Information Processing and Management 63 (2026) 104910

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

# Evaluating the effectiveness of fine-tuning in financial NLP: The case of Social Trading Action Detection☆

Simone D'Amico [a,*], Andrea Maurino [b], Francesco Osborne [c,d], Giancarlo Sperlì [a]

[a] Department of Electrical Engineering and Information Technology (DIETI), University of Naples "Federico II", Via Claudio 21, Naples, Italy
[b] Department of Informatics, Systems and Communication (DISCo), University of Milan-Bicocca, Viale Sarca 336, Milano, Italy
[c] Department of Business and Law, University of Milan-Bicocca, Via Bicocca degli Arcimboldi 8, Milano, Italy
[d] Knowledge Media Institute, The Open University, Berrill Building, Milton Keynes, United Kingdom

ARTICLE INFO

*Keywords:*
Fine-tuning
Large language models
Financial NLP
Social trading

ABSTRACT

Financial Natural Language Processing crucially leverages social media for market insights. However, most existing methods for this purpose rely on simple sentiment analysis models, which fail to capture the concrete trading intentions expressed in these discussions. While Large Language Models (LLMs) offer a promising alternative to simplistic sentiment analysis, the actual benefits of fine-tuning across different model families remain unclear in noisy, domain-specific contexts like online forums. To address this gap, we present a comprehensive assessment of the advantages and limitations of fine-tuning for Social Trading Action Detection (STAD), a novel task that aims to classify online posts into actionable categories, namely buy, sell, or other. In addition, we introduce FinReddit-2K, a manually annotated dataset consisting of 2123 Reddit posts, designed to serve as a benchmark for this task. Our experimental analysis goes beyond standard performance metrics and identifies both the types of errors that fine-tuning can successfully mitigate and those that it may inadvertently introduce. Through a systematic evaluation of 57 models, comparing 14 traditional models with 23 zero-shot LLMs and 20 fine-tuned variants, our results show that fine-tuning yields an average F1-score improvement of +15.1%. The best-performing model, a fine-tuned Mistral-7B, achieves an F1-score of 86.0%, although our analysis reveals that fine-tuning fails to produce meaningful performance gains in several scenarios.

## 1. Introduction

Large Language Models (LLMs) have demonstrated strong effectiveness across many NLP tasks in the financial domain (Ferraro & Sperlì, 2024; Shang et al., 2023; Zhang ,Zhang, Bao et al., 2024), particularly when adapted through parameter-efficient fine-tuning methods such as LoRA (Hu et al., 2022). The majority of existing studies, however, focus on identifying the best-performing model, which may rapidly change given the velocity of progress in this field, rather than developing a deeper understanding of how fine-tuning influences model behaviour in specific domains.

☆ This article is part of a Special issue entitled: 'FLLM' published in Information Processing and Management.

* Correspondence to: Department of Electrical Engineering and Information Technology (DIETI), University of Naples "Federico II", Via Claudio 21, 80125, Naples, Italy

*E-mail addresses:* simone.damico@unina.it (S. D'Amico), andrea.maurino@unimib.it (A. Maurino), francesco.osborne@unimib.it (F. Osborne), giancarlo.sperli@unina.it (G. Sperlì).

https://doi.org/10.1016/j.ipm.2026.104910
Received 29 October 2025; Received in revised form 19 February 2026; Accepted 14 May 2026
Available online 23 May 2026

In business contexts, organisations may not always be able to adopt the model that has been reported in the literature as the top performer for a specific task. This limitation may arise from several factors, such as licensing constraints, costs, resource availability, or the coexistence of different model families. In such cases, a company may attempt to specialise a model it already uses, with the expectation that fine-tuning should improve its performance compared to the base model. However, this assumption has not been consistently confirmed, as several studies have shown that fine-tuning does not always lead to significant gains (Barnett et al., 2024a; Macháček et al., 2025; Pu et al., 2023).

In this work, we perform a detailed analysis of the effect of fine-tuning on a variety of LLMs for a novel financial NLP task involving the analysis of noisy social media posts. In addition to identifying the best-performing architectures, we examine how fine-tuning interacts with different model characteristics and analyse the types of errors it alleviates or amplifies. The results of our study provide deeper insights into the strengths and limitations of fine-tuning when applied to a financial NLP task.

### 1.1. Social trading action detection

It is widely recognised that the dissemination and exchange of online information can have a substantial impact on economic dynamics, particularly on stock prices and trading volumes (Zhang ,Zhang, Bao et al., 2024). The rapid growth of digital platforms has provided numerous arenas where investors, traders, and other stakeholders actively discuss stocks, financial markets, and investment strategies (Zhuang et al., 2025). Platforms such as Yahoo Finance, StockTwits, InvestorHub, and Reddit have become central venues for these interactions (Ferraro & Sperlì, 2024). These dynamics enable large groups of retail investors to coordinate their behaviour through online platforms. Such groups may emerge spontaneously or coalesce around influential social media figures, whose posts act as focal points for collective decision-making. This coordinated behaviour can exert a substantial impact on financial markets. The most prominent example is the GameStop case (Betzer & Harries, 2022), which showed how coordinated retail investors could challenge traditional market mechanisms. To fully harness the informational potential of online platforms, it is essential to accurately identify and interpret investment-related recommendations within user-generated content. This task is commonly addressed as Financial Sentiment Analysis (FSA), where each post about a particular stock is assigned a positive or negative label. Consequently, automated sentiment analysis tools have become widely adopted in this domain due to their accessibility, ease of deployment, and the growing body of supporting research (Gentzkow et al., 2019; Loughran & McDonald, 2011; Wankhade et al., 2022).

However, despite their popularity, FSA methods have a strong limitation: they extract the general sentiment of a text rather than capturing the specific action performed or recommended by the user with respect to a particular stock. Although a positive sentiment may sometimes align with a recommendation to purchase a stock, this correspondence is not consistent. For instance, a post may express positive sentiment towards a company's values without suggesting a purchase, or it may criticise a stock while still recommending buying it for strategic reasons. A notable example occurred during the GameStop case, where negatively worded posts could indicate either a "buy" action as a form of protest against institutional investors or a "sell" recommendation to avoid anticipated losses. Consequently, identical sentiment labels could correspond to opposing investment strategies depending on the context.

In order to capture actionable signals emerging from financial online communities without reducing them to the oversimplified categories of sentiment analysis, it is necessary to directly interpret the concrete suggestions expressed in user posts. To this end, this paper introduces and investigates a novel task, *Social Trading Action Detection* (STAD), which involves classifying online posts according to the user's expressed action or recommendation regarding the intention to buy or sell a specific stock. We formalise STAD as a multi-class classification problem in which each post is assigned to one of three categories: *buy*, *sell*, or *other*. This task poses significant challenges due to the complexity, high level of noise, and pervasive use of domain-specific jargon in online financial discussions. Moreover, it cannot be effectively addressed using standard sentiment analysis techniques, as evidenced by the evaluation presented in this paper.

LLMs have demonstrated remarkable performance on tasks that require analysing financial documents (Li et al., 2023; Phogat et al., 2023; Wu et al., 2023). This capability makes them promising candidates for addressing STAD. Nonetheless, these models also exhibit limitations that may hinder their effectiveness in interpreting financial discussions. Notable weaknesses include hallucinations, difficulties in handling specialised jargon, and reduced robustness when processing noisy, ambiguous, or sarcastic language (Kalai et al., 2025; Zhang et al., 2025). These shortcomings are particularly relevant to STAD, since online financial discourse often relies on idiomatic expressions, irony, and community-specific terminology.

For these reasons, we argue that STAD represents a rigorous and practically relevant testbed for evaluating both the strengths and the limitations of next-generation LLMs. The task requires fine-grained linguistic understanding, resilience to noisy and domain-specific language, and the ability to distinguish actionable recommendations from general commentary. It therefore provides an opportunity to assess LLM performance in a challenging real-world setting, while also paving the way towards enhancing their capacity to support decision-making in complex social and financial environments.

### 1.2. Research questions

In this paper, we present a comprehensive study on the application of LLMs to the STAD task. Our objective is not limited to determining whether LLMs can outperform alternative methods or identifying the most effective LLM-based architecture. Rather, we provide a critical assessment of the performance exhibited by the LLMs under consideration, with special attention to the extent to which fine-tuning contributes to performance gains. Furthermore, we investigate their ability to handle linguistic phenomena that frequently occur in online posts, such as implicit language and sarcasm.

The analysis was driven by the following research questions:

- *RQ1:* How effective are modern LLMs at performing STAD when compared with traditional models, and which LLM achieves the best performance?
- *RQ2:* Does fine-tuning LLMs on high-quality data significantly enhance their performance on this task?
- *RQ3:* To what extent does the effectiveness of fine-tuning depend on the characteristics of the underlying model?
- *RQ4:* What types of errors occur most frequently, and how does fine-tuning influence their distribution, severity, and nature?

To address these research questions, we introduce FinReddit-2K, a new dataset including 2123 Reddit posts that have been manually classified by domain experts into the three predefined categories. The data were gathered from multiple subreddits and span a broad range of stocks, ensuring both thematic diversity and close alignment with real discussions in the financial domain.

To answer *RQ1*, we compared the performance of LLMs against 14 alternative approaches. These include nine neural network classifiers, such as the Multilayer Perceptron (MLP), Long Short-Term Memory (LSTM), and bidirectional LSTM (Bi-LSTM), three zero-shot sequence classifiers, and two encoder-based models built on BERT. While some of these approaches may appear relatively simple by current standards, they constitute strong and nontrivial baselines that are challenging to outperform. Indeed, our experiments show that two of these models still surpassed all LLMs in zero-shot settings and also outperformed several fine-tuned LLMs.

*RQ2* and *RQ3* motivated a comprehensive evaluation of multiple models. To address these questions, we conducted a systematic assessment of LLMs on FinReddit-2K, testing 23 models in zero-shot settings and 20 models in fine-tuned settings. These experiments enable a critical analysis of model performance and provide insights into the extent to which fine-tuning enhances results across different architectures.

Finally, to answer *RQ4*, we conducted an in-depth error analysis of the misclassifications produced by the models and the underlying factors contributing to them. We also grouped the errors into distinct categories that reveal the current limitations of foundational models in this domain. The main limitations include: (1) errors arising from indirect or context-dependent statements; (2) difficulty in distinguishing between generic opinions and explicitly recommended actions; (3) challenges in correctly interpreting sarcasm and slang; and (4) failure to differentiate between the reporting of news and the expression of an actual recommendation.

### 1.3. Main contributions

This paper advances research in financial NLP for information extraction from social media in two main ways. First, it formalises the STAD task and introduces a comprehensive benchmark that enables systematic model evaluation. Second, it reports an extensive evaluation of current LLM-based solutions, focusing not only on performance but also on the impact of fine-tuning and the typical errors observed in this domain.

The experimental evidence indicates that contemporary LLMs can substantially outperform traditional models on STAD when fine-tuned on high-quality datasets. In contrast, zero-shot LLMs underperform compared with both a strong MLP baseline and BERT-based models, reinforcing the conclusion that fine-tuning is essential to fully realise their capabilities. The extent of the improvements, however, varies substantially across architectures. Instruction-tuned medium-sized models (e.g., Mistral-7B) yield the most effective trade-off between accuracy and efficiency, whereas certain architectures fail to benefit and may even experience degradation after fine-tuning. The error analysis indicates that fine-tuning systematically mitigates critical errors such as action inversions and biased false positives, shifting the overall error profile towards less harmful misclassifications. Yet, ambiguity, irony, and implicit language remain persistent difficulties, highlighting the need for more sophisticated contextual reasoning. In conclusion, our findings show that carefully designed fine-tuning strategies, together with the selection of robust model families, are fundamental for achieving state-of-the-art results in this complex task.

In summary, the main contributions of this paper are as follows:

- We adopt the task of *Social Trading Action Detection* as a testbed for evaluating a wide range of LLMs and provide an in-depth investigation of the impact of fine-tuning and the types of errors produced.
- We introduce and release FinReddit-2K, a novel dataset for financial post classification that contains 2123 manually annotated Reddit posts categorised into three trading actions. The dataset is publicly available as a repository[1] .
- We benchmark LLMs against 14 traditional baselines, showing that they achieve superior performance on this task when fine-tuned.
- We conduct a detailed study on the effect of fine-tuning, demonstrating both its contribution to performance improvement and its influence on model behaviour and error distributions.
- We provide an in-depth analysis that identifies, quantifies, and discusses the typical limitations of LLMs in interpreting online posts.

### 1.4. Paper structure

The remainder of the paper is organised as follows. Section 2 reviews the relevant literature. Section 3 defines the task, introduces the proposed dataset, and details the experimental setup. Section 4 reports the main findings, while Section 5 answers the research questions and outlines their theoretical and practical implications. Finally, Section 6 summarises the main contributions and outlines potential directions for future research.

[1] https://github.com/Simone-Damico/FinReddit-2K-STAD

## 2. Related work

Stock market forecasting has garnered increasing interest from the research community, leading to the design of various approaches (Nazareth & Ramana Reddy, 2023; Tang et al., 2022), although several methodological and practical challenges remain unresolved (Ge et al., 2022; Thakkar & Chaudhari, 2021a). The inherently dynamic and highly non-linear nature of stock prices is driven by different factors, including corporate earnings reports, government policies, actions of influential stakeholders, and expert interpretations of current events (Thakkar & Chaudhari, 2021b; Wankhade et al., 2022), as further underlined in recent studies (Choi et al., 2024; Olorunnimbe & Viktor, 2023).

Despite the rapid advancement of Artificial Intelligence methodologies in financial analysis (Birti et al., 2025; Huang et al., 2023; Ma et al., 2022), their predictive accuracy remains constrained in the presence of unexpected or anomalous events (Gangopadhyay & Majumder, 2023; Gjerstad et al., 2021). To address these limitations, both academic and industry communities have increasingly explored unstructured textual content, such as news articles (Motta et al., 2024) and social media posts (Schmitz et al., 2023). In this context, the recent emergence of LLMs has demonstrated strong capabilities for analysing large-scale textual data originating from heterogeneous sources, including research papers (Bolanos et al., 2024; Buscaldi et al., 2024), medical records (Omiye et al., 2024), social media posts (Yang et al., 2024), user reviews (Chessa et al., 2023), domain taxonomies (Aggarwal et al., 2026), legal texts (Savelka & Ashley, 2023), domain-specific knowledge graphs (Tsaneva et al., 2025), and a wide range of financial documents (Li et al., 2023; Phogat et al., 2023; Wu et al., 2023). Therefore, the increasing availability of unstructured textual data, coupled with the evolution of NLP approaches, led to the development of novel methodologies for capturing investor sentiment in social media. It, however, adds further complexity to stock prediction tasks, which now require the integration and interpretation of high-volume, heterogeneous content to infer user opinions and market-relevant emotions (Choi et al., 2024). While traditional news media tend to provide more technical and structured insights (Motta et al., 2025), content from social media has demonstrated a stronger influence on public opinion (Angioni et al., 2024; Ferraro & Sperlì, 2024; Zhang ,Zhang, Bao et al., 2024). Dong et al. (2022) have demonstrated that news content yields higher predictive reliability on short-term (one-day) horizons, whereas social media exhibits greater utility in capturing market signals over extended periods ranging from two to five days.

Most of the existing approaches to financial market analysis have mainly focused on the processing of news articles (Li et al., 2024). Although (Xie et al., 2024) provide an extensive open-source evaluation benchmark — including 42 datasets spanning 24 financial tasks and covering eight critical dimensions, namely information extraction (IE), textual analysis, question answering (QA), text generation, risk management, forecasting, decision-making, and bilingual tasks — their focus remains largely constrained to news-based data. This limitation overlooks the variability and complementary perspectives introduced by social media platforms, which have increasingly attracted attention in financial contexts, as exemplified by the GameStop case (Ferraro & Sperlì, 2024).

For this reason, the analysis of user sentiment inferred from social textual content has garnered increasing attention in recent years, falling under the umbrella of Financial Sentiment Analysis (FSA) (Du et al., 2024).

The need to explore the relationship between investor sentiment and informed investment decisions, particularly through the integration of sentiment signals, has been emphasised by Qin et al. (2024), who incorporate this information into a machine learning framework to rank stocks and generate investment recommendations. However, their approach relies on pre-labelled data sourced from finance-oriented social platforms (e.g., StockTwits), without addressing the inherent challenges of identifying and classifying relevant posts within noisy and unstructured textual environments. In a different vein, (Zhuang et al., 2025) propose a two-stage method for enabling sentiment-informed decision-making in real-world financial contexts. Initially, a Llama-2-13b model is fine-tuned on a manually annotated binary sentiment dataset. The resulting labelled outputs are then used to construct sentiment indices, which serve as input features in a subsequent quantitative investment analysis. While the methodology illustrates the feasibility of leveraging large language models for sentiment extraction, it is constrained by the limited size of the annotated dataset (approximately 200 samples) the reliance on a single large language model for analysis, and the narrow scope of the data source, which is derived from a single investment-focused discussion forum.

