Interpreting News for Volatility: An Attention-Weighted HAR Framework

Younwoo Jeong[*] Bong-Gyu Jang[†‡]

November 28, 2025

Abstract

We test whether attention-weighted embeddings of after-hours news add interpretable and economically meaningful information to next-day realized volatility forecasts. We propose AW-HAR, which augments a per-stock HAR baseline with an additive correction $c(N_t;\theta)$ built from transformer embeddings pooled by an attention layer. Using S&P 500 constituents (continuous data availability) and restricting evaluation to news-days, we run an expanding-window backtest in 2023–2024 with dynamic HAR refits and standard forecast metrics. Across overlapping out-of-sample windows, AW-HAR lifts R^2 to 0.5710 and reduces QLIKE and SFE by 26.21% and 26.46% relative to HAR; improvements are statistically significant by Diebold–Mariano tests. Gains concentrate in high-volatility tails ($Z > 1$), where linear history-only models fail. Diagnostics show attention weights focus on earnings and uncertainty-laden phrases, and an ablation indicates extended-hours news (16:00–09:30 ET) is slightly more informative than full-day aggregation. By keeping the fusion in log space and applying loss on the natural scale after exponentiation, the approach remains transparent and deployment-friendly. Results support integrating attention-weighted news semantics into interpretable volatility models used in risk management.

Keywords: Realized Volatility, Volatility Forecasting, HAR Model, Natural Language Processing, Deep Learning, Interpretability

JEL Codes: C53, C55, C58, G17

[*] POSTECH, younwoo48@postech.ac.kr

[†] POSTECH, bonggyujang@postech.ac.kr

[‡] We would like to thank Jae Wook Song, Jeonggyu Huh, Jihoon Kim (discussant), and seminar participants from KAFE2025.

1 Introduction

Volatility forecasts are central to asset allocation, derivatives risk management, and capital planning. Linear history–based benchmarks in the HAR family (Corsi, 2008) are strong on persistence, but they often underreact to event-driven spikes concentrated in the extended hours when prices cannot immediately adjust. This paper asks whether semantically rich, attention-weighted representations of after-hours news can provide an interpretable correction to HAR-style forecasts for next-day realized volatility (RV).

The literature has established time-varying volatility and its autoregressive structure, from ARCH and GARCH (Engle, 1982; Bollerslev, 1986) to the HAR model that captures multi-horizon dynamics via daily, weekly, and monthly RV components (Corsi, 2008). Complementary strands study how to construct and evaluate RV and related loss functions: realized measures from high-frequency data (Barndorff-Nielsen and Shephard, 2004; Andersen et al., 2003; Barndorff-Nielsen, 2004), and forecast evaluation tools such as QLIKE and robust model comparison tests (Hansen and Lunde, 2006; Diebold and Mariano, 1995; Patton, 2011). Despite their empirical success, these linear, history-based approaches principally extrapolate from past volatility, limiting responsiveness to new information.

Extensions incorporate exogenous drivers. HAR-X models add macro and other covariates (Degiannakis and Filis, 2017), and implied volatility contributes predictive content beyond realized measures (Kambouroudis et al., 2021; Christensen and Prabhala, 1998; Poon and Granger, 2003). Theory links volatility to information flow through the mixture-of-distributions hypothesis (Clark, 1973; Epps and Epps, 1976; Tauchen and Pitts, 1983), while practice documents sharp volatility responses to macro releases and firm announcements (e.g., Patell and Wolfson, 1984; Anderson et al., 2003). Importantly, a large fraction of corporate disclosures arrive outside regular trading hours, shaping next-day dynamics when markets reopen (French and Roll, 1986; Bamber et al., 1997).

Text has become a first-class signal in finance. Early work connected online discussions and media tone to volatility and returns (Antweiler and Frank, 2004; TETLOCK, 2007; TETLOCK et al., 2008), domain lexicons improved tone measurement in filings

(LOUGHRAN and MCDONALD, 2011), and modern transformers (Vaswani et al., 2017; Devlin et al., 2019) with finance-specific adaptation (FinBERT) enhanced contextual relevance (Yang et al., 2020). Recent pipelines fine-tune large language models (LLMs) on news flows for return prediction (Guo and Hauptmann, 2024). Closely related to our focus, Audrino et al. (2020) augment HAR with sentiment and attention measures and document forecast gains. However, scalar sentiment can compress away event semantics that matter most for shock-sensitive RV prediction. At the same time, purely deep neural RV forecasters often outperform linear baselines yet face adoption frictions in regulated settings due to limited transparency (Moreno-Pino and Zohren, 2024; Li and Tang, 2025).

We propose Attention-Weighted HAR (AW–HAR), an interpretable fusion that preserves the transparent HAR backbone and adds an additive, news-driven correction built from attention pooled transformer embeddings of after-hours articles. The HAR component is estimated stock-by-stock to capture heterogeneous persistence, while a single cross-sectionally shared neural module maps shared news semantics to absolute shock magnitudes. This separation mitigates overfitting, enables transfer across firms, and keeps the fusion legible by design.

Our contributions are threefold. **First**, we introduce a transparent fusion that targets information shocks not contained in historical volatility by adding an attention-weighted news correction drawn specifically from extended-hours disclosures. **Second**, we operationalize attention pooling to compress each night's article set into a salience-weighted vector, so learned weights provide text-level interpretability for forecast adjustments. **Third**, we adopt a hybrid estimation scheme—dynamically refitting HAR by stock while sharing the news module cross-sectionally—and evaluate it in an expanding-window backtest (2023–2024) alongside ablations on fusion architecture, embedding family, and the timing of news aggregation.

Across overlapping out-of-sample windows in 2023–2024, AW–HAR lifts R^2 over both HAR and sentiment-augmented HAR, while reducing QLIKE and Squared Forecast Error (SFE) by about 25% relative to HAR over the full period ($R^2 = 0.5710$; QLIKE improvement 26.21%; SFE improvement 26.46%). Gains concentrate in high-volatility tails where linear baselines fail; attention weights place mass on earnings and uncertainty-laden phrases; and an ablation shows extended-hours news is slightly more informative than full-day aggregation for next-day RV.

The remainder of the paper details data and the AW–HAR design, presents empirical results and diagnostics, reports robustness exercises and ablations, and concludes with implications for risk management and deployment. For reproducibility, we keep the fusion in log space and compute loss on the natural scale after exponentiation, preserving transparency and implementation simplicity.

2 Data and Methodology

We forecast next-day realized volatility by fusing a transparent econometric baseline with information extracted from financial news using modern NLP. To ensure a fair assessment, we use an expanding-window backtest in which the HAR baseline is refit on every training window in lockstep with the fused model, thereby avoiding look-ahead and specification advantages.

