Evaluating Model Approaches to Extracting Financial Sentiment from MAG7 Earnings Calls: Natural Language Processing (FinBERT) vs Large Language Models (DeepSeek)

Teanna Kilia Puthucheary
Supervisor: Dr. Ahmad Abu-Kazneh
Bayes Business School

This report is submitted as part of the requirements for the award of the MSc in Business Analytics 2024-25

Abstract

This paper evaluates the effectiveness of two sentiment analysis approaches, an NLP-based model (FinBERT) and a large language model (DeepSeek), in determining the financial sentiment of earnings call transcripts from different technology companies. The objective is to assess which model most accurately predicts sentiment scores compared to a manually annotated baseline and to examine how closely each model's outputs align with external indicators of market sentiment, using short-term stock price reactions. The study also considers whether there are differences in performance across companies with varying product portfolios.

Earnings call transcripts were segmented by speaker and then into text blocks to increase granularity, since aggregating at speaker level alone would hinder the evaluation of classification accuracy. Each segment was scored by FinBERT and DeepSeek using a consistent five-bin sentiment scale from Strong Positive to Strong Negative.

To benchmark the findings, we compared model outputs to (1) manually annotated sentiment labels, (2) percentage changes in stock prices before and after earnings announcements, categorised into the same five sentiment bins.

This study concludes that FinBERT achieves higher fidelity to transcript sentiment, while DeepSeek more closely aligns with market reactions; their complementary strengths suggest that integrating both approaches can deliver more robust sentiment analysis and stronger explanatory power in linking corporate communications to financial outcomes.

Contents

List of Figures

1 Introduction

Natural Language Processing (NLP) has become a core component of financial market analysis due to the vast amounts of unstructured textual data available. Financial news, analyst reports, social media discussions, and earnings call transcripts can greatly influence market sentiment and subsequently stock price movements (Konstantinidis et al., 2024). This is consistent with the Efficient Market Hypothesis (EMH), which states that stock prices adjust rapidly in response to new and unexpected information (Fama, 1970). While traditional financial analysis focuses on quantitative data, markets often react just as strongly to qualitative information from company communications, which investors also use in their decision-making processes (Tonin, 2021). NLP enables the systematic extraction of these qualitative signals from both mandatory reports and voluntary disclosures, such as earnings calls.

Earnings calls hold significant weight in financial markets as they provide management teams with the opportunity to discuss business performance and contain forward-looking information, going beyond the numerical results disclosed in financial statements (Konstantinidis et al., 2024). A study from Chen Liang (2013), highlights that the tone and language used during these calls can materially affect investor confidence and market reactions. Subtle differences in phrasing or confidence levels can impact stock prices, particularly for growth-focused firms that trade at high valuations (McGugan, 2024), such as the Magnificent Seven (MAG7) technology firms. This study examines three MAG7 firms, Apple, Amazon, and Nvidia, across different sectors examining whether model performance is consistent across industry segments.

Recent advancements in NLP have expanded the tools available for financial sentiment analysis. Models such as FinBERT provide robust classification capabilities tailored to financial text. More recently, general purpose large language models (LLMs) have gained prominence because of their deeper contextual understanding and ability to interpret nuance in complex language. This study employs Deepseek, an advanced LLM, with prompt-engineered strategies to enhance classification accuracy.

Despite these developments, few studies directly compare FinBERT with generalist LLMs like DeepSeek in the context of earnings calls, and even fewer benchmark results against multiple measures. This research addresses this gap by evaluating model outputs against two benchmarks: manually annotated sentiment labels and short-term stock price reactions. By assessing alignment with both human judgment and market behaviour, this study provides insights into refining sentiment analysis methods for financial forecasting. As sentiment continues to be a major driver of short-term price movements, advancing model accuracy has the potential to improve predictive power and strengthen the tools available to investors and analysts.

2 Literature Review

Financial sentiment analysis seeks to anticipate market reactions to information disclosed in financial text (Araci, 2019). Consistent with the Efficient Market Hypothesis (Fama, 1970), markets adjust rapidly to new qualitative and quantitative signals, making sentiment extraction a valuable tool. This study compares the performance of FinBERT, a domain-specific sentiment classifier, and DeepSeek, a generalist large language model (LLM), in analysing earnings call transcripts.

FinBERT is based on the BERT architecture and fine-tuned specifically for financial text classification. It performs well in structured contexts but can oversimplify implicit financial language (Araci, 2019). Chen et al. (2023) also show its performance deteriorates in syntactically complex sentences, a common feature of earnings calls. Small shifts in tone may therefore be missed, limiting effectiveness in real-world contexts.

FinBERT, built on the BERT architecture and fine-tuned for finance, performs well on structured language but struggles with implicit or complex phrasing, a frequent feature of earnings calls (Araci, 2019; Chen et al., 2023). Subtle tonal shifts may therefore be overlooked, limiting predictive value. DeepSeek, by contrast, is a generative model capable of processing extended narrative contexts across multiple speakers. Although not trained specifically for finance, its ability to interpret broader discourse offers potential advantages. Kang and Choi (2025) note the growing application of LLMs in financial analysis, yet direct comparisons with FinBERT models remain scarce. This study addresses that gap by testing whether DeepSeek can complement or substitute FinBERT.

Earnings calls provide an ideal setting, containing forward-looking statements, strategy, and spontaneous QA. Chen and Liang (2013) highlight that tone can materially affect investor confidence. Stock price movements serve as a quantitative market reaction measure. Kang, Park, and Han (2018) show that call tone strongly correlates with short-term response, especially when expressed by senior executives.

2.1 Why MAG7 Companies?

The three MAG7 companies, Apple, Amazon, and Nvidia, are highly influential yet operate in distinct segments: consumer hardware, e-commerce/cloud, and semiconductors/AI (Fidelity, 2024). This diversity tests whether models maintain accuracy across industries or show sectoral variation. As members of the SP 500, they provide high-quality transcripts and attract heavy analyst scrutiny (Guest, 2021). Their inclusion in major indices and elevated P/E multiples mean prices are especially sensitive to call tone (McGugan, 2024). Thus, performance differences are more likely due to sectoral or linguistic nuance than unrelated macro factors.

2.2 Rationale for Using a Generalist LLM

The aim is to evaluate whether a generalist LLM can match or outperform a finance-specific model. Prior research by Kirtac and Germano (2024) found that a general-purpose LLM outperformed FinBERT on financial news sentiment. Using DeepSeek as-is preserves the cross-domain comparison and mirrors practice, as many firms adopt general LLMs for API accessibility, multitask capability, and lower infrastructure needs. If DeepSeek achieves results comparable to FinBERT without tuning, it offers strong evidence that general LLMs can complement or even replace niche financially tuned sentiment models.

3 Data Variables

The dataset for this study consists of earnings call transcripts and corresponding stock price data for three major technology companies: NVIDIA, Amazon, and Apple. The uniform processing and modelling pipeline was used across firms for consistency.

3.1 Data Sources

- **Earning Call Transcripts:**
  Quarterly earnings call transcripts for the most recent eight quarters of each company were scraped from *The Motley Fool*, a publicly available source chosen due to its uniform formatting and comprehensive coverage of both management and analyst sections. The transcripts were formatted to present each speaker's name and position alongside their spoken content, structured in a clear line-by-line format.
- **Financial Market Data:**
  Daily closing prices for each company's stock were retrieved from Yahoo Finance using the `yfinance` Python package, selected for its wide coverage and free API-accessible format. This facilitated efficient alignment of market price movements with sentiment analysis results, using earnings announcement dates as the anchor points.

3.2 Data Structure

(A) Transcript Dataset

This dataset is a row-level representation of individual transcript segments. Each row corresponds to a single spoken block from a specific speaker during the earnings call, paired with sentiment scores from multiple models and manual annotations. The variables are:

Variable Name	Description
ticker	Stock ticker symbol of the company (e.g., NVDA, AMZN, AAPL).
quarter	Fiscal quarter of the earnings call (e.g., Q2 2025).
content	Full text of the spoken segment from the transcript.
role_category	Position or role of the person speaking (e.g., CEO, CFO, Analyst).
FinBERT_sentiment	Sentiment classification output from FinBERT, a financial domain-specific NLP model.
LLM_sentiment	Sentiment classification output from DeepSeek LLM.
manual_sentiment	Sentiment label assigned by human annotators, used as a baseline for benchmarking model performance.

Table 3.1: Earnings call transcript variables.

(B) Stock Price Benchmark Dataset

This dataset contains daily stock prices for the period surrounding each earnings call, enabling the calculation of percentage price changes two days before and after the announcement. These changes were categorised into the same sentiment bins as the transcript data to facilitate benchmarking.