Despite recent efforts to build financial-domain LLMs — e.g., FinMA (Xie, Han, Zhang et al., 2024) and Fin-GPT (Yang et al., 2023) — the available models remain relatively small in terms of number of parameters and their performance degrades when applied outside of the specific tasks for which they were optimised. Large foundation models such as BloombergGPT (Wu et al., 2023), trained with 50 billion parameters on internal Bloomberg datasets, show promising performance, although they are withheld from public release, limiting researchers in assessing domain-specific LLMs at scale in financial applications (Dong et al., 2024).

As highlighted in the literature, the analysis of opinions expressed in posts is predominantly addressed using FSA techniques, also focusing on a specific use case (Ferraro & Sperlì, 2024). This approach is reasonable in the context of news analysis, where positive sentiment is typically associated with good news about companies — encouraging investors to buy shares — while negative sentiment reflects bad news that may prompt investors to sell. However, FSA is not applicable in the context of individual investors' opinions because, as discussed in the introduction, the sentiment of a post is orthogonal to the author's trading intentions. The only work that recognises the need to measure trading intentions rather than sentiment is Zhuang et al. (2025), but it essentially employs standard FSA techniques.

Furthermore, the majority of existing approaches focuses on the analysis of news articles, whose content is typically written in a formal language, while other studies investigate sentiment analysis on social media. Although social media content are often

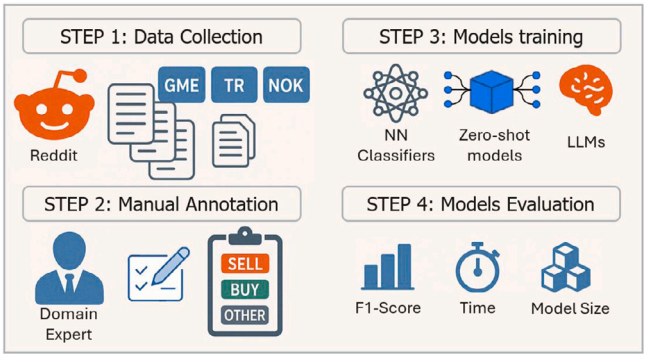


**Fig. 1.** Workflow of the proposed methodology.

characterised by informal and non-standard language, state-of-the-art approaches generally rely on pre-cleaned datasets, as for instance BigData22,[2] ACL18,[3] and CIKM18.[4]

Our approach is therefore original and, to the best of our knowledge, has not been explored in the literature except by Gupta et al. (2014), where the task is not specifically focused on stock trading.

## 3. Methodology

In this section, we describe the methodology employed to investigate LLMs on STAD. Our approach, illustrated in Fig. 1, involves the construction of a benchmark dataset followed by the evaluation of multiple families of classification methods on this benchmark. We first provide a formal definition of the STAD tasks, and then describe each stage of the experimental methodology.

### 3.1. Task definition

We formulate STAD as a multi-class classification problem: for a given natural language input, the objective is to categorise it into one of three classes: *buy*, *sell*, or *other*. More, formally, we define STAD as a function $f$ that maps a textual input $x$ (a social media post) to one of three predefined class labels $y$ in the set $\mathcal{Y}$:

$$f : x \to y \quad \text{where} \quad y \in \mathcal{Y} = \{\texttt{buy}, \texttt{sell}, \texttt{other}\} \tag{1}$$

The labels are defined as follows:

- `buy`: The post explicitly suggests, implicitly encourages, or affirms the author's intention to *purchase* the stock.
- `sell`: The post explicitly suggests, implicitly encourages, or affirms the author's intention to *divest* or *short* the stock.
- `other`: The post discusses the stock but does not advocate for a specific trading action. This includes neutral commentary, news reporting, factual questions, or off-topic discussion.

Although several approaches (Fatouros et al., 2024; Papasotiriou et al., 2024) have been proposed to classify social media posts into five discrete categories (i.e., strong buy, moderate buy, strong sell, moderate sell, and hold), such formulations impose a rigid assignment of posts to predefined classes. In contrast, our interactions with financial domain experts suggest that aggregating posts into continuous buy/sell signals, whose values are the probability with which each post is assigned to the underlying classes, is more appropriate. This probabilistic and continuous formulation defines an indicator that can be naturally integrated into portfolio optimisation and algorithmic trading models.

This formalisation allows us to tackle the problem with standard classification methodologies. Nevertheless, the task introduces distinct challenges stemming from the informal, context-dependent, and frequently nuanced nature of language used on social media.

### 3.2. The FinReddit-2K dataset

The proposed task is highly challenging, primarily because it requires linking social data, which are dynamic, unstructured, and expressed using non-standard vocabulary, to trading actions. To support the training and evaluation of classifiers for STAD (step 2 in Fig. 1), we introduce FinReddit-2K, a manually annotated benchmark. This dataset includes 2123 Reddit posts labelled into three actionable categories: *buy*, *sell*, or *other*. In the following, we describe the procedure used to identify and select the data, and then discuss the annotation process.

[2] https://github.com/deeptrade-public/slot
[3] https://github.com/yumoxu/stocknet-dataset
[4] https://github.com/wuhuizhe/CHRNN

**Table 1**
Distribution of posts collected from Reddit by stock in the FinReddit-2K dataset.

| Stock | Ticker | # Posts | Coverage (%) |
|---|---|---|---|
| GameStop | GME | 979 | 46.114% |
| Nokia | NOK | 166 | 7.819% |
| Tootsie Roll | TR | 132 | 6.218% |
| Aurora Cannabis | ACB | 125 | 5.888% |
| AMC Entertainment Holdings | AMC | 124 | 5.841% |
| Hims & Hers Health | HIMS | 123 | 5.794% |
| General Motors | GM | 123 | 5.794% |
| First Majestic Silver | AG | 120 | 5.652% |
| BlackBerry | BB | 118 | 5.558% |
| Rocket Companies | RKT | 113 | 5.323% |
| **Total** | 10 | **2123** | **100%** |

*Data source identification.* We selected Reddit as the data source because it hosts several communities focused on market discussion and analysis. Reddit is a well-known social media platform organised into thematic communities known as *subreddits*, which facilitate focused discussions and structured content sharing among users. For this study, we targeted eight finance-oriented subreddits: *wallstreetbets*, *finance*, *economics*, *investing*, *pennystocks*, *StockMarket*, *Stocks*, and *Dividends*. These communities represent central hubs for financial discourse, where users actively exchange views on stock price movements, macroeconomic trends, and investment strategies. Consequently, they represent a relevant and rich source of data for investigating retail investor sentiment and collective decision-making processes in digital environments.

To ensure a comprehensive analysis of the relationship between online discourse and investor sentiment, we selected ten stocks (listed in Table 1) that capture both the distinctive dynamics of the meme stock phenomenon and a diverse cross-section of market sectors. The selection was informed by two domain experts, both professors of finance with extensive expertise in financial markets and meme stocks. Specifically, *GameStop* (GME) and *AMC Entertainment* (AMC) were selected as prominent representative examples of meme stocks, given their high volatility and large retail investor communities, which make them particularly suitable for studying social media-driven market effects (Aggarwal et al., 2022). *BlackBerry* (BB), *Nokia* (NOK), and *Rocket Companies* (RKT) have also been strongly associated with online community discussions (Lee et al., 2025), but did not reach the same level of sustained attention or market disruption as GME and AMC. We can therefore consider them as secondary meme stocks. *BlackBerry* (BB) and *Nokia* (NOK) also represent legacy technology firms undergoing strategic transitions, while *Rocket Companies* (RKT) reflects the fintech sector. The remaining stocks were selected to ensure diversity in industry coverage and investment characteristics. *First Majestic Silver* (AG) provides exposure to the commodities sector; *Hims & Hers Health* (HIMS) reflects the emerging telehealth industry; *General Motors* (GM) exemplifies a traditional automotive manufacturer adapting to technological innovation; *Aurora Cannabis* (ACB) represents the speculative cannabis market; and *Tootsie Roll* (TR) serves as a contrast as a relatively stable consumer goods company.

Overall, this selection is balanced and representative of both meme stocks, which tend to attract heightened online attention and may be particularly sensitive to online events, and a diverse set of well-known companies operating across multiple market sectors.

*Data collection and selection.* To collect posts related to these ten stocks, we developed a Python-based module that leverages the PRAW (Python Reddit API Wrapper)[5] library to retrieve Reddit posts that met specific criteria. Posts were included if either the title or the body text contained the full name or ticker symbol of any of the selected stocks. Each Reddit post was associated with a single stock. We selected a total of 3000 posts spanning the period from *March 20, 2008* to *July 9, 2024*. The dataset was balanced so that 50% of the posts focused on GameStop, while the remaining 50% covered the other nine stocks.

*Annotation process.* Two domain experts manually annotated the posts. The annotators were asked to perform two tasks. First, they were instructed to identify and exclude posts that were off-topic, duplicates or near-duplicates of previously selected posts, clearly generated by bots, or too ambiguous to be reliably classified due to low relevance or high uncertainty (Li et al., 2020; Wang et al., 2025). Second, they were asked to assign each remaining post to one of three predefined categories. In particular, they were instructed to apply the *buy* and *sell* labels only when a post expressed an explicit or implicit trading intention. Posts expressing generic positive or negative sentiment without an investment-related action were instead to be assigned to the *other* category. This category was also to include informational or descriptive content, neutral discussions, and cases in which no actionable recommendation could be inferred.

Each annotator labelled the posts independently. After completing the independent labelling phase, the agreement between the experts was 87%. To account for agreement expected by chance, we also computed Cohen's Kappa, which provides a more robust measure of annotation reliability in the presence of class imbalance. The resulting value of $\kappa = 0.818$ indicates very strong agreement between annotators, confirming the high quality and consistency of the annotation process. When the assigned labels did not match, the annotators discussed the discrepancies to reach consensus and determine the appropriate shared label from among the initially proposed options.

[5] PRAW - https://praw.readthedocs.io/en/stable/

*Dataset statistics and characteristics.* The final version of the dataset consists of 2123 annotated Reddit posts. The label distribution is inherently imbalanced, with *buy*, *sell*, and *other* accounting for 54.0%, 9.5%, and 36.5% of the instances, respectively, reflecting realistic patterns of investor communication in online financial forums. Post length exhibits substantial variability, with an average of 160 words and a long-tailed distribution ranging from very short comments to extended multi-paragraph discussions (minimum 6 words, maximum 5032 words); the full length distribution is reported in Appendix A.1 (Fig. A.6).

The dataset spans more than 16 years of financial discussions, covering the period from 2008 to 2024, and includes content posted during different market phases and levels of volatility. As previously discussed, posts were collected from multiple finance-related subreddits, capturing heterogeneous discussion styles and investor behaviours; the distribution of posts across subreddits is detailed in Appendix A.1 (Table A.15). Stock-level coverage is reported in Table 1. Linguistically, the dataset is characterised by a high prevalence of informal language, community-specific slang, abbreviations, emojis, and domain-specific financial jargon, which contribute to its linguistic complexity and pose additional challenges for automatic classification.

*Limitations.* While FinReddit-2K was carefully constructed to ensure annotation quality and ecological validity, some limitations should be acknowledged. First, the dataset reflects a strongly imbalanced distribution of both sources and labels. Posts from the *wallstreetbets* subreddit account for the majority of the corpus, and discussions related to GameStop represent a substantial fraction of the data. Similarly, the *sell* class constitutes a relatively small portion of the dataset compared to *buy* and *other*. This imbalance is not a defect of the data collection process, but rather a direct consequence of the underlying dynamics of retail investor discourse on Reddit, where speculative and buy-oriented narratives are far more prevalent than explicit sell recommendations. Second, the dominance of a small number of highly active subreddits may limit the generalisability of the models to other forms of financial discourse, such as long-term investment discussions or institutional communication. However, capturing this concentration was a deliberate design choice, aimed at preserving the natural structure of social media-driven financial conversations.

Future work may address these limitations by expanding the dataset to additional platforms, increasing coverage of underrepresented trading actions, or incorporating temporal and user-level context to improve the detection of rare but critical signals such as sell intentions.

### 3.3. Adopted models

In this section, we discuss in detail the five categories of approaches evaluated on the FinReddit-2K dataset.

*Neural network classifiers.* We evaluated the performance of three traditional architectures on STAD: MLP, LSTM, and Bi-LSTM. Before training, we applied a preprocessing step to the texts to remove irrelevant tokens. To identify an optimal text representation strategy, we leveraged the Massive Text Embedding Benchmark (MTEB) (Muennighoff et al., 2022), a comprehensive framework for evaluating text embedding models across a diverse suite of NLP tasks. MTEB facilitates comparison of over 300 models based on criteria including performance, runtime efficiency, and embedding dimensionality. Based on this evaluation, we selected three open models from the Hugging Face platform to serve as encoders for the Reddit posts, prioritising a balance between model performance and computational efficiency. The selected models are: all-MiniLM-L6-v2[6] (22.7M params), all-mpnet-base-v2[7] (109M params), and gte-large-en-v1.5[8] (Zhang, Zhang, Long et al., 2024) (434M params). These encoders were used to transform the preprocessed Reddit posts into fixed-dimensional vector representations, which subsequently served as input features for the neural network classifiers.

*Pretrained zero-shot sequence classifiers.* Breakthroughs in transfer learning enabled the use of large pre-trained language models for downstream tasks with minimal or no additional training (Bonfigli et al., 2024; Brown et al., 2020). In this context, zero-shot sequence classification has emerged as a powerful methodology, allowing models to perform categorisation tasks without requiring domain-specific labelled data.

In our study, we employed three widely used transformer-based models for zero-shot learning. The first, mDeBERTa-v3-base-mnli-xnli[9] (279M params), incorporates disentangled attention mechanisms to enhance contextual understanding. The second, bart-large-mnli[10] (407M params), is a sequence-to-sequence model pre-trained on large-scale corpora and fine-tuned on the Multi-Genre Natural Language Inference (MNLI) dataset. The last model is xlm-roberta-large-xnli[11] (561M parameters), which extends the xlm-roberta-large model by fine-tuning it on a combination of natural language inference datasets. All models are designed for Natural Language Inference, enabling them to classify posts by evaluating their semantic similarity to predefined category descriptions.

*Encoder-only transformers.* We also evaluated two encoder-only transformer models fine-tuned on the FinReddit-2K dataset. In particular, we selected two widely adopted models from the BERT family: BERT-base uncased[12] (110M parameters) and BERT-large uncased[13] (340M parameters). Both models were fine-tuned using a standard sequence classification head applied to the encoder output. These models represent an intermediate category that bridges traditional neural classifiers and full-scale large language models.

[6] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2
[7] https://huggingface.co/sentence-transformers/all-mpnet-base-v2
[8] https://huggingface.co/Alibaba-NLP/gte-large-en-v1.5
[9] https://huggingface.co/MoritzLaurer/mDeBERTa-v3-base-mnli-xnli
[10] https://huggingface.co/facebook/bart-large-mnli
[11] https://huggingface.co/joeddav/xlm-roberta-large-xnli
[12] https://huggingface.co/google-bert/bert-base-uncased
[13] https://huggingface.co/google-bert/bert-large-uncased

**Table 2**
Overview of selected LLMs with release year, family, size category, and training methodology. All the models are available on the Hugging Face platform.

| Model | Size | Family | Release year | Training methodology |
|---|---|---|---|---|
| GPT-2-XL | Small (1.6B) | GPT-2 | Pre-2024 | Base |
| Falcon3-7B | Small (7B) | Falcon | 2024 | Base |
| Gemma-2-7B | Small (7B) | Gemma | 2024 | Instruction-Tuned |
| Mistral-7B | Small (7B) | Mistral | 2024 | Instruction-Tuned |
| Neural-chat-7b-v3-1 | Small (7B) | Mistral | Pre-2024 | Instruction-Tuned |
| Nous-Hermes-2-7B | Small (7B) | Mistral | 2024 | Instruction-Tuned |
| OLMo-7B (Groeneveld et al., 2024) | Small (7B) | OLMo | 2024 | Instruction-Tuned |
| Qwen-Distill-DeepSeek-R1 | Small (7B) | Qwen | 2025 | Distilled |
| Qwen2.5-7B | Small (7B) | Qwen | 2024 | Instruction-Tuned |
| Starling-LM-7B (Zhu et al., 2023) | Small (7B) | Mistral | 2024 | Instruction-Tuned |
| Zephyr-7B (Tunstall et al., 2023) | Small (7B) | Mistral | Pre-2024 | Instruction-Tuned |
| Llama-3.1-8B | Small (8B) | Llama | 2024 | Instruction-Tuned |
| Llama-Distill-DeepSeek-R1 (8B) | Small (8B) | Llama | 2024 | Distilled |
| Falcon3-10B | Medium (10B) | Falcon | 2024 | Instruction-Tuned |
| SOLAR-10.7B | Medium (10.7B) | Mistral | Pre-2024 | Instruction-Tuned |
| Gemma-3-12B | Medium (12B) | Gemma | 2025 | Instruction-Tuned |
| FinGPT-v3.3-13B (Wang et al., 2023) | Medium (13B) | Llama | Pre-2024 | Instruction-Tuned |
| Llama-2-13B | Medium (13B) | Llama | Pre-2024 | Instruction-Tuned |
| Qwen1.5-14B | Medium (14B) | Qwen | 2024 | Base |
| Phi-4-14B | Medium (14B) | Phi | 2024 | Base |
| Gemma-3-27B | Large (27B) | Gemma | 2025 | Instruction-Tuned |
| QwQ-32B | Large (32B) | Qwen | 2025 | Instruction-Tuned |
| Llama-3.3-70B | Large (70B) | Llama | 2024 | Instruction-Tuned |

*LLMs in zero-shot setting.* We assessed the efficacy of LLMs for zero-shot learning (ZSL) text classification utilising a prompt-based inference paradigm. Each Reddit post was presented to the model alongside a fixed instructional prompt that directed it to categorise the content into one of the three predefined labels. This methodology capitalises on the robust generalisation abilities inherent in LLMs, which, by virtue of their pre-training on extensive and diverse corpora, can interpret and execute natural language instructions without requiring any task-specific fine-tuning. The full prompt is reported in Appendix A.5.