2.1 Data Construction

2.1.1 Realized Volatility

Our prediction target is the logarithm of daily realized volatility. Following standard practice, we compute realized volatility (RV) from high-frequency intraday prices sampled at five-minute intervals during regular trading hours (09:30–16:00 ET). For day t, let $P_{t,i}$ denote the last price in interval i and

$$r_{t,i} = \log P_{t,i} - \log P_{t,i-1}.$$

Daily realized volatility is then

$$\mathrm{RV}_t = \sqrt{\sum_{i=1}^{M} r_{t,i}^2}, \qquad M = 78.$$

We model $\log(\text{RV}_t)$ rather than RV_t because the log transform stabilizes the scale, down-weights occasional spikes, and facilitates linear baselines and neural components. Inference and reported losses are on the natural scale: forecasts are exponentiated before scoring so that performance metrics are interpretable in units of volatility.

Two details are worth highlighting. First, we focus on returns within the trading session (open–close) and exclude close to open gaps to isolate the intraday response to information, consistent with our interest in how after-hours disclosures shape the next day's trading dynamics. Second, the five-minute sampling strikes the usual balance between capturing intraday variation and mitigating microstructure noise (Anderson et al., 2003).

2.1.2 News Corpus and Processing

We collect headlines and editor-written summaries for U.S. equities from *CNBC*, *The Wall Street Journal*, and *Yahoo Finance*. All timestamps are parsed from the source feeds and converted to US/Eastern. For each trading day, we define an extended-hours window (16:00–09:30 ET). News items in this window are assigned to the next day's open ($t+1$) so that they enter the forecast before prices can react. On weekends and market holidays, we aggregate items over the full calendar days and carry them forward to the next trading day at 09:30 ET (e.g., Fri 16:00 → Mon 09:30).

At the document level, we concatenate each article's title and summary to form the input text. Articles associated with a given stock and night are embedded with a finance-domain encoder (FinBERT), producing vectors in $\mathbb{R}^d$ (d=768) (Yang et al., 2020). Intuitively, FinBERT maps each piece of text to a point in a high-dimensional 'meaning space' where semantically similar articles lie close together. This lets us treat news as quantitative covariates: two articles that use different words but convey similar information (e.g., 'missed earnings' vs. 'below estimates') generate similar embeddings. Stacking the n vectors from the same extended-hours window yields an $n \times d$ matrix that is intentionally time-agnostic: it treats after-hours releases as a single information influx that will be processed when the market reopens. An attention-pooling layer then converts this set into a single, salience-weighted representation used by the news module (details in §2.2).

This construction aligns with the paper's design choices for two reasons. First, it targets information that cannot be fully incorporated while the market is closed; our ablation later shows that extended-hours aggregation is slightly more informative than pooling over the full 24 hour day. Second, it enables per-stock pooling while sharing the semantic mapping across firms, which we find reduces overfitting and improves transfer.

2.2 Forecasting Models

Let $\ell_t \equiv \log(RV_t)$ denote log realized volatility. We forecast in log space and map predictions back to the natural scale by $\widehat{RV}_{t+1} = \exp(\widehat{\ell}_{t+1})$ for evaluation.

2.2.1 Baseline: HAR

The Heterogeneous Autoregressive (HAR) model (Corsi, 2008) is a transparent, widely used baseline for volatility forecasting. In log space, the daily, weekly, and monthly components enter as moving averages of ℓ_t:

$$\ell_{t+1} \;=\; \beta_0 \;+\; \beta_d\,\ell_t \;+\; \beta_w\,\ell_t^{(w)} \;+\; \beta_m\,\ell_t^{(m)} \;+\; \varepsilon_{t+1},$$

with

$$\ell_t^{(w)} \;=\; \frac{1}{5}\sum_{k=0}^{4}\ell_{t-k}, \qquad \ell_t^{(m)} \;=\; \frac{1}{22}\sum_{k=0}^{21}\ell_{t-k}.$$

Estimating HAR stock-by-stock captures persistence and heterogeneity in each asset's volatility process while keeping the specification interpretable.

2.2.2 Benchmark: HAR–Sentiment

To test whether embeddings add information beyond coarse polarity, we augment HAR with a one-factor sentiment proxy constructed from after-hours articles:

$$\ell_{t+1} \;=\; \beta_0 \;+\; \beta_d\,\ell_t \;+\; \beta_w\,\ell_t^{(w)} \;+\; \beta_m\,\ell_t^{(m)} \;+\; \beta_s\,S_t \;+\; \varepsilon_{t+1}.$$

We use the same underlying FinBERT model in two ways: as a classifier to obtain sentiment probabilities for HAR–Sentiment, and as an encoder to produce continuous embeddings for the AW–HAR news module.

For article i in the extended-hours window mapped to day $t+1$, a FinBERT classifier outputs class probabilities for $\{\text{pos}, \text{neu}, \text{neg}\}$; we convert them to a scalar

$$S_{i,t} \;=\; (+1)\,P(\text{pos})_{i,t} \;+\; 0 \cdot P(\text{neu})_{i,t} \;+\; (-1)\,P(\text{neg})_{i,t},$$

and average across the N_t articles,

$$S_t \;=\; \frac{1}{N_t}\sum_{i=1}^{N_t} S_{i,t}.$$

Comparing HAR–Sentiment with both HAR and the embedding based fusion isolates the marginal value of full-text semantics beyond polarity (cf. Audrino et al., 2020). Scalar sentiment compresses each article into a single number and discards distinctions among different types of events (earnings, guidance changes, regulatory actions, etc.). Embeddings retain this richer structure, allowing the news module to treat earnings miss with weak guidance differently from earnings miss with strong guidance, even if both look similarly negative on a one-dimensional sentiment scale.

2.2.3 AW–HAR: HAR with an Additive News Correction

Let $\ell_t \equiv \log(RV_t)$. For each stock i on day t, the HAR model produces a log-RV baseline $\widetilde{\ell}^{(i)}_{t+1} = \text{HAR}(i,t)$ and a corresponding baseline on the natural scale

$$\widehat{RV}^{\,\text{HAR},(i)}_{t+1} \;=\; \exp\big(\widetilde{\ell}^{(i)}_{t+1}\big).$$

The news module takes the attention pooled embedding $N^{(i)}_t$ (Section 2) and outputs a same-day correction $\delta^{(i)}_t = c(N^{(i)}_t;\theta)$ on the natural RV scale. The fused forecast first

