Financial News and Short-Horizon Market Reactions: FinBERT Baselines on IV-Normalized Price Movements and Volatility

Columbia COMS4705 Final Project Report

Keywords: *financial NLP, FinBERT, implied volatility, intraday prediction*

Yifei Fu
Department of Computer Science
Columbia University
yf2711@columbia.edu

Zhonghao Liu
Department of Computer Science
Columbia University
zl3593@columbia.edu

Shilun Dai
Department of Computer Science
Columbia University
sd3839@columbia.edu

Abstract

Text models are increasingly deployed in financial markets, including ultra-low-latency systems where news feeds can trigger trades at millisecond timescales. In this project we ask a simpler but fundamental question: given realistic minute-level news and price data, how well can a pretrained financial language model, FinBERT, predict short-horizon market reactions when labels are carefully normalized by option-implied volatility?

We build a dataset of news for the 50 largest U.S. equities by market capitalization (S&P 500 constituents), aligned to 1–60 minute post-news returns and 5–60 minute realized volatility, and scale both by previous-day 30-day at-the-money (ATM) implied volatility using either an approximate straddle price (for price movements) or direct IV normalization (for volatility). Our first stage evaluates a *head-only* FinBERT classifier on straddle-normalized price movements across multiple horizons and label granularities (2/3/7-class). Despite clear improvements over a zero-shot FinBERT sentiment baseline, normalized price moves remain very hard to predict, with accuracies only modestly above chance. Motivated by the fact that signed returns at minute horizons are driven by many non-text factors, we shift to IV-normalized realized volatility and add a lightweight LoRA adapter on top of FinBERT. On these volatility tasks, LoRA-tuned FinBERT delivers consistent and non-trivial gains over the zero-shot baseline (e.g., roughly +10 percentage points for 30-minute 2-class volatility), suggesting that IV-normalized volatility is a more stable and learnable signal of "how much" the market reacts to news than the signed return itself.

1 Key Information

- **Mentor:** Daniel Zhang
- **External Collaborators:** None
- **Sharing project:** Yes (course staff and future COMS4705 students)

2 Introduction

Text-based prediction of asset price movements is a classic topic at the intersection of NLP and finance. A large body of work uses news or social media to predict next-day returns, earnings surprises, or medium-horizon volatility [Tetlock, 2007, Engelberg and Parsons, 2011, Ke et al., 2019, Ding et al., 2015]. High-frequency trading firms operate specialized infrastructures to react to machine-readable news at millisecond timescales, colocated with exchanges and using proprietary NLP pipelines. In contrast, our setting is more realistic for many institutional investors: we work at 1–60 minute horizons on standard news APIs and minute bars, and ask what is achievable with off-the-shelf financial language models and modest fine-tuning.

Option-implied volatility (IV) and at-the-money (ATM) straddle prices provide a natural scale for how large a price move is "expected" to be. IV-based measures are standard tools for studying volatility risk premia and option richness [Carr and Wu, 2009, Bollerslev et al., 2009, Hull, 2018, Gatheral, 2006], but are less commonly treated as a design choice for *text prediction labels*. In particular, scaling realized returns or realized volatility by IV yields dimensionless targets that are more comparable across tickers, which should in principle make the learning problem easier for a single text model.

Concretely, we ask three questions:

1. How strong is the *zero-shot* FinBERT baseline on short-horizon price-movement and volatility prediction when labels are defined purely from price and IV data?
2. If we only change the *classification head* of FinBERT and train it on straddle-normalized price movements, how much can we improve across horizons and class granularities?
3. Does switching the target from normalized price movements to IV-normalized volatility, and adding a small LoRA adapter, lead to more robust gains?