Our evaluation considered a diverse set of 23 contemporary open-weight LLMs, selected to span a wide range of architectural families and model scales. This breadth enables a systematic analysis of how different model properties affect zero-shot learning (ZSL) performance on the STAD task. The primary characteristics used to describe the models, and reported in Table 2, are as follows:

1. **Model Size:** Categorised as Small (< 10B), Medium (10B–20B), and Large (> 20B), reflecting parameter count and the corresponding representational capacity.
2. **Architectural Family:** Grouped according to model lineage, including widely adopted families such as Llama, Mistral, Gemma, Falcon, Qwen, and Phi.
3. **Release Era:** Classified as Pre-2024, 2024, or 2025, capturing temporal advances in training strategies and pre-training data.
4. **Training Methodology:** Distinguished as Base (pre-trained only), Instruction-Tuned (supervised fine-tuning on instruction-response data), and Distilled (student models trained from larger teacher models).

*Fine-tuned LLMs.* In addition to evaluating LLMs in a zero-shot setting, we fine-tuned most of the models listed in Table 2 on the FinReddit-2K dataset. We adopted a LoRA-based adaptation strategy, which enables efficient fine-tuning of large architectures by updating a limited set of low-rank parameters while keeping the original model weights frozen. During fine-tuning, each Reddit post was processed using an instruction-style prompt.

Only three of the LLMs evaluated in ZST were excluded from the fine-tuning experiments. Llama-3.3-70B could not be fine-tuned due to computational constraints, while two other models exhibited severe performance degradation after fine-tuning and were therefore discarded.

### 3.4. Experimental analysis

In this section, we describe the comparative evaluation of the previously presented methods using the FinReddit-2K dataset. All experiments were conducted on a machine equipped with an NVIDIA A100-SXM4 with 80 GB of GPU memory.

#### 3.4.1. Experimental protocol

We conducted a 5-fold cross-validation on FinReddit-2K using the models described in the previous section. These models include 9 neural network classifiers (MLP, LSTM, and Bi-LSTM, each employing the 3 encoder models), 3 zero-shot sequence classifiers, 2 encoder-based models, 23 LLMs in ZSL setting, and 20 fine-tuned LLMs.

The experimental protocol comprises the following steps:

1. To prepare the textual data for downstream analysis and enhance the models' ability to distinguish between the three classes, we applied a pre-processing pipeline aimed at reducing noise and standardising input. Specifically, we removed HTML tags, URLs, Reddit-specific markers (e.g., /r/, u/, [deleted]), and all type of whitespace characters.
2. For neural network classifiers and BERT-based models, we conducted a grid search over their model-specific hyperparameters. These hyperparameters include batch size, hidden layer dimensions, dropout rate, maximum sequence length, and the number of training epochs. Other models were trained only once, using specific values for parameters such as training epochs, optimiser, or temperature for generation.
3. For each model and each hyperparameter combination, we performed a stratified $k$-fold cross-validation with $k = 5$. This procedure ensures reliable performance estimation while maintaining the original class distribution within each fold. During training, 10% of the training data was reserved as a validation set.

Given the imbalance among the target classes, we employ weighted versions of standard evaluation metrics to ensure a more accurate assessment of the model's performance. Specifically, we compute the weighted *precision*, *recall*, and *F1-score*, which account for the number of samples in each class. For selected models, we also report the macro-averaged precision, recall, and F1-score (see Tables 6 and 8). In addition, we report the overall *accuracy*.

While these metrics provide a general indication of quality, they do not reveal *which types of errors* the model makes or *how costly* those errors are in practical terms. For instance, misclassifying a clear *sell* signal as *buy* may have significantly worse consequences than confusing it with a neutral *other*. Hence, we perform a structured error analysis from different points of view. First, we analyse consistent patterns of misclassification. Second, we conduct a qualitative analysis of the best model predictions, considering both correctly and incorrectly classified posts. For each case, we also asked the model to generate an explanation of its own prediction, which we used to assess the consistency between the predicted label, the underlying text, and the model's reasoning.

#### 3.4.2. Evaluation settings

In the following, we report the evaluation settings used for the four families of approaches.

- **Neural network classifiers:** For each model, we perform a grid search over the following parameter combinations: *encoder* $\in$ {*all-MiniLM-L6-v2, all-mpnet-base-v2, gte-large-en-v1.5*}$\times$*model* $\in$ {*MLP, LSTM, Bi-LSTM*}$\times$ *hidden dimension layer* $\in \{64, 128\}\times$*batch size* $\in \{16, 32, 64\}\times$ *dropout* $\in \{0, 0.05, 0.1\}\times$*max sequence length* $\in \{64, 128, 256\} \times$ *training epochs* $\in \{5, 10\}$ for a total of 972 generated models.
- **Pretrained zero-shot sequence classifiers:** We use the three models described in Section 3.3 directly in their original form, without additional training or fine-tuning.
- **Encoder-only transformers:** A grid search was conducted to fine-tune BERT using the following parameters: *model* $\in$ {*BERT-base, BERT-large*}$\times$ *batch size* $\in \{8, 16, 32\}\times$ *max sequence lenght* $\in \{128, 256, 512\}\times$ *training epochs* $\in \{2, 4\}$, for a total of 36 models.
- **LLMs in zero-shot setting:** All models were deployed using an 8-bit quantisation technique specifically designed to reduce memory consumption and speed up inference. For the generation step, to ensure reproducibility, we set the temperature to 0.001.
- **Fine-tuned LLMs:** For the Fine-tuned LLMs, as for zero-shot, we used 8-bit quantisation and a temperature of 0.001. Additionally, for the training phase, the following parameters were set: LoRA attention dimension ($r$ parameter) = 64, batch size= 8, learning rate = 2e-4, number of training epochs = 2 and AdamW optimiser. We adopted commonly used hyperparameter values in order to assess the performance improvements that can realistically be expected under typical fine-tuning settings. Given the large number of LLMs evaluated, conducting an exhaustive hyperparameter grid search would be computationally infeasible and beyond the scope of this work.

## 4. Results

In the following, we present and discuss the results of the experiments. For clarity, the analysis is organised according to the four research questions defined in the introduction.

### 4.1. RQ1 - performance of LLMs against traditional models

Concerning RQ1, which investigates how the performance of LLMs compares with traditional solutions, Table 3 presents the results of the 57 evaluated models. Appendix A.2 reports the full list of optimal hyperparameter configurations for the neural and encoder-only classifiers. The pre-trained zero-shot sequence classifiers, together with the LSTM and Bi-LSTM baselines, achieve relatively low F1-scores, only slightly above 50%. In contrast, the MLP model combined with the gte-large-en-v1.5 encoder attains a substantially higher F1-score (69.6%). This result outperforms all large language models in the zero-shot setting and suggests that, without fine-tuning, LLMs are not necessarily the most effective solution. Table A.16 presents the optimal parameter configurations for the baseline models identified through a grid search. The encoder-only transformers also achieve strong performance. In particular, the BERT base model outperforms all other baselines and all LLMs in the zero-shot setting, achieving an F1 score of 73.7%. Nevertheless, its performance remains substantially lower than that of the best-performing fine-tuned LLMs.

**Table 3**
Model evaluation results. Each metric is reported along with its 95% confidence interval, computed over the 5-fold cross-validation. Models are ranked by F1 for each family. The table shows the best model for each family and the **overall best model**.

| Family | Model | Accuracy | | Precision | | Recall | | F1-score | |
|---|---|---|---|---|---|---|---|---|---|
| | | Value | C.I.95% | Value | C.I.95% | Value | C.I.95% | Value | C.I.95% |
| Neural network classifiers | gte-large-en-v1,5 + MLP | 0.729 | 0.724–0.734 | 0.673 | 0.647–0.699 | 0.729 | 0.724–0.734 | 0.696 | 0.685–0.707 |
| | all-MiniLM-L6-v2 + MLP | 0.674 | 0.667–0.681 | 0.628 | 0.623–0.633 | 0.674 | 0.667–0.681 | 0.625 | 0.616–0.634 |
| | all-mpnet-base-v2 + MLP | 0.687 | 0.676–0.698 | 0.631 | 0.621–0.641 | 0.687 | 0.676–0.698 | 0.644 | 0.632–0.656 |
| | all-mpnet-base-v2 + LSTM | 0.540 | 0.539–0.541 | 0.292 | 0.291–0.293 | 0.540 | 0.539–0.541 | 0.379 | 0.378–0.380 |
| | all-MiniLM-L6-v2 + LSTM | 0.535 | 0.518–0.552 | 0.447 | 0.389–0.505 | 0.535 | 0.518–0.552 | 0.419 | 0.397–0.441 |
| | gte-large-en-v1,5 + LSTM | 0.540 | 0.539–0.541 | 0.292 | 0.291–0.293 | 0.540 | 0.539–0.541 | 0.379 | 0.378–0.380 |
| | gte-large-en-v1,5 + Bi-LSTM | 0.605 | 0.561–0.649 | 0.544 | 0.418–0.670 | 0.605 | 0.561–0.649 | 0.530 | 0.432–0.628 |
| | all-MiniLM-L6-v2 + Bi-LSTM | 0.540 | 0.539–0.541 | 0.292 | 0.291–0.293 | 0.540 | 0.539–0.541 | 0.379 | 0.378–0.380 |
| | all-mpnet-base-v2 + Bi-LSTM | 0.540 | 0.539–0.541 | 0.292 | 0.291–0.293 | 0.540 | 0.539–0.541 | 0.379 | 0.378–0.380 |
| Pretrained zero-shot sequence classifiers | bart-large-mnli | 0.535 | 0.522–0.548 | 0.617 | 0.598–0.636 | 0.535 | 0.522–0.548 | 0.549 | 0.536–0.562 |
| | xlm-roberta-large-xnli | 0.508 | 0.496–0.520 | 0.636 | 0.619–0.653 | 0.508 | 0.496–0.520 | 0.516 | 0.502–0.530 |
| | mDeBERTa-v3-base-mnli-xnli | 0.513 | 0.494–0.532 | 0.643 | 0.609–0.677 | 0.513 | 0.494–0.532 | 0.496 | 0.474–0.518 |
| Encoder-only transformers | BERT-base-uncased | 0.746 | 0.739–0.753 | 0.741 | 0.731–0.751 | 0.746 | 0.739–0.753 | 0.737 | 0.728–0.746 |
| | BERT-large-uncased | 0.709 | 0.692–0.726 | 0.649 | 0.614–0.684 | 0.709 | 0.692–0.726 | 0.669 | 0.640–0.698 |
| LLMs in zero-shot setting | Gemma-3-27B | 0.694 | 0.681–0.707 | 0.704 | 0.691–0.717 | 0.694 | 0.681–0.707 | 0.682 | 0.668–0.696 |
| | Gemma-3-12B | 0.613 | 0.587–0.639 | 0.651 | 0.615–0.687 | 0.613 | 0.587–0.639 | 0.600 | 0.573–0.627 |
| | Falcon3-10B | 0.589 | 0.574–0.604 | 0.560 | 0.544–0.576 | 0.589 | 0.574–0.604 | 0.548 | 0.530- 0.566 |
| | SOLAR-10.7B | 0.586 | 0.572–0.600 | 0.586 | 0.565–0.607 | 0.586 | 0.572–0.600 | 0.543 | 0.523–0.563 |
| | Llama-2-13b | 0.592 | 0.575–0.609 | 0.592 | 0.567–0.617 | 0.592 | 0.575–0.609 | 0.538 | 0.512–0.564 |
| | Falcon3-7B | 0.582 | 0.563–0.601 | 0.566 | 0.542–0.590 | 0.582 | 0.563–0.601 | 0.538 | 0.521–0.555 |
| | Starling-LM-7B | 0.584 | 0.571–0.597 | 0.582 | 0.562–0.602 | 0.584 | 0.571–0.597 | 0.535 | 0.520–0.550 |
| | Llama-3.1-8B | 0.569 | 0.558–0.580 | 0.570 | 0.548–0.592 | 0.569 | 0.558–0.580 | 0.526 | 0.515–0.537 |
| | OLMo-7B | 0.515 | 0.488–0.542 | 0.525 | 0.504–0.546 | 0.515 | 0.488–0.542 | 0.513 | 0.490–0.536 |
| | Llama-3.3-70B | 0.497 | 0.469–0.525 | 0.558 | 0.526–0.590 | 0.497 | 0.469–0.525 | 0.510 | 0.482–0.538 |
| | Mistral-7B | 0.558 | 0.542–0.574 | 0.542 | 0.509–0.575 | 0.558 | 0.542–0.574 | 0.505 | 0.483–0.527 |
| | Nous-Hermes-2-7B | 0.558 | 0.543–0.573 | 0.539 | 0.510–0.568 | 0.558 | 0.543–0.573 | 0.504 | 0.487–0.521 |
| | Llama-Distill-DeepSeek-R1-8B | 0.517 | 0.497–0.537 | 0.497 | 0.469–0.525 | 0.517 | 0.497–0.537 | 0.482 | 0.459–0.505 |
| | Qwen2.5-7B | 0.514 | 0.493–0.535 | 0.547 | 0.521–0.573 | 0.514 | 0.493–0.535 | 0.480 | 0.460–0.500 |
| | GPT-2-XL-1.6B | 0.525 | 0.515–0.535 | 0.520 | 0.492–0.548 | 0.525 | 0.515–0.535 | 0.472 | 0.458–0.486 |
| | Phi-4-14B | 0.492 | 0.480–0.504 | 0.542 | 0.508–0.576 | 0.492 | 0.480–0.504 | 0.431 | 0.423–0.439 |
| | QwQ-32B | 0.428 | 0.404–0.452 | 0.533 | 0.513–0.553 | 0.428 | 0.404–0.452 | 0.424 | 0.404–0.444 |
| | Gemma-7b | 0.358 | 0.330–0.386 | 0.538 | 0.514–0.562 | 0.358 | 0.330–0.386 | 0.384 | 0.358–0.410 |
| | Neural-chat-7b | 0.349 | 0.335–0.363 | 0.642 | 0.623–0.661 | 0.349 | 0.335–0.363 | 0.383 | 0.366–0.400 |
| | Qwen-Distill-DeepSeek-R1-7B | 0.407 | 0.393–0.421 | 0.487 | 0.468–0.506 | 0.407 | 0.393–0.421 | 0.360 | 0.341–0.379 |
| | Qwen1.5-14B | 0.272 | 0.253–0.291 | 0.558 | 0.48–0.636 | 0.272 | 0.253–0.291 | 0.225 | 0.206–0.244 |
| | Zephyr-7B | 0.227 | 0.218–0.236 | 0.371 | 0.329–0.413 | 0.227 | 0.218–0.236 | 0.160 | 0.155–0.165 |
| | FinGPT-v3.3-13B | 0.096 | 0.094–0.098 | 0.225 | 0.000–0.484 | 0.096 | 0.094–0.098 | 0.018 | 0.015–0.021 |
| Fine-tuned LLMs | Mistral-7B | 0.860 | 0.844–0.876 | 0.862 | 0.846–0.878 | 0.860 | 0.844–0.876 | **0.860** | 0.844–0.876 |
| | Neural-chat-7B | 0.848 | 0.833–0.863 | 0.849 | 0.832–0.866 | 0.848 | 0.833–0.863 | 0.847 | 0.831–0.863 |
| | Phi-4–14B | 0.847 | 0.835–0.859 | 0.847 | 0.835–0.859 | 0.847 | 0.835–0.859 | 0.846 | 0.834–0.858 |
| | Zephyr-7B | 0.840 | 0.811–0.869 | 0.844 | 0.821–0.867 | 0.840 | 0.811–0.869 | 0.840 | 0.812–0.868 |
| | Llama-3.1-8B | 0.821 | 0.805–0.837 | 0.832 | 0.815–0.849 | 0.821 | 0.805–0.837 | 0.822 | 0.807–0.837 |
| | SOLAR-10.7B | 0.581 | 0.554–0.608 | 0.604 | 0.571–0.637 | 0.581 | 0.554–0.608 | 0.574 | 0.545–0.603 |
| | FinGPT-v3.3-13B | 0.566 | 0.536–0.596 | 0.593 | 0.554–0.632 | 0.566 | 0.536–0.596 | 0.548 | 0.513–0.583 |
| | Nous-Hermes-2-7B | 0.572 | 0.554–0.590 | 0.588 | 0.557–0.619 | 0.572 | 0.554–0.590 | 0.548 | 0.525–0.571 |
| | Llama-2-13B | 0.569 | 0.537–0.601 | 0.588 | 0.547–0.629 | 0.569 | 0.537–0.601 | 0.546 | 0.504–0.588 |
| | Starling-LM-7B | 0.570 | 0.547–0.593 | 0.581 | 0.546–0.616 | 0.570 | 0.547–0.593 | 0.531 | 0.507–0.555 |
| | OLMo-7B | 0.568 | 0.286–0.850 | 0.629 | 0.320–0.938 | 0.568 | 0.286–0.850 | 0.529 | 0.259–0.799 |
| | Llama-Distill-DeepSeek-R1-8B | 0.567 | 0.551–0.583 | 0.555 | 0.532–0.578 | 0.567 | 0.551–0.583 | 0.513 | 0.495–0.531 |
| | Qwen1.5-14B | 0.504 | 0.100–0.908 | 0.504 | 0.100–0.908 | 0.504 | 0.100–0.908 | 0.504 | 0.101–0.907 |
| | Falcon3-10B | 0.575 | 0.560–0.590 | 0.570 | 0.545–0.595 | 0.575 | 0.560–0.590 | 0.492 | 0.470–0.514 |
| | Falcon3-7B | 0.564 | 0.552–0.576 | 0.552 | 0.508–0.596 | 0.564 | 0.552–0.576 | 0.475 | 0.462–0.488 |
| | Qwen2.5-7B | 0.491 | 0.479–0.503 | 0.574 | 0.549–0.599 | 0.491 | 0.479–0.503 | 0.442 | 0.429–0.455 |
| | Qwen-Distill-DeepSeek-R1-7B | 0.530 | 0.518–0.542 | 0.522 | 0.477–0.567 | 0.530 | 0.518–0.542 | 0.437 | 0.422–0.452 |
| | GPT-2-XL-1.6B | 0.539 | 0.531–0.547 | 0.488 | 0.460–0.516 | 0.539 | 0.531–0.547 | 0.425 | 0.415–0.435 |
| | QwQ-32B | 0.487 | 0.467–0.507 | 0.549 | 0.525–0.573 | 0.487 | 0.467–0.507 | 0.422 | 0.403–0.441 |
| | Gemma-7b | 0.366 | 0.314–0.418 | 0.569 | 0.529–0.609 | 0.366 | 0.314–0.418 | 0.383 | 0.34–0.426 |

