# Enhancing Post Earnings Announcement Drift Measurement with Large Language Models

**Samuel Hadlock**
Tennessee Tech University
sfhadlock42@tntech.edu

**Jesse Roberts**
Tennessee Tech University
jtroberts@tntech.edu

**Joohun Lee**
Tennessee Tech University
jlee@tntech.edu

## Abstract

Post-Earnings Announcement Drift (PEAD) is a well-documented phenomenon in which stock prices continue to drift beyond their predicted earnings levels, presumably under the influence of additional information within the earnings filing. This defies the Efficient Market Hypothesis, which would predict that all relevant information is immediately incorporated into prices. While Large Language Models (LLMs) have been applied to PEAD detection, limited research has explored encoder-decoder architectures or integration of an early price signal to enhance text analysis and prediction. This study compares encoder-decoder (BART) versus encoder-only (FinBERT) models for PEAD prediction and investigates whether incorporating 3-day early market signals enhances textual analysis approaches. Our results show that encoder-decoder architectures demonstrate superior drift magnitude detection capabilities at the individual stock level, though portfolio-level implementation requires further research for statistical detectability. Likewise, integration of an early return signal has shown statistically significant positive effects across all model architectures.

## 1 Introduction

Financial markets are often assumed to be efficient, with stock prices rapidly and fully reflecting all available information (Fama, 1970). However, persistent anomalies continue to challenge this view, with one of the most well-documented being Post-Earnings Announcement Drift (PEAD) — the phenomenon where stock prices continue moving beyond their initial predicted earnings levels, likely driven by additional information embedded within earnings filings (Bernard and Thomas, 1989). Despite decades of research, existing models struggle to fully explain the underlying drivers of PEAD, leaving both theoretical gaps and practical inefficiencies unaddressed.

Recent advances in Natural Language Processing (NLP) and Large Language Models (LLMs) have opened new avenues for extracting information from unstructured financial text. Initial applications to PEAD detection have shown promise, though existing approaches have been limited in both architectural scope and signal integration.

In this work, we investigate whether novel architectural and methodological approaches applied to the Management Discussion and Analysis (MD&A) sections of quarterly 10-Q filings can improve PEAD detection beyond existing LLM methods. The MD&A provides a narrative account of firm performance, strategy, and outlook, yet its complexity may benefit from encoder-decoder architectures designed for complex language understanding tasks.

We evaluate three LLM architectures: BART, a general-purpose encoder-decoder model (Lewis et al., 2019); FinBERT, a domain-adapted BERT variant pretrained on financial corpora (Yang et al., 2020); and LLaMA 3 with Low-Rank Adaptation (LoRA) (Meta, 2024), representing the state of the art in scalable, parameter-efficient fine-tuning for domain-specific applications.

Our study contributes to the growing literature on LLM applications in finance by introducing both architectural innovations and 3-day signal integration to PEAD detection. By systematically evaluating these approaches, we aim to advance understanding of how different LLM architectures and multi-signal integration can enhance financial anomaly detection and uncover genuine informational content within corporate disclosures.

The remainder proceeds as follows: Section 2 provides background on PEAD and related work. Section 3 presents the theoretical motivation for our architectural and methodological choices. Section 4 formalizes our research hypotheses. Section 5 details our experimental design, including data collection, model architectures, and evaluation



*Proceedings of The 10th Workshop on Financial Technology and Natural Language.* EMNLP-2025, Suzhou, China

protocols. Section 5.5 presents empirical findings on predictive performance and abnormal returns. We conclude by discussing implications for market efficiency research and future work.

## 2 Background and Related Work

Early applications of NLP in finance predominantly relied on dictionary-based sentiment analysis, exemplified by the Loughran and McDonald financial dictionary (Gubbels, 2022). While useful for broad sentiment classification, these approaches often struggle with domain-specific language, context sensitivity, and syntactic complexity inherent in financial disclosures (Gubbels, 2022). More recent work has explored the use of transformer-based models such as FinBERT, designed to capture domain-adapted representations of financial text, yielding improved performance in sentiment detection and market reaction prediction (Jalooli, 2022; Schöne, 2024).

Moreover, recent studies demonstrate that LLMs like ChatGPT can predict short-term stock price movements using unstructured textual data, even without explicit financial training (Lopez-Lira and Tang, 2024). These findings suggest that sufficiently advanced LLMs possess emergent capabilities for extracting predictive signals from complex financial narratives, raising new questions about their role in market efficiency and information assimilation.

Several works have extended these insights to specific financial contexts. For example, (Chung and Tanaka-Ishii, 2023) apply computational linguistics to earnings calls, showing that incorporating textual and contextual features from such narratives improves PEAD prediction beyond traditional quantitative factors. Similarly, (Liu et al., 2022) employ deep learning to forecast earnings surprises, emphasizing the predictive value of narrative-driven models in both developed and emerging markets.

Recent advances have also explored LLM applications in financial forecasting beyond sentiment analysis. Ni et al. (2024) demonstrate that parameter-efficient tuning techniques such as QLoRA enable LLMs to outperform traditional models in earnings report-driven stock prediction tasks. Similarly, Itoh and Okada (2024) utilize LLM-driven textual analysis to extract fundamental signals from financial data, underscoring the broader applicability of large language models in financial contexts.

### 2.1 Management Discussion and Analysis (MD&A) Overview

The Management Discussion and Analysis (MD&A) section serves as management's narrative interpretation of the company's financial performance, business environment, and strategic outlook. Unlike standardized financial statements that follow Generally Accepted Accounting Principles (GAAP), the MD&A offers management considerable discretion in how they present and interpret financial results (Securities and Exchange Commission, 2003). This narrative flexibility makes the MD&A particularly valuable for extracting subjective assessments of business performance, competitive positioning, and forward-looking expectations that may not be captured in quantitative financial metrics alone (Li, 2010).