Our workflow is: (1) construct IV-based labels for price movements and volatility; (2) benchmark the pretrained FinBERT sentiment head as a zero-shot classifier; (3) fine-tune a head-only FinBERT classifier on straddle-normalized price movements; (4) build a LoRA-tuned FinBERT classifier on IV-normalized realized volatility. We find that: (i) directional price movements at minute horizons remain extremely hard to predict, with both zero-shot and head-only FinBERT performing only modestly above chance; (ii) head-only classification on normalized price movements yields limited but consistent gains without fundamentally changing this difficulty; and (iii) IV-normalized volatility is substantially more predictable, with LoRA bringing meaningful improvements. Overall, our results highlight volatility—rather than direction—as a more promising and operationally realistic target for news-based intraday prediction.

3 Related Work

Text and asset returns. Tetlock [2007] show that negative media tone predicts temporary underperformance in the Dow Jones index, while Engelberg and Parsons [2011] use local newspapers to establish a causal impact of news on retail trading and returns. Subsequent work scales to firm-level news and social media, using sentiment or event features to explain cross-sectional returns and earnings surprises [Ke et al., 2019, Ding et al., 2015], mostly at daily frequency or around discrete event windows such as earnings announcements. These papers demonstrate that text contains economically meaningful information, but they typically evaluate at relatively coarse horizons and in unnormalized return space. They do not ask how well text can predict *minute-level* reactions, nor do they explicitly control for cross-sectional differences in volatility or option-implied risk. Our work can be viewed as a next step in this line: we keep the high-level idea of return/volatility prediction from text, but move to short horizons and introduce IV-based label design in order to isolate what a language model can learn about intraday reactions.

Pretrained language models in finance. Domain-adapted transformers such as FinBERT [Araci, 2019, Yang and Matthews, 2020] and BloombergGPT [Wu et al., 2023] extend BERT-style architectures to financial text and achieve strong performance on supervised tasks including sentiment, risk disclosure, and question answering. Most applications in this literature focus on relatively long-horizon outcomes (e.g., quarterly earnings surprises, credit ratings, or daily volatility) and treat the models as general-purpose encoders that are fine-tuned on downstream labels. Importantly, the released FinBERT models are trained on sentence-level corpora such as Financial PhraseBank and do *not* condition explicitly on ticker symbols or firm-specific context, which limits their ability to capture heterogeneity across names. Existing work rarely analyzes (i) how the original FinBERT sentiment head behaves as a zero-shot baseline on new targets such as short-horizon returns or volatility, or (ii) how much incremental gain comes from modest architectural changes (e.g., replacing only the classification head, or adding LoRA adapters) as opposed to full fine-tuning. By systematically comparing zero-shot FinBERT, head-only fine-tuning, and LoRA on IV-normalized minute-horizon labels, our study fills this gap and provides an analysis of when simple adaptations of financial LMs do—and do not—deliver meaningful predictive improvements.

4 Approach

We now describe our models and training objectives. Throughout, we use the ProsusAI FinBERT encoder as the base transformer.

4.1 Source-aware Transformer baseline

News quality varies substantially across providers, and some outlets (e.g., Reuters, Bloomberg) tend to produce short, factual summaries that are much easier to align with price moves and volatility. Rather than hard-filtering by source, we use these "good authority" outlets as *prototypes* of high-quality news summaries and train a small *source-aware transformer* to score the quality of each item.

Let D denote the text consisting of the source token and headline. We pass D through a small transformer encoder with parameters ψ to obtain

$$h = \mathrm{Transformer}_{\psi}(D) \in \mathbb{R}^{d}. \tag{1}$$

We then define a scalar score

$$f(D) = \sigma(h^{\top} g), \tag{2}$$

where $g \in \mathbb{R}^{d}$ are classifier parameters and $\sigma(\cdot)$ is the logistic sigmoid. Items from curated sources form a positive set $\mathcal{D}_{\text{auth}}$, and the remaining items form a background pool $\mathcal{D}$. We optimize

$$\min_{\psi, g} \mathbb{E}_{D\sim\mathcal{D}_{\text{auth}}}[-\log f(D)] - \mathbb{E}_{D\sim\mathcal{D}}[-\log f(D)], \tag{3}$$

so that $f(D)$ can be interpreted as a weakly-supervised "summary quality" score: high values indicate that the style and content of D resemble those of authoritative sources, even if the article itself comes from a different provider. In experiments we use this score to downweight or filter low-scoring summaries and as an additional baseline for robustness checks.