The results of the LLMs in the zero-shot scenario show considerable variation. Gemma-3-27B delivers the best performance, reaching an F1-score of 68.2%. Conversely, several systems, such as QwQ-32B, score below 50%. A particularly striking case is FinGPT 13B, a model explicitly designed for financial sentiment classification, which performs very poorly. This outcome reinforces the observation that STAD cannot be adequately tackled by relying on standard sentiment analysis approaches.

The fine-tuned LLMs include both the overall best-performing model and several underperforming ones. The highest performance is achieved by Mistral-7B, which attains an F1-score of 86.0%. Notably, the top five fine-tuned LLMs (Mistral-7B, Neural-chat-7b, Phi-4, Zephyr-7B, and Llama-3.1-8B) are the only models to exceed the 80% threshold among the 57 evaluated solutions. This result confirms the effectiveness of high-performing fine-tuned LLMs for this task.

These results indicate that both the complexity of the model and the application of task-specific fine-tuning play a crucial role in attaining strong performance. Although lightweight architectures like MLPs can achieve competitive outcomes when combined with high-quality sentence embeddings, LLMs show clear benefits once fine-tuned to the target task. In addition, parameter-efficient strategies such as LoRA allow effective fine-tuning without incurring excessive computational overhead, reinforcing recent directions towards scalable and widely accessible fine-tuning approaches (Hu et al., 2022). The performance gains observed through fine-tuning

**Table 4**
Stock-level Mistal-7B performance metrics and aggregated results for meme and non-meme stocks.

| Stock | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| ACB | 0.776 | 0.763 | 0.744 | 0.752 |
| AG | 0.858 | 0.831 | 0.844 | 0.837 |
| AMC | 0.823 | 0.571 | 0.560 | 0.565 |
| BB | 0.899 | 0.798 | 0.866 | 0.826 |
| GM | 0.797 | 0.779 | 0.720 | 0.737 |
| GME | 0.884 | 0.876 | 0.879 | 0.877 |
| HIMS | 0.878 | 0.854 | 0.788 | 0.814 |
| NOK | 0.855 | 0.814 | 0.801 | 0.807 |
| RKT | 0.858 | 0.836 | 0.789 | 0.808 |
| TR | 0.811 | 0.812 | 0.809 | 0.807 |

are consistent with earlier studies (Zhuang et al., 2025), underscoring the importance of domain adaptation for fully exploiting the potential of pre-trained LLMs (Gururangan et al., 2020; Howard & Ruder, 2018).

*Stock-level analysis.* To further investigate within-dataset generalisation, we conducted a stock-level analysis of the best model by evaluating classification metrics separately for each stock included in the FinReddit-2K dataset. Table 4 reports accuracy, precision, recall, and F1-score for each individual stock.

The results reveal variability across stocks, which can be partially attributed to differences in linguistic style and community engagement. The two primary meme stocks (GME and AMC) obtain an average F1 score of 72.1%, but with notably different results. In particular, GME exhibits the highest performance among all stocks, likely benefiting from both its large representation in the dataset and the presence of recurring and highly stereotyped linguistic patterns characteristic of meme stock discourse. In contrast, AMC, another well-known meme stock with a well-documented turbulent history (Vasileiou & Tzanakis, 2024), shows lower performance. This may reflect the more heterogeneous and evolving narratives associated with AMC discussions, resulting in greater linguistic variability that can reduce classification accuracy, particularly during periods of high price volatility. The three secondary meme stocks (BB, NOK, RKT) introduced in Section 3.2 also exhibit strong performance, achieving an average F1 score of 81.4%. This result is noteworthy because these stocks are representative of assets frequently discussed in online investor communities, but are not associated with market dynamics as extreme as those observed for GME. Evaluating the proposed approach on these stocks, therefore, provides evidence that the method generalises beyond highly anomalous events, offering a realistic testbed for automatic methods in community-driven financial discussions. Traditional non-meme stocks (ACB, AG, GM, HIMS, TR) achieve stable and competitive performance, with an average F1 score of 78.9%. This suggests that the model generalises well beyond meme-driven language.

Overall, the model maintains reasonable generalisation across different types of financial discourse. Although meme stocks are characterised by distinctive lexical cues and recurrent narratives that can facilitate the identification of actions, they also pose challenges due to the frequent use of irony, contextual language, and the highly speculative and volatile nature of the discussions surrounding them. We further discuss the relevant error patterns in Section 4.4.

### 4.2. RQ2 - benefits of fine-tuning

Table 5 presents a comparison of different LLMs by reporting their average inference and fine-tuning times, together with the variation in F1-score when transitioning from zero-shot to fine-tuned configurations. As previously discussed, fine-tuning proved to be highly effective, increasing the average F1-score by approximately 15.1%. Seven models achieve an F1 improvement of more than 25%. Among these, five (Zephyr-7B, Neural-chat-7b, Phi-4-14B, Mistral-7B, and Llama-3.1-8B) also stand out as the top overall performers. Notably, Mistral-7B achieves the highest F1-score (86.0%) while maintaining one of the fastest inference times (37 s) and requiring only a moderate fine-tuning time (33 min). Interestingly, a few models (e.g., Falcon3-7B, QwQ-32B) perform worse than in the zero-shot scenario.

The fine-tuning process adapts a model's internal representations to a specific downstream task by reinforcing task-relevant patterns in the training data. However, consistent with the bias–variance trade-off, excessive specialisation may increase model bias and reduce generalisation, potentially leading to catastrophic forgetting (Barnett et al., 2024b; Mai et al., 2024). In particular, post-training optimisation can reshape model representations in ways that favour the fine-tuning objective while partially degrading capabilities acquired during pretraining.

In our experiments, this effect appears to be model-dependent. Mistral-based models, Phi, and LLaMA show consistent improvements. In contrast, performance degradation is observed for Falcon, Qwen2.5, and GPT-2 models. This pattern suggests that the impact of fine-tuning depends on factors such as model architecture, pretraining mixture, and alignment strategy, rather than on the dataset alone. Recent work supports this interpretation, showing that instruction tuning or supervised fine-tuning can occasionally reduce robustness and impair reasoning performance outside the fine-tuning distribution (Zhang et al., 2026; Zhou et al., 2023). Furthermore, alignment-oriented post-training procedures, such as supervised fine-tuning (SFT) and reinforcement learning from human feedback (RLHF), which are used in Falcon and Qwen training pipelines, introduce trade-offs between helpfulness, safety, and raw task accuracy. Ouyang et al. (2022) also show that alignment can shift output distributions in ways that improve agreement

**Table 5**
Performance of fine-tuning across different LLMs, showing average inference/training times and F1-score improvement.

| Model | Avg. Inference Time (m:s) | Avg. Training Time (m:s) | Improved F1-score |
|---|---|---|---|
| Zephyr-7B | 01:07 | 33:31 | +0.680 |
| FinGPT-v3.3-13B | 01:03 | 27:32 | +0.530 |
| Neural-chat-7b | 01:07 | 33:48 | +0.464 |
| Phi-4-14B | 01:14 | 52:02 | +0.415 |
| Mistral-7B | 00:37 | 33:44 | +0.355 |
| Llama-3.1-8B | 00:40 | 28:47 | +0.296 |
| Qwen1.5-14B | 01:21 | 49:33 | +0.279 |
| Qwen-Distill-DeepSeek-R1-7B | 00:38 | 10:39 | +0.077 |
| Nous-Hermes-2-7B | 00:40 | 19:12 | +0.044 |
| SOLAR-10.7B | 00:44 | 27:32 | +0.031 |
| Llama-Distill-DeepSeek-R1-8B | 00:40 | 11:23 | +0.031 |
| OLMo-7B | 00:53 | 27:07 | +0.016 |
| Llama-2-13b | 01:07 | 26:47 | +0.008 |
| Gemma-7b | 00:26 | 22:41 | −0.001 |
| QwQ-32B | 02:33 | 63:18 | −0.002 |
| Starling-LM-7B | 00:35 | 19:10 | −0.004 |
| Qwen2.5-7B | 00:37 | 15:33 | −0.038 |
| GPT-2-XL-1.6B | 00:15 | 06:36 | −0.047 |
| Falcon3-10B | 00:51 | 23:03 | −0.056 |
| Falcon3-7B | 00:26 | 16:11 | −0.063 |

with human preferences while occasionally degrading benchmark performance. Similar non-monotonic behavioural shifts following post-training adaptation are reported by OpenAI et al. (2024) and Touvron et al. (2023).

In summary, this analysis indicates that the benefits of fine-tuning vary substantially across models, and extended post-training does not necessarily translate into improved performance.

*Statistical significance.* We used McNemar's test (Rainio et al., 2024) to determine the statistical significance of performance differences between zero-shot (ZS) and fine-tuned (FT) models on the same test set. This non-parametric test is designed for paired nominal data and is particularly suited for evaluating classifiers on the same test set and under the assumption of an existing better-than-random reference classifier (Riezler & Hagmann, 2024).

The test indicates that fine-tuning produced highly significant improvements ($p < 0.001$) for the three models that achieved the largest gains in F1 (Zephyr-7B, FinGPT-v3.3-13B, and Neural-chat-7b). In contrast, among the three models with the smallest improvements (GPT-2-XL-1.6B, Falcon3-10B, and Falcon3-7B), only Falcon3-7B exhibited a statistically significant change ($p = 0.02075$). These findings suggest that fine-tuning consistently benefits models with stronger baseline performance, while its impact on weaker architectures is limited and often not statistically significant. The complete results of the statistical tests are reported in Appendix A.3.

Furthermore, comparing the best fine-tuned model (Mistral-7B) with the worst (Gemma-7B), McNemar's test ($\chi^2 = 888.338$, $p < 0.00001$) confirms that the performance difference between the two models is statistically significant, reinforcing the conclusions drawn from the standard evaluation metrics.

*Overall and per-class metrics.* Table 6 presents the comparison of accuracy, precision, recall, and F1-score, reported in both macro-averaged and weighted forms, before and after fine-tuning. As in the previous analysis, we focus on the six models that exhibit the highest and lowest performance gains during fine-tuning.

The results indicate that the three models benefiting most from fine-tuning achieved consistent improvements across all reported metrics, with both precision and recall increasing simultaneously. In contrast, the models that exhibited performance degradation behaved more erratically, with some showing greater losses in recall and others in precision. For instance, the two Falcon models demonstrated a marked decline in recall after fine-tuning.

Table 7 reports the precision, recall, and F1 scores for the three classification categories: buy, sell, and other. We observe several noteworthy patterns. For the models that achieve the largest F1 improvements after fine-tuning, recall consistently increases for the *buy* category and precision improves for the *sell* category, although this gain is partially counterbalanced by a decrease in recall for *sell*. In contrast, the *other* category shows consistent improvements across all metrics.

For the models that exhibited a performance decline after fine-tuning, the behaviour differs considerably. They consistently show a reduction in the precision of the *buy* class, which is only partly compensated by an increase in recall. This indicates that the model becomes more prone to predicting the *buy* category even when it is not appropriate. The *other* class is most strongly affected in terms of recall.

Across all models, the *buy* category benefits the most from fine-tuning, especially in recall, while *sell* remains the most challenging class, with low precision and recall even for strong models.

The comparison between the best fine-tuned model (Mistral-7B) and the weakest one (Gemma-7B) highlights how two models that are comparable in size and overall performance across all tasks can exhibit a substantial difference in this specific context.