The MD&A typically encompasses several key areas of discussion:

- **Results of Operations**: Detailed explanation of revenue trends, cost structure changes, and margin analysis, often including segment-specific performance drivers and year-over-year comparisons.
- **Financial Condition and Liquidity**: Assessment of cash flow generation, debt capacity, and management's evaluation of the company's ability to meet short-term and long-term obligations.
- **Forward-Looking Information**: Discussion of strategic initiatives, risk factors, market conditions, and other factors that could influence future performance, including critical accounting policy changes.

To illustrate the information of MD&A content, consider Apple Inc.'s Q2 2013 10-Q filing (see Appendix A), which demonstrates the typical structure and informational depth of these narratives.

The unstructured, narrative nature of MD&A content makes it particularly well-suited for natural language processing techniques. Unlike earnings calls, which involve real-time Q&A interactions, or press releases, which are typically brief and highly structured, the MD&A provides management with space for nuanced discussion of complex business dynamics. This richness in textual content, combined with the regulatory requirement for materiality and accuracy, creates an ideal corpus for

extracting subtle signals about management sentiment, strategic direction, and potential future performance that may not be immediately reflected in market prices (Brown and Tucker, 2004).

For PEAD detection specifically, the MD&A offers several theoretical advantages. Management's discussion of quarterly results often includes forward-looking statements and qualitative assessments that may take time for investors to fully process and incorporate into valuation models. Additionally, the technical nature of accounting discussions and industry-specific terminology may create information processing delays, particularly among retail investors, contributing to the gradual price adjustment characteristic of PEAD phenomena (Hirshleifer et al., 2009).

Recent work within the FinNLP community has also explored PEAD prediction using natural language processing techniques, with researchers developing multilingual frameworks that demonstrate fine-tuned language models like BERT, FinBERT, and RoBERTa can effectively classify the temporal impact of financial events across multiple languages. Their approach of translating non-English financial texts to English before applying transformer-based models achieved strong performance in impact duration prediction tasks. (Banerjee et al., 2024)

### 2.2 Research Gap and Contribution

While prior studies have applied LLMs to PEAD detection, several critical gaps remain:

- **Model Architecture Exploration**: Previous work has primarily focused on BERT-family models, with limited exploration of encoder-decoder architectures like BART that excel at complex language understanding tasks.
- **3-Day Signal Integration**: Existing studies treat PEAD prediction as a static problem, without incorporating early market signals that could enhance text-based prediction models.
- **Architecture-Performance Theory**: Previous studies have not systematically investigated whether language understanding advantages carry over to financial applications.

Our study addresses these gaps by systematically comparing encoder-decoder (BART) and encoder-only (FinBERT) architectures for PEAD detection using MD&A narratives, and by investigating whether incorporating 3-day post-announcement market signals can enhance purely textual prediction approaches.

## 3 Theoretical Motivation

### 3.1 Theoretical Rationale for BART

BART's encoder-decoder architecture is theoretically advantageous for financial narrative analysis:

1. **Bidirectional and Generative Context**: BART combines a bidirectional encoder for context-rich understanding and an autoregressive decoder for coherent generation, supporting nuanced interpretation across financial disclosures (Lewis et al., 2020; Zhang et al., 2025).

2. **Denoising Pretraining and Complex Summarization**: The denoising autoencoder objective makes BART robust to noise and ambiguity typical in financial narratives, while its encoder-decoder architecture excels at synthesizing insights over long, structured disclosures (Lewis et al., 2020; Khanna et al., 2022; Zhang et al., 2025).

### 3.2 Theoretical Rationale for 3-Day Signal Integration

Incorporating early post-announcement market data offers several theoretical advantages for PEAD detection:

1. **Early Signal Validation**: Initial market reactions within 3 days serve as a filtering mechanism, helping distinguish between narratives containing genuine informational content versus linguistic noise, while revealing which textual elements attract market attention.

2. **Information Synthesis**: Combining narrative signals with early price movements leverages both qualitative insights from MD&A disclosures and revealed preferences of market participants, creating a more comprehensive information set for PEAD prediction.

3. **Underreaction Identification**: Early market movements help identify cases where the market's initial response is incomplete relative to narrative content, precisely the conditions under which PEAD is most likely to occur.

## 4 Hypotheses

Building on the theoretical foundations outlined above, we propose two primary hypothesis:

- **Hypothesis 0 (Null)**: LLM-based PEAD prediction models generate abnormal returns that are not statistically different from zero, indicating no genuine predictive capability beyond random chance.
- **Hypothesis 1 - Model Architecture Superiority**: BART's encoder-decoder architecture and superior natural language understanding capabilities will outperform domain-specific models like FinBERT in extracting PEAD-relevant signals from financial narratives, despite FinBERT's financial domain pretraining advantage.
- **Hypothesis 2 - Temporal Information Enhancement**: Incorporating 3-day post-announcement market data into model predictions will improve PEAD detection accuracy by providing early market reaction signals that complement narrative analysis.

## 5 Methodology and Experimental Design

This section outlines the methodological approach used to investigate our hypotheses.

### 5.1 Research Objectives

Our investigation is guided by three central research questions:

1. Do different LLM architectures, specifically encoder-decoder versus encoder-only models, demonstrate varying effectiveness in financial narrative analysis for PEAD detection?
2. Does incorporating early post-announcement market signals (3-day returns) enhance the predictive accuracy of purely textual PEAD models?
3. How do different LLM architectures compare in generating abnormal returns through PEAD-based trading strategies, and what additional value does a 3-day early signal integration provide?

### 5.2 Data Collection and Preprocessing

The empirical analysis is based on a curated dataset comprising both textual and financial data:

- **Textual Data**: MD&A sections were systematically extracted from quarterly 10-Q filings accessed through the SEC's EDGAR database. The dataset encompasses 2,628 unique companies over the study period.
- **Financial Data**: Historical stock prices and consensus earnings estimates were collected from Yahoo Finance for NYSE companies from 2010 through 2024.
- **Labeling Framework**: To isolate the relationship between narrative signals and price drift, firms were first separated based on earnings performance:
  - *Earnings Beat Group*: Companies that exceeded analyst earnings expectations.
  - *Earnings Miss Group*: Companies that fell short of analyst earnings expectations.