4.2 FinBERT encoder and notation

Given a news item with cleaned text x (headline + snippet), ticker i, and timestamp τ, we form a structured prompt

`[CLS] [TICKER] TICK [SRC] SRC [TEXT] SUMMARY [SEP]`,

where `TICK` is the ticker symbol, `SRC` the news source, and `SUMMARY` the text. Let θ denote all FinBERT parameters. The encoder output for the `[CLS]` token is

$$h_{[\mathrm{CLS}]} = \mathrm{FinBERT}_{\theta}(x) \in \mathbb{R}^{d}. \tag{4}$$

FinBERT includes a 3-way sentiment head (negative/neutral/positive).

4.3 Zero-shot FinBERT baseline

Our first baseline keeps FinBERT exactly as released and uses it in a purely *zero-shot* way, without seeing any of our labels during training. Given $h_{[\mathrm{CLS}]}$ from (4), the pretrained sentiment head produces

$$u = W^{(\mathrm{sent})} h_{[\mathrm{CLS}]} + b^{(\mathrm{sent})}, \qquad p = \mathrm{softmax}(u), \tag{5}$$

where $p = (p_{\text{neg}}, p_{\text{neu}}, p_{\text{pos}})$ are the negative / neutral / positive sentiment probabilities.

For price-movement tasks we use a *directional* score

$$s_{\text{ret}} = p_{\text{pos}} - p_{\text{neg}}, \tag{6}$$

which is large and positive when the news is strongly positive and large and negative when it is strongly negative. For volatility tasks we instead use a *non-neutrality* score

$$s_{\text{vol}} = 1 - p_{\text{neu}}, \tag{7}$$

which measures how confident FinBERT is that the news is not neutral (either positive or negative).

To turn these one-dimensional scores into $C \in \{2, 3, 7\}$ classes, we simply cut the score axis into quantile bins based on the *training* split. For $C = 3$ and $C = 7$ we use equal-probability bins (tertiles or septiles). For $C = 2$ we focus on the tails: we use the bottom 30% vs. top 30% of scores as "low" and "high", discarding the middle 40%. In this way, the released FinBERT sentiment head becomes a zero-shot multi-class classifier for our IV-based labels, with no additional parameters or fine-tuning.

4.4 Head-only FinBERT classifiers

Our first learned models keep the encoder θ fixed and replace the sentiment head with a task-specific linear classifier. For any classification task t (straddle-normalized price movements or IV-normalized volatility) and horizon/window index k, we define

$$\ell_k^{(t)} = W_k^{(t)} h_{[\text{CLS}]} + b_k^{(t)}, \qquad q_k^{(t)} = \text{softmax}\big(\ell_k^{(t)}\big), \tag{8}$$

where $q_k^{(t)} \in \mathbb{R}^C$ are class probabilities over labels derived from z_h (price movements) or ν_w (volatility), as defined in Section 5.1. Given a dataset $\mathcal{D}_k^{(t)} = \{(x_n, y_n^{(t)})\}$, we minimize the standard cross-entropy loss

$$\mathcal{L}_{t,k}^{\text{head}}(\theta, W_k^{(t)}, b_k^{(t)}) = -\mathbb{E}_{(x,y)\sim\mathcal{D}_k^{(t)}}\big[\log q_k^{(t)}(y \mid x)\big], \tag{9}$$

with θ frozen in all head-only runs. In practice we train head-only models for both straddle-normalized price movements and IV-normalized volatility.