**Table 6**
Performance comparison of LLMs in ZS and after FT. Metrics include accuracy, precision (PR), recall (RC), F1-score (F1), in both macro-averaged and weighted versions.

| Model | Accuracy | | Macro PR | | Macro RC | | Macro F1 | | Weighted-PR | | Weighted-RC | | Weighted-F1 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ZS | FT | ZS | FT | ZS | FT | ZS | FT | ZS | FT | ZS | FT | ZS | FT |
| *Models with the largest F1 improvement after fine-tuning.* | | | | | | | | | | | | | | |
| Zephyr-7B | 0.227 | **0.840** | 0.304 | **0.823** | 0.401 | **0.819** | 0.212 | **0.821** | 0.368 | **0.840** | 0.227 | **0.840** | 0.160 | **0.840** |
| FinGPT-v3.3-13B | 0.096 | **0.566** | 0.365 | **0.502** | 0.334 | **0.501** | 0.059 | **0.467** | 0.549 | **0.594** | 0.096 | **0.566** | 0.018 | **0.548** |
| Neural-chat-7b | 0.349 | **0.848** | 0.486 | **0.824** | 0.491 | **0.799** | 0.348 | **0.811** | 0.643 | **0.847** | 0.349 | **0.848** | 0.383 | **0.847** |
| *Models with the lowest F1 improvement after fine-tuning.* | | | | | | | | | | | | | | |
| GPT-2-XL-1.6B | 0.525 | **0.539** | **0.414** | 0.388 | **0.374** | 0.346 | **0.352** | 0.286 | **0.519** | 0.485 | 0.525 | **0.539** | **0.472** | 0.425 |
| Falcon3-10B | **0.589** | 0.575 | 0.479 | **0.518** | **0.421** | 0.389 | **0.416** | 0.360 | 0.560 | **0.571** | **0.589** | 0.575 | **0.548** | 0.493 |
| Falcon3-7B | **0.582** | 0.564 | **0.479** | 0.472 | **0.422** | 0.371 | **0.417** | 0.331 | **0.565** | 0.544 | **0.582** | 0.564 | **0.538** | 0.475 |

**Table 7**
ZS vs FT performance metrics for the best three and worst three models across the three classes. Best values per row are highlighted in bold.

| Model | buy PR | | buy RC | | buy F1 | | sell PR | | sell RC | | sell F1 | | other PR | | other RC | | other F1 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ZS | FT | ZS | FT | ZS | FT | ZS | FT | ZS | FT | ZS | FT | ZS | FT | ZS | FT | ZS | FT |
| *Models with the largest F1 improvement after fine-tuning.* | | | | | | | | | | | | | | | | | | |
| Zephyr-7B | 0.479 | **0.884** | 0.020 | **0.873** | 0.039 | **0.878** | 0.178 | **0.797** | **0.802** | 0.777 | 0.292 | **0.787** | 0.254 | **0.788** | 0.382 | **0.808** | 0.305 | **0.798** |
| FinGPT-v3.3-13B | **1.000** | 0.634 | 0.002 | **0.781** | 0.004 | **0.700** | 0.095 | **0.247** | **1.000** | 0.446 | 0.174 | **0.318** | 0.000 | **0.625** | 0.000 | **0.278** | 0.000 | **0.385** |
| Neural-chat-7b | 0.857 | **0.880** | 0.262 | **0.897** | 0.401 | **0.888** | 0.147 | **0.775** | **0.866** | 0.683 | 0.252 | **0.726** | 0.455 | **0.817** | 0.344 | **0.818** | 0.392 | **0.817** |
| *Models with the lowest F1 improvement after fine-tuning.* | | | | | | | | | | | | | | | | | | |
| GPT-2-XL-1.6B | **0.565** | 0.548 | 0.834 | **0.943** | 0.673 | **0.694** | 0.124 | **0.133** | **0.114** | 0.020 | **0.119** | 0.035 | **0.553** | 0.483 | **0.174** | 0.075 | **0.265** | 0.130 |
| Falcon3-10B | **0.605** | 0.572 | 0.847 | **0.936** | 0.706 | **0.710** | 0.259 | **0.355** | **0.074** | 0.055 | **0.115** | 0.094 | 0.573 | **0.627** | **0.341** | 0.176 | **0.428** | 0.275 |
| Falcon3-7B | **0.599** | 0.564 | 0.867 | **0.933** | **0.709** | 0.703 | 0.240 | **0.267** | **0.119** | 0.020 | **0.159** | 0.037 | **0.599** | 0.585 | **0.282** | 0.160 | **0.383** | 0.251 |

**Table 8**
Comparison between the best and worst fine-tuned models in terms of classification performance.

| Model | Accuracy | Macro Avg | | | Weighted Avg | | | Buy | | | Sell | | | Other | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Precision | Recall | F1 | Precision | Recall | F1 | Precision | Recall | F1 | Precision | Recall | F1 | Precision | Recall | F1 |
| Gemma-7B | 0.366 | 0.459 | 0.427 | 0.323 | 0.570 | 0.366 | 0.386 | 0.480 | 0.560 | 0.520 | 0.190 | 0.250 | 0.220 | 0.210 | 0.270 | 0.240 |
| Mistral-7B | 0.860 | 0.842 | 0.829 | 0.835 | 0.861 | 0.860 | 0.860 | 0.880 | 0.920 | 0.900 | 0.770 | 0.800 | 0.780 | 0.840 | 0.820 | 0.830 |

Interestingly, Mistral-7B shows unbalanced performance across tasks, achieving an F1 score of 90.0% for the *buy* category, while the *sell* category performs comparatively worse, with an F1 of 78.0%. This discrepancy may be due to an imbalance in the training data or to the tendency of users to more frequently discuss purchases than sales in online discourse. However, this difference warrants further investigation.

### *4.3. RQ3 - influence of model characteristics on fine-tuning gains*

To examine how the effectiveness of fine-tuning varies with the characteristics of the underlying model, Fig. 2 reports the distribution of model performance, measured by the F1 score, across three dimensions: model family, training methodology, and release year.

When comparing the different model families, the Mistral models achieve the best overall performance, reaching a maximum F1 score of 86.0% and consistently outperforming the other groups. In contrast, the Llama models exhibit lower accuracy; only one model reaches good performance. The Qwen and Falcon models perform less effectively for the considered task, yielding only limited results.

If we consider the training methodology, instruction-tuning clearly emerges as the most effective approach. By contrast, base models obtain only moderate results, while distilled models perform the worst, exhibiting both the lowest minimum and maximum scores. These results indicate that, while distillation reduces model complexity, it simultaneously limits the model's capacity to generalise in this task.

Finally, when considering the release year, models from 2024 and earlier achieve broadly similar results, with a slight advantage for the most recent ones. In contrast, models released in 2025 perform less well. However, this outcome may simply depend on the small sample size of 2025.

Overall, the figure highlights two clear trends: the superiority of the Mistral family and the central role of instruction-tuning in boosting performance.

Fig. 3 illustrates the relationship between F1 scores, inference time, model family, and training methodology. Among the five models with the highest F1-to-inference-time ratio, three belong to the Mistral family. Together with Llama-3.1-8B, these four models were also fine-tuned using instruction-based techniques. The plot further indicates that compact fine-tuned models (7–8B parameters) with relatively short inference times can match, and in some cases surpass, the performance of substantially larger

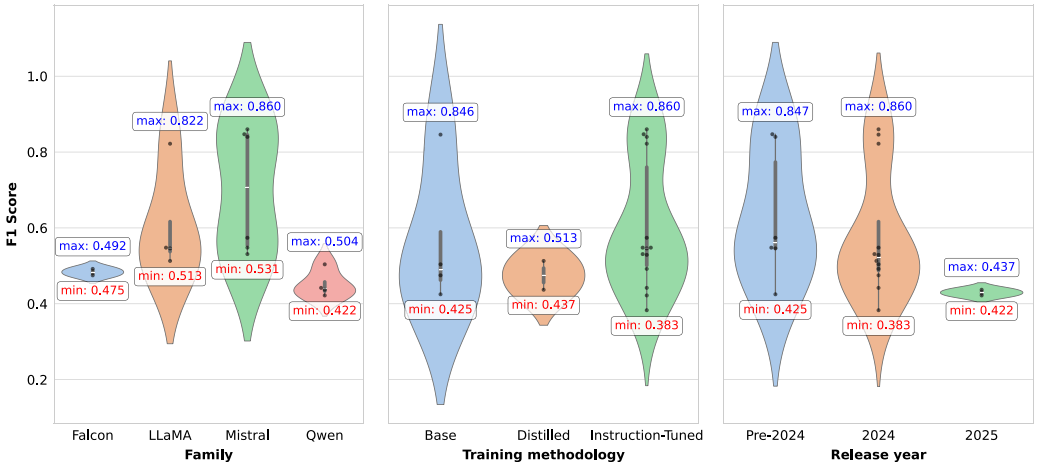


**Fig. 2.** F1-score distribution of fine-tuned LLMs by model family, training methodology, and release year (only groups with at least two observations are included).

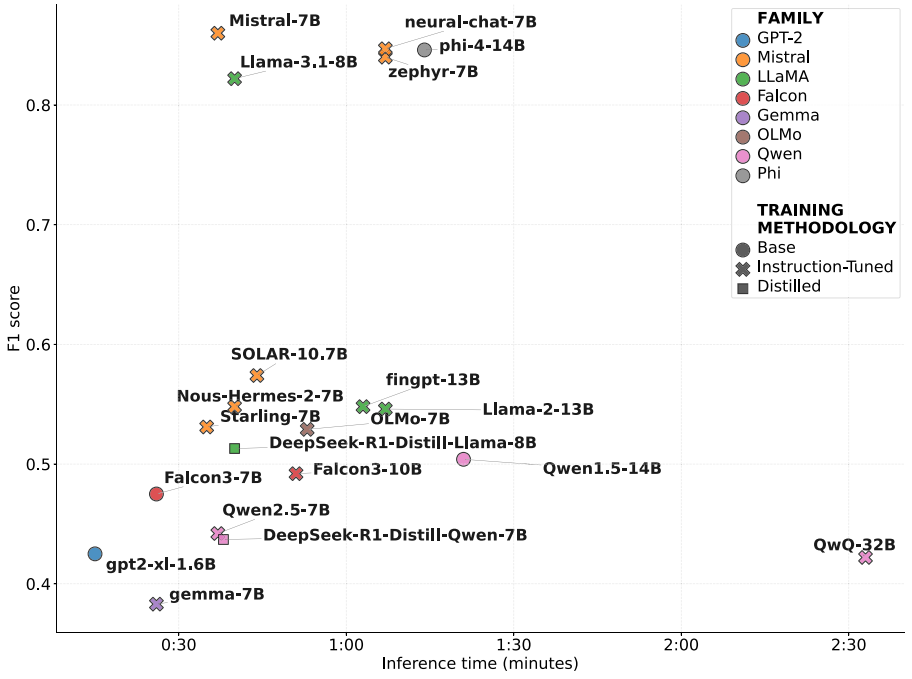


**Fig. 3.** F1 scores of the fine-tuned LLMs as a function of their inference time. Marker colour indicates the model family, marker shape represents the training methodology.

models. In particular, Mistral-7B and Llama-3.1-8B distinguish themselves by combining strong predictive accuracy with highly efficient inference. These results suggest that such models are well-suited for powering classification frameworks that are both robust and scalable, supporting the effective processing of the large volume of market-related posts generated daily.

### *4.4. RQ 4 - error types and limitations*

In this section, we present a comprehensive error analysis that includes a detailed breakdown of error types, such as classical Type I/II errors, domain-specific misclassifications (e.g., confusing a buy with a sell), and a qualitative examination of their underlying causes (e.g., sarcasm or linguistic ambiguity). We then focus on the best-performing model (Mistral-7B) and analyse its capacity to produce coherent explanations for its own errors.

In line with the procedure adopted for RQ2, we first evaluate the effect of fine-tuning by comparing the three models that achieved the largest F1 improvements after fine-tuning (Zephyr-7B, FinGPT-v3.3-13B, and Neural-chat-7b) with the three models

**Table 9**
Number of Type I errors (false positives) for each class in both configurations.

| Model | Buy | | Sell | | Other | |
|---|---|---|---|---|---|---|
| | FP(ZS) | FP(FT) | FP(ZS) | FP(FT) | FP(ZS) | FP(FT) |
| *Models with the largest F1 improvement after fine-tuning.* | | | | | | |
| Zephyr-7B | 25 | 132 | 746 | 40 | 871 | 168 |
| FinGPT-v3.3-13B | 0 | 518 | 1918 | 275 | 1 | 129 |
| Neural-chat-7b | 50 | 141 | 1013 | 40 | 319 | 142 |
| *Models with the lowest F1 improvement after fine-tuning.* | | | | | | |
| GPT-2-XL-1.6B | 737 | 891 | 163 | 26 | 109 | 62 |
| Falcon3-10B | 633 | 802 | 43 | 20 | 197 | 81 |
| Falcon3-7B | 665 | 826 | 76 | 11 | 146 | 88 |

**Table 10**
Number of Type II errors (false negatives) for each class in both configurations.

| Model | Buy | | Sell | | Other | |
|---|---|---|---|---|---|---|
| | FN(ZS) | FN(FT) | FN(ZS) | FN(FT) | FN(ZS) | FN(FT) |
| *Models with the largest F1 improvement after fine-tuning.* | | | | | | |
| Zephyr-7B | 1124 | 146 | 40 | 45 | 478 | 149 |
| FinGPT-v3.3-13B | 1145 | 251 | 0 | 112 | 774 | 559 |
| Neural-chat-7b | 847 | 118 | 27 | 64 | 508 | 141 |
| *Models with the lowest F1 improvement after fine-tuning.* | | | | | | |
| GPT-2-XL-1.6B | 191 | 65 | 179 | 198 | 639 | 716 |
| Falcon3-10B | 176 | 74 | 187 | 191 | 510 | 638 |
| Falcon3-7B | 153 | 77 | 11 | 198 | 88 | 650 |

**Table 11**
Type I and Type II errors for fine-tuned Gemma-7B and Mistral-7B.

| Model | Type I (FP) | | | Type II (FN) | | |
|---|---|---|---|---|---|---|
| | Buy | Sell | Other | Buy | Sell | Other |
| Gemma-7B | 379 | 910 | 56 | 616 | 66 | 663 |
| Mistral-7B | 114 | 36 | 146 | 121 | 49 | 126 |

that exhibited the smallest, and in some cases negative, F1 gains (GPT-2-XL-1.6B, Falcon3-10B, and Falcon3-7B). When compatible with the analysis, we also contrast the best-performing model overall (Mistral-7B) with the worst-performing one (Gemma-7B).

*False positives and false negatives.* In classification tasks, the analysis of Type I (false positives) and Type II (false negatives) errors offers a more nuanced understanding than overall accuracy alone.

Table 9 reports the number of Type I errors (false positives) produced by the models under evaluation. For the models that showed performance gains after fine-tuning, the process significantly reduced false positives in the *sell* and *other* classes, at the cost of a limited increase in false positives for the *buy* class. For instance, Zephyr-7B decreases its false positives from 746 to 40 for the *sell* class and from 871 to 168 for the *other* class, highlighting a strong corrective effect of fine-tuning.

The models that experienced a performance decline after fine-tuning displayed a consistent pattern. Specifically, false positives increased slightly for the *buy* classes, while they decreased marginally for the *sell* and *other* classes.

Table 10 reports the number of Type II errors (false negatives) produced by the evaluated models. The models that exhibited performance gains after fine-tuning show a substantial reduction in false negatives for the *buy* class and a moderate reduction for the *other* class. In contrast, the *sell* class experienced an increase in false negatives, although this effect was less pronounced for the best-performing models. The models that do not benefit from fine-tuning still exhibit a reduction in false negatives for the *buy* class, but they suffer from an increased number of such errors in the *sell* and *other* classes.

Table 11 reports the distribution of Type I and Type II errors for the best and the worst fine-tuned models. The error patterns differ substantially between the two models. For Gemma-7B, false positives are dominated by the *sell* class (910), indicating a strong bias towards predicting *sell* even when incorrect. Moreover, false negatives are particularly high for both *buy* (616) and *other* (663), suggesting that the model often fails to detect these classes correctly. In contrast, Mistral-7B presents a more balanced error profile. Both false positives and false negatives remain relatively low across all classes.

*Other types of errors.* In addition to standard Type I and Type II errors, it is important to examine specific semantic misclassifications that are particularly relevant in the financial domain. We identify three types of errors that are especially critical. The first is **action inversions**, which occur when the model predicts *sell* instead of *buy* (*buy* → *sell*) or *buy* instead of *sell* (*sell* → *buy*). These errors reverse the intended trading signals and can therefore be the most severe when the predictions are used for market forecasting or similar applications. The second is **non-action misclassifications** (*buy/sell* → *other*), where actual trading signals are classified as

**Table 12**
Count of action misclassifications across models.

| Model | Action inversions | | | | Non-action misclassifications | | | | False action misclassifications | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | buy → sell | | sell → buy | | buy → other | | sell → other | | other → buy | | other → sell | |
| | ZS | FT | ZS | FT | ZS | FT | ZS | FT | ZS | FT | ZS | FT |
| *Models with the largest F1 improvement after fine-tuning.* | | | | | | | | | | | | |
| Zephyr-7B | 277 | 14 | 16 | 9 | 847 | 132 | 24 | 36 | 9 | 123 | 469 | 26 |
| FinGPT-v3.3-13B | 1144 | 154 | 0 | 80 | 1 | 97 | 0 | 32 | 0 | 438 | 774 | 121 |
| Neural-chat-7b | 554 | 14 | 1 | 26 | 293 | 104 | 26 | 38 | 49 | 115 | 459 | 26 |
| *Models with the lowest F1 improvement after fine-tuning.* | | | | | | | | | | | | |
| GPT-2-XL-1.6B | 96 | 13 | 165 | 188 | 95 | 52 | 14 | 10 | 572 | 703 | 67 | 13 |
| Falcon3-10B | 24 | 11 | 142 | 173 | 152 | 63 | 45 | 18 | 491 | 629 | 19 | 9 |
| Falcon3-7B | 40 | 10 | 145 | 177 | 113 | 67 | 33 | 21 | 520 | 649 | 36 | 1 |

**Table 13**
Semantic error analysis for fine-tuned Gemma-7B and Mistral-7B.

| Error category | Error type | Gemma-7B | Mistral-7B |
|---|---|---|---|
| Action Inversions | buy → sell | 579 | 16 |
| | sell → buy | 52 | 8 |
| Non-action Misclassifications | buy → other | 42 | 105 |
| | sell → other | 14 | 41 |
| False Action Misclassifications | other → buy | 327 | 106 |
| | other → sell | 336 | 20 |

non-actionable. Such errors lead to missed opportunities and can be particularly problematic when they cause weaker signals from social media or other sources to be overlooked. The third is **false action misclassifications** (*other → buy/sell*), where non-actionable content is incorrectly classified as a trading signal. These errors introduce spurious signals and may provoke misleading indications that confound subsequent applications relying on the predictions.