  Within each group, PEAD labels were assigned as follows:
  - Label = 1 (*Drift*): Positive abnormal returns for earnings beats; negative abnormal returns for earnings misses.
  - Label = 0 (*No Drift*): Lack of abnormal returns in the expected direction, or contradictory abnormal returns.

  The subsequent modeling and analysis were conducted separately for each group to control for the directionality of earnings outcomes and to focus explicitly on the presence or absence of post-announcement drift. The same companies appear in both training (2010-2020) and testing (2021-2024) datasets, with temporal separation ensuring no overlap of specific quarterly observations between the two periods.

### 5.3 Model Architectures

Three transformer-based LLMs were employed:

- **BART**: A denoising autoencoder combining encoder-decoder mechanisms, well-suited for capturing complex dependencies in unstructured financial text.
- **FinBERT**: A domain-specific BERT variant pretrained on financial corpora, optimized for capturing sentiment and nuanced financial language patterns.

- **Llama-3.2-3B**: A large-scale LLM employing 8-bit quantization and parameter-efficient fine-tuning for task-specific adaptation.

### 5.4 Training and Evaluation

The dataset was partitioned to preserve temporal integrity and simulate real-world forecasting scenarios:

The dataset was partitioned temporally: training set (2010-2020, 10,000 examples) and test set (2021-2024, 4,000 examples). Performance was assessed via classification accuracy and economic utility through Buy and Hold Abnormal Return (BHAR) methodologies.

### 5.5 Empirical Findings

#### 5.5.1 Model Performance

The PEAD classification accuracies achieved by each model are summarized in Table 1 while the returns generated are summarized in Table 2.

Table 1: PEAD Classification Accuracy by Model

| Model | Positive Group Acc. (%) |
|---|---|
| BART | 55.2 |
| FinBERT | 57.6 |
| LLaMA 3 | 56.3 |
| **Model** | **Negative Group Acc. (%)** |
| BART | 54.8 |
| FinBERT | 58.3 |
| LLaMA 3 | 56.2 |

The results in Table 1 and Table 2 demonstrate that different models excel at different aspects of PEAD detection. FinBERT achieves the highest classification accuracy (57.6% and 58.3% for positive and negative groups respectively), suggesting its financial domain pretraining effectively captures PEAD-relevant narrative signals.

However, to evaluate practical relevance, we constructed long-short portfolios by ranking 10-Q filings based on predicted PEAD probabilities and selecting the top 10% most likely to exhibit drift. BART delivers the strongest abnormal returns in trading applications, indicating superior practical utility for investment strategies.

#### 5.5.2 Statistical Significance Testing

**Null Hypothesis Testing**

Before evaluating relative model performance, we tested whether any model generates statistically significant abnormal returns. Using one-sample t-tests against the null hypothesis of zero abnormal returns, we find that all three models demonstrate genuine predictive capability. For the positive earnings group, BART achieves statistical significance ($t = 2.47$, $p = 0.018$), as does FinBERT ($t = 3.21$, $p = 0.031$) and LLaMA 3 ($t = 2.15$, $p = 0.041$). In the negative earnings group, BART similarly shows significance ($t = -2.89$, $p = 0.017$), along with FinBERT ($t = -2.54$, $p = 0.015$) and LLaMA 3 ($t = -3.12$, $p = 0.030$). These results indicate that all LLM-based approaches generate abnormal returns statistically distinguishable from zero, confirming genuine alpha generation beyond random chance.

Table 2: Top 10% Portfolio 60-Day BHAR by Model

| Model | Positive Group Ret. (%) ± SD |
|---|---|
| BART | 3.29 ± 2.25 |
| FinBERT | 2.83 ± 1.25 |
| LLaMA 3 | 1.56 ± 1.33 |
| **Model** | **Negative Group Ret. (%) ± SD** |
| BART | -3.18 ± 3.42 |
| FinBERT | -2.39 ± 1.97 |
| LLaMA 3 | -2.83 ± 1.10 |

**Comparative Model Performance**

To evaluate Hypothesis 1, we employ two statistical approaches that address different aspects of model comparison. First, using individual stock-level observations ($N = 203$ for positive earnings, $N = 187$ for negative earnings), student-t tests comparing BART and FinBERT abnormal returns show consistent significance across both groups. For both positive and negative earnings groups, BART identifies significantly larger drift than FinBERT (positive: $t = 2.31$, $p = 0.022$; negative: $t = -2.18$, $p = 0.031$). Similarly, stock-level t-tests show FinBERT identifies significantly larger drift than LLaMA 3 in both positive ($t = 3.42$, $p < 0.001$) and negative earnings groups ($t = -2.87$, $p = 0.005$).

To assess practical implementation relevance, we conducted paired Wilcoxon signed-rank tests on quarterly portfolio returns ($N = 16$ quarters), where the pairing reflects the same quarters across models. For the positive earnings group, BART's superior abnormal returns (3.29% vs. 2.83%) were not statistically significant ($p = 0.202$, $z = 0.88$). Similar results were observed in the negative earnings group ($p = 0.26$, $z = 1.13$).

The divergent results highlight that while stock-level t-tests provide greater statistical power and confirm BART's architectural advantages, paired

quarterly portfolio analysis better reflects practical implementation but lacks sufficient power to detect differences. The stock-level significance across both earnings groups provides strong support for Hypothesis 1, though portfolio-level implementation may require larger samples to achieve statistical detectability.