4.5 LoRA-tuned FinBERT for volatility

To allow a slightly richer adaptation for volatility we add LoRA [Hu et al., 2022] adapters to the self-attention layers while keeping the original weights frozen. For each attention block we modify the query and value projections as

$$W_q' = W_q + A_q B_q, \tag{10}$$

$$W_v' = W_v + A_v B_v, \tag{11}$$

where $A_\cdot \in \mathbb{R}^{d\times r}$ and $B_\cdot \in \mathbb{R}^{r\times d}$ are trainable low-rank matrices (rank $r = 8$), and W_q, W_v are the frozen base weights. Let $\tilde{\theta}$ denote the union of all LoRA parameters and the volatility classification head $(W_k^{(\text{vol})}, b_k^{(\text{vol})})$. The encoder output $h_{[\text{CLS}]}(\tilde{\theta})$ now depends on the adapters, and we reuse the same softmax classifier as in (8) for volatility tasks. The LoRA objective for IV-normalized volatility is

$$\mathcal{L}_{\text{vol},k}^{\text{LoRA}}(\tilde{\theta}) = -\mathbb{E}_{(x,y)\sim\mathcal{D}_k^{(\text{vol})}}\big[\log q_k^{(\text{vol})}(y \mid x; \tilde{\theta})\big]. \tag{12}$$

We only apply LoRA to volatility tasks; price-movement experiments use the head-only models described above.

5 Experiments

5.1 Data

News and prices (Alpha Vantage). We obtain both news and prices from Alpha Vantage for the 50 largest U.S. equities by market capitalization. The news endpoint provides, for each item, a ticker symbol, timestamp, source, short textual summary, and a *relevance score* indicating how closely the article is related to a given ticker. According to the provider, this relevance score is generated by prompting large language models on the semantic relevance between the company and the news content; it is based on text and metadata only and does not use the stock price. We filter to English news, remove markup, and truncate the cleaned summary to 2,000 characters. There are around 22000 news available for training after filtration.

Example news item. To illustrate the structure of the raw news data, we show a representative item from our processed `news_df` table (fields include ticker, relevance, source, authority level, timestamp, summary, and URL):

> **Ticker:** MSFT
> **Relevance:** 0.600980
> **Source:** Benzinga (**authority_level:** 1)
> **Time:** 2025-11-12 18:39:51+00:00
> **Summary:** "Microsoft Corp (NASDAQ:MSFT) is accelerating ..."
> **URL:** `https://www.benzinga.com/markets/tech/25/11/...`

This format is typical: each news item contains ticker metadata, a model-provided semantic relevance score, a coarse authority category based on source reliability, a precise timestamp, and a short textual summary (headline + snippet). These structured fields allow us to align each news event with minute-level prices and with previous-day ATM implied volatility to construct the IV-normalized labels described above.

Intraday prices are taken from Alpha Vantage's 1-minute time-series endpoint and aligned to U.S. trading hours to construct minute-by-minute mid prices P_t for each ticker.

Implied volatility (Bloomberg). For options-implied information we use daily 30-day ATM implied volatilities for the same top-50 ticker universe, obtained from Bloomberg's listed options data. For each trading day we extract the 30-day ATM IV for each underlying, interpolate or forward-fill across missing calendar dates when necessary, and define $\sigma^{\text{ATM}}_{\text{prev}}(d, i)$ as the previous trading day's 30-day ATM IV for ticker i on date d.

Straddle-normalized price movements. For a news event at time τ and horizon $h \in \{1, 5, 30, 60\}$ minutes we define the simple return

$$r_h = \frac{P_{\tau+h} - P_\tau}{P_\tau}, \qquad \Delta P_h = P_{\tau+h} - P_\tau.$$

For an at-the-money straddle with maturity $T = 30$ days, the Black–Scholes approximation gives an expected straddle price

$$A^{\text{ATM}}_h \approx \sqrt{\frac{2}{\pi}}\, S_\tau\, \sigma \sqrt{\frac{t_h}{T}},$$

where $S_\tau = P_\tau$, σ is the previous-day 30-day ATM implied volatility, and t_h is the fraction of T represented by h minutes. For small moves we have $\Delta P_h \approx S_\tau r_h$, so the price movement measured in *units of an ATM straddle* is approximately

$$M_h \approx \frac{\Delta P_h}{A^{\text{ATM}}_h} \approx \frac{r_h}{\sqrt{\frac{2}{\pi}}\, \sigma \sqrt{\frac{t_h}{T}}}.$$