Table 12 presents the error types for the six models that exhibit the largest and smallest F1 improvements after fine-tuning.

Overall, the three models that benefit the most from fine-tuning show a substantial reduction in action inversion errors, which is highly positive given that these represent the most severe type of mistake. In addition, non-action misclassifications also tend to decrease, indicating that the fine-tuned models are more sensitive to distinguishing between buy and sell decisions rather than defaulting to other when uncertain. Nevertheless, the behaviour of false action misclassifications remains more nuanced. While the frequency of *other → sell* errors decreases after fine-tuning, *other → buy* errors increase, suggesting that the model can sometimes become overly confident in predicting *buy*.

Table 13 summarises the distribution of semantic errors for the best and worst fine-tuned models. Gemma-7B is dominated by inversion errors, with 579 cases of misclassifying *buy* as *sell*. Conversely, Mistral-7B rarely exhibits inversion errors (16 and 8, respectively). However, it demonstrates a certain vulnerability to non-action misclassifications, with 105 instances of *buy → other* and 41 instances of *sell → other*, suggesting a tendency to underestimate actionable decisions.

*Corrections and new errors introduced by fine-tuning.* It is crucial to analyse not only the errors corrected relative to the zero-shot setting but also the new errors introduced.

Figs. 4 and 5 show the distribution of semantic errors that were corrected (blue) and those that were introduced (orange) after fine-tuning. The results indicate that the types of errors corrected or introduced are strongly influenced by the characteristics of each model. A common pattern is the correction of *sell → other* errors in the model that benefited the most from the fine-tuning process. This is likely due to fine-tuning enabling the model to better handle the *sell* category, which is the rarest in our dataset. Another interesting pattern is the emergence of new *other → buy* errors in the model that did not benefit from fine-tuning. This behaviour appears to stem from the model's tendency to conflate a general, potentially positive sentiment towards a stock with an explicit trading signal.

*Qualitative analysis of underlying error causes.* In order to better understand the limitations of LLMs in financial text classification, we performed a qualitative analysis of their errors. Specifically, we examined misclassified posts produced by different models, which allowed us to identify recurring patterns and common linguistic challenges. Based on this analysis, we developed a simple taxonomy of errors grounded in manual inspection by the authors. To support the systematic assignment of misclassified instances to these categories, we used OpenAI's GPT-5 as an assistive tool to automatically label each error according to the proposed taxonomy. All GPT-5 assignments were subsequently reviewed by domain experts, who validated the classifications and corrected them when necessary, particularly in ambiguous or borderline cases. During this validation step, overlapping or weakly defined categories were also merged. This human-in-the-loop, iterative process resulted in a consolidated representation of the most frequently observed error types, as summarised below.

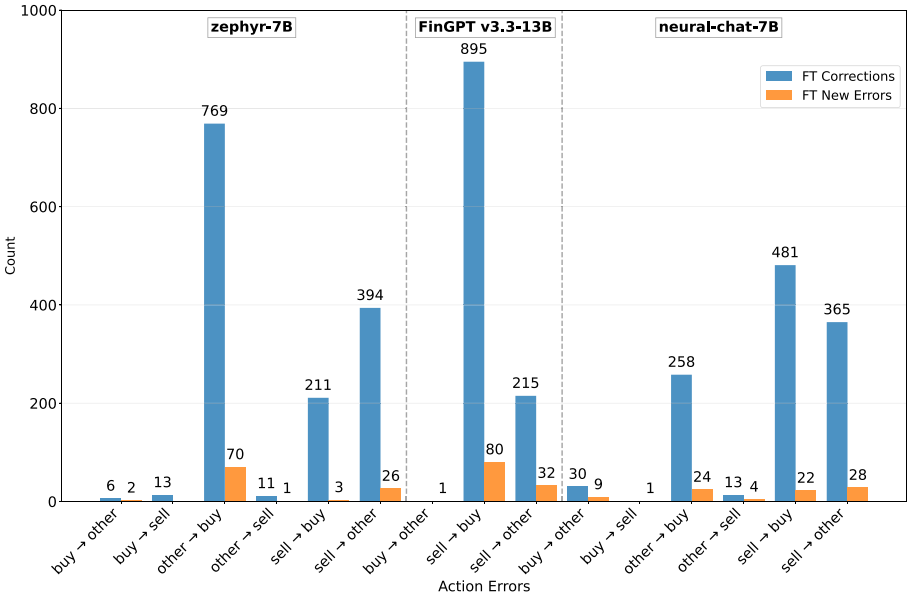


**Fig. 4.** Errors corrected and new errors introduced by fine-tuning in the models that achieved the largest improvement in F1 score after fine-tuning.

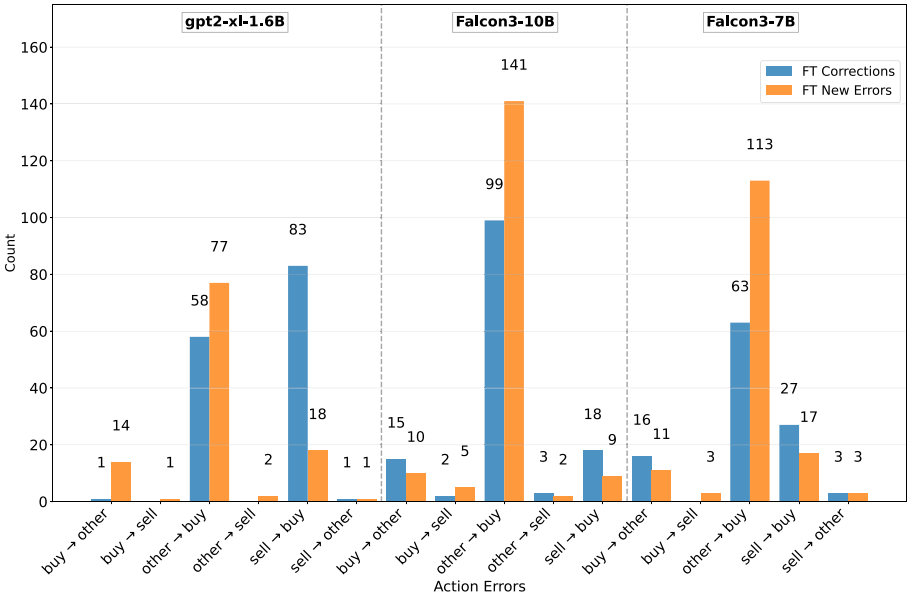


**Fig. 5.** Corrections and New Errors Introduced by Fine-Tuning in the models with the lowest F1 improvement after fine-tuning.

- **Category 1 - Ambiguity and Implicit Language:** Errors arising from vague, indirect, or context-dependent statements, where the intended action is not explicitly stated.
- **Category 2 - Generic Opinions vs. Actual Actions:** Misclassifications between texts expressing general opinions or sentiments and those describing concrete trading actions.
- **Category 3 - Noise, Memes, Irony, and Sarcasm:** Errors due to informal or humorous language, often including memes, exaggerations, or irony that obscure the underlying meaning.
- **Category 4 - Informational or News-Like Content:** Cases where factual reporting of news or events is mistaken for an actionable trading signal.

**Table 14**
Error distribution across categories for different LLMs.

| Model | Ambiguity and Implicit language | Generic opinions vs. Actual actions | Informational or News-like content | Noise, Memes, Irony, and Sarcasm |
|---|---|---|---|---|
| **Models with the largest improvement after FT:** | | | | |
| Zephyr-7B | 271 | 7 | 20 | 42 |
| FinGPT-v3.3-13B | 774 | 19 | 39 | 88 |
| Neural-chat-7b | 266 | 6 | 18 | 33 |
| **Models with the lowest F1 improvement after fine-tuning:** | | | | |
| GPT-2-XL-1.6B | 823 | 24 | 42 | 88 |
| Falcon3-10B | 758 | 23 | 43 | 76 |
| Falcon3-7B | 786 | 23 | 41 | 72 |
| **Overall Worst:** Gemma-7B | 1017 | 24 | 41 | 260 |
| **Overall Best:** Mistral-7B | 228 | 7 | 21 | 41 |

Each category captures a distinct linguistic or semantic challenge. For instance, ambiguous or ironic posts are often difficult to map to a clear trading intent, while news-like content may lead models to predict *buy* or *sell* even when no explicit recommendation is expressed (see Table 14).

*Ambiguity and Implicit Language* is by far the most prevalent source of errors across all models, confirming that vague and context-dependent financial texts remain particularly challenging to interpret correctly. To address these cases, it may be beneficial to incorporate additional contextual information, for instance from surrounding posts or user profiles. *Noise, Memes, Irony, and Sarcasm* represent the second major source of errors and remain difficult to handle even for the strongest models. Notably, the particularly high number of such errors in Gemma-7B indicates that this is one of the primary points of failure for models that are weaker on this task. *Informational or News-Like Content* also remains problematic, although the best-performing model shows a clear improvement in handling these cases. Finally, errors related to *Generic Opinions vs. Actual Actions* are relatively rare, especially in the best model. This suggests that fine-tuned models are able to discriminate effectively between generic statements and personal intentions or suggestions regarding trade actions.

*Analysis of the model explanations.* We further evaluated the explanatory capabilities of the best-performing model, the fine-tuned *Mistral-7B*, by instructing it to generate natural language justifications for its predictions. This analysis helped differentiate correct predictions supported by coherent reasoning from those that may be coincidentally accurate. Through examination of multiple prediction errors across the dataset, we identified several recurring failure patterns that reveal systematic limitations in the model's understanding of financial discourse.

The error analysis identifies three persistent challenges: (i) *Over-sensitivity to optimism*, where general positive sentiment is incorrectly interpreted as a specific *buy* recommendation; (ii) *Sequential action confusion*, in which attention to initial verbs overlooks final trading intentions (e.g., "sold X for buying Y"), leading to label inversion; (iii) *Informality and irony*, where the model fails to detect the underlying intent in posts containing slang or humorous expressions.

These patterns emerged consistently across various error cases and underscore the model's difficulty in interpreting community-specific communication patterns, understanding the strategic intent embedded in transactional language, and accounting for emotional cues that influence directive meaning.

To illustrate these challenges concretely, Appendix A.4 presents six examples of challenging positive predictions and six representative error cases with model-generated explanations. In the positive examples, the model demonstrates a nuanced understanding of financial sentiment by accurately identifying urgent selling pressure expressed through bearish terminology (examples 1 to 3), distinguishing non-actionable financial commentary (example 4), and correctly interpreting informal trading jargon related to buying strategies (examples 5 and 6).

The negative cases highlight the most significant and recurring failure modes identified in our analysis. For instance, example 1 shows a typical case of oversensitivity to optimistic language, example 2 illustrates confusion in interpreting sequential actions, and example 3 demonstrates the mishandling of sarcasm. We also observe less frequent and more subtle errors, such as negative example 4, which is misclassified as sell because the user criticises market dynamics using domain terminology ("covered", "squeeze"), even though the statement does not directly express a personal trading action.

Overall, the findings highlight the need for future work to strengthen contextual reasoning and improve the model's handling of pragmatic aspects in financial discourse, particularly for social media content where informal language and implicit intentions are prevalent.

## 5. Discussion

In the following sections, we discuss the main results obtained from the experiments, explain how they address the research questions, and outline their theoretical and practical implications.

*RQ1 - How effective are modern LLMs at performing STAD when compared with traditional models, and which LLM achieves the best performance?*

The experiments demonstrate that LLMs can outperform traditional models in STAD, but their effectiveness critically depends on fine-tuning. Simple baselines such as LSTM and Bi-LSTM exhibit limited performance, whereas an MLP combined with high-quality embeddings provides a strong non-LLM baseline (69.6% F1) that surpasses all zero-shot LLMs. Similarly, BERT base (73.7% F1) outperforms all zero-shot LLMs. Among the latter, Gemma-3-27B achieves the best zero-shot performance (68.2% F1). By contrast, fine-tuned LLMs, led by Mistral-7B with an F1 of 86.0%, achieve substantial gains, clearly outperforming both traditional models and zero-shot LLMs.

The weak performance of FinGPT, a language model specifically optimised for financial sentiment analysis, supports one of the main hypotheses discussed in the introduction: sentiment analysis methods, even when tailored to the financial domain, are not well-suited to address STAD. Although sentiment analysis is often used as a proxy for this task (Gentzkow et al., 2019; Loughran & McDonald, 2011; Wankhade et al., 2022), our results suggest that it fails to capture the underlying structure of action-oriented discourse. This indicates the need to focus more on detecting the actions implied or suggested by the text rather than on assessing its general sentiment. This paper represents an important first step in this direction, introducing a high-quality dataset and providing a detailed analysis of the behaviour of LLMs in this context.

Furthermore, these findings have several additional implications for the effectiveness of current NLP methods on STAD: (1) zero-shot LLMs are not inherently superior to well-designed traditional pipelines; (2) fine-tuning is essential for achieving strong performance with LLMs; and (3) parameter-efficient approaches, such as LoRA, make such adaptation practically feasible.

*RQ2 - Does fine-tuning LLMs on high-quality data significantly enhance their performance on this task?*

The experiments provide clear evidence that fine-tuning on high-quality data can significantly enhance performance, but with strong variation across model families. On average, fine-tuning yielded an F1 increase of over 15%, with several modern, instruction-tuned 7–8B models (e.g., Mistral-7B, Zephyr-7B, Neural-chat-7b) achieving improvements above 25% and reaching competitive accuracy levels with relatively efficient inference times. Statistical testing confirmed that these gains were highly significant. Conversely, a few models showed marginal or even negative effects, indicating that fine-tuning is not uniformly beneficial and may exacerbate weaknesses in certain architectures.

These findings have important practical implications, as many companies in this domain often specialise existing models with the expectation that fine-tuning will enhance performance relative to the base version. Our results in this field are instead consistent with a few prior studies showing that fine-tuning does not always lead to substantial performance improvements (Barnett et al., 2024a; Macháček et al., 2025; Pu et al., 2023). Overall, the evidence suggests that fine-tuning can be a powerful approach, particularly when applied to instruction-tuned models; however, its effectiveness strongly depends on the underlying model architecture.

*RQ3 - To what extent does the effectiveness of fine-tuning depend on the characteristics of the underlying model?*

The experiments show that the effectiveness of fine-tuning is strongly shaped by the characteristics of the underlying model. Model family plays a central role, with Mistral models consistently outperforming others and demonstrating the best balance between accuracy and efficiency. Training methodology is equally critical: instruction-tuned models yield markedly higher gains than base or distilled variants. Finally, efficiency analyses reveal that smaller models such as Mistral-7B and Llama-3.1-8B can rival or exceed much larger counterparts.

Overall, these findings offer a comprehensive overview of the current state of the art for this challenging and relatively unexplored task, while also identifying an effective baseline configuration that can be directly adopted by organisations aiming to detect user actions in this domain. The recommended setup is based on open-weight, instruction-tuned language models of moderate size (7–8B parameters), drawn from the most suitable model families (e.g., Mistral, Llama). When fine-tuned on our domain-specific datasets, these models demonstrate strong and consistent performance, achieving a favourable balance between accuracy and computational efficiency. Furthermore, their open-weight nature facilitates transparency, reproducibility, and ease of integration into existing industrial pipelines.

*RQ4 - What types of errors occur most frequently, and how does fine-tuning influence their distribution, severity, and nature?*

The error analysis shows that, prior to fine-tuning, models are dominated by ambiguity and implicit-language failures, with irony and sarcasm representing the second major source of errors. In the zero-shot setting, weaker models also tend to over-predict *buy*, generating both false actions (*other* → *buy/sell*) and harmful action inversions.

Fine-tuning recalibrates the stronger models by substantially reducing false positives on sell/other, lowering false negatives on buy, and decreasing both inversion and false-action errors. The remaining mistakes are mostly shifted towards non-action misclassifications (*buy/sell* → *other*), which are safer because they avoid trades rather than triggering incorrect ones. By contrast, models that do not benefit from fine-tuning exhibit an increase in buy-biased false positives and introduce *other* → *buy* errors, indicating an over-correction rather than an improved understanding.