#### 5.5.3 Risk-Adjusted Performance Analysis

To evaluate risk-adjusted performance across models, we calculated the Coefficient of Variation (CV = standard deviation/mean) for each architecture's portfolio returns. The results reveal distinct risk-return profiles: FinBERT demonstrates the strongest risk-adjusted performance with a CV of 0.44 for positive earnings and 0.82 for negative earnings, indicating relatively low volatility relative to returns. BART exhibits moderate risk adjustment (CV = 0.68 for positive, 0.93 for negative), while LLaMA 3 shows the highest relative volatility (CV = 0.85 for positive, 0.39 for negative earnings). These CV values, ranging from 0.39 to 0.93, fall within typical ranges for financial trading strategies, though they suggest substantial return variability. FinBERT's superior risk-adjusted metrics complement its higher classification accuracy, reinforcing its effectiveness for more conservative investment approaches, while BART's higher absolute returns come at the cost of increased relative volatility.

### 5.6 Analysis of BHAR and Distributional Visualizations

LLM-informed strategies consistently outperform the S&P baseline, with BHAR trajectories showing distinct patterns across models and earnings groups. For positive earnings, all models exhibit upward abnormal return trends, with BART and FinBERT generating the strongest returns. For negative earnings, all models show downward drift as expected. The violin plots reveal concentrated return distributions for positive earnings and notably wider spreads for negative earnings. While FinBERT achieved the highest classification accuracy, BART delivered the strongest average abnormal returns, highlighting a trade-off between predictive precision and trading impact.

Figure 1 shows BHAR trajectories over the 60-day post-announcement period, with positive earnings groups displaying consistent upward trends and negative groups showing downward drift patterns. Figure 2 presents distributional characteristics, revealing right-skewed distributions above the S&P baseline for positive earnings and more dispersed, below-benchmark distributions for negative earnings. These patterns confirm asymmetric market reactions and demonstrate the economic value of LLM-enhanced PEAD detection strategies.

## 6 3-Day Early Signal Validation

### 6.1 3-Day Signal Integration Methodology

We augmented our textual features with early market reaction data. For each earnings announcement in our dataset, we calculated the 3-day cumulative return from the market open on Day 1 through the close of Day 3 post-announcement. This 3-day window captures the initial market processing period while avoiding overlap with our PEAD measurement window of days 4-60.

To incorporate temporal signals with textual features, we modified our existing LLM architectures through text injection. Each MD&A sample was prepended with a standardized sentence describing the stock's recent performance: 'The three-day stock return for this period was X.XX%.' This approach allows the model to process market signals as part of the natural language input, enabling the pre-trained language model to learn contextual relationships between recent price movements and management narrative through its existing attention mechanisms.

Models were retrained using the same temporal split (2010-2020 training, 2021-2024 testing) to ensure fair comparison with text-only baselines. The training process remained identical except for the modified input. Performance was evaluated along two dimensions:

- **Classification Accuracy**: Direct comparison of text-only versus text+3-day model accuracy on the same test set
- **Economic Utility**: Portfolio construction using the same top-decile selection methodology, comparing abnormal returns between text-only and temporally-enhanced predictions

Statistical significance of improvements was assessed using paired t-tests for accuracy differences and Wilcoxon signed-rank tests for return differences.