Let

$$c_h = \sqrt{\frac{2}{\pi}}\, \sigma \sqrt{\frac{t_h}{T}},$$

so that $M_h \approx r_h / c_h$. Our final label for price movement is the signed log "number of straddles":

$$z_h = \operatorname{sign}(r_h)\, \log \frac{|r_h|}{c_h}.$$

In implementation we compute z_h directly from returns r_h; the stock price S_τ cancels because both ΔP_h and A^{ATM}_h are proportional to S_τ.

Model	Tuned parameters	Learning rate	Batch (train/eval)	Epochs
Head-only (price move)	classifier only	2×10^{-5}	16 / 32	5
Head-only (volatility)	classifier only	2×10^{-5}	16 / 32	5
LoRA (volatility)	classifier + LoRA W_q, W_v	10^{-4}	16 / 32	5

Table 1: Training configurations for learned models. All runs use cross-entropy loss and the AdamW optimizer.

IV-normalized realized volatility. For a window length $w \in \{5, 30, 60\}$ minutes we anchor each news event to the first minute bar at or after τ, denoted t^*, and compute a realized volatility based on minute log returns $\Delta\ell_t = \log(P_t/P_{t-1})$:

$$\sigma_w^{\text{real}}(\tau, i) = \operatorname{Std}\big(\Delta\ell_{t^*}, \ldots, \Delta\ell_{t^*+w-1}\big). \tag{13}$$

We then define the IV-normalized volatility target

$$\nu_w = \frac{\sigma_w^{\text{real}}(\tau, i)}{\sigma_{\text{prev}}^{\text{ATM}}(d, i)}, \tag{14}$$

which measures "realized volatility in units of IV". We again clip ν_w to its $[1\%, 99\%]$ range.

Discretization. Both z_h and ν_w are discretized via empirical quantiles on the training split:

- 2-class: use the 30% and 70% quantiles, discard the middle 40%, and label the tails as "low" vs. "high" (balanced).
- 3-class: use 33% and 67% quantiles (three equiprobable bins).
- 7-class: split into seven equiprobable bins ($\approx 14.3\%$ each).

5.2 Evaluation method

For each configuration (task, horizon/window, number of classes) we split the labeled dataset into train/validation/test with ratios 80/10/10 using stratified sampling. Accuracy is our primary metric because the discretization yields approximately balanced classes. We also compute per-class precision/recall/F1 using `sklearn`'s `classification_report` (omitted for space).

5.3 Experimental details

All models are implemented in PyTorch using HuggingFace Transformers and the `peft` library for LoRA. We use the official `ProsusAI/finbert` weights and tokenizer without modification. Zero-shot FinBERT uses the frozen release model with no additional training. The learned models share a small set of hyperparameters summarized in Table 1.

5.4 Results

5.4.1 Straddle-normalized price movements

We first study straddle-normalized price movements z_h. Table 2 reports 2-class results across horizons, comparing zero-shot FinBERT to the head-only classifier.

Horizon	Zero-shot	Head-only	Δ
1 min	0.5233	0.5126	-1.07%
5 min	0.4867	0.4981	$+1.14\%$
30 min	0.5202	0.5393	$+1.91\%$
60 min	0.4973	0.5195	$+2.22\%$

Table 2: Straddle-normalized price-movement prediction (2-class). Test accuracy of zero-shot FinBERT sentiment vs. head-only FinBERT classifier.

At the 1-minute horizon, head-only fine-tuning on straddle-normalized price movements slightly underperforms the zero-shot sentiment head. From 5 to 60 minutes, the head-only classifier yields small but consistent gains (roughly 1–2 percentage points), suggesting that somewhat longer windows average out some noise and make the directional signal slightly more learnable. Nevertheless, absolute accuracies remain in the low-to-mid-50% range despite balanced labels.