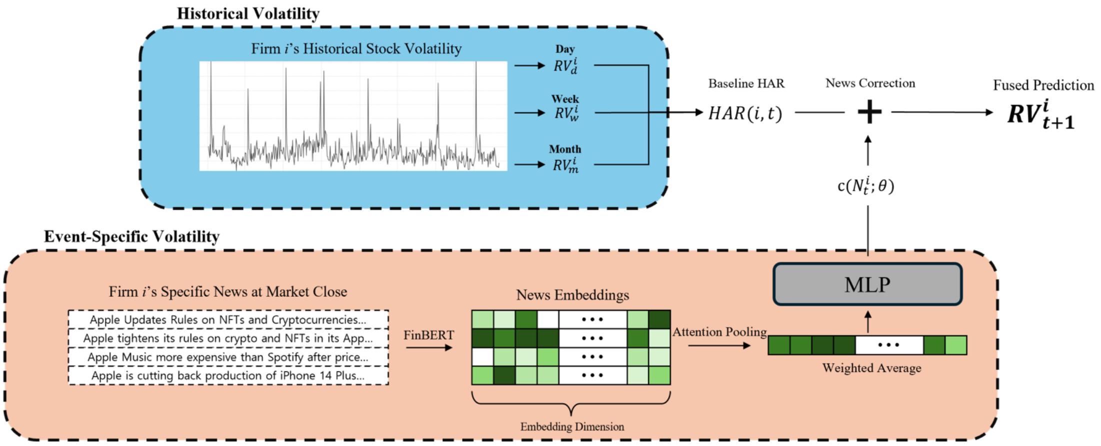

Historical Volatility
Firm i's Historical Stock Volatility
Day
RV_d^i
Week
RV_w^i
Month
RV_m^i
Baseline HAR
$HAR(i,t)$
News Correction
+
Fused Prediction
RV_{t+1}^i
$c(N_t^i;\theta)$
Event-Specific Volatility
Firm i's Specific News at Market Close
Apple Updates Rules on NFTs and Cryptocurrencies...
Apple tightens its rules on crypto and NFTs in its App...
Apple Music more expensive than Spotify after price...
Apple is cutting back production of iPhone 14 Plus...
FinBERT
News Embeddings
Embedding Dimension
Attention Pooling
MLP
Weighted Average

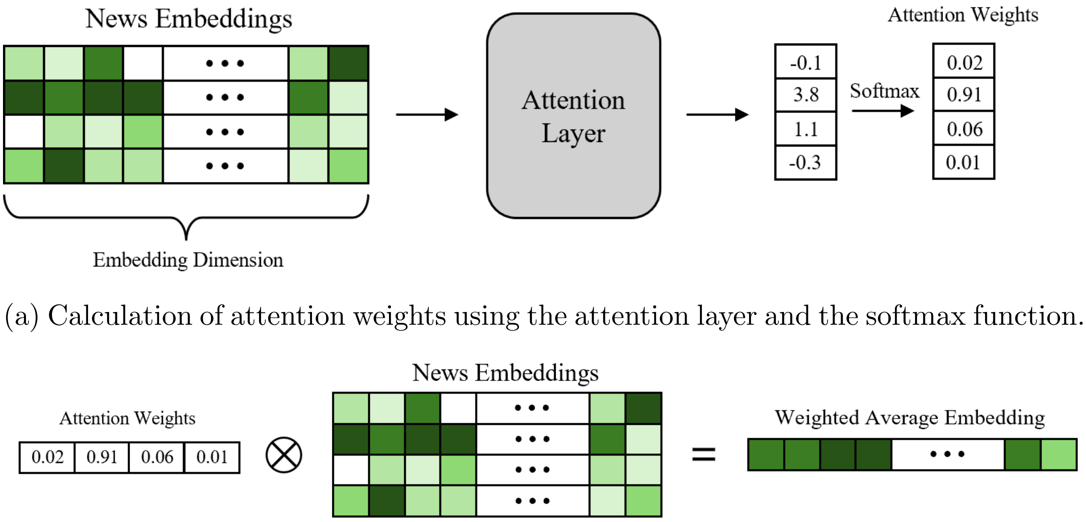
Figure 1: Architecture overview: a per-stock HAR baseline (log space) and a cross-sectionally shared news module producing a correction on the natural (RV) scale. The two are summed before taking logs for the final prediction.

sums on the natural scale and then returns to log space:

$$\widehat{RV}_{t+1}^{(i)} = \widehat{RV}_{t+1}^{\mathrm{HAR},(i)} + \delta_t^{(i)}, \tag{1}$$

$$\widehat{\ell}_{t+1}^{(i)} = \log\Big(\widehat{RV}_{t+1}^{\mathrm{HAR},(i)} + \delta_t^{(i)}\Big). \tag{2}$$

The small floor ε guarantees positivity inside the logarithm and has no practical effect away from zero.



(a) Calculation of attention weights using the attention layer and the softmax function.

(b) Creation of the final embedding via a weighted average.

Figure 2: The two-step attention pooling mechanism. First (a), embeddings are fed through an attention layer and raw scores are normalized into attention weights. Second (b), these weights are used to compute a weighted average of the original embeddings.

The HAR component is estimated per stock; the news module $c(\cdot;\theta)$ is a small feed-forward network with parameters θ shared across stocks to pool statistical strength in high-dimensional embeddings (d=768). Same-night articles are embedded with FinBERT and aggregated by attention pooling. Given embeddings $E_{t,1}^{(i)}, \ldots, E_{t,n}^{(i)}$, attention weights α_j are formed by softmax (Bridle, 1989) and the pooled vector is $N_t^{(i)} = \sum_{j=1}^{n} \alpha_j E_{t,j}^{(i)}$ (see Figure 9). The weights α_j serve as text-level saliency scores, aiding interpretability.

Equations (1)–(2) make the role of news explicit: on nights with informative disclosures, the network produces a signed absolute shock $\delta_t^{(i)}$ that is added to the HAR baseline on the volatility scale; on quiet nights $\delta_t^{(i)} \approx 0$ and the model collapses to HAR. This design matches the behavior observed in our diagnostics and ablations—larger corrections on high-volatility days and near-zero (or slightly negative) adjustments when attention deems the news uninformative.

For the extended-hours set of n articles linked to stock i on night t, let $E_1, \ldots, E_n \in \mathbb{R}^d$ be FinBERT embeddings (title+summary). An attention layer scores each item and forms weights

$$\alpha_i \; = \; \frac{\exp(s_i)}{\sum_{j=1}^{n} \exp(s_j)}\,, \qquad \sum_{i=1}^{n} \alpha_i = 1,\; \alpha_i > 0,$$

yielding a salience-weighted representation

$$N_t \; = \; \sum_{i=1}^{n} \alpha_i E_i.$$

This pooling treats after-hours disclosures as a single information arrival processed at the next open; it also furnishes α_i as text-level importance weights, supporting interpretability. The fused design is additive by construction—news contributes an absolute correction to the baseline—consistent with the role of discrete information shocks (see Figure 1 for the full flow). Economically, attention pooling lets the model read all same-night articles and assign each a weight between 0 and 1, summarizing how important that article is for volatility. The pooled vector N_t then behaves like a news index for night t, but one whose composition is learned from the data rather than fixed ex ante. The attention weights α_i can therefore be interpreted as text-level saliency scores that indicate which articles are driving the forecast adjustment.