Variable Name	Description
Earnings Date	Company-reported date of the earnings release.
T-2 Date	Trading day two days before the earnings date.
T+2 Date	Trading day two days after the earnings date.
T-2 Price	Closing price on the T-2 date.
T+2 Price	Closing price on the T+2 date.
% Change	Percentage change in closing price from T-2 to T+2.
Sentiment	Market reaction sentiment label derived from % Change (Strong Positive, Positive, Neutral, Negative, Strong Negative).
EPS Estimate	Analyst consensus estimate for earnings per share before release.
Reported EPS	Actual earnings per share reported by the company.
Surprise (%)	Percentage difference between reported and estimated EPS.

Table 3.2: Financial market sentiment variables.

3.2.1 Sentiment Categories

All sentiment scores generated by FinBERT, DeepSeek and manual annotation were standardised into five distinct categories: Strong Positive, Slightly Positive, Neutral, Slightly Negative, and Strong Negative. This standardisation was applied to ensure a consistent framework for interpreting sentiment across all sources, enabling direct comparison between model outputs and the manual annotations benchmark.

4 Methodology

This study evaluates the performance of a domain-specific NLP model (FinBERT) and a general-purpose LLM (DeepSeek) in classifying sentiment from quarterly earnings call transcripts of three MAG7 companies: Apple, Amazon, and NVIDIA. A consistent preprocessing and evaluation pipeline is applied across firms to test whether performance generalises or varies with sector-specific language and communication styles. Model outputs are assessed against two benchmarks: manually annotated sentiment labels, providing a classification ground truth, and short-term stock price reactions, capturing quantitative market responses.

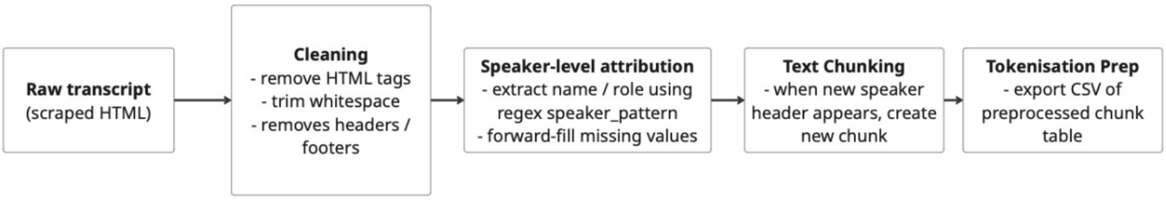

Raw transcript
(scraped HTML)
Cleaning
- remove HTML tags
- trim whitespace
- removes headers / footers
Speaker-level attribution
- extract name / role using regex speaker_pattern
- forward-fill missing values
Text Chunking
- when new speaker header appears, create new chunk
Tokenisation Prep
- export CSV of preprocessed chunk table

Figure 4.1: Methodology Steps

4.1 Data Preprocessing Pipeline

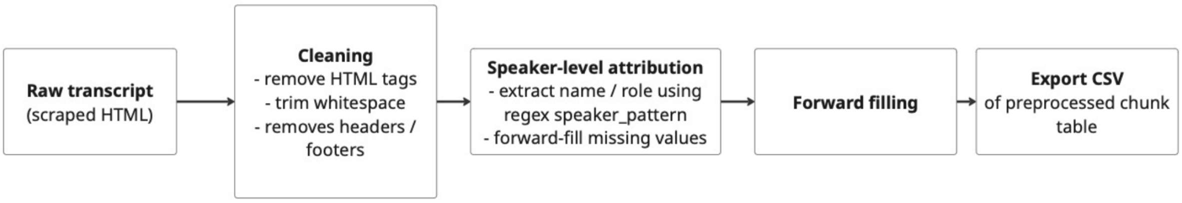

Raw transcript
(scraped HTML)
Cleaning
- remove HTML tags
- trim whitespace
- removes headers / footers
Speaker-level attribution
- extract name / role using regex speaker_pattern
- forward-fill missing values
Forward filling
Export CSV
of preprocessed chunk table

Figure 4.2: Preprocessing Pipeline

To enhance data quality and reduce potential biases, multiple preprocessing steps were undertaken. Text was cleaned by removing HTML tags, excess whitespace, and extraneous symbols, while missing speaker names were forward-filled. Sentiment was initially aggregated at the speaker level and then refined to the paragraph-per-speaker level, increasing granularity and capturing shifts in tone. This process expanded the number of observations and improved the robustness of the findings

4.2 Experimental Design

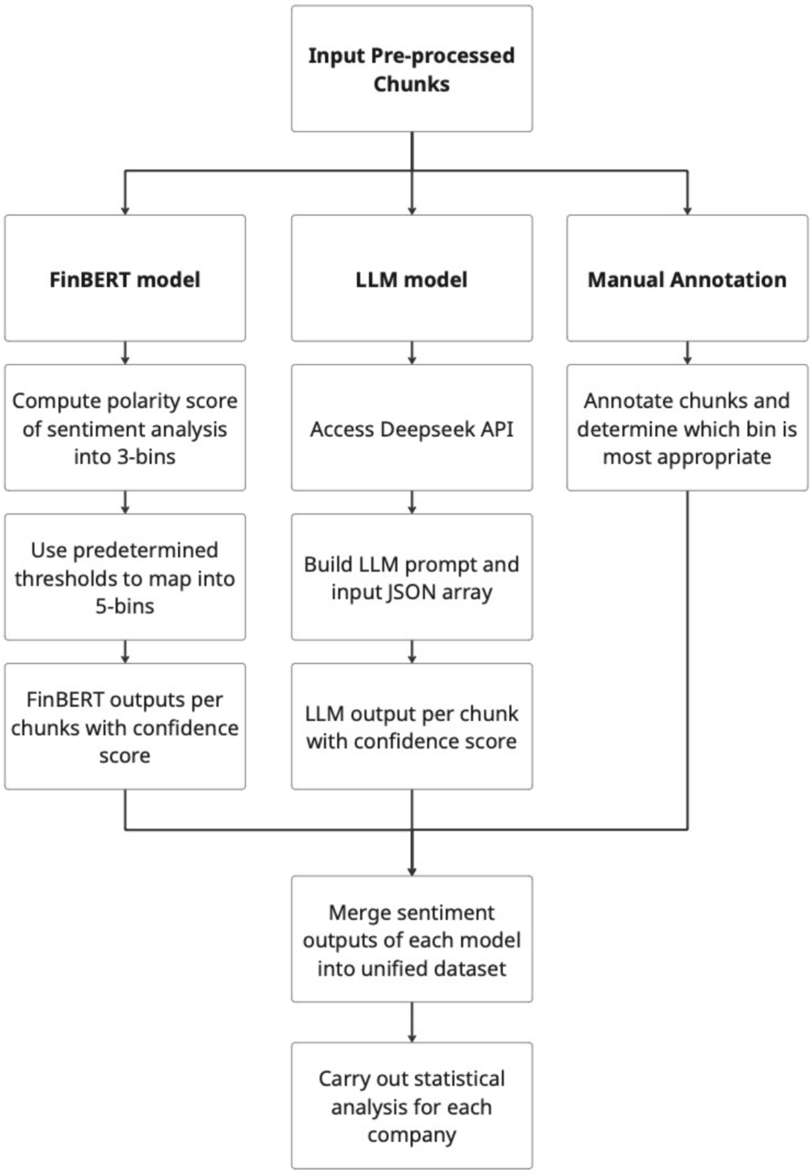

Input Pre-processed Chunks
FinBERT model
LLM model
Manual Annotation
Compute polarity score of sentiment analysis into 3-bins
Access Deepseek API
Annotate chunks and determine which bin is most appropriate
Use predetermined thresholds to map into 5-bins
Build LLM prompt and input JSON array
FinBERT outputs per chunks with confidence score
LLM output per chunk with confidence score
Merge sentiment outputs of each model into unified dataset
Carry out statistical analysis for each company

Figure 4.3: Experimental Design

This study evaluates sentiment classification using three parallel approaches, FinBERT model (NLP), DeepSeek model (LLM) and manual annotation, on identical text chunks extracted from pre-processed earnings call transcripts.

- **FinBERT**
  Transcript chunks were processed with FinBERT's tokenizer, which converts text into the model's required input format. FinBERT then produced class probabilities across a 3-bin sentiment scheme (Negative, Neutral, Positive). These probabilities were transformed into a polarity score ranging from –1 to +1 and subsequently

mapped into a 5-bin scale (Strong Positive, Slightly Positive, Neutral, Slightly Negative, Strong Negative) using thresholds based on the equal-width discretisation method (Putri et al., 2023). This method divides the score range into evenly sized intervals, providing a straightforward and consistent way to convert continuous sentiment values into discrete categories. Such uniformity facilitates cross-company comparison and reflects the importance of interpretability in trading contexts, where clarity of market sentiment is more valuable than strict statistical balance. For each transcript chunk, FinBERT outputs both a sentiment label.