The qualitative analysis confirms that ambiguity and implicit language remain the leading source of misclassification, underscoring the persistent difficulty of interpreting vague or context-dependent financial texts. Noise, memes, irony, and sarcasm form the second main error category, particularly affecting weaker models such as Gemma-7B. Although informational or news-like content remains challenging, stronger models achieve significant improvements, and errors involving generic opinions versus explicit actions are comparatively rare.

The main implication of these findings is the need for further research into the ability of LLMs to navigate and understand complex online language. Future work should also investigate more sophisticated processing pipelines that combine LLMs with additional contextual information extracted from relevant online communities. Such integration, which we plan to explore in future work, could enhance the models' capacity to interpret nuanced discourse patterns, social dynamics, and domain-specific conventions that characterise these environments.

Finally, the analysis of Mistral-7B explanations indicates that requiring models to generate justifications can provide valuable insights into their reasoning in this domain. Specifically, we observed cases where the model misclassified optimism as buy signals, confused sequential trading actions, or failed with irony and slang. These errors reflect an occasional reliance on surface cues rather than deeper pragmatic reasoning. Such qualitative analyses can also serve as a practical tool for stakeholders to quickly assess a model's understanding of complex and noisy online environments.

Overall, our experiments reveal persistent weaknesses in the current generation of language models when dealing with text types that demand extensive contextual comprehension and discourse-level interpretation. In particular, the ability to analyse a conversation within the broader community context, and to understand how collective dynamics, memes, and irony shape online discourse, remains limited. The analysis presented in this paper represents an initial step in this direction. Future work should further investigate these aspects across multiple domains, exploring methods to enhance contextual grounding, pragmatic reasoning, and the robustness of model explanations in diverse social media environments.

## 6. Conclusion

Since online information exchange has been shown to influence market trends (Zhang ,Zhang, Bao et al., 2024), monitoring the behaviour of retail investors has become increasingly important (Ferraro & Sperlì, 2024; Zhuang et al., 2025). Theoretical models that account for the role of social networks in shaping asset prices require reliable indicators of investors' online intentions towards specific stocks. However, traditional financial sentiment analysis techniques are limited in their ability to capture explicit trading intentions or direct recommendations to buy or sell individual stocks.

This paper explores how LLMs can be employed to address challenges in financial NLP tasks, with particular attention to the role of fine-tuning and to the common limitations and errors that arise. To this end, we introduce a new task, *STAD* (Social Trading Action Detection), which aims to infer users' trading intentions by classifying online posts into three categories: *buy*, *sell*, or *other*. We also release FinReddit-2K, a dataset of 2123 manually annotated Reddit posts. We then conduct a systematic study on the effects of fine-tuning across a range of LLMs. In addition to identifying the best-performing architectures, we analyse how fine-tuning interacts with different model characteristics and the extent to which it mitigates or exacerbates specific types of errors.

Our findings, structured around four research questions, indicate that LLMs provide significant advantages over alternative approaches, particularly when fine-tuned for the target task. The experiments show that modern LLMs can substantially outperform traditional models in STAD, but only when trained on high-quality data. In contrast, zero-shot LLMs underperform compared to a strong MLP baseline, confirming that fine-tuning is essential to fully exploit their potential. The performance gains, however, vary considerably across architectures. Instruction-tuned mid-sized models (e.g., Mistral-7B) achieve the best balance between accuracy and efficiency, while some models may fail to benefit and can even degrade after fine-tuning. The error analysis further reveals that fine-tuning consistently reduces severe mistakes such as action inversions and biased false positives, shifting the error distribution towards safer misclassifications. Nonetheless, ambiguity, irony, and implicit language remain persistent challenges, underscoring the need for more advanced contextual reasoning. Future studies should therefore explore in greater depth how conversations evolve within broader community contexts and how to enhance the understanding of texts shaped by collective dynamics and ironic expression.

Future research will proceed along two main directions. First, we will investigate additional features to improve prediction accuracy, including social signals derived from user interaction graphs, user activity history, and discussion threads. We will also explore multimodal models capable of extracting information from images and videos, which are frequently included in relevant online posts and may provide valuable contextual cues. Second, we will integrate the results of the STAD task into asset price prediction frameworks with the aim of improving the reliability of stock price movement forecasts.

## CRediT authorship contribution statement

**Simone D'Amico:** Writing – review & editing, Writing – original draft, Methodology, Data curation, Conceptualization. **Andrea Maurino:** Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Francesco Osborne:** Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Giancarlo Sperlì:** Writing – review & editing, Writing – original draft, Methodology, Conceptualization.

## Acknowledgments

This work is supported by the Italian Ministry of University and Research (MUR) within the PRIN2022–ISALDI: Interpretable Stock Analysis Leveraging Deep multI-modal models (CUP: E53D23008150006).

## Appendix

In this appendix, we provide additional information on five aspects of the study: (1) dataset composition and descriptive statistics (Appendix A.1); (2) the optimal hyperparameters for the classifiers (Appendix A.2); (3) the results of McNemar's tests comparing ZS and FT models (Appendix A.3); (4) representative examples of model behaviour (Appendix A.4); and (5) the prompt template used for the LLMs (Appendix A.5).

**Table A.15**
Distribution of posts across subreddits in the FinReddit-2Kdataset.

| subreddit | # posts | percentage |
|---|---|---|
| wallstreetbets | 1672 | 78.76% |
| Stocks | 203 | 9.56% |
| investing | 112 | 5.28% |
| pennystocks | 60 | 2.83% |
| StockMarket | 41 | 1.93% |
| Dividends | 31 | 1.46% |
| economics | 2 | 0.09% |
| finance | 2 | 0.09% |

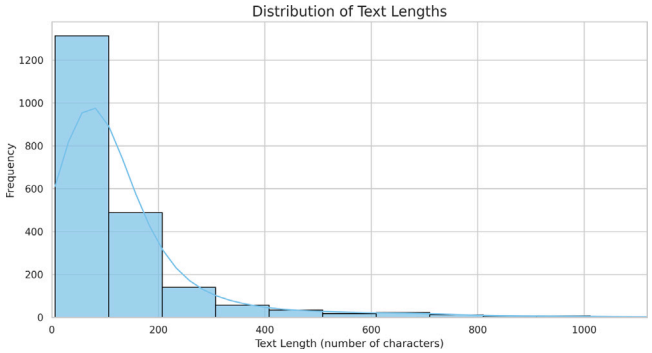


**Fig. A.6.** Distribution of text lengths in the dataset.

**Table A.16**
Best parameter combinations for each neural network classifier and BERT models.

| Model | Epochs | Batch size | Dropout | Input size | Hidden size |
|---|---|---|---|---|---|
| all-MiniLM-L6-v2+MLP | 10 | 16 | 0 | 256 | 128 |
| all-mpnet-base-v2+MLP | 10 | 16 | 0 | 128 | 128 |
| gte-large-en-v1,5+MLP | 10 | 16 | 0 | 64 | 64 |
| all-MiniLM-L6-v2+LSTM | 5 | 32 | 0.1 | 64 | 128 |
| all-mpnet-base-v2+LSTM | 5 | 64 | 0 | 256 | 128 |
| gte-large-en-v1,5+LSTM | 10 | 64 | 0 | 256 | 64 |
| all-MiniLM-L6-v2+Bi-LSTM | 5 | 32 | 0.1 | 128 | 128 |
| all-mpnet-base-v2+Bi-LSTM | 10 | 32 | 0.1 | 256 | 128 |
| gte-large-en-v1,5+Bi-LSTM | 10 | 32 | 0.1 | 256 | 128 |
| BERT-base | 4 | 16 | 0.1 | 256 | 768 |
| BERT-large | 4 | 32 | 0.1 | 256 | 1024 |

### *A.1. Dataset composition and descriptive statistics*

This section presents additional descriptive statistics for the FinReddit-2K dataset. Specifically, Table A.15 reports the distribution of posts across Reddit communities, while Fig. A.6 illustrates the distribution of text lengths.

### *A.2. Optimal hyperparameter configuration for the baseline model*

Table A.16 provides the optimal hyperparameter configurations for neural and transformer-based classifiers.

### *A.3. Statistical comparison between ZS and FT models*

Table A.17 presents the results of McNemar's test comparing zero-shot and fine-tuned LLMs.

**Table A.17**
McNemar's test results comparing Zero-Shot (ZS) vs Fine-Tuned (FT) models on the test set. Statistically significant results ($p < 0.05$) are highlighted in bold.

| Model | Test statistic | $p$-value |
|---|---|---|
| *Best Three Models* | | |
| Zephyr-7B | 1123.905 | **0.00000** |
| FinGPT-v3.3-13B | 811.133 | **0.00000** |
| Neural-chat-7b | 906.368 | **0.00000** |
| *Worst Three Models* | | |
| GPT-2-XL-1.6B | 3.285 | 0.06991 |
| Falcon3-10B | 2.766 | 0.09626 |
| Falcon3-7B | 5.348 | **0.02075** |

### *A.4. Qualitative analysis of the model explanations*

This subsection provides a qualitative analysis of model behaviour through representative prediction examples. We report both correctly classified instances and systematic error cases, together with the model-generated explanations, in order to illustrate how different linguistic phenomena affect classification decisions.

## Positive Prediction with Explanation

**Example 1:** *DUMP IT BEFORE IT CRASHES*
**Real Label:** sell
**Predicted Label:** sell
**Model Explanation:** The financial sentiment expressed in the post is negative, indicating a desire to sell an asset. The use of the word "crashes" further emphasises this sentiment, suggesting a belief that the asset's value is about to decrease significantly. Therefore, the label assigned to this post is "sell."

**Example 2:** *Just get rid of it NOW*
**Real Label:** sell
**Predicted Label:** sell
**Model Explanation:** The post explicitly states "Just get rid of it NOW", which is a clear indication that the author intends to sell their holdings. The urgency expressed in the post further supports this interpretation.

**Example 3:** *JUST SOLD AND MADE SOME TENDIES*
**Real Label:** sell
**Predicted Label:** sell
**Model Explanation:** The post title itself is an indication of a sell transaction. The phrase "JUST SOLD" is a clear signal that the author has recently sold some securities. The term "made tendies" is a reference to tendies, a popular term in the stock market community for gains or profits. The post does not contain any language that would suggest a buy transaction, so the "sell" label is appropriate.

**Example 4:** *Just a reminder of what could have been*
**Real Label:** other
**Predicted Label:** other
**Model Explanation:** The given post does not express any clear financial sentiment. It appears to be a reflective statement about a missed opportunity rather than an expression of optimism or pessimism regarding the financial markets. Therefore, it falls under the "other" category in financial sentiment classification.

**Example 5:** *Lowest dip it will hit today. When is a good time buy a lot I just put into AMC at a pop*
**Real Label:** buy
**Predicted Label:** buy
**Model Explanation:** The financial sentiment expressed in this post is bullish, indicating a positive outlook towards the financial markets or a particular asset. The user is expressing their intention to buy more of a specific stock (AMC) and is optimistic about its future price movement. This sentiment can be classified as "buy" or "bullish."

**Example 6:** *Live footage of me and the guys buying the dips these days*
**Real Label:** buy
**Predicted Label:** buy
**Model Explanation:** The post title mentions "buying the dips" which is a clear indication of a bullish sentiment towards the market. The author is expressing their intention to purchase stocks during a market downturn, indicating a belief that the market will recover and prices will rise in the future. The use of the word "guys" also implies a sense of community and shared belief among investors. Overall, the post expresses a positive outlook towards buying stocks during market dips, making the 'buy' label an appropriate classification.

*These examples illustrate correct predictions across different linguistic phenomena, including imperative forms, community-specific slang, and narrative descriptions of trading actions.*

## Negative Prediction with Explanation

**Example 1:** *were cool guys this whole week was a huge step forward. It will get even better next week because we will hold together*
**Real Label:** other
**Predicted Label:** buy
**Model Explanation:** The financial sentiment expressed in the post is positive and optimistic. The author expresses confidence in the market and their investment decisions, using phrases like "huge step forward" and "it'll get even better next week." This sentiment is consistent with a buy signal, as it indicates a bullish outlook and a willingness to hold onto investments.
**Error Analysis:** The error arises from a confusion between positive sentiment and trading intent. While the post uses optimistic language, it does not express any explicit or implicit investment action. The model overinterprets generic encouragement and group cohesion as a buy signal, illustrating a sentimentintent mismatch.

**Example 2:** *sold all my bag holding weed stocks for another gme*
**Real Label:** buy
**Predicted Label:** sell
**Model Explanation:** The financial sentiment expressed in this post is one of selling, likely due to the author's decision to sell their holdings in weed stocks and invest in GameStop (GME) instead. This sentiment is often associated with a bearish outlook on the market or specific securities, as the author is reducing their exposure to certain assets.
**Error Analysis:** The model focuses on the selling action in the first part of the post while ignoring the subsequent reinvestment into GME. Although assets are sold, the overall intent is to reallocate capital into another stock, which corresponds to a buy decision. This reflects difficulty in handling multi-action statements involving portfolio rebalancing.

**Example 3:** *yes now TRY TO SELL IT LOLOL*
**Real Label:** sell
**Predicted Label:** other
**Model Explanation:** The given text "TRY TO SELL IT LOLOL" does not express a clear financial sentiment. The use of the word "sell" could be interpreted as a bearish sentiment, but the inclusion of "LOLOL" suggests that the author may be joking or expressing excitement rather than fear or pessimism. Without additional context, it is difficult to definitively classify this text as bullish, bearish, or neutral. Therefore, it is labelled as "other" to reflect its ambiguous nature.
**Error Analysis:** The post uses sarcasm and exaggerated tone, which reduces the models confidence in assigning an action label. Despite the imperative form suggesting a sell intent, the presence of humour and mockery leads the model to classify the post as ambiguous, highlighting challenges in interpreting ironic language.

**Example 4:** *The squeeze is squoze the shares have been covered these guys need to come back to reality*
**Real Label:** other
**Predicted Label:** sell
**Model Explanation:** The post expresses a clear intention to sell, as indicated by the phrase "The squeeze is squoze" which is a term used to describe a situation where a short squeeze has occurred, meaning that the short sellers have been forced to buy back their shares to cover their positions, driving up the price. The author then states that "the shares have been covered" which implies that they believe the price has peaked and it is time to sell. The overall tone of the post is bearish and suggests that the author is looking to take profits and exit their position.
**Error Analysis:** This post is informational and descriptive, using community-specific jargon to comment on market dynamics rather than expressing a personal trading action. The model incorrectly infers a sell intention from domain terminology (covered, squeeze), demonstrating sensitivity to technical cues without explicit action markers.

**Example 5:** *I thought i was gonna be able to buy the dip*
**Real Label:** other
**Predicted Label:** buy
**Model Explanation:** The post expresses the author's regret for not buying during a dip, indicating a bullish sentiment towards the market. The use of the word 'buy' in the text further supports this classification.
**Error Analysis:** The post describes a missed opportunity and expresses regret about a past action that did not occur. The model incorrectly interprets the reference to buy the dip as a current or future intent, failing to capture the counterfactual and retrospective nature of the statement.

**Example 6:** *When my OTM PUTs are printing*
**Real Label:** sell
**Predicted Label:** buy
**Model Explanation:** The post suggests that the author is currently experiencing profits from their option trades. The term "OTM PUT's" refers to out-of-the-money put options, which are bought with the expectation that the underlying asset's price will decrease. The fact that these options are "printing" means they are generating profits for the author. The overall sentiment expressed in the post is positive, as the author is experiencing gains."
**Error Analysis:** Although the post expresses positive outcomes (printing), the underlying financial instrument indicates a bearish position. The model associates profitability with a buy signal, overlooking the directional semantics of put options, which implicitly correspond to a sell-oriented market stance.

*These error cases highlight recurring challenges for the models, difficulty in handling sarcasm and counterfactual statements, and limited understanding of domain-specific financial semantics.*

### A.5. Prompt template

This subsection reports the prompt employed to apply LLMs to the STAD task.

**Prompt Template**

```
You are a financial sentiment classifier. Classify the following post into three categories:

"buy" if the post expresses intentions to buy or hold stocks, "sell" if the post expresses intentions to
sell stocks, and "other" if the post does not express intentions to sell or buy/hold stocks.
```

## Data availability

We have shared the link to our data in the manuscript.

## References


Aggarwal, D., Choi, A. H., & Lee, Y.-H. A. (2022). The meme stock frenzy: Origins and implications. *Southern California Law Review*, *96*, 1387.

Aggarwal, T., Salatino, A., Osborne, F., & Motta, E. (2026). Large language models for scholarly ontology generation: An extensive analysis in the engineering field. *Information Processing & Management*, *63*(1), Article 104262.

Angioni, S., Consoli, S., Dessì, D., Osborne, F., Recupero, D. R., & Salatino, A. (2024). Exploring environmental, social, and governance (esg) discourse in news: An ai-powered investigation through knowledge graph analysis. *IEEE Access*, *12*, 77269–77283.