2.3 Evaluation Framework

2.3.1 Data Coverage

Due to data availability constraints, our analysis utilizes realized volatility and news data spanning from January 1, 2016, to December 31, 2024. The stock universe consists of S&P 500 constituents that maintained continuous data availability throughout the training period, resulting in approximately 450 stocks. While this selection introduces survivorship bias, we expect limited impact on our volatility forecasting objective because conditional variance dynamics and news–volatility linkages are less sensitive to constituent turnover than return predictability. To ensure model training focuses on news-driven volatility patterns, we restrict our analysis to trading days with available news coverage, excluding days without news from the training set to prevent potential model confusion during optimization.

2.3.2 Expanding Window Training

To ensure a rigorous and unbiased evaluation of performance, we employ an expanding-window backtest from 2023 to 2024. The most critical aspect of our methodology is the dynamic baseline refitting. In each iteration of the expanding window, the HAR baseline models are completely re-trained using the exact same expanding training dataset that is available to our proposed neural network model. This eliminates any potential for lookahead bias and ensures a fair comparison.

The expanding window procedure operates with a validation and test period fixed at 6 months each. After each training session, the validation, and test windows are advanced by one quarter (3 months), while the train window increases by a quarter, creating overlapping evaluation periods that provide robust out-of-sample performance assessment. This walk-forward approach mimics real-world trading conditions where models are periodically retrained as new data becomes available. The whole training process can be seen in Figure 3

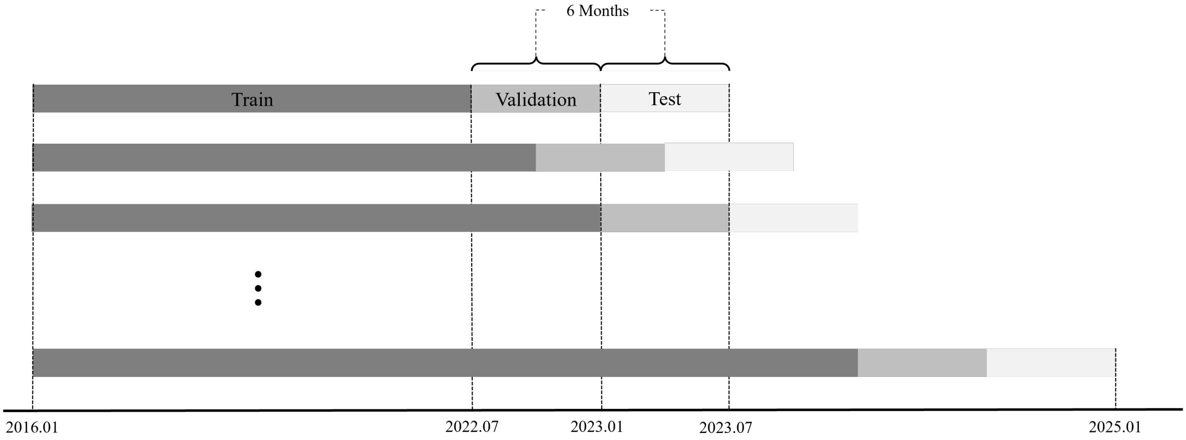


Figure 3: Expanding Window Outline

2.3.3 Evaluation Metrics

We evaluate model performance using three complementary criteria—out-of-sample R^2, Squared Forecast Error (SFE), and Quasi-likelihood (QLIKE)—defined as:

$$R^2 = 1 - \frac{\sum_{t=1}^{T}(RV_t - \widehat{RV}_t)^2}{\sum_{t=1}^{T}(RV_t - \overline{RV})^2}$$

$$SFE = \frac{1}{T}\sum_{t=1}^{T}(RV_t - \widehat{RV}_t)^2$$

$$QLIKE = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{RV_t}{\widehat{RV}_t} - \log\left(\frac{RV_t}{\widehat{RV}_t}\right) - 1\right)$$

Higher R^2 indicates greater explained variation in realized volatility, SFE measures mean squared predictive error, and QLIKE is tailored to volatility forecasting and is robust to outliers.

2.3.4 Training Challenges and Mitigation Strategies

Volatility series exhibit extreme spikes that can dominate the training loss and induce overfitting to rare events, degrading typical-day performance (Shi et al. (2025)). We employ the following mitigations:

- **Log-transformed target and inputs:** We train the network to predict log-realized volatility, which compresses the scale of large spikes and stabilizes optimization. For evaluation and loss computation, predictions are mapped back to the natural scale via $\widehat{RV}_t = \exp(\log(\widehat{RV}^i_{t+1}))$.

- **Huber loss on RV with $\delta = 0.3$:** Let the error on the natural scale be $e_t = RV_t - \widehat{RV}_t$. The per-sample Huber loss is

$$L_\delta(e_t) = \begin{cases} \frac{1}{2}e_t^2, & \text{if } |e_t| \le \delta, \\ \delta\left(|e_t| - \frac{1}{2}\delta\right), & \text{otherwise,} \end{cases} \qquad \delta = 0.3,$$

  and the training objective is $\mathcal{L}_{\text{Huber}} = \frac{1}{T}\sum_{t=1}^{T} L_\delta(e_t)$. This retains quadratic sensitivity for small errors while limiting the influence of large outliers.

- **Gradient clipping:** We clip gradient norms during backpropagation to prevent unstable updates triggered by extreme events, improving convergence stability.

2.3.5 Model and Optimization Setup

We train a feedforward architecture with three residual blocks, each block containing two fully connected layers of width 1024 with dropout 0.2. Optimization uses AdamW with learning rate 5×10^{-5}, batch size 1024, for up to 75 epochs. Early stopping with patience 15 and gradient clipping are applied to reduce overfitting and promote stable training.

3 Empirical Results

3.1 Out-of-Sample Performance of AW-HAR

Table 1: Out-of-sample for next day realized volatility forecasting results

Period	HAR R^2_{oos}	HAR+Sent. R^2_{oos}	QLIKE	SFE	AW-HAR R^2_{oos}	QLIKE	SFE
2023-01 to 2023-06	0.3973	0.4057	3.42***	1.39***	**0.5121**	**24.51*****	**19.04*****
2023-04 to 2023-09	0.4168	0.4283	3.76***	1.97***	**0.5839**	**30.07*****	**28.65*****
2023-07 to 2023-12	0.4331	0.4427	2.95***	1.70***	**0.5928**	**26.68*****	**28.17*****
2023-10 to 2024-03	0.4308	0.4411	2.93***	1.82***	**0.5862**	**24.86*****	**27.30*****
2024-01 to 2024-06	0.3861	0.3977	3.45***	1.89***	**0.5664**	**27.90*****	**29.36*****
2024-04 to 2024-09	0.4304	0.4418	3.53***	2.00***	**0.5868**	**25.96*****	**27.46*****
2024-07 to 2024-12	0.4074	0.4187	3.44***	1.91***	**0.5645**	**25.69*****	**26.53*****
Full Period	0.4186	0.4290	3.36***	1.80***	**0.5710**	**26.21*****	**26.46*****

Note: Performance of HAR, HAR-Sentiment, and Fused models using news-day observations. The QLIKE and SFE columns report the percentage improvement relative to the baseline HAR model (higher is better). Best results are in **bold**. *** indicates statistically significance at the 1% level of QLIKE and SFE loss compared to the baseline HAR model, according to the Diebold-Mariano test.