- **DeepSeek**
  Using the DeepSeek API, each transcript chunk was classified into the 5-label sentiment framework through prompt engineering. Constrained decoding was applied, restricting outputs so that the model could only select one of the five valid sentiment labels, with the probability of the chosen label recorded as its confidence.

- **Manual Annotation**
  Human annotators independently labelled the same transcript chunks using the 5-bin sentiment framework. These labels provided the ground truth against which FinBERT and DeepSeek outputs were evaluated.

Outputs from all three approaches were consolidated into a single dataset, with sentiment distributions aggregated at the company, quarterly, and speaker levels to detect patterns and trends. Model performance was then evaluated by comparing FinBERT and DeepSeek predictions against manual labels using accuracy, precision, recall, and F1-score, ensuring both overall correctness and sensitivity to class imbalances. This evaluation was conducted separately for each company, allowing sector-level variation in model performance to be identified

4.3 Evaluation Method

To evaluate the model's ability to classify sentiment accurately, we used four key metrics: accuracy, precision, recall, and F1-score. In a classification task, each prediction is compared to the actual class, labelled as true (correct) or false (incorrect). For multi-class evaluation, we adopted a one-vs-rest approach, treating each sentiment category in turn as the "positive" class while grouping all others as "negative".

- Accuracy reflects the proportion of all correct predictions
- Precision measures how many predicted positives are truly positive
- Recall captures how many actual positives are correctly identified
- F1-score is the harmonic mean of precision and recall, balancing both metrics

Given the multi-class structure of the task, weighted F1-scores were reported to account for class imbalance by averaging across categories in proportion to their support. In addition, per-class F1-scores were examined to provide insight into how well each sentiment category was classified, allowing us to assess whether FinBERT or DeepSeek performed better at detecting both extreme sentiments (strong positive/negative) and subtler tonal shifts.

Although this analysis treats all sentiment classes equally, future evaluations could apply custom weights to false positives and false negatives to reflect asymmetric business costs. In financial contexts, correctly identifying strong positive or strong negative sentiment is particularly critical, as these less frequent extremes often signal significant opportunities or risks. A missed strong negative, for example, may carry far greater consequences than overlooking a slightly positive signal.

4.4 Stock Market Analysis

An additional verification step assessed each model's quarterly sentiment classifications against market sentiment derived from financial data. Market sentiment was defined by stock price changes within a two-day window around earnings releases, capturing immediate reactions while minimising intraday volatility. Price movements were mapped into five categories, from Strong Positive (¿+5%) to Strong Negative (¡ –5%), and EPS surprise was calculated as a complementary indicator of bullish or bearish signals. These benchmarks were then compared with FinBERT and DeepSeek outputs at the quarterly level to evaluate alignment with market behaviour.

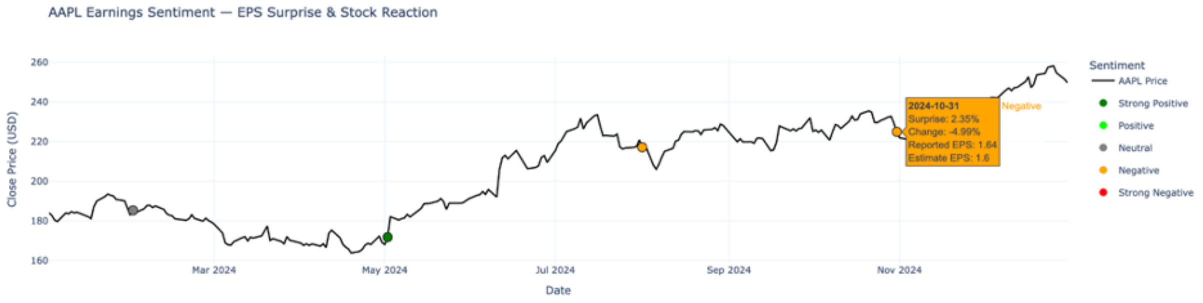


Figure 4.4: Apple Market Sentiment Graph

4.5 Limitations of Methodology

Rationales were not requested from the LLM in order to maintain a fair comparison with FinBERT and avoid variability introduced by prompt design. While this ensured consistency across models, it limited interpretability by excluding qualitative insights into misclassifications. Future work could incorporate concise rationales to identify systematic weaknesses and improve model transparency.

5 Results

Model performance is evaluated against two complementary benchmarks. The first, linguistic fidelity, measures how closely FinBERT and DeepSeek align with human-annotated sentiment labels, capturing each model's ability to interpret language in a manner consistent with human judgment. The second, predictive validity, assesses alignment with market sentiment derived from stock price reactions around earnings releases, indicating how well model outputs reflect investor behaviour. These benchmarks address different dimensions of performance and are interpreted separately depending on whether the intended use is linguistic analysis or financial forecasting.

5.1 Amazon

Metric	**FinBERT**	**DeepSeek**
Accuracy	0.68	0.69
Weighted F1-score	0.64	0.66
Macro F1-score	0.45	0.36
Macro Precision	0.47	0.34
Macro Recall	0.50	0.39

Table 5.1: Amazon: Overall performance comparison

As shown in Table 5.1, DeepSeek marginally outperforms FinBERT on both overall accuracy and weighted F1. Given the class imbalance, where extreme sentiments are relatively rare, greater emphasis is placed on weighted and per-class F1 scores, as these better capture performance under real-world class distributions.

The normalised confusion matrices present the percentage distribution of predictions for each true class, reducing the influence of class imbalance on interpretation and enabling clearer within-class comparisons. Results show that FinBERT is more effective at detecting extreme cases, correctly identifying instances of strong positive sentiment, though no examples of 'Strong Negative' were present in the test set. DeepSeek, by contrast, achieves higher recall for 'Slightly Positive' (83% vs. 56% for FinBERT, a

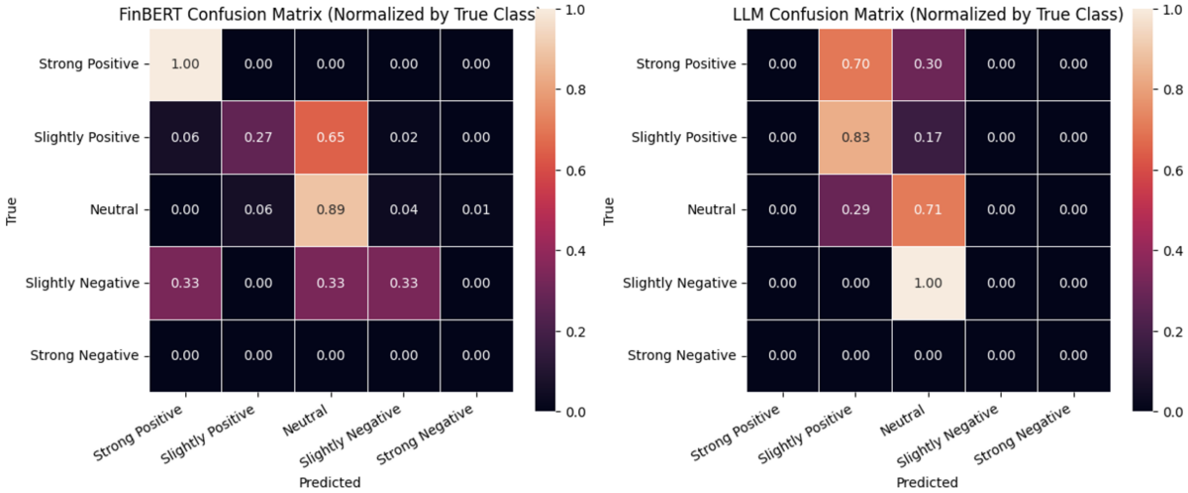


Figure 5.1: Amazon: Normalised Confusion Matrices

27% improvement), indicating stronger performance in recognising moderately positive sentiment. However, DeepSeek also displays a positivity bias, frequently misclassifying ‘Neutral’ cases as ‘Slightly Positive.

Sentiment Class	FinBERT	DeepSeek
Strong Positive	0.83	0.00
Slightly Positive	0.39	0.67
Neutral	0.78	0.75
Slightly Negative	0.25	0.00
Strong Negative	0.00	0.00

Table 5.2: Amazon: Per-class F1-score comparison

The per-class F1 scores in Table 5.2 summarise the confusion matrix patterns into single performance measures, reinforcing our earlier observations. FinBERT outperforms DeepSeek in classifying ‘Strong Positive’ (0.35 vs. 0.22) and ‘Slightly Negative’ sentiment, demonstrating more balanced performance across extreme categories. By contrast, DeepSeek shows stronger results in the middle classes, reflecting its tendency to capture general optimism. These results suggest that FinBERT may be more reliable for stress-testing upside and downside scenarios, while DeepSeek is better suited to identifying sentiment shifts in the broader mid-range where markets often anchor their reactions.