Appendix Tables 6 and 7 report 3-class and 7-class results. For 3-class tasks, head-only FinBERT improves accuracy by roughly 0–2.8 percentage points, with the largest gains at the 30-minute horizon. For 7-class tasks, gains are larger in relative terms (about 2–4.8 points, strongest at 5–30 minutes) but overall accuracies remain below about 0.18.

5.4.2 IV-normalized volatility

We now turn to IV-normalized realized volatility ν_w, where both head-only and LoRA-tuned FinBERT models are trained. Here we move all volatility tables into the main text and focus on the most informative comparisons.

Two-class volatility tasks. Table 3 shows 2-class results (low vs. high IV-normalized volatility) for windows $w \in \{5, 30, 60\}$. The labels are derived from ν_w via the 30/70 quantiles.

Window	Zero-shot	Head-only	LoRA	Head-only − Zero	LoRA − Zero
5 min	0.5063	0.5370	**0.5732**	$+3.1\%$	$+6.7\%$
30 min	0.5134	0.5784	**0.6118**	$+6.5\%$	$+9.8\%$
60 min	0.5254	0.5515	**0.5894**	$+2.6\%$	$+6.4\%$

Table 3: IV-normalized volatility prediction (2-class). Test accuracy of zero-shot FinBERT sentiment, head-only FinBERT, and LoRA-tuned FinBERT.

Even without training, the zero-shot FinBERT sentiment scores achieve accuracies around 0.51–0.53, comparable to what we saw on normalized price movements. Replacing the sentiment head and training on IV-normalized volatility adds roughly 3–7 percentage points. Allowing LoRA adapters delivers the largest improvements: for the 30-minute window, accuracy jumps from 0.5134 to 0.6118, a gain of almost 10 percentage points. This is a substantially stronger effect than anything observed on price movements and indicates that IV-normalized volatility is much more aligned with what FinBERT can extract from news.

Three-class volatility tasks. Table 4 presents 3-class results, where labels correspond to low/medium/high IV-normalized volatility.

Window	Zero-shot	Head-only	LoRA	Head-only − Zero	LoRA − Zero
5 min	0.3289	0.3728	**0.3896**	$+4.4\%$	$+6.1\%$
30 min	0.3203	0.3893	**0.4076**	$+6.9\%$	$+8.7\%$
60 min	0.3320	0.3696	**0.4114**	$+3.8\%$	$+7.9\%$

Table 4: IV-normalized volatility prediction (3-class). LoRA consistently outperforms both zero-shot and head-only models.

The zero-shot baseline hovers around random-guess performance ($\approx 1/3$). Head-only FinBERT adds about 4–7 percentage points, while LoRA adds around 6–9 points over zero-shot. The 30-minute window again stands out as the sweet spot, reaching 0.408 accuracy.

Seven-class volatility tasks. Table 5 reports 7-class volatility results, where labels correspond to seven equiprobable bins of IV-normalized volatility.

Window	Zero-shot	Head-only	LoRA	Head-only − Zero	LoRA − Zero
5 min	0.1478	0.1644	**0.1696**	+1.7%	+2.2%
30 min	0.1358	0.1882	**0.1974**	+5.2%	+6.2%
60 min	0.1443	0.1770	**0.2043**	+3.3%	+6.0%

Table 5: IV-normalized volatility prediction (7-class). LoRA improves accuracy for all windows.

Absolute accuracies are lower in this 7-way setting, but the gains relative to the zero-shot baseline are still substantial: LoRA improves accuracy by about 2–6 percentage points, with the 60-minute window achieving the best performance (0.2043).