The empirical results from our rigorous out-of-sample expanding-window backtest provide strong evidence for the superior predictive power of the proposed AW-HAR framework. As summarized in Table 1, our model consistently and significantly outperforms both the baseline HAR model and the sentiment-augmented HAR model across all seven evaluation periods from 2023 to 2024. Over the full test period, the AW-HAR model achieves an out-of-sample R^2 of 0.5710, a substantial improvement over the 0.4186 and 0.4290 recorded by the HAR and HAR-Sentiment models, respectively. It is worth noting that augmenting the baseline with a simple sentiment score (HAR-Sentiment) does yield a consistent, though modest, improvement across all periods. However, this gain is dwarfed by the substantial leap in performance achieved by the AW-HAR model. This indicates that our news-driven component explains a significantly larger portion of the variance in next-day realized volatility than just simple sentiment.

The performance gains are not only limited to the R^2 metric but are also evident in the key error metrics designed for volatility forecasting. For the full period, the AW-HAR model reduces the SFE by 26.46% and improves the QLIKE loss by 26.21% relative to the baseline HAR. It is particularly noteworthy that the HAR-Sentiment model offers only a marginal improvement over the standard HAR. This finding underscores our central hypothesis: a sophisticated, attention-weighted framework that processes the full semantic content of news embeddings can extract a much richer predictive signal than what is available from simple sentiment polarity alone.

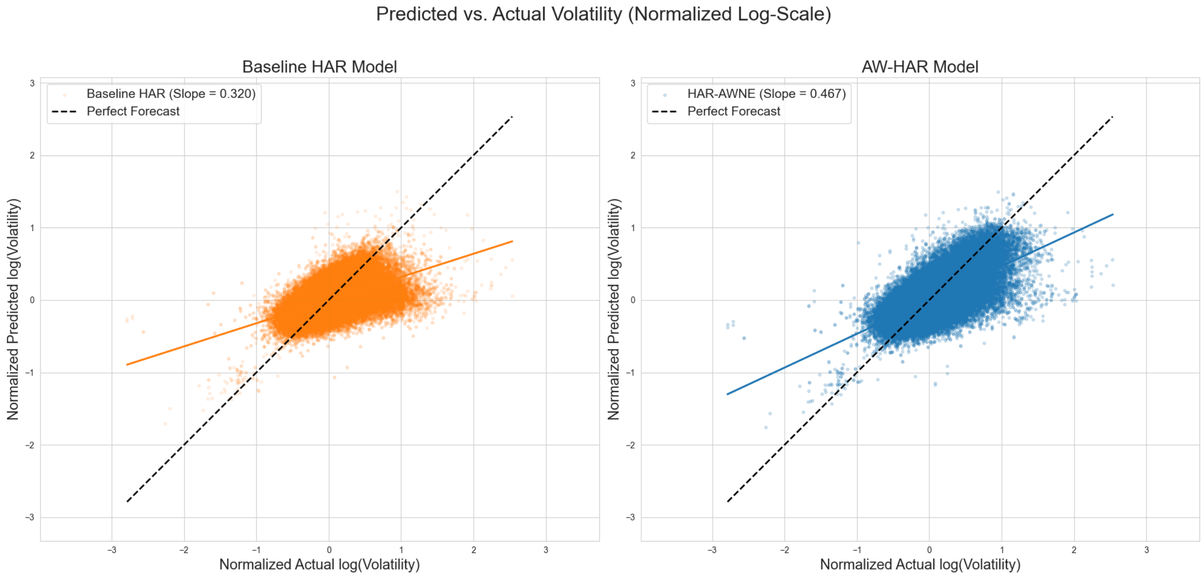


Figure 4: Predicted vs. Actual Realized Volatility: AW-HAR can more closely follow the actual forecast.

Visual evidence further corroborates the quantitative findings. Figure 4 plots the predicted versus actual realized volatility on a normalized log-scale for both the baseline HAR and the AW-HAR models. The predictions from the AW-HAR model (right panel) cluster far more tightly around the 45-degree perfect forecast line, with a regression slope of 0.467 that is substantially closer to the ideal of 1 compared to the baseline's 0.320. A closer inspection reveals that the model's predictive accuracy is particularly strong for higher volatility values. This is logical, as high-volatility days are often driven by significant, market-moving news events, which the news-embedding component of our model is specifically designed to interpret and quantify. In contrast, on days with average or low volatility, where news may be less impactful or simply add noise, the baseline HAR component already performs well, and the added news signal provides less of an edge. This visual confirmation aligns with our hypothesis that the model's primary strength lies in capturing the dynamics of high-impact events.

To illustrate the model's practical advantage, Figure 5 presents a time-series forecast for a single ticker KR (Kroger) during 2024. The plot highlights several instances of large, sudden volatility spikes. As circled, the AW-HAR model successfully anticipates

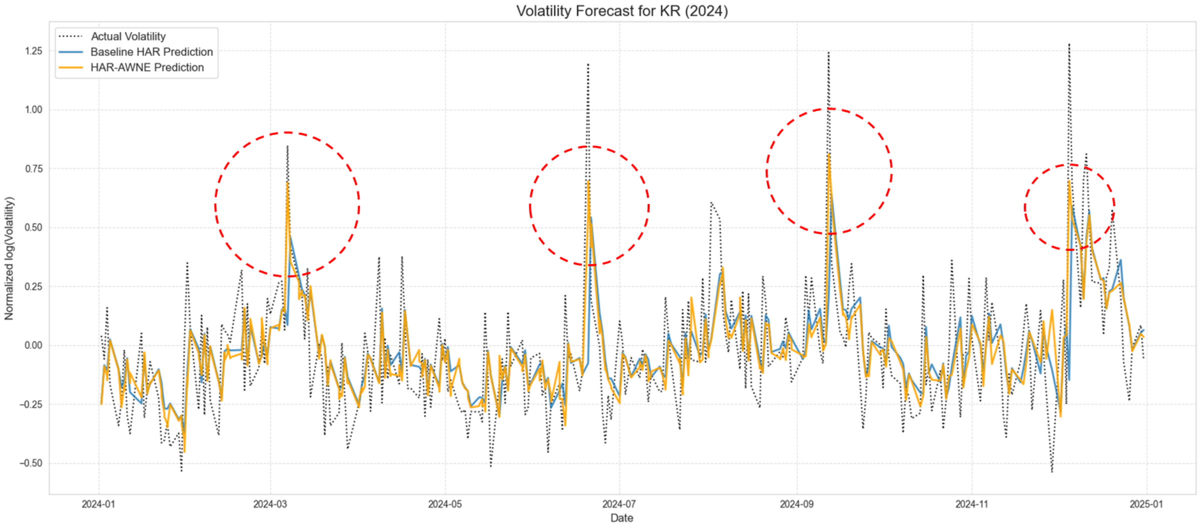


Figure 5: Log-volatility predictions for ticker KR (Kroger) in 2024: The AW-HAR model anticipates event-driven spikes (red circles), while the baseline HAR lags behind.

and captures these event-driven movements with remarkable accuracy. In contrast, the baseline HAR model, which relies solely on historical volatility patterns, consistently lags and fails to predict the magnitude of these sharp increases. This case study demonstrates the AW-HAR model's key strength: its ability to incorporate real-time, forward-looking information from news to forecast high-impact events where traditional autoregressive models fall short.