5.2 Apple

Metric	FinBERT	DeepSeek
Accuracy	0.74	0.72
Weighted F1-score	0.74	0.70
Macro F1-score	0.42	0.38
Macro Precision	0.43	0.60
Macro Recall	0.40	0.41

Table 5.3: Apple: Overall performance comparison

Table 5.3 shows that FinBERT marginally outperforms DeepSeek on both accuracy and weighted F1, indicating closer overall alignment with manual classification for Apple's transcripts.

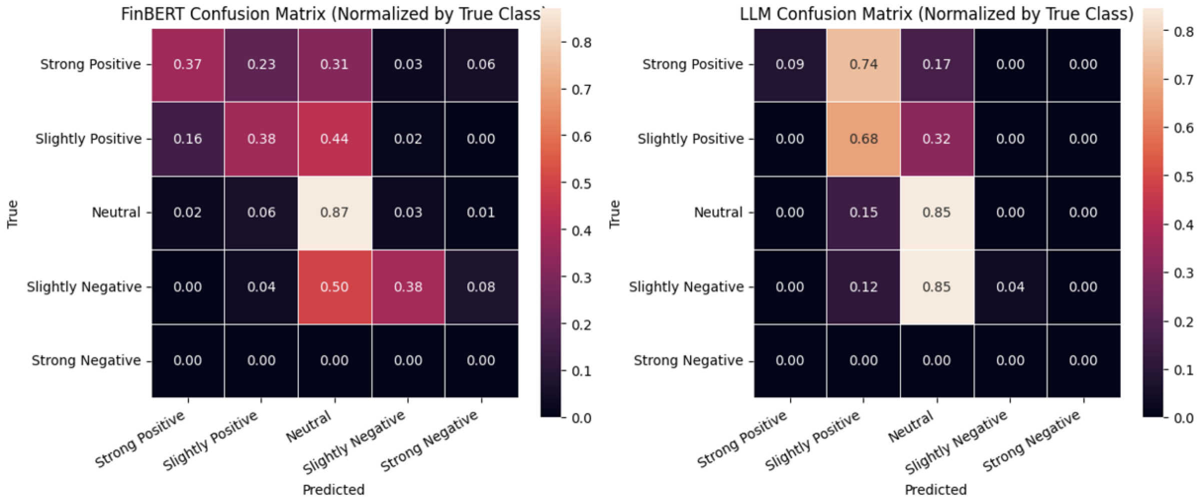


Figure 5.2: Apple: Normalised Confusion Matrices

The error patterns, however, reveal important differences in model behaviour. FinBERT distributes its misclassifications across adjacent classes — for instance, often confusing 'Strong Positive' with 'Slightly Positive' or 'Neutral.' While this reduces precision at the margins, it suggests that FinBERT is at least sensitive to the presence of extremes, even if it occasionally underestimates their strength. By contrast, DeepSeek displays a pronounced bias towards 'Neutral' and 'Slightly Positive,' correctly identifying neutral language 85% of the time but collapsing much of the variation in sentiment into these middle categories. This leads to severe under-detection of downside cues, with 85% of 'Slightly Negative' cases classified as neutral.

Sentiment Class	**FinBERT**	**DeepSeek**
Strong Positive	0.41	0.15
Slightly Positive	0.39	0.43
Neutral	0.86	0.85
Slightly Negative	0.42	0.07
Strong Negative	0.00	0.00

Table 5.4: Apple: Per-class F1-score comparison

Table 5.4 shows that FinBERT clearly outperforms DeepSeek on F1-scores, highlighting its advantage in financial sentiment analysis. Its domain-specific training enables it to interpret nuanced financial language more effectively, reflected in substantially higher scores for ‘Strong Positive’ (0.41 vs. 0.15) and ‘Slightly Negative’ (0.42 vs. 0.07). This consistent outperformance across classes suggests that domain-specialised tuning provides a critical edge in capturing subtle sentiment shifts, making FinBERT better suited for the complex discourse of financial markets

5.3 Nvidia

Metric	**FinBERT**	**DeepSeek**
Accuracy	0.58	0.56
Weighted F1-score	0.49	0.56
Macro F1-score	0.31	0.57
Macro Precision	0.56	0.78
Macro Recall	0.30	0.58

Table 5.5: Nvidia: Overall performance comparison

For Nvidia, DeepSeek holds a clear overall advantage, outperforming FinBERT on key aggregate metrics such as accuracy and weighted F1-score.

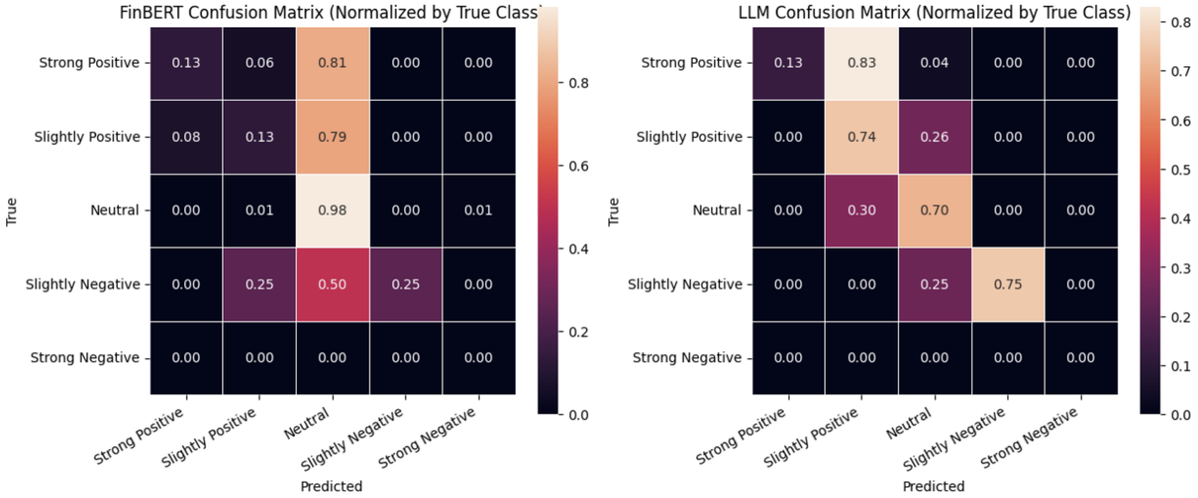


Figure 5.3: Nvidia: Normalised Confusion Matrices

Where FinBERT collapses most predictions into the ‘Neutral’ class (78% of test cases), DeepSeek demonstrates greater differentiation, achieving much higher recall for both ‘Slightly Negative’ (86% vs. 40%) and ‘Slightly Positive’ (72% vs. 44%). This indicates that, in NVIDIA’s context, the generalist LLM’s broader training enabled it to capture more nuanced sentiment shifts. FinBERT’s tendency to over-predict neutrality instead underscores its limitations when dealing with technical but non-financial language, such as chip design and AI development.

Sentiment Class	**FinBERT**	**DeepSeek**
Strong Positive	0.22	0.23
Slightly Positive	0.20	0.41
Neutral	0.73	0.77
Slightly Negative	0.40	0.86
Strong Negative	0.00	0.00

Table 5.6: Nvidia: Per-class F1-score comparison

Table 5.6 highlights DeepSeek’s clear advantage over FinBERT in capturing Nvidia’s sentiment, particularly in nuanced categories. Its F1-score for ‘Slightly Negative’ is more than twice as high (0.86 vs. 0.40), indicating a markedly stronger ability to detect this key sentiment. These results suggest that FinBERT’s financial specialisation does not fully generalise to highly technical industries, whereas DeepSeek’s broader training enables it to interpret sentiment in complex, non-financial contexts

5.4 Alignment with Financial Market Sentiment

We compare the average quarterly sentiment scores from DeepSeek, FinBERT, and manual annotations against a market-based benchmark. Market sentiment is calculated as the percentage change in share price within a two-day window before and after each earnings release. In Figure 5.7, the x-axis applies the same numeric encoding used across all model outputs (2 = Strong Negative, 0 = Neutral, 2 = Strong Positive).