6 Analysis

The empirical results highlight a clear gap between price-direction and volatility prediction. On straddle-normalized price movements, both the zero-shot FinBERT sentiment head and the head-only classifiers deliver accuracies only slightly above chance, even with carefully designed labels. At minute horizons, intraday returns are dominated by order-flow, index moves, microstructure effects, and prior information, so the incremental signal from a single headline is extremely noisy. By contrast, IV-normalized volatility is substantially more learnable: for the 30-minute 2-class task, adding LoRA adapters on top of FinBERT improves accuracy from about 0.51 to 0.61, a gain of roughly 10 percentage points, with consistent improvements across 3- and 7-class settings as well. This suggests that the model is better at predicting the *size* of the reaction than its *sign*.

Price direction versus volatility. From a market-microstructure perspective, this pattern is natural. At short horizons the sign of the move is sensitive to transient order imbalances and market-wide shocks that may offset or even reverse the "natural" reaction implied by the text. The magnitude of volatility, however, is more closely tied to how surprising, complex, or contentious the information is. Because we scale by previous-day ATM IV, our labels ask whether realized volatility exceeds what was already priced into options. The fact that a modest LoRA adapter can exploit this signal indicates that FinBERT's latent representations already encode useful information about uncertainty and disagreement that becomes visible once the model is trained on volatility-based targets.

Case studies on IV-normalized volatility. To better understand what the model has learned, we examine two representative news items from the 30-minute 2-class task. Both are drawn from the test split and are correctly classified by the LoRA-tuned model.

The first headline is attached to a large-cap technology stock and reads: "On another difficult Friday afternoon for the stock market, the broad U.S. indexes ended with significant declines — with tech stocks standing out." The label for this item belongs to the high-volatility class (ν_{30} in the upper 30% tail), and the model predicts high volatility as well. The text describes a market-wide sell-off, with broad index declines and concentrated underperformance in the technology sector, rather than a firm-specific event. In such conditions, short-horizon realized volatility for an individual tech stock is driven largely by index and sector flows. The correct high-volatility prediction suggests that the LoRA-tuned FinBERT has learned to associate macro and sector-wide stress—even when not tied to a specific ticker—with elevated IV-normalized volatility at the single-name level. This is a subtler signal than raw sentiment: the headline is clearly negative, but what matters for volatility is that the shock is broad-based and affects a correlated sector.

The second headline is attached to a large U.S. bank and reads: "The S&P 500 Has a New Price Target. The Stock Market Is Getting Dicey." Here the true label falls in the low-volatility class (ν_{30} in the lower 30% tail), and the model again predicts low volatility. Unlike the previous example, this is essentially market commentary: it does not introduce a new firm-specific catalyst for the bank or describe a discrete macro event, but instead offers an opinion on index valuation and uses colloquial language ("getting dicey") to characterize conditions. Such pieces can sound dramatic yet rarely trigger additional trading in a particular constituent over a 30-minute window. By classifying this item as low volatility, the model appears to distinguish between analysis of the overall market and genuinely new information about a specific firm or sector, downweighting generic commentary even when the tone is mildly bearish.

Taken together, these two case studies illustrate how the LoRA-tuned FinBERT goes beyond raw sentiment. Sector-wide and index-level stress events tend to generate high IV-normalized volatility for individual names, whereas opinion-style market commentary without firm-specific content is more likely to coincide with low short-horizon volatility.

7 Limitations and Future Work

Our study has several limitations that suggest directions for future work:

- **Limited ticker conditioning.** The released FinBERT is trained primarily on sentence-level corpora such as Financial PhraseBank and does not explicitly condition on ticker symbols or firm identifiers. With roughly 440 labeled examples per ticker in our dataset, the model cannot learn rich ticker-specific patterns. Future work could explore meta-learning or explicit ticker embeddings conditioned on fundamentals, options data, or sector information.
- **Single aggregated news source.** Alpha Vantage aggregates stories from multiple providers into a single feed. We treat this as a homogeneous source, but source-specific differences in reliability, timeliness, and style almost certainly matter. Analyzing per-source performance, or using the source-aware transformer more aggressively to filter or reweight items, could improve signal quality.
- **Short horizons and limited event types.** We focus on 1–60 minute horizons where idiosyncratic order flow and index moves dominate. Textual signals may accumulate over longer horizons, especially around quarterly earnings or M&A announcements where options markets reprice volatility over days rather than minutes. Extending IV-normalized labels to daily or multi-day windows is a natural next step.
- **Attention visualization and interpretability.** Our qualitative analysis is based on a small set of case studies. A more systematic approach would quantify attention entropy and token-level attribution scores, and relate these to volatility outcomes. This could validate—or challenge—our intuition about which linguistic features drive predictions.