Barnett, S., Brannelly, Z., Kurniawan, S., & Wong, S. (2024a). Fine-tuning or fine-failing? Debunking performance myths in large language models. arXiv: 2406.11201. URL https://arxiv.org/abs/2406.11201.

Barnett, S., Brannelly, Z., Kurniawan, S., & Wong, S. (2024b). Fine-tuning or fine-failing? debunking performance myths in large language models. arXiv preprint arXiv:2406.11201.

Betzer, A., & Harries, J. P. (2022). How online discussion board activity affects stock trading: the case of GameStop. *Financial Markets and Portfolio Management*, *36*(4), 443–472.

Birti, M., Maurino, A., & Osborne, F. (2025). Optimizing large language models for esg activity detection in financial texts. In *Proceedings of the 6th ACM international conference on AI in finance* (pp. 856–863).

Bolanos, F., Salatino, A., Osborne, F., & Motta, E. (2024). Artificial intelligence for literature reviews: Opportunities and challenges. arXiv preprint arXiv: 2402.08565.

Bonfigli, A., Bacco, L., Merone, M., & Dell'Orletta, F. (2024). From pre-training to fine-tuning: An in-depth analysis of large language models in the biomedical domain. *Artificial Intelligence in Medicine*, *157*, Article 103003. http://dx.doi.org/10.1016/j.artmed.2024.103003.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., .... Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Buscaldi, D., Dessí, D., Motta, E., Murgia, M., Osborne, F., & Recupero, D. R. (2024). Citation prediction by leveraging transformers and natural language processing heuristics. *Information Processing & Management*, *61*(1), Article 103583.

Chessa, A., Fenu, G., Motta, E., Osborne, F., Recupero, D. R., Salatino, A., & Secchi, L. (2023). Data-driven methodology for knowledge graph generation within the tourism domain. *IEEE Access*, *11*, 67567–67599.

Choi, M., Lee, H. J., Park, S. H., Jeon, S. W., & Cho, S. (2024). Stock price momentum modeling using social media data. *Expert Systems with Applications*, *237*, Article 121589. http://dx.doi.org/10.1016/j.eswa.2023.121589.

Dong, Z., Fan, X., & Peng, Z. (2024). FNSPID: A comprehensive financial news dataset in time series. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining* (pp. 4918–4927). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3637528.3671629.

Dong, H., Ren, J., Padmanabhan, B., & Nickerson, J. V. (2022). How are social and mass media different in relation to the stock market? A study on topic coverage and predictive value. *Information & Management*, *59*(2), Article 103588. http://dx.doi.org/10.1016/j.im.2021.103588.

Du, K., Xing, F., Mao, R., & Cambria, E. (2024). Financial Sentiment Analysis: Techniques and Applications. *ACM Computing Surveys*, *56*(9), http://dx.doi.org/10.1145/3649451.

Fatouros, G., Metaxas, K., Soldatos, J., & Kyriazis, D. (2024). Can large language models beat wall street? Evaluating GPT-4's impact on financial decision-making with MarketSenseAI. *Neural Computing and Applications*, 1–26.

Ferraro, A., & Sperlì, G. (2024). How does user-generated content on Social Media affect stock predictions? A case study on GameStop. *Online Social Networks and Media*, *43–44*, Article 100293. http://dx.doi.org/10.1016/j.osnem.2024.100293.

Gangopadhyay, S., & Majumder, P. (2023). Text representation for direction prediction of share market. *Expert Systems with Applications*, *211*, Article 118472. http://dx.doi.org/10.1016/j.eswa.2022.118472.

Ge, W., Lalbakhsh, P., Isai, L., Lenskiy, A., & Suominen, H. (2022). Neural Network–Based Financial Volatility Forecasting: A Systematic Review. *ACM Computing Surveys*, *55*(1), http://dx.doi.org/10.1145/3483596.

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature*, *57*(3), 535–574.

Gjerstad, P., Meyn, P. F., Molnár, P., & Næss, T. D. (2021). Do President Trump's tweets affect financial markets? *Decision Support Systems*, *147*, Article 113577. http://dx.doi.org/10.1016/j.dss.2021.113577.

Groeneveld, D., Beltagy, I., Walsh, P., Bhagia, A., Kinney, R., Tafjord, O., Jha, A. H., Ivison, H., Magnusson, I., Wang, Y., Arora, S., Atkinson, D., Authur, R., Chandu, K., Cohan, A., Dumas, J., Elazar, Y., Gu, Y., Hessel, J., .... Hajishirzi, H. (2024). OLMo: Accelerating the science of language models. Preprint.

Gupta, V., Varshney, D., Jhamtani, H., Kedia, D., & Karwa, S. (2014). Identifying purchase intent from social posts. In *Proceedings of the international AAAI conference on web and social media*: *Vol. 8*, (1), (pp. 180–186).

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. arXiv preprint arXiv:2004.10964.

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. arXiv preprint arXiv:1801.06146.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). Lora: Low-rank adaptation of large language models.. *International Conference on Learning Representation*, *1*(2), 3.

Huang, W.-C., Chen, C.-T., Lee, C., Kuo, F.-H., & Huang, S.-H. (2023). Attentive gated graph sequence neural network-based time-series information fusion for financial trading. *Information Fusion*, *91*, 261–276. http://dx.doi.org/10.1016/j.inffus.2022.10.006.

Kalai, A. T., Nachum, O., Vempala, S. S., & Zhang, E. (2025). Why language models hallucinate. arXiv preprint arXiv:2509.04664.

Lee, S., Lee, Y., Lee, J., & Kim, H. (2025). A statistical analysis of the relationship between meme stocks and social media. *IEEE Access*.

Li, X., Shen, X., Zeng, Y., Xing, X., & Xu, J. (2024). FinReport: Explainable stock earnings forecasting via news factor analyzing model. In *Companion proceedings of the ACM web conference 2024* (pp. 319–327). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3589335.3648330.

Li, Y., Wang, S., Ding, H., & Chen, H. (2023). Large language models in finance: A survey. In *Proceedings of the fourth ACM international conference on AI in finance* (pp. 374–382).

Li, X., Wu, P., & Wang, W. (2020). Incorporating stock prices and news sentiments for stock market prediction: A case of Hong Kong. *Information Processing & Management*, *57*(5), Article 102212. http://dx.doi.org/10.1016/j.ipm.2020.102212.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, *66*(1), 35–65.

Ma, X., Zhao, T., Guo, Q., Li, X., & Zhang, C. (2022). Fuzzy hypergraph network for recommending top-K profitable stocks. *Information Sciences*, *613*, 239–255. http://dx.doi.org/10.1016/j.ins.2022.09.010.

Macháček, R., Grishina, A., Hort, M., & Moonen, L. (2025). The impact of fine-tuning large language models on automated program repair. arXiv:2507.19909. URL https://arxiv.org/abs/2507.19909.

Mai, Z., Chowdhury, A., Zhang, P., Tu, C.-H., Chen, H.-Y., Pahuja, V., Berger-Wolf, T., Gao, S., Stewart, C., Su, Y., & Chao, W.-L. (2024). Fine-tuning is fine, if calibrated. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, & C. Zhang (Eds.), *Advances in neural information processing systems*: *vol. 37*, (pp. 136084–136119). Curran Associates, Inc., http://dx.doi.org/10.52202/079017-4323, URL https://proceedings.neurips.cc/paper_files/paper/2024/file/f573c36434796efe066d2f4cf3349e7f-Paper-Conference.pdf.

Motta, E., Daga, E., Gangemi, A., Gjelsvik, M. L., Osborne, F., & Salatino, A. (2025). The epistemology of fine-grained news classification. *Semantic Web*, *16*(3), Article 22104968251344461.

Motta, E., Osborne, F., Pulici, M. M., Salatino, A., & Naja, I. (2024). Capturing the viewpoint dynamics in the news domain. In *International conference on knowledge engineering and knowledge management* (pp. 18–34). Springer.

Muennighoff, N., Tazi, N., Magne, L., & Reimers, N. (2022). MTEB: Massive text embedding benchmark. arXiv preprint arXiv:2210.07316.

Nazareth, N., & Ramana Reddy, Y. V. (2023). Financial applications of machine learning: A literature review. *Expert Systems with Applications*, *219*, Article 119640. http://dx.doi.org/10.1016/j.eswa.2023.119640.

Olorunnimbe, K., & Viktor, H. (2023). Deep learning in the stock market—a systematic survey of practice, backtesting, and applications. *Artificial Intelligence Review*, *56*(3), 2057–2109. http://dx.doi.org/10.1007/s10462-022-10226-0.

Omiye, J. A., Gui, H., Rezaei, S. J., Zou, J., & Daneshjou, R. (2024). Large language models in medicine: the potentials and pitfalls: a narrative review. *Annals of Internal Medicine*, *177*(2), 210–220.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., .... Zoph, B. (2024). GPT-4 technical report. arXiv:2303.08774. URL https://arxiv.org/abs/2303.08774.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in neural information processing systems*: *vol. 35*, (pp. 27730–27744). Curran Associates, Inc., URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

Papasotiriou, K., Sood, S., Reynolds, S., & Balch, T. (2024). AI in investment analysis: LLMs for equity stock ratings. In *Proceedings of the 5th ACM international conference on AI in finance* (pp. 419–427). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3677052.3698694.

Phogat, K. S., Harsha, C., Dasaratha, S., Ramakrishna, S., & Puranam, S. A. (2023). Zero-shot question answering over financial documents using large language models. arXiv preprint arXiv:2311.14722.

Pu, G., Jain, A., Yin, J., & Kaplan, R. (2023). Empirical analysis of the strengths and weaknesses of PEFT techniques for LLMs. arXiv:2304.14999. URL https://arxiv.org/abs/2304.14999.

Qin, C., Chang, J., Tu, W., & Yu, C. (2024). FollowAKOInvestor: Stock recommendation by hearing voices from all kinds of investors with machine learning. *Expert Systems with Applications*, *249*, Article 123522. http://dx.doi.org/10.1016/j.eswa.2024.123522.

Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, *14*(1), 6086.

Riezler, S., & Hagmann, M. (2024). *Validity, reliability, and significance: empirical methods for NLP and data science*. Springer Nature.

Savelka, J., & Ashley, K. D. (2023). The unreasonable effectiveness of large language models in zero-shot semantic annotation of legal texts. *Frontiers in Artificial Intelligence*, *6*, Article 1279794.

Schmitz, H. C., Lutz, B., Wolff, D., & Neumann, D. (2023). When machines trade on corporate disclosures: Using text analytics for investment strategies. *Decision Support Systems*, *165*, Article 113892. http://dx.doi.org/10.1016/j.dss.2022.113892.

Shang, L., Xi, H., Hua, J., Tang, H., & Zhou, J. (2023). A Lexicon Enhanced Collaborative Network for targeted financial sentiment analysis. *Information Processing & Management*, *60*(2), Article 103187. http://dx.doi.org/10.1016/j.ipm.2022.103187.

Tang, Y., Song, Z., Zhu, Y., Yuan, H., Hou, M., Ji, J., Tang, C., & Li, J. (2022). A survey on machine learning models for financial time series forecasting. *Neurocomputing*, *512*, 363–380. http://dx.doi.org/10.1016/j.neucom.2022.09.003.

Thakkar, A., & Chaudhari, K. (2021a). A comprehensive survey on deep neural networks for stock market: The need, challenges, and future directions. *Expert Systems with Applications*, *177*, Article 114800. http://dx.doi.org/10.1016/j.eswa.2021.114800.

Thakkar, A., & Chaudhari, K. (2021b). Fusion in stock market prediction: A decade survey on the necessity, recent developments, and potential future directions. *Information Fusion*, *65*, 95–107. http://dx.doi.org/10.1016/j.inffus.2020.08.019.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., .... Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv:2307.09288. URL https://arxiv.org/abs/2307.09288.

Tsaneva, S., Dessì, D., Osborne, F., & Sabou, M. (2025). Knowledge graph validation by integrating LLMs and human-in-the-loop. *Information Processing & Management*, *62*(5), Article 104145.

Tunstall, L., Beeching, E., Lambert, N., Rajani, N., Rasul, K., Belkada, Y., Huang, S., von Werra, L., Fourrier, C., Habib, N., Sarrazin, N., Sanseviero, O., Rush, A. M., & Wolf, T. (2023). Zephyr: Direct distillation of LM alignment. arXiv:2310.16944.

Vasileiou, E., & Tzanakis, P. (2024). The impact of google searches, put-call ratio, and trading volume on stock performance using wavelet coherence analysis: The AMC case. *Journal of Behavioral Finance*, *25*(1), 111–119.

Wang, N., Yang, H., & Wang, C. D. (2023). Fingpt: Instruction tuning benchmark for open-source large language models in financial datasets. arXiv preprint arXiv:2310.04793.

Wang, R., Zhang, Z., Siau, K. L., & Zhang, Z. (2025). Managing rumors on electronic interaction platforms: How management responses affect investor reaction. *Information Processing & Management*, *62*(5), Article 104162. http://dx.doi.org/10.1016/j.ipm.2025.104162.

Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, *55*(7), 5731–5780. http://dx.doi.org/10.1007/s10462-022-10144-1.

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., & Mann, G. (2023). Bloomberggpt: A large language model for finance. arXiv preprint arXiv:2303.17564.

Xie, Q., Han, W., Chen, Z., Xiang, R., Zhang, X., He, Y., Xiao, M., Li, D., Dai, Y., Feng, D., Xu, Y., Kang, H., Kuang, Z., Yuan, C., Yang, K., Luo, Z., Zhang, T., Liu, Z., Xiong, G., .... Huang, J. (2024). FinBen: A holistic financial benchmark for large language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, & C. Zhang (Eds.), *37*, *Advances in neural infor mation processing systems* (pp. 95716–95743). Curran Associates, Inc..

Xie, Q., Han, W., Zhang, X., Lai, Y., Peng, M., Lopez-Lira, A., & Huang, J. (2024). PIXIU: a large language model, instruction data and evaluation benchmark for finance. In *Proceedings of the 37th international conference on neural information processing systems*. Red Hook, NY, USA: Curran Associates Inc..

Yang, H., Liu, X.-Y., & Wang, C. D. (2023). FinGPT: Open-Source Financial Large Language Models. *FinLLM Symposium At IJCAI 2023*.

Yang, K., Zhang, T., Kuang, Z., Xie, Q., Huang, J., & Ananiadou, S. (2024). MentaLLaMA: interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM web conference 2024* (pp. 4489–4500).

Zhang, S., Dong, L., Li, X., Zhang, S., Sun, X., Wang, S., Li, J., Hu, R., Zhang, T., Wang, G., & Wu, F. (2026). Instruction tuning for large language models: A survey. *ACM Computing Surveys*, *58*(7), http://dx.doi.org/10.1145/3777411.

Zhang, Q., Zhang, Y., Bao, F., Liu, Y., Zhang, C., & Liu, P. (2024). Incorporating stock prices and text for stock movement prediction based on information fusion. *Engineering Applications of Artificial Intelligence*, *127*, Article 107377. http://dx.doi.org/10.1016/j.engappai.2023.107377.

Zhang, X., Zhang, Y., Long, D., Xie, W., Dai, Z., Tang, J., Lin, H., Yang, B., Xie, P., Huang, F., Zhang, M., Li, W., & Zhang, M. (2024). mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval. In F. Dernoncourt, D. Preo¸tiuc Pietro, & A. Shimorina (Eds.), *Proceedings of the 2024 conference on empirical methods in natural language processing: industry track* (pp. 1393–1412). Miami, Florida, US: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2024.emnlp-industry.103, URL https://aclanthology.org/2024.emnlp-industry.103/.

Zhang, Y., Zou, C., Lian, Z., Tiwari, P., & Qin, J. (2025). Sarcasmbench: Towards evaluating large language models on sarcasm understanding. *IEEE Transactions on Affective Computing*.

Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., Ma, X., Efrat, A., Yu, P., YU, L., Zhang, S., Ghosh, G., Lewis, M., Zettlemoyer, L., & Levy, O. (2023). LIMA: Less is more for alignment. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in neural information processing systems*: *vol. 36*, (pp. 55006–55021). Curran Associates, Inc., URL https://proceedings.neurips.cc/paper_files/paper/2023/file/ac662d74829e4407ce1d126477f4a03a-Paper-Conference.pdf.

Zhu, B., Frick, E., Wu, T., Zhu, H., & Jiao, J. (2023). Starling-7B: Improving LLM helpfulness & harmlessness with RLAIF.

Zhuang, Y., Wang, F., Chiu, D. K., & Ho, K. K. (2025). Leveraging large language models to examine the interaction between investor sentiment and stock performance. *Engineering Applications of Artificial Intelligence*, *150*, Article 110602. http://dx.doi.org/10.1016/j.engappai.2025.110602.