5.4.1 Amazon

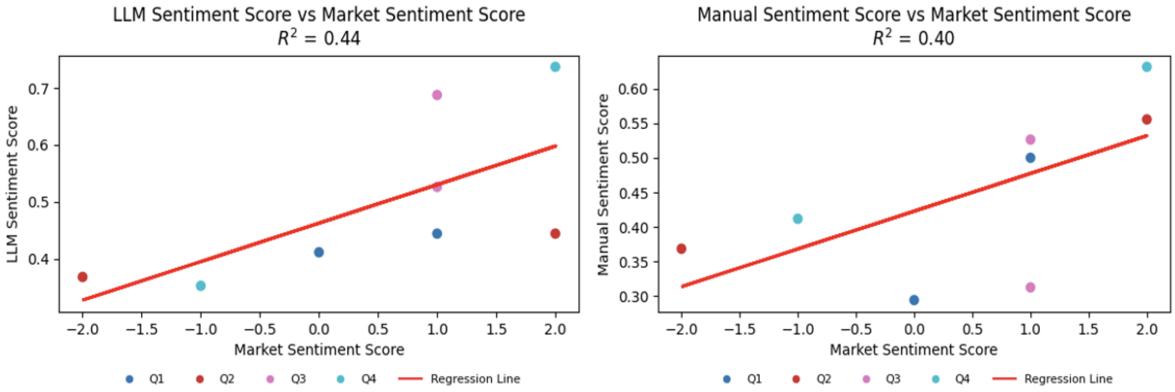


Figure 5.4: Amazon: Model Comparison with Market Sentiment Score

A notable contradiction emerges when comparing model performance across benchmarks. The model that best aligned with market sentiment was not the one most consistent with human-annotated classifications. For Amazon, DeepSeek's average sentiment scores exhibited the strongest linear relationship with market movements (R^2 = 0.44), despite its weaker confusion matrix results. This indicates that while FinBERT more closely mirrors human judgments, DeepSeek better reflects how investors interpret and trade on earnings calls. The contrast underscores that model evaluation depends on the intended objective: linguistic fidelity versus predictive validity.

5.4.2 Apple

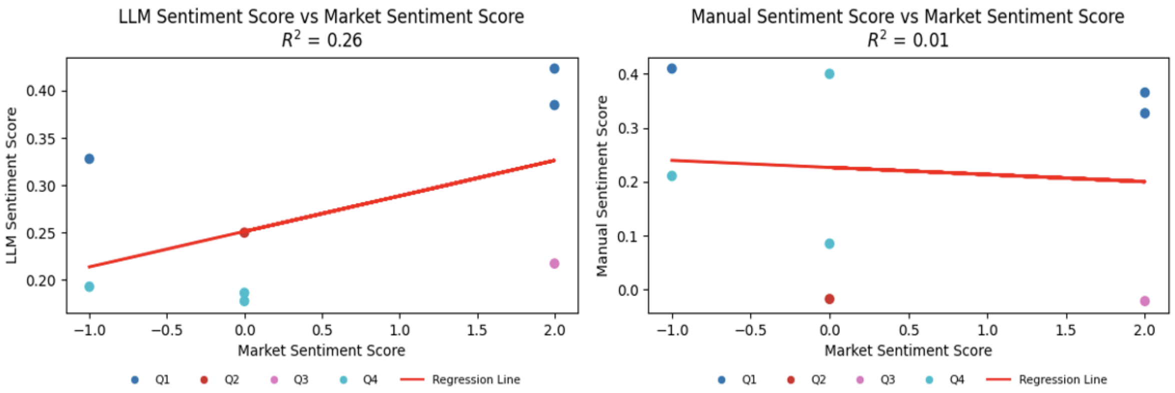


Figure 5.5: Apple: Model Comparison with Market Sentiment Score

When predicting market sentiment, DeepSeek aligns more closely with Apple's stock price reactions, despite FinBERT achieving higher agreement with manual annotations. DeepSeek's broader training corpus may enable it to capture the wider market context and investor psychology that shape price movements, rather than focusing solely on financial phrasing. This highlights the distinction between theoretical accuracy (FinBERT on annotations) and practical relevance (DeepSeek on market outcomes), emphasising that the optimal model depends on whether the objective is analyst interpretation or anticipating investor behaviour.

5.4.3 Nvidia

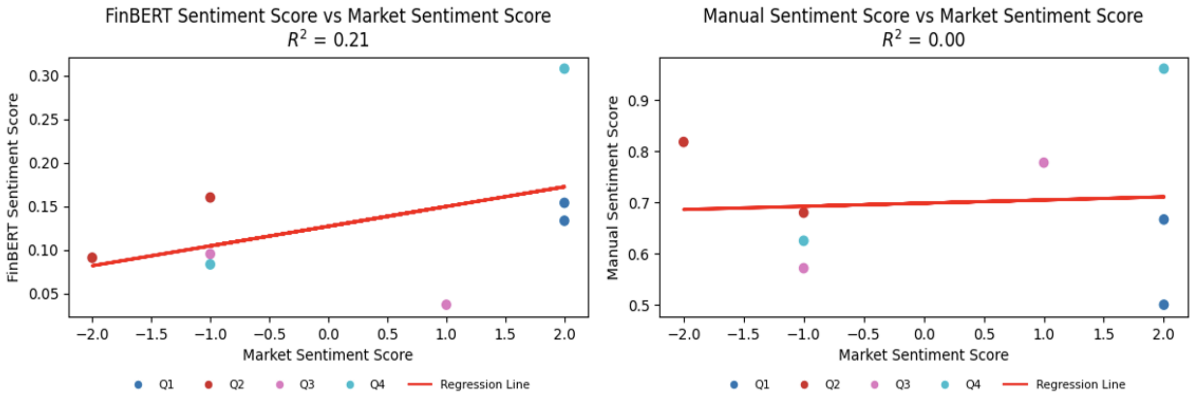


Figure 5.6: Nivida: Model Comparison with Market Sentiment Score

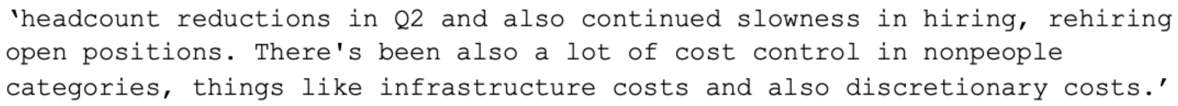
For Nvidia, FinBERT shows the strongest alignment with market sentiment, with an R^2 of 0.21 compared to virtually no relationship between manual sentiment and market movements (R^2 = 0). This suggests that Nvidia's earnings calls, which are highly technical and focused on areas such as chip design and AI, align more closely with FinBERT's domain-specific training in financial and corporate language. In this context, specialised tuning proves critical: FinBERT captures the nuanced signals that investors price in, whereas manual interpretation fails to translate the technical discourse into market-relevant sentiment.

Amazon's relatively high R^2 underscores that, despite being consumer-facing, its earnings calls are priced more directly on fundamentals such as AWS growth and retail margins. By contrast, Apple's weaker alignment suggests its price reactions are more expectation-driven, consistent with many consumer-facing firms where sentiment reflects broader market narratives as much as company performance. Nvidia also shows weak alignment between manual sentiment and market outcomes, likely due to the highly technical nature of its calls. In this case, domain-specific signals such as chip design or AI development are difficult for manual annotation to capture, yet still priced by investors.

6 Analysis

6.1 Where the model fails to classify correctly

- **FinBERT failed, DeepSeek succeeded**

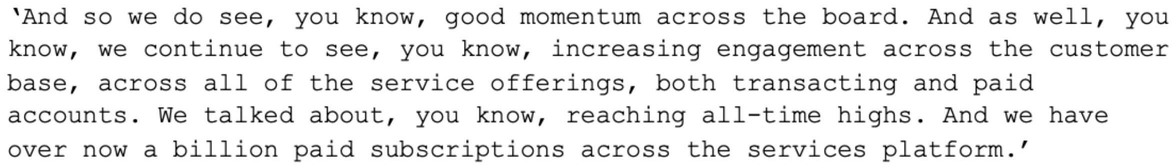


'headcount reductions in Q2 and also continued slowness in hiring, rehiring open positions. There's been also a lot of cost control in nonpeople categories, things like infrastructure costs and also discretionary costs.'

Figure 6.1: Amazon: Text Extract

A consistent pattern is FinBERT's tendency to over-assign the Neutral class when operational challenges or cost-cutting are discussed. For instance, in Amazon's call, a passage on headcount reduction was misclassified as Neutral rather than Strong Negative. This reflects FinBERT's financial training corpus, where cost-cutting is often framed as a neutral or even positive strategic action. By contrast, DeepSeek's broader training makes it more attuned to everyday language use, where headcount reductions are typically perceived negatively.