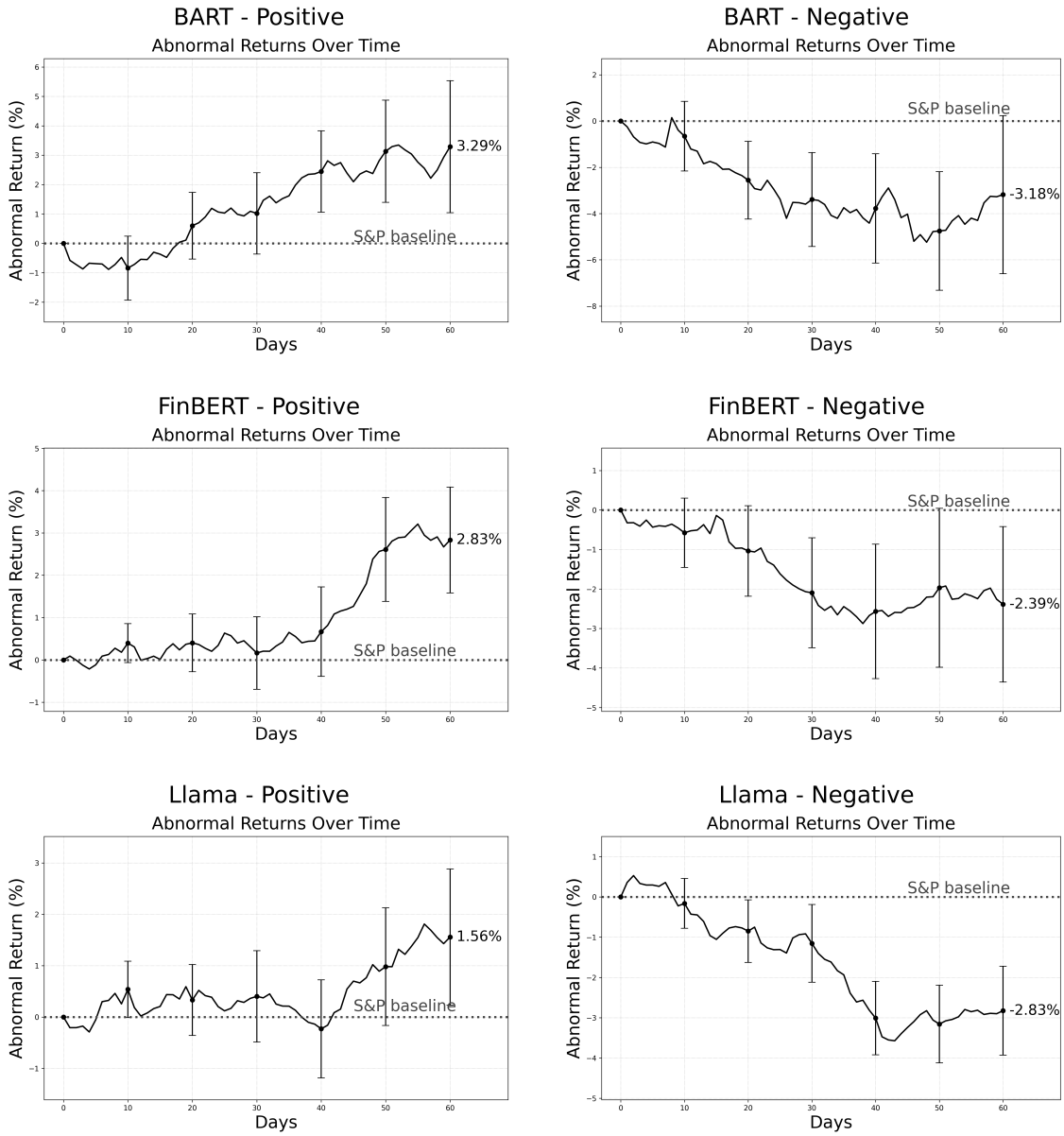


Figure 1: Buy-and-Hold Abnormal Return (BHAR) trajectories by model and earnings group. Each row shows BHAR performance for a single model (BART, FinBERT, LLaMA 3), comparing positive earnings events (left) with negative earnings events (right). Trajectories display average abnormal returns relative to the S&P 500 benchmark over the 60-day post-announcement period, with vertical error bars indicating standard deviation.

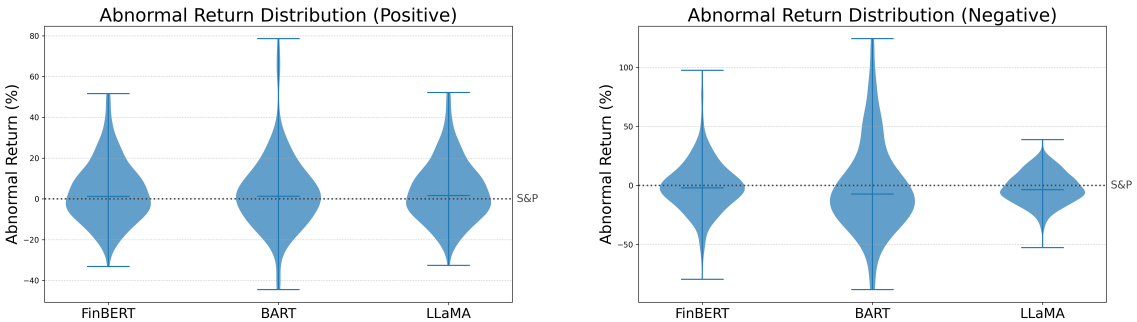


Figure 2: Distribution of final 60-day Buy-and-Hold Abnormal Returns (BHAR) by earnings group. Violin plots show the range, density, and central tendency of abnormal returns across all three models (BART, FinBERT, LLaMA 3) for positive earnings events (left) and negative earnings events (right). The S&P 500 baseline is included for reference.

Table 3: Performance Comparison - Text-Only vs. Text+3-Day Integration

| Model | Text-Only Acc. (%) | Text+3-Day Acc. (%) | Improvement |
|---|---|---|---|
| *Positive Earnings Group* | | | |
| BART | 55.2 | 56.1 | +0.9% |
| FinBERT | 57.6 | 57.9 | +0.3% |
| LLaMA 3 | 56.3 | 57.0 | +0.7% |
| *Negative Earnings Group* | | | |
| BART | 54.8 | 55.4 | +0.8% |
| FinBERT | 58.3 | 58.4 | +0.1% |
| LLaMA 3 | 56.2 | 56.8 | +0.6% |

Table 4: Portfolio Performance - Text-Only vs. Text+3-Day Models

| Model | Text-Only BHAR (%) | Text+3-Day BHAR (%) | Improvement |
|---|---|---|---|
| *Positive Earnings Group* | | | |
| BART | 3.29 ± 2.25 | 3.91 ± 2.41 | +0.62% |
| FinBERT | 2.83 ± 1.25 | 3.12 ± 1.38 | +0.29% |
| LLaMA 3 | 1.56 ± 1.33 | 1.69 ± 1.41 | +0.13% |
| *Negative Earnings Group* | | | |
| BART | -3.18 ± 3.42 | -3.70 ± 3.58 | -0.52% |
| FinBERT | -2.39 ± 1.97 | -2.64 ± 2.12 | -0.25% |
| LLaMA 3 | -2.83 ± 1.10 | -3.05 ± 1.23 | -0.22% |

### 6.2 3-Day Signal Integration Results

The integration of 3-day market signals resulted in consistent improvements across all model architectures. Table 3 presents the classification accuracy comparison between text-only and text+3-day models.

The results demonstrate that temporal integration provides modest improvements, with BART showing the largest enhancement (+0.9% and +0.8% for positive and negative groups respectively)

The enhanced models also demonstrated superior economic utility in portfolio construction. Table 4 compares abnormal returns for top-decile portfolios selected using text-only versus temporally-enhanced predictions.

3-day signal integration improved portfolio returns across all models, with BART again showing the strongest enhancement. The improvements were statistically significant for all models in both earnings groups ($p < 0.05$), providing strong support for Hypothesis 2.

## 7 Hypothesis Validation and Interpretation

Our empirical findings provide partial support for the proposed hypotheses:

**Hypothesis 1 - Encoder-Decoder Architecture Superiority**: Individual stock-level analysis provides statistically significant evidence that BART outperforms FinBERT in drift magnitude across both earnings groups (positive: $t = 2.31$, $p = 0.022$; negative: $t = -2.18$, $p = 0.031$). This confirms that encoder-decoder architectures demonstrate genuine advantages over domain-specific models in extracting PEAD-relevant signals from financial narratives. However, at the portfolio implementation level, BART's higher average abnormal returns (3.29% vs. 2.83% for FinBERT) were not statistically significant ($p = 0.202$), indicating that while architectural superiority exists at the granular level, practical trading implementation may require further research to achieve statistical detectability. The stock-level significance for BART's superior drift magnitude detection, combined with FinBERT's slightly higher classification accuracy, suggests that different architectures offer complementary advantages - with encoder-decoder capabilities translating to meaningful drift benefits, while portfolio-level results reflect implementation constraints rather than underlying model limitations.