3.2 Z-Score Sorted Results

Table 2: Out-of-Sample Performance by Volatility Z-Score Bin

Z-Score Bin	Observations	HAR R^2_{oos}	AW-HAR R^2_{oos}	QLIKE	SFE
$-3 \leq Z < -2$	27	-16.0196	-15.8808	-0.39	0.82
$-2 \leq Z < -1$	12,783	-1.2364	-1.2282	1.24	0.37
$-1 \leq Z < 0$	110,586	0.4825	0.4632	-3.72	-2.46
$0 \leq Z < 1$	53,940	0.4799	0.4557	-2.00	-4.65
$1 < Z \leq 2$	12,601	-0.2298	-0.0144	21.40	17.51
$2 < Z \leq 3$	4,530	-1.5820	-0.4287	52.88	44.67
$Z > 3$	4,112	-1.8617	-0.7431	58.27	39.09

Note: This table shows the out-of-sample performance of the baseline HAR and the fused AW-HAR model, binned by the per-stock Z-score of the target volatility. The QLIKE and SFE improvement columns report the percentage improvement of the fused model relative to the HAR baseline for each bin (higher is better).

To rigorously evaluate the model's performance under different market conditions, we bin the out-of-sample results based on the per-stock Z-score of the target volatility. Z-score is computed per stock using the mean and standard deviation estimated over the current training window (no look-ahead). The results, presented in Table 2, reveal a distinct and asymmetric U-shaped performance curve for the fused AW-HAR model.

The model's primary contribution is evident in the positive tail of the volatility distribution. For days experiencing significant positive shocks ($Z > 1$), the baseline HAR model fails completely, as indicated by its large negative R^2. In sharp contrast, the AW-HAR model delivers substantial and increasing improvements, with the QLIKE improvement reaching a remarkable 58.27% for the most extreme events ($Z > 3$). This demonstrates that the news-driven component successfully captures the high-impact, non-linear dynamics of volatility spikes that the linear HAR model cannot.

In the central regime, performance is asymmetric across adjacent bins: the fused model modestly underperforms for $-1 \leq Z < 0$ (QLIKE -3.72%, SFE -2.46%) and for $0 \leq Z < 1$ (QLIKE -2.00%, SFE -4.65%) as reported in Table 2. This pattern is consistent with a learned behavior in which the news-driven component reduces the predicted volatility when the attention pooled embedding conveys weak or noisy after-hours signal, effectively shrinking forecasts toward calmer regimes. Such shrinkage is beneficial when realized volatility is below average ($Z < 0$), improving accuracy by correcting the HAR baseline's occasional overprediction on quiet days; however, it induces underprediction when realized volatility is slightly above average ($0 < Z < 1$), leading to the observed performance give-back in that bin. Formally, under the fusion $\widehat{RV}_{t+1} = HAR(i, t) + c(N_t; \theta)$, weak-signal nights tend to yield negative $c(\cdot)$, lowering $\widehat{RV}_{t+1}$ and improving fit when RV_{t+1} is modestly below trend but hurting when it is modestly above.

In the negative tail ($-2 \leq Z < -1$), the fused model improves QLIKE and SFE despite negative R^2 for both models, suggesting the news channel helps correct HAR's tendency to overpredict during unusually calm episodes.

Overall, these results validates the hybrid approach. The AW-HAR framework demonstrates its strength precisely when it is most needed: by successfully interpreting the semantic content of news to forecast statistically rare, high-impact events where traditional time-series models are known to fail.

3.3 Economic Interpretation

To gain qualitative insight into the model's behavior, we connect the learned news-driven correction c_t to observable text patterns and volatility regimes. Recall that c_t is defined on the natural (RV) scale as an additive adjustment to the per-stock HAR forecast: on quiet nights $c_t \approx 0$, whereas positive (negative) values correspond to upward (downward) revisions of next-day volatility. The goal of this section is to verify that these adjustments are systematically related to economically meaningful information rather than reflecting random noise.

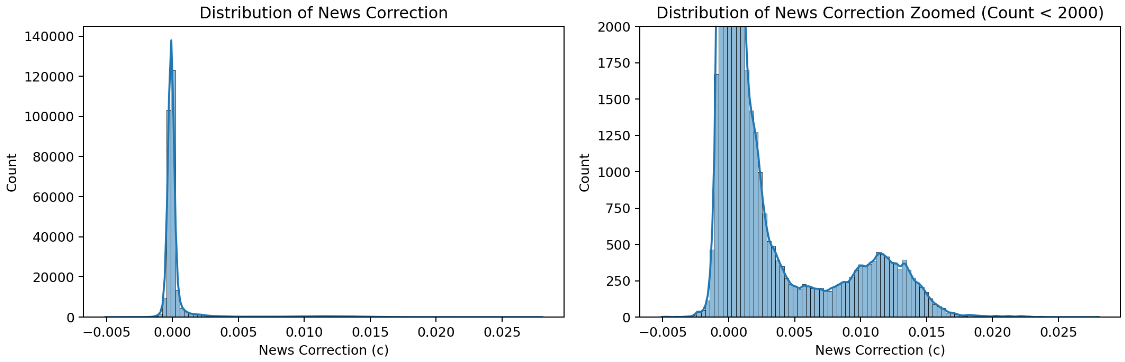


Figure 6: Distribution of news correction values c_t. The right panel zooms in on the left panel with the y–axis capped at 2,000 to reveal finer detail. Most nights receive corrections close to zero, with a secondary mode around $c_t \approx 0.012$, which we interpret as the typical magnitude of additive shocks on market–moving nights.

Figure 6 shows the empirical distribution of c_t across stocks and nights. The mass concentrated near zero confirms that, by construction, the model largely reverts to the HAR baseline on typical trading days. The secondary mode around $c_t \approx 0.012$ highlights a smaller set of "event nights" for which the model adds a sizeable, positive correction, consistent with discrete information shocks.