- **DeepSeek failed, FinBERT succeeded**



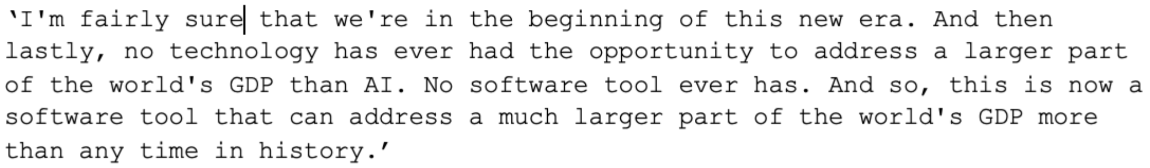
'And so we do see, you know, good momentum across the board. And as well, you know, we continue to see, you know, increasing engagement across the customer base, across all of the service offerings, both transacting and paid accounts. We talked about, you know, reaching all-time highs. And we have over now a billion paid subscriptions across the services platform.'

Figure 6.2: Apple: Text Extract

Conversely, DeepSeek often struggled with highly optimistic yet technical financial language that FinBERT interprets more effectively. In Apple's call, for example, DeepSeek classified a statement as only Slightly Positive, whereas the true label was Strong Positive.

FinBERT’s exposure to domain-specific financial expressions such as ‘all-time highs’ and ‘billion paid subscriptions’ enables it to capture the full intensity of such positive market signals.

- **Both DeepSeek and FinBERT failed**



‘I'm fairly sure that we're in the beginning of this new era. And then lastly, no technology has ever had the opportunity to address a larger part of the world's GDP than AI. No software tool ever has. And so, this is now a software tool that can address a much larger part of the world's GDP more than any time in history.’

Figure 6.3: Nvidia: Text Extract

For Nvidia, both models struggled with long technical passages that blended neutral descriptions with forward-looking optimism. While manual annotators consistently labelled these as Strong Positive due to their visionary tone, FinBERT frequently assigned Neutral and DeepSeek only Slightly Positive. As shown in the excerpts above, both models understate the optimism and future-oriented sentiment that humans readily perceive. This highlights the particular difficulty models face in capturing visionary language

6.2 Where sentiment diverges from market movement?

This section explores cases where transcript-based sentiment diverges from actual market reactions, highlighting the limits of language-only analysis. Model sentiment scores reflect average quarterly outputs, which generally cluster between Neutral and Slightly Positive (0–1). By contrast, market sentiment is derived from stock price movements around earnings announcements and mapped onto the same five-bin scale, ranging from +2 (Strong Positive) to –2 (Strong Negative).

6.2.1 Amazon

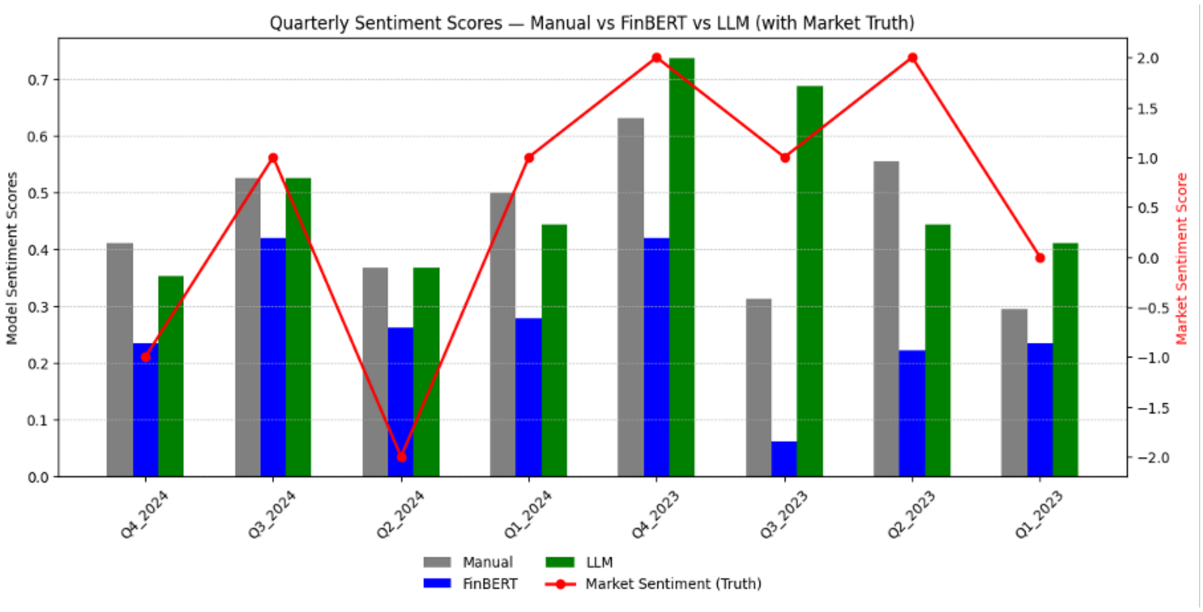


Figure 6.4: Amazon: Comparing Model Sentiment with Market Truth

In Q2 2024, all three models produced broadly similar sentiment scores from Amazon's earnings call, reflecting mild positivity. Yet the market reacted negatively after CFO Brian Olsavsky warned that 'consumers are continuing to be cautious with their spending, trading down' (Reuters, 2024). This illustrates that transcript-based sentiment alone is an incomplete predictor of market movements, as external factors and forward-looking guidance can decisively shape investor reactions.

6.2.2 Apple

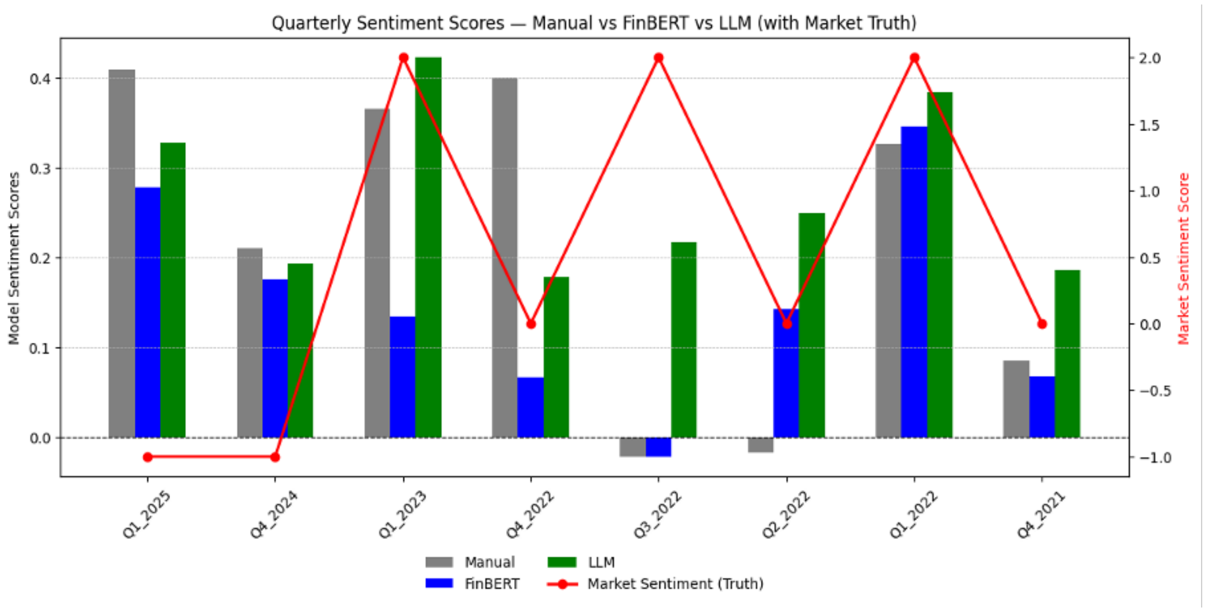


Figure 6.5: Apple: Comparing Model Sentiment with Market Truth

In Q4 2024, Apple's earnings call conveyed mildly positive sentiment across all models, yet the market reacted negatively in response to weak forward guidance and supply-chain pressures in China. Wider industry challenges, including elevated freight rates and strained logistics noted by UNCTAD (2024), further explain investor caution. This divergence illustrates that call tone does not always translate directly into price movements, as external factors and macro trends often carry greater weight. It also highlights management's ability to frame information strategically, softening negative signals in ways that may mask underlying concerns