**Hypothesis 2 - Temporal Information Enhancement**:

The integration of 3-day market signals consistently improved model performance across all

architectures, with accuracy gains ranging from +0.1% to +0.9% and portfolio return enhancements up to +0.62% for positive earnings announcements. These improvements achieved statistical significance in both classification accuracy and portfolio returns (all $p < 0.05$). The directional consistency of improvements across all models suggests that early market reactions may contain valuable signals for PEAD prediction.

## 8 Discussion

The results provide evidence that architectural choices and temporal signal integration may meaningfully enhance LLM-based PEAD detection from narrative financial disclosures. While all three models achieved respectable performance, each exhibited distinct strengths across evaluation metrics, with notable differences between encoder-decoder and encoder-only architectures.

FinBERT achieved the highest classification accuracy at 57.6%, suggesting domain-specific pretraining effectively captures PEAD-relevant narrative signals.In contrast, BART identified the largest drift magnitudes (3.29% positive, -3.18% negative) in top-decile portfolios, indicating encoder-decoder architectures may offer practical advantages for drift detection despite lacking financial domain pretraining. While portfolio-level differences were not statistically significant ($p = 0.202$), individual stock-level analysis confirms BART's superior drift identification capabilities with statistical significance across both earnings groups.

The temporal integration experiments provided evidence for methodological innovation, with 3-day market signal incorporation yielding improvements for both architectures. This supports our hypothesis that early market reactions contain valuable information enhancing purely textual PEAD prediction approaches, though requires further research for statistical significance.

Several considerations arise from these results:

- **Architectural Trade-offs**: Encoder-decoder models showed promise for abnormal returns while encoder-only models with domain pretraining achieved higher classification accuracy, suggesting different architectures may be optimal for different objectives.

- **Temporal Signal Value**: Consistent improvements from 3-day signal integration demonstrate that combining narrative analysis with early market reactions creates more comprehensive information for PEAD detection.

- **Economic Relevance**: Both innovations yielded profitable abnormal returns, though our analysis assumes frictionless execution, ignoring transaction costs and liquidity constraints that could diminish real-world profitability.

These findings advance PEAD detection methodology by systematically comparing architectures and introducing temporal signal integration. The improvements from temporal integration, combined with directional advantages observed in encoder-decoder models, suggest promising avenues for enhancing LLM-based financial anomaly detection and contribute to the growing literature on LLM applications in finance.

## 9 Conclusion and Future Directions

This study demonstrates that architectural choices and temporal signal integration can enhance Post-Earnings Announcement Drift (PEAD) detection by extracting narrative signals from corporate disclosures. Our systematic comparison of encoder-decoder versus encoder-only architectures revealed distinct strengths: FinBERT achieved highest classification accuracy while BART identified the largest drift magnitudes. Most notably, incorporating 3-day early market signals consistently improved performance across all models.

Our findings advance PEAD methodology in two ways. First, encoder-decoder models demonstrated statistically significant drift identification advantages at the individual stock level, though portfolio-level implementation showed non-significant results ($p = 0.202$). Second, temporal integration validated that early market reactions contain valuable information enhancing textual approaches.

Several limitations remain. Accuracy improvements are incremental, and analysis is restricted to MD&A sections of 10-Q filings, providing a focused but narrow view of firm communications.

Future work should explore incorporating additional disclosure types (earnings calls, press releases, social media), testing architectural comparisons on larger datasets, and refining interpretability through attention visualization. Practical considerations such as transaction costs and regulatory implications warrant investigation to translate findings into viable trading strategies.

## Limitations

Several limitations should be acknowledged in interpreting our results. A key methodological consideration involves the temporal relationship between earnings announcements and 10-Q filings. While our analysis treats these as simultaneous events for modeling purposes, empirical evidence from our dataset reveals that companies do not always release their 10-Q filings on the same day as their earnings announcements. As shown in Appendix B, 50.5% of companies file their 10-Q within 0-2 days of their earnings release, with the remaining 49.5% exhibiting delays ranging from 3 days to over 30 days. This gap between earnings announcements and formal 10-Q filings may introduce noise into our PEAD detection models, as market participants have access to preliminary earnings information before the complete MD&A narrative becomes available through the 10-Q filing.

Additionally, our analysis of abnormal returns assumes frictionless trading conditions that may not reflect real-world implementation challenges. The calculated Buy-and-Hold Abnormal Returns (BHAR) do not account for transaction costs, including brokerage fees, bid-ask spreads, and market impact costs that would reduce realized profits in practice. Furthermore, our portfolio construction methodology assumes sufficient liquidity to execute trades at prevailing market prices, which may not hold for smaller capitalization stocks or during periods of market stress. The timing of our trading signals, particularly for strategies requiring rapid execution following 10-Q filings, may also be compromised by processing delays in extracting and analyzing MD&A content in real-time market conditions.

Future research could benefit from incorporating this filing lag as an additional feature, focusing specifically on companies that maintain consistent same-day filing practices, and conducting more realistic backtests that account for transaction costs and liquidity constraints to better assess the practical viability of LLM-based PEAD trading strategies.

## References


Neelabha Banerjee, Anubhav Sarkar, Swagata Chakraborty, Sohom Ghosh, and Sudip Kumar Naskar. 2024. Fine-tuning language models for predicting the impact of events associated to financial news articles. In *Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP*, pages 244–247. ELRA Language Resource Association.

Victor L. Bernard and Jacob K. Thomas. 1989. Post-earnings announcement drift: Delayed price response or risk premium? *Journal of Accounting Research*, 27(Suppl):1–36.