We first examine whether the correction term c_t scales systematically with next day realized volatility. To visualize this relationship, we plot mean centered log realized volatility (normalized per stock to have zero mean) against the model's news correction c_t. Values of the normalized log volatility above zero correspond to days that are more volatile than usual for that stock, while values below zero correspond to relatively calm days. To focus on nights where the network identifies strong news impact, we restrict the sample to the top decile of c_t.

The Locally Weighted Scatterplot Smoothing (LOWESS) trend in Figure 7 displays a clear upward slope. Higher (demeaned) log realized volatility coincides with larger news corrections, indicating a monotonic relationship in which the news module scales the forecast according to the economic significance of the information shock. This scaling behavior is consistent with our earlier Z–score stratification, where the largest forecast gains occurred in the high–volatility tail ($Z > 1$). Taken together with the distribution in Figure 6, these results suggest that the attention–weighted news channel activates primarily on a subset of nights associated with elevated risk and amplifies forecasts precisely when realized volatility turns out to be high, rather than overfitting to typical trading days.

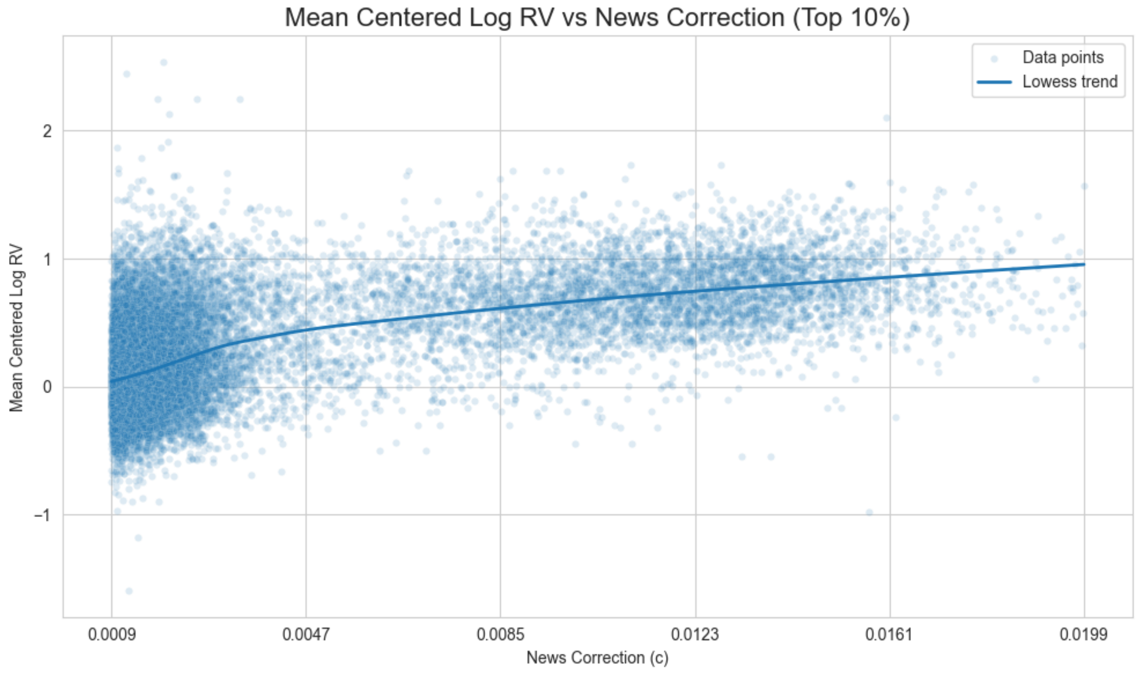


Figure 7: **Mean-centered log RV vs. news correction (top decile of c_t).** Each point represents a stock–day observation; the solid line represents a LOWESS fit. Higher (demeaned) log volatility coincides with larger news corrections, indicating that the neural module's output is systematic—not random—and scales with the magnitude of volatility shocks.

We next investigate which specific terms drive the model's attention on volatile days.

To do so, we apply a Term Frequency–Inverse Document Frequency (TF–IDF) analysis to the news corpus. The weight for a term t in document d is defined as

$$\text{TF-IDF}(t,d) = \underbrace{\text{TF}(t,d)}_{\text{Term Frequency}} \cdot \underbrace{\log\left(\frac{N}{1+\text{df}(t)}\right)}_{\text{Inverse Document Frequency}},$$

where N is the number of documents and $\text{df}(t)$ is the document frequency of term t. High TF–IDF values therefore highlight terms that are frequent in a given document but relatively rare in the overall corpus.

As illustrated in Figure 8, terms associated with earnings announcements—such as “earnings revenue,” “surprises,” and “quarter ended”—receive the highest TF–IDF weights on days with elevated volatility. In addition, uncertainty-laden phrases like “clues lies” also rank prominently. These patterns mirror our attention diagnostics, which allocate substantial probability mass to earnings- and uncertainty-related language. This alignment reinforces the interpretation that the model tends to increase c_t when the incoming text explicitly signals heightened risk or the resolution of important information.

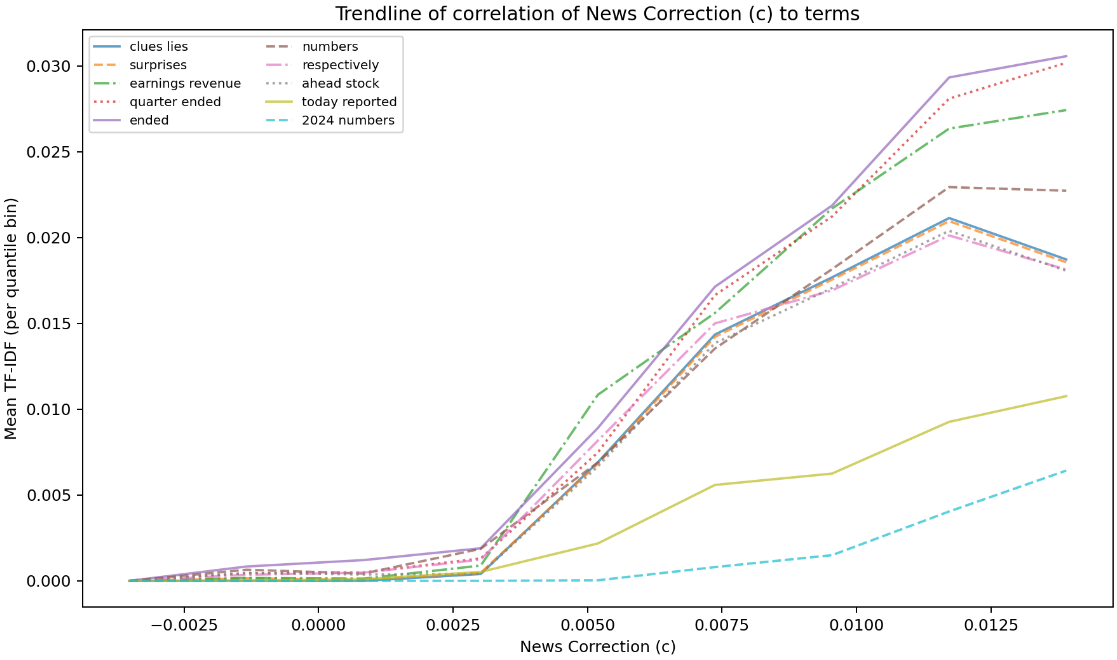
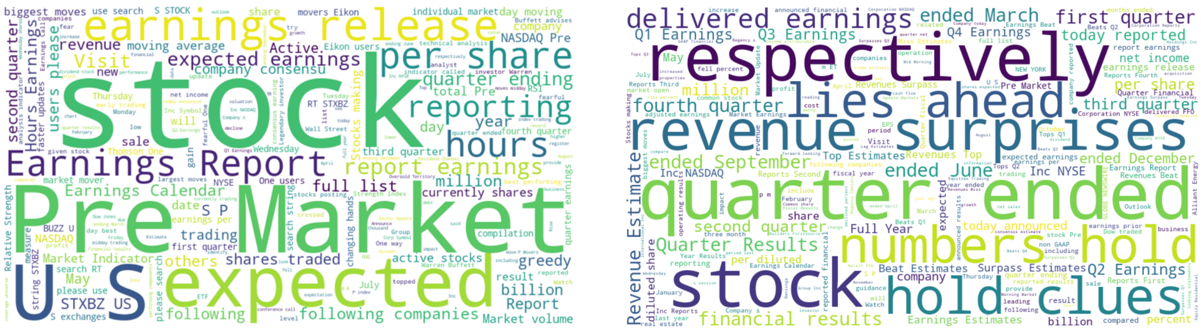