6.2.3 Nvidia

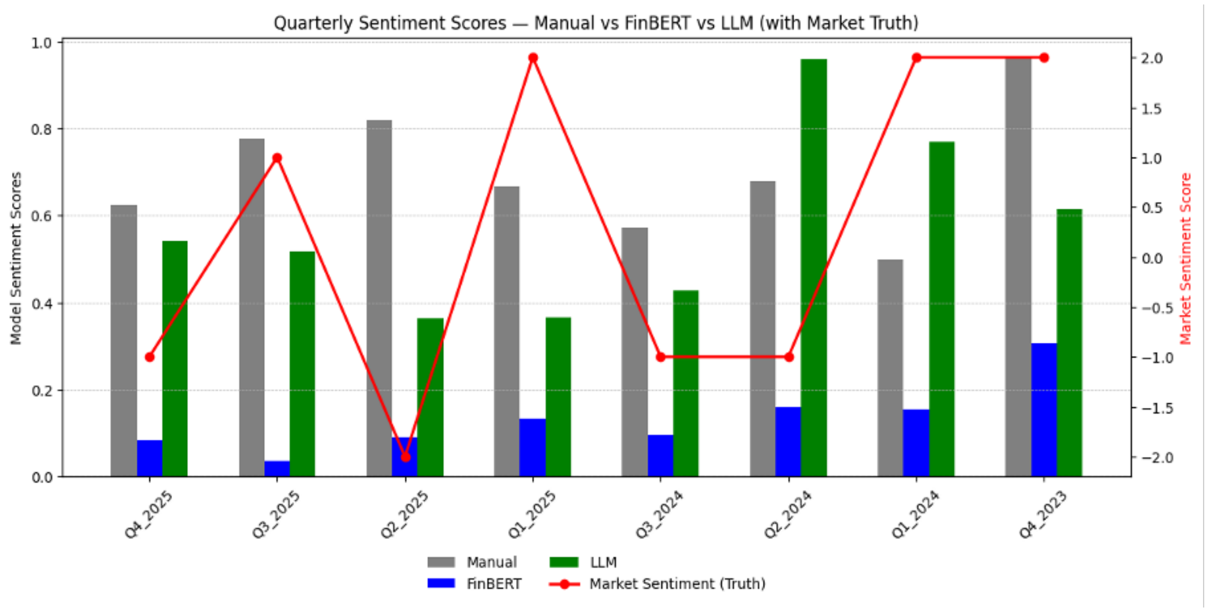


Figure 6.6: Nvidia: Comparing Model Sentiment with Market Truth

Q2 2025 illustrates the limits of transcript-based sentiment analysis. Manual annotations (0.82), DeepSeek (0.36), and FinBERT (0.07) all signalled a positive tone in Nvidia's call, yet the stock dropped sharply (Strong Negative, –2.0) in the two-day window. The decline was triggered by exogenous factors—Super Micro Computer, a key Nvidia partner, delayed its annual report, causing a 19% sell-off and sparking fears of an AI bubble (Wall Street Journal, 2024). This case highlights how external shocks can override transcript sentiment, constraining the predictive power of language-based models.

Company	**Linguistic Fidelity** (Manual Annotation)	**Predictive Validity** (Market Sentiment)
Amazon	FinBERT	DeepSeek
Apple	FinBERT	DeepSeek
NVIDIA	DeepSeek	FinBERT

Table 6.1: Summary of Model Performance

Table 6.1 summarises the comparative results across benchmarks. FinBERT demonstrates stronger alignment with manual annotations for Amazon and Apple, reflecting higher linguistic fidelity, while DeepSeek outperforms in capturing market sentiment for both firms. In contrast, Nvidia shows the opposite pattern: DeepSeek aligns more closely with manual sentiment labels, but FinBERT better predicts market reactions

7 Limitations

Manual sentiment annotation was performed by multiple team members, introducing subjectivity, particularly in distinguishing nuanced categories such as ‘Strong Positive’ versus ‘Slightly Positive.’ These inconsistencies may have influenced benchmarking outcomes when comparing models to human judgments. Best practices for subjective labelling recommend multiple annotators assessing the same segments and using majority agreement or multi-annotator modelling to capture consensus while mitigating individual bias (Davani et al., 2022). This would improve consistency in benchmark labels and reduce the influence of individual perspectives.

LLM outputs are inherently probabilistic, meaning repeated runs on identical inputs can produce different sentiment classifications even with fixed prompts. This non-determinism complicates direct comparisons with human annotations. Prior work shows such variation is common (Yadkori et al., 2024), and recommends providing richer contextual information to the model to reduce epistemic uncertainty and improve reliability. One way to provide richer contextual information is by including speaker roles and identities (e.g., CEO, CFO, Analyst), which can help the model distinguish between financially material guidance and analyst questions. Although this metadata was available in the dataset, it was not incorporated into the LLM prompts

8 Conclusion

This study evaluated the effectiveness of two approaches to financial sentiment analysis: FinBERT, a domain-specific NLP model, and DeepSeek, a general-purpose large language model, using earnings call transcripts from Amazon, Apple, and Nvidia. The objective was to assess which model best aligns with human-annotated sentiment (linguistic fidelity) and which better reflects investor reactions through short-term stock price movements (predictive validity).

The findings show a clear divergence between these criteria. FinBERT outperformed DeepSeek against manual annotations for Amazon and Apple, highlighting its strength in recognising nuanced financial expressions such as caution around spending or enthusiasm about subscriptions. However, DeepSeek proved more effective for Nvidia, where FinBERT often defaulted to Neutral and underplayed subtler cues, showing that specialisation is not consistently advantageous across contexts.

By contrast, DeepSeek aligned more closely with market sentiment for Amazon and Apple, while FinBERT performed better for Nvidia, where technical language directly shaped investor responses. This highlights the importance of sectoral context: consumer-facing firms are influenced by demand expectations and macro narratives, while technical firms display a closer link between call tone and stock price movements.

For Amazon, DeepSeek captured market-consistent sentiment even though FinBERT was more accurate linguistically, while manual annotations also correlated with price, suggesting efficient information processing. Apple revealed a larger gap, with positive call tone diverging from negative price reactions, reflecting the role of expectations and external factors. Nvidia showed the opposite pattern: FinBERT's financial lexicon enabled it to parse technical content such as chip cycles and data centre sales, aligning with both human labels and market outcomes.

Overall, the results demonstrate that there is no single best model. FinBERT excels at transcript-faithful classification, whereas DeepSeek better anticipates investor behaviour. Model selection should therefore depend on the intended objective and industry context.

This study contributes by showing how financial-domain and generalist LLMs complement one another, and future work could enhance performance by integrating contextual cues such as speaker roles or guidance metrics into model evaluation.

9 Recommendations

Use Both Models Complementarily
FinBERT and DeepSeek should be deployed together to capture both transcript-faithful sentiment and likely market response. FinBERT can strengthen internal reporting and audit processes by ensuring accuracy in interpreting financial language, while DeepSeek's outputs may offer greater value for portfolio managers seeking trading signals and anticipating investor behaviour. Using both models in tandem provides a balanced perspective that supports decision-making across research and trading functions.

Future Research on Tuned LLMs
Given that an untuned DeepSeek model already demonstrated comparable performance to FinBERT in several cases, future research into a domain-tuned DeepSeek could plausibly compete directly with specialised models such as FinBERT. Fine-tuning may narrow the gap in linguistic fidelity while retaining DeepSeek's broader contextual advantages, making it a potential standalone alternative for financial sentiment analysis.

Sector-Specific Strategy
Model selection should reflect industry context. FinBERT proves more effective for technical sectors such as semiconductors, where precise financial terminology dominates, while DeepSeek is better suited to consumer-facing firms where macro narratives and investor psychology play a stronger role. Adopting a sector-tailored strategy maximises predictive accuracy and ensures resources are applied where each model adds the most value.

10 Further Work

An important direction for future research is the integration of cultural weighting into speaker-level sentiment analysis to improve cross-market accuracy. Brockman et al. (2015) show that cultural and market contexts influence which speakers carry the most weight: for example, Hong Kong markets often react more strongly to analysts, whereas U.S. markets prioritise CEOs. This suggests identical sentiment can have different market impact depending on role and culture. A role-weighted sentiment scheme could therefore improve alignment between transcript sentiment and observed market moves, potentially raising explanatory power (e.g., higher R^2 scores). For instance, in QA sessions, analyst questions could be weighted less heavily than management responses.

A complementary extension would be to incorporate news-headline sentiment as an external benchmark of market mood. Headlines around earnings release dates capture macroeconomic and geopolitical influences beyond the transcript itself. As Doey and de Jong (2025) highlight, call tone often shapes subsequent media coverage, meaning headline sentiment not only reflects market reactions but also reframes corporate communication for wider audiences. Incorporating this benchmark would help estimate firm-level sensitivity to transcript sentiment and identify companies whose prices are more responsive to calls versus external narratives.

Finally, future work could involve rational generation from LLMs. Concise explanations of why a sentiment label was assigned could enhance interpretability and highlight systematic model biases. This would not only improve transparency but also help practitioners trust and refine model outputs in real-world applications.