8 Conclusion

We presented a systematic evaluation of FinBERT on short-horizon, IV-based price-movement and volatility prediction tasks. By explicitly measuring dollar moves in units of an ATM straddle and realized volatility in units of IV, and by comparing zero-shot, head-only, and LoRA-tuned variants, we showed that: (i) directional price movements at minute horizons remain extremely difficult to predict, even after sophisticated scaling; but (ii) IV-normalized volatility is substantially more learnable, with LoRA adapters delivering meaningful gains over strong zero-shot baselines. These findings support volatility—rather than direction—as a more promising target for news-based intraday prediction, and highlight ATM-straddle and IV scaling as useful design principles for label construction in this setting.

9 Team Contributions

Yifei Fu: Led data collection (Alpha Vantage API integration, minute-bar alignment), designed IV-based volatility adjustment mechanism after TA discussion, performed layer-freezing experiments, and contributed to report writing.

Zhonghao Liu: Developed synthetic dataset prototype, implemented the LoRA fine-tuning pipeline, researched FinBERT training data (Financial PhraseBank analysis revealing lack of company context, leading to the volatility-normalization insight), performed evaluation, and contributed to report writing.

Shilun Dai: Conducted FinBERT background research, developed real-data fine-tuning scripts, tested multiple horizon configurations revealing 1-minute vs. longer-horizon patterns, performed baseline comparisons, and contributed to report writing.

All team members participated in project planning, TA discussions that led to the volatility reformulation, and iterative refinement throughout the project.

References

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

Tim Bollerslev, George Tauchen, and Hao Zhou. Expected stock returns and variance risk premia. *Review of Financial Studies*, 22(11):4463–4492, 2009.

Peter Carr and Liuren Wu. Variance risk premiums. *Review of Financial Studies*, 22(3):1311–1341, 2009.

Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. Deep learning for event-driven stock prediction. In *IJCAI*, pages 2327–2333, 2015.

Joseph Engelberg and Christopher A Parsons. Journalists and the stock market. *Review of Financial Studies*, 24(3):795–834, 2011.

Jim Gatheral. *The Volatility Surface: A Practitioner's Guide*. Wiley, 2006.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2022.

John Hull. *Options, Futures, and Other Derivatives*. Pearson, 10th edition, 2018.

Shuyi Ke, Bryan Kelly, and Dacheng Xiu. Predicting cross-sectional stock returns using text mining. *The Review of Financial Studies*, 32(9):3723–3759, 2019.

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

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*, 2023.

Yi Yang and Scott Matthews. Finbert: A pre-trained financial language representation model for financial text mining. *arXiv preprint arXiv:2006.08097*, 2020.

A Additional price-movement results

Table 6: Straddle-normalized price-movement prediction (3-class).

Horizon	Zero-shot	Head-only	Δ
1 min	0.3297	0.3361	$+0.64\%$
5 min	0.3396	0.3487	$+0.91\%$
30 min	0.3278	0.3558	$+2.80\%$
60 min	0.3397	0.3402	$+0.05\%$

Table 7: Straddle-normalized price-movement prediction (7-class).

Horizon	Zero-shot	Head-only	Δ
1 min	0.1387	0.1589	$+2.02\%$
5 min	0.1318	0.1762	$+4.44\%$
30 min	0.1245	0.1722	$+4.77\%$
60 min	0.1232	0.1442	$+2.10\%$