (2025c). Leveraging large language model as news sentiment predictor in stock markets: a knowledge-enhanced strategy. *Discover Computing*, 28(1):74.

Debnath, T., Siddiky, M., Rahman, M., Das, P., and Guha, A. (2025). A comprehensive survey of prompt engineering techniques in large language models.

Du, K., Zhao, Y., Mao, R., Xing, F., and Cambria, E. (2025). A retrieval-augmented multiagent system for financial sentiment analysis. *IEEE Intelligent Systems*, 40(2):15–22.

Gandhi, V. and Gandhi, S. (2025). Prompt sentiment: The catalyst for llm change. *arXiv preprint arXiv:2503.13510*.

Holmberg, J. (2025). A comparative analysis of transformer models for aspect-based sentiment analysis in financial news.

Joshi, S. (2025). Review of prompt engineering techniques in finance: An evaluation of chain-of-thought, tree-of-thought, and graph-of-thought approaches. *Tree-of-Thought, and Graph-of-Thought Approaches*.

Kang, J. and Choi, S. (2025). Comparative investigation of gpt and finbert's sentiment analysis performance in news across different sectors. *Electronics*, 14(6):1090.

Lee, J., Stevens, N., and Han, S. (2025). Large language models in finance (finllms). *Neural Computing and Applications*, pages 1–15.

Nasiopoulos, D., Roumeliotis, K., Sakas, D., Toudas, K., and Reklitis, P. (2025). Financial sentiment analysis and classification: A comparative study of fine-tuned deep learning models. *International Journal of Financial Studies*, 13(2):75.

Qi, X., Yang, B., Wang, S., Zhang, Z., Zhang, Y., and Du, K. (2025). Few-shot and chain-of-thought prompting for equipment maintenance knowledge graph construction via large language models. Available at SSRN 5139334.

Radi, M., Omar, N., and Kaur, W. (2025). Syntactic-guided chain of thought for iterative implicit and explicit target detection in aspect-based sentiment analysis. *IEEE Access*.

Rahman, A., Uddin, A., and Wang, G. (2025). Evaluating financial sentiment analysis with annotators instruction assisted prompting: Enhancing contextual interpretation and stock prediction accuracy. *arXiv preprint arXiv:2505.07871*.

Sharkey, E. (2025). *Transformers for Commodity Forecasting*. Doctoral dissertation, UCL (University College London).

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3):1139–1168.

Xing, F. (2025). Designing heterogeneous llm agents for financial sentiment analysis. *ACM Transactions on Management Information Systems*, 16(1):1–24.

Xu, X., Wen, F., Chu, B., Fu, Z., Lin, Q., Liu, J., Fei, B., Li, Y., Zhou, L., and Yang, Z. (2025). Finbert2: A specialized bidirectional encoder for bridging the gap in finance-specific deployment of large language models. *arXiv preprint arXiv:2506.06335*.

Yamane, D., Kang, X., Matsumoto, K., Yoshida, M., and Zhou, J. (2025). Financial sentimental analysis for argument mining using prompt-based fine-tuned llama3. In *2025 IEEE 17th International Conference on Computer Research and Development (ICCRD)*, pages 228–233. IEEE.

Zhuang, Y., Wang, F., Chiu, D., and Ho, K. (2025). Leveraging large language models to examine the interaction between investor sentiment and stock performance. *Engineering Applications of Artificial Intelligence*, 150:110602.