Bibliography

[1] Araci, D.T., 2019. FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. MSc thesis. University of Amsterdam. Available at: `https://arxiv.org/abs/1908.10063` (Accessed: 4 August 2025).

[2] Brockman, P., Li, X. and Price, S.M., 2015. Differences in conference call tones: Managers vs. analysts. *Financial Analysts Journal*, 71(4), pp.24–42.

[3] Chen, Z., Gössi, S., Kim, W., Bermeitinger, B. and Handschuh, S., 2023. FinBERT-FOMC: Fine-Tuned FinBERT Model with Sentiment Focus Method for Enhancing Sentiment Analysis of FOMC Minutes. In *Proceedings of the 4th ACM International Conference on AI in Finance*, pp. 357–364.

[4] Davani, A.M., Díaz, M. and Prabhakaran, V., 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10, pp.92–110.

[5] Doey, B. and de Jong, P., 2025. How negative tones in earnings calls shape media narratives. *Review of Behavioral Finance*, 17(3), pp.406–423.

[6] Du, K., Xing, F., Mao, R. and Cambria, E., 2024. Financial sentiment analysis: Techniques and applications. *ACM Computing Surveys*, 56(9), pp.1–42.

[7] Fama, E.F., 1970. Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance*, 25(2), pp.383–417.

[8] Fidelity, 2024. The Magnificent Seven stocks: why they matter to investors. Available at: `https://www.fidelity.com/learning-center/smart-money/magnificent-7-stocks` (Accessed: 29 July 2025).

[9] Guest, N.M., 2021. The information role of the media in earnings news. *Journal of Accounting Research*, 59(3), pp.1021–1076.

[10] Kang, T., Park, D.H. and Han, I., 2018. Beyond the numbers: The effect of 10-K tone on firms' performance predictions using text analytics. *Telematics and Informatics*, 35(2), pp.370–381.

[11] Kang, J.W. and Choi, S.Y., 2025. Comparative investigation of GPT and FinBERT's sentiment analysis performance in news across different sectors. *Electronics*, 14(6), p.1090.

[12] Kirtac, K. and Germano, G., 2024. Sentiment trading with large language models. *Finance Research Letters*, 62, p.105227.

[13] Konstantinidis, T., Iacovides, G., Xu, M. and Mandic, D., 2024, November. Finllama: LLM-based financial sentiment analysis for algorithmic trading. In *Proceedings of the 5th ACM International Conference on AI in Finance*, pp.134–141.

[14] Krause, D., 2025. DeepSeek's Potential Impact on the Magnificent 7: A Valuation Perspective. Available at SSRN 5117909.

[15] McGugan, I., 2024. The Magnificent Seven are more affordable than you might think. *Globe & Mail (Toronto, Canada)*, pp.B1–B1.

[16] Putri, P.A.R., Prasetiyowati, S.S. and Sibaroni, Y., 2023. The performance of the equal-width and equal-frequency discretization methods on data features in classification process. *Sinkron: Jurnal dan Penelitian Teknik Informatika*, 7(4), pp.2082–2098.

[17] Reuters, 2024. Amazon projects quarterly revenue below estimates. 1 August. Available at: `https://www.reuters.com/technology/amazon-projects-quarterly-revenue-below-estimates-2024-08-01/` (Accessed: 20 August 2025).

[18] UNCTAD, 2024. High freight rates strain global supply chains, threaten vulnerable economies. 13 March. Available at: `https://unctad.org/news/high-freight-rates-strain-global-supply-chains-threaten-vulnerable-economies` (Accessed: 20 August 2025).

[19] Wall Street Journal, 2024. Nvidia earnings and stock market live coverage. 28 August. Available at: `https://www.wsj.com/livecoverage/nvidia-earnings-stock-market-today-08-28-2024` (Accessed: 20 August 2025).

[20] Yadkori, Y.A., Kuzborskij, I., György, A. and Szepesvári, C., 2024. To believe or not to believe your LLM. *arXiv preprint*, arXiv:2406.02543.

11 Appendix

Financial Stock Price Filtering and Dynamic Charting

As part of this project, the financial stock market dataset was transformed into an interactive chart with a date-range filter, allowing users to set custom start and end dates to visualise stock price movements. This functionality supports different trading styles, from long-term investors tracking trends to intraday traders focusing on short-term shifts. Figure 12.1 illustrates data spanning 01-01-2024 to 01-01-2025.

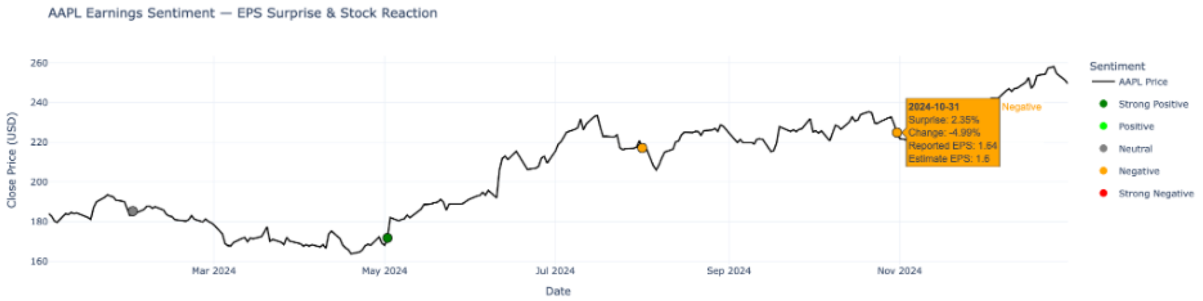


Figure 11.1: Apple Market Sentiment Chart

The tool also highlights intraday price movements on earnings release days, providing a focused view of how markets respond in real time to company announcements. Due to limitations of the yfinance package, intraday data is available only for the most recent 730 days. This feature directly supports the project's aim of linking sentiment analysis to market behaviour: by isolating key earnings days, we can benchmark model-generated sentiment more precisely against actual price reactions. For example, Figure 12.2 presents a granular intraday view for 31-10-2024 (Q3 earnings release, categorised as negative sentiment), including movements in the preceding and following trading sessions (±1 day).

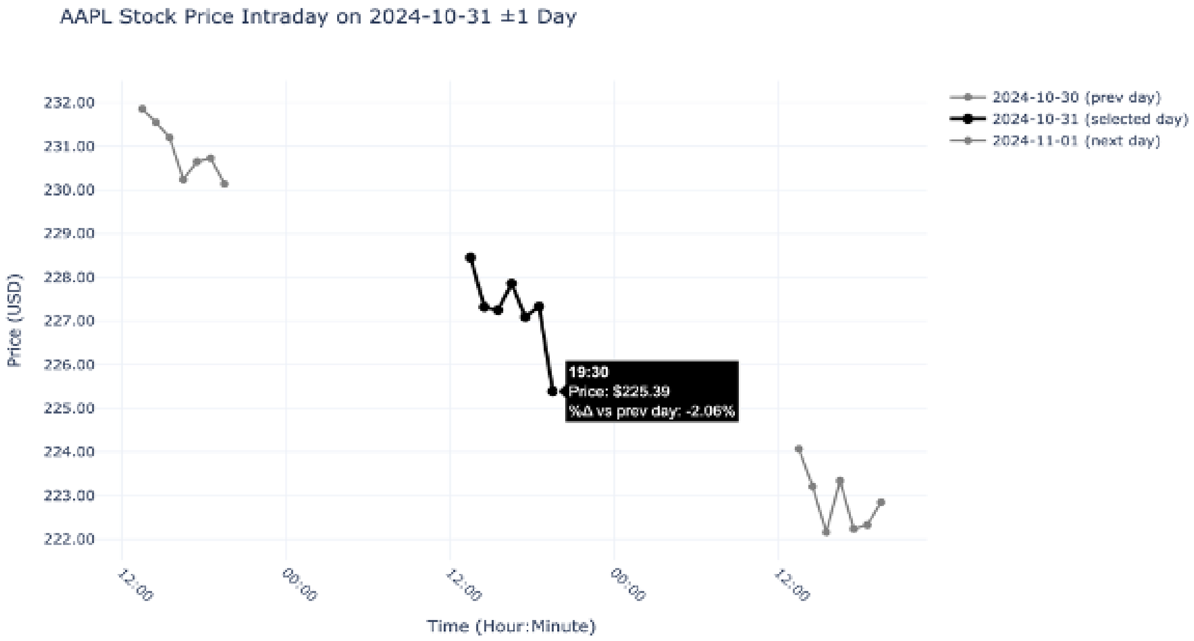


Figure 11.2: Apple Intraday Dynamic Chart