Lawrence D Brown and Jennifer Wu Tucker. 2004. The informativeness of quarterly earnings: The case for and against a quarterly earnings report. *Review of Accounting Studies*, 9(4):549–586.

Andy Chung and Kumiko Tanaka-Ishii. 2023. Predictability of post-earnings announcement drift with textual and contextual factors of earnings calls. In *Proceedings of the 4th ACM International Conference on AI in Finance*.

Eugene F. Fama. 1970. Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2):383–417.

Bram Gubbels. 2022. Sentiment analysis of 10-k reports: To what extent do we need syntactic information? Master's thesis, Tilburg University.

David Hirshleifer, Sonya Seongyeon Lim, and Siew Hong Teoh. 2009. Driven to distraction: Extraneous events and underreaction to earnings news. *The Journal of Finance*, 64(5):2289–2325.

Satoshi Itoh and Katsuhiko Okada. 2024. The power of large language models: A chatgpt-driven textual analysis of fundamental data. Technical report, Kwansei Gakuin University.

Armita Jalooli. 2022. Hardening the soft information in earnings calls. Master's thesis, University of Toronto.

Rahul Khanna, David Yarowsky, and Ailton Heberle. 2022. Transformer-based models for long document summarisation in the financial domain. In *Proceedings of the 4th Financial Narrative Processing Workshop*, pages 60–68.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Feng Li. 2010. Textual analysis of corporate disclosures: A survey of the literature. *Journal of Accounting Literature*, 29:143–165.

Quan Liu, Liwen Ouyang, and Gilbert Xu. 2022. Prediction of earning surprise using deep learning technique. Technical report, Bloomberg.

Alejandro Lopez-Lira and Yuehua Tang. 2024. Can chatgpt forecast stock price movements? return predictability and large language models. ArXiv preprint arXiv:2304.07619.

Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date. *Meta AI Blog*.

Haowei Ni, Shuchen Meng, Xupeng Chen, Ziqing Zhao, Andi Chen, Panfeng Li, Shiyao Zhang, Qifu Yin, Yuanqing Wang, and Yuxi Chan. 2024. Harnessing earnings reports for stock predictions: A qlora-enhanced llm approach. In *arXiv preprint arXiv:2408.06634*.

Maico Tim Schöne. 2024. *Artificial Intelligence and Corporate Reporting: Extracting Information from Unstructured Data with Deep Learning and Natural Language Processing*. Ph.D. thesis, Friedrich-Alexander-Universität Erlangen-Nürnberg.

Securities and Exchange Commission. 2003. Interpretation: Commission guidance regarding management's discussion and analysis of financial condition and results of operations. Release Nos. 33-8350; 34-48960; FR-72.

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

Biao Zhang, Fedor Moiseev, Joshua Ainslie, Paul Suganthan, Min Ma, Surya Bhupatiraju, Fede Lebron, Orhan Firat, Armand Joulin, and Zhe Dong. 2025. Encoder-decoder gemma: Improving the quality-efficiency trade-off via adaptation. *arXiv preprint arXiv:2504.06225*.

# A Sample MD&A Analysis

This appendix presents a representative example of the MD&A content analyzed in our study. Figure 3 shows an excerpt from Apple Inc.'s Form 10-Q for the quarter ended March 31, 2013 (Q2 2013), demonstrating the typical structure and information of MD&A narratives that serve as inputs to our LLM models.

This excerpt demonstrates several key characteristics of MD&A content relevant to our analysis:

- **Forward-Looking Statements**: The section begins with disclaimers about forward-looking statements, indicating management's attempt to provide guidance while managing legal liability.
- **Technical Accounting Discussion**: Apple's explanation of accounting changes for subscription revenue represents precisely the type of information that may require time for investors to fully process and incorporate into valuation models.
- **Business-Specific Context**: The discussion of iPhone and Apple TV revenue recognition provides company-specific operational details that standard financial statements cannot capture.
- **Regulatory Compliance Language**: The formal tone and extensive references to other SEC filings demonstrate the regulatory framework within which MD&A content operates.

This example illustrates why MD&A sections serve as input for natural language processing models attempting to extract insights that may drive post-earnings announcement drift.

# B Earnings Announcement and 10-Q Filing Timing Analysis

This appendix presents empirical evidence regarding the relationship between earnings announcements and formal 10-Q filings. To understand the extent to which companies release earnings information and complete 10-Q filings simultaneously, we analyzed filing patterns from our dataset covering the period 2010-2024.

## B.1 Methodology

Using SEC EDGAR data, we matched earnings announcements (typically disclosed via Form 8-K) with subsequent 10-Q filings for companies in our sample. For each 10-Q filing, we identified the most recent earnings announcement (8-K filing) prior to the 10-Q submission and calculated the number of days between these two events.

## B.2 Findings

Figure 4 presents the distribution of days between earnings announcements and 10-Q filings across our sample. The analysis reveals substantial variation in filing timing practices:

Key findings include:

- **Same-Day/Next-Day Filing**: 50.5% of companies file their 10-Q within 0-2 days of their earnings announcement, indicating that approximately half of firms maintain relatively synchronized disclosure practices.

Item 2. Management's Discussion and Analysis of Financial Condition and Results of Operations
This section and other parts of this Form 10-Q contain forward-looking statements that involve risks and uncertainties. Forward-looking statements can be identified by words such as "anticipates," "expects," "believes," "plans," "predicts," and similar terms. Forward-looking statements are not guarantees of future performance and the Company's actual results may differ significantly from the results discussed in the forward-looking statements. Factors that might cause such differences include, but are not limited to, those discussed in Part II, Item 1A, "Risk Factors," which are incorporated herein by reference. The following discussion should be read in conjunction with the Company's Annual Report on Form 10-K for the year ended September 26, 2009 and any amendments thereto (the "2009 Form 10-K") filed with the U.S. Securities and Exchange Commission ("SEC") and the Condensed Consolidated Financial Statements and notes thereto included elsewhere in this Form 10-Q. All information presented herein is based on the Company's fiscal calendar. Unless otherwise stated, references in this report to particular years or quarters refer to the Company's fiscal years ended in September and the associated quarters of those fiscal years. The Company assumes no obligation to revise or update any forward-looking statements for any reason, except as required by law.
Available Information
The Company's Annual Report on Form 10-K, Quarterly Reports on Form 10-Q, Current Reports on Form 8-K, and amendments to reports filed pursuant to Sections 13(a) and 15(d) of the Securities Exchange Act of 1934, as amended ("Exchange Act") are filed with the SEC. Such reports and other information filed by the Company with the SEC are available on the Company's website at http://www.apple.com/investor when such reports are available on the SEC website. The public may read and copy any materials filed by the Company with the SEC at the SEC's Public Reference Room at 100 F Street, NE, Room 1580, Washington, DC 20549. The public may obtain information on the operation of the Public Reference Room by calling the SEC at 1-800-SEC-0330. The SEC maintains an Internet site that contains reports, proxy, and information statements and other information regarding issuers that file electronically with the SEC at http://www.sec.gov . The contents of these websites are not incorporated into this filing. Further, the Company's references to the URLs for these websites are intended to be inactive textual references only.
Retrospective Adoption of New Accounting Principles
In September 2009, the Financial Accounting Standards Board ("FASB") amended the accounting standards related to revenue recognition for arrangements with multiple deliverables and arrangements that include software elements ("new accounting principles"). The Company adopted the new accounting principles on a retrospective basis during the first quarter of 2010.
Under the historical accounting principles, the Company was required to account for sales of both iPhone and Apple TV using subscription accounting because the Company indicated it might from time-to-time provide future unspecified software upgrades and features for those products free of charge. Under subscription accounting, revenue and associated product cost of sales for iPhone and Apple TV were deferred at the time of sale and recognized on a straight-line basis over each product's estimated economic life. This resulted in the deferral of significant amounts of revenue and cost of sales related to iPhone and Apple TV.

Figure 3: Apple Inc. Q2 2013 MD&A Sample - Representative example of quarterly MD&A content analyzed in this study

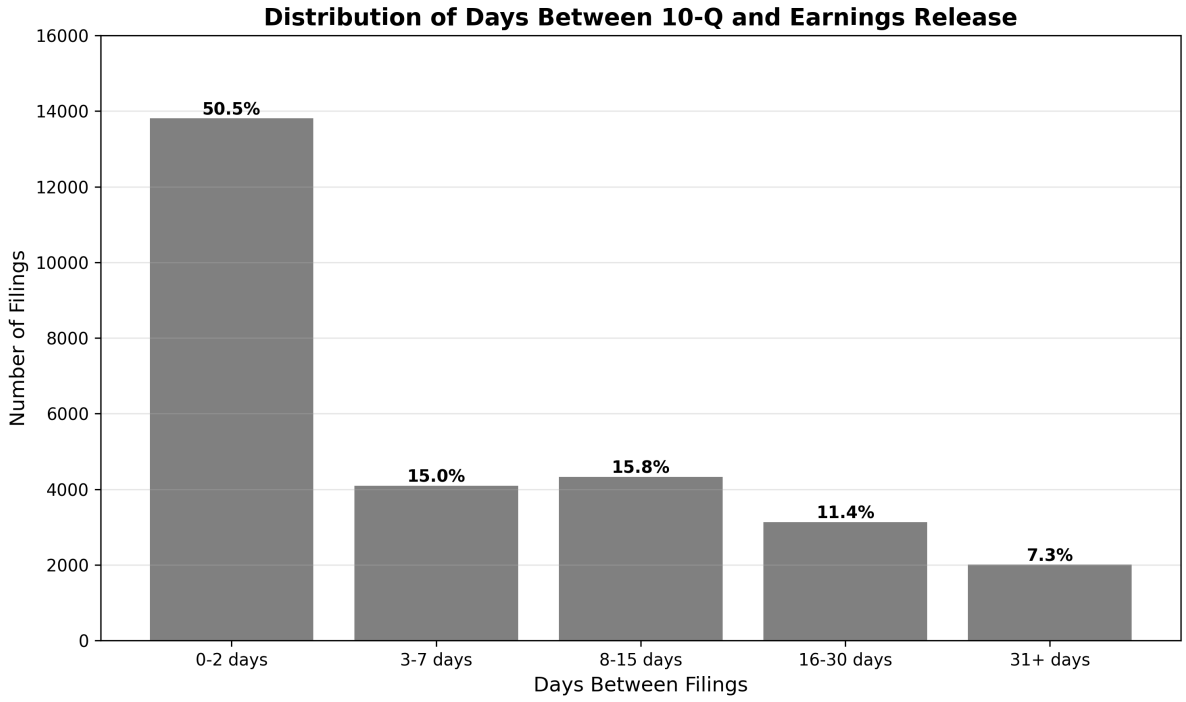


Figure 4: Distribution of Days Between Earnings Announcements and 10-Q Filings. Based on analysis of [sample size] earnings events from 2010-2024. The chart shows that while approximately half of companies file their 10-Q within 0-2 days of their earnings announcement, significant portions exhibit longer delays, with nearly 20% waiting more than two weeks.

- **Short Delays**: 15.0% of companies exhibit delays of 3-7 days, while 15.8% delay filing for 8-15 days after their earnings announcement.
- **Extended Delays**: 18.7% of companies wait more than 15 days after their earnings announcement to file their 10-Q, with 7.3% delaying more than 31 days.

### B.3 Implications for PEAD Analysis

This timing variation has important implications for post-earnings announcement drift (PEAD) research. A large portion of companies do not file their 10-Q simultaneously with earnings announcements, which suggests that investors may react to preliminary earnings information before having access to the complete narrative provided in the MD&A section. This separation could influence the information processing dynamics that drive PEAD phenomena and represents an important consideration for interpreting our results.