Figure 8: Top TF–IDF–ranked unigrams and bigrams on high–volatility days. Earnings–related terms dominate, and uncertainty phrases also rank highly, consistent with the model’s focus on economically salient events.

Finally, we visualize how the semantic content of news varies with the magnitude of the correction. To isolate the specific items most influential for the model, we filter for articles with attention weights exceeding 0.3, ensuring that only highly salient pieces of news are considered. We then partition the corresponding correction values c_t into two regimes, motivated by the empirical distribution in Figure 6: a “Medium Correction” regime ($0.002 \leq c_t \leq 0.005$) and a “High Correction” regime ($c_t > 0.005$).

Figure 9 presents word clouds for these two regimes. The “Medium Correction” cloud contains a mix of generic financial and firm-specific terminology, reflecting moderate but non-trivial information updates. By contrast, the “High Correction” cloud is dominated by explicit earnings-related descriptors such as “revenue,” “surprises,” “estimates,” and “misses.” This shift in vocabulary confirms that large positive adjustments c_t are associated with semantically rich disclosures that plausibly drive tail-risk events, rather than generic or low-information news. In other words, the attention mechanism not only pinpoints which articles matter but also ties large forecast corrections to interpretable, economically meaningful topics.



(a) Medium Correction ($0.002 \leq c_t < 0.005$)

(b) High Correction ($c_t > 0.005$)

Figure 9: Word clouds illustrating the semantic content of news items with high attention weights (> 0.3), grouped by the magnitude of the resulting model correction. High corrections are associated with strong earnings-related terminology and other semantically rich financial disclosures.

4 Robustness Studies

We assess the stability of the main findings by re-estimating the full pipeline under targeted variations in modeling choices and data construction, attributing observed gains to specific components while verifying invariance to reasonable perturbations. All exercises retain the expanding-window protocol, dynamic HAR refits, and the evaluation metrics established in the main analysis to ensure strict comparability.

4.1 News Embedding Type

To benchmark the text encoder choice, we compare four representative families under an identical estimation and evaluation setup: (i) BERT (Base/Large), the canonical general-domain encoder pretrained on large corpora by Devlin et al. (2019); (ii) ModernBERT (Base/Large), a modernized BERT family with architectural and pretraining enhancements by Warner et al. (2024); (iii) FinBERT, a domain-adapted encoder specialized for financial text used as our default model; and (iv) SimCSE, a contrastive sentence-embedding approach that produces semantically meaningful representations by Brinner and Zarriess (2025). Results are reported in Table 3 under the same expanding-window partitions and scoring metrics for one-to-one comparability across encoders.

Table 3: Full Period Out-of-Sample Forecasting Results by Embedding Model

Model Name	Emb. Dim. (d)	$R^2_{\mathbf{oos}}$	QLIKE	SFE
ModernBERT-Large	1024	*0.5647*	*25.90*	25.12
ModernBERT-Base	768	0.5658	25.59	*25.31*
FinBERT (default)	768	**0.5710**	**26.21**	**26.46**
BERT-Large	1024	0.5599	24.94	24.30
BERT-Base	768	0.5614	25.07	24.57
SimCSE	768	0.5538	23.92	23.26

Note: This table summarizes the full-period out-of-sample forecasting performance of the AW-HAR framework using different embedding models. The evaluation period is from 2023-01 to 2024-12. R^2 values are reported directly. QLIKE and SFE columns report the percentage improvement of each model relative to the baseline HAR model (higher is better). The best results for each metric is reported in **bold** and second best in *italic*.

As reported in Table 3, FinBERT attains the highest out-of-sample R^2 and the largest QLIKE/SFE improvements relative to the HAR baseline, with ModernBERT-Large as a close second, while general-domain BERT and SimCSE trail but remain competitive under the same protocol. The stable ranking across metrics and the modest performance dispersion indicate that the attention-weighted fusion is effective across encoder choices, suggesting that the main gains are not driven by a single "magic" embedding model.

4.2 Full-Day vs. Extended-Hours News

We conduct an ablation to test whether after-hours news carries the strongest signal for next-day volatility. We compare AW-HAR trained on extended-hours news (16:00–09:30 ET) with an alternative trained on full-day news (24 hours). Both models use identical expanding-window estimation and are evaluated out of sample across seven overlapping test periods from 2023–2024. For comparability, we restrict the evaluation to trading days with at least one extended-hours news item; days with only intraday news are excluded from both variants.

Table 4: Full Day News vs. Extended-Hours News Ablation Study

News Type	**Full-Day**			**Extended-Hours**		
Period	R^2_{oos}	QLIKE	SFE	R^2_{oos}	QLIKE	SFE
2023-01 to 2023-06	0.5090	24.24	18.53	**0.5121**	**24.51**	**19.04**
2023-04 to 2023-09	0.5821	29.52	28.35	**0.5839**	**30.06**	**28.65**
2023-07 to 2023-12	0.5903	25.70	27.74	**0.5928**	**26.67**	**28.17**
2023-10 to 2024-03	**0.5932**	**26.03**	**28.53**	0.5862	24.86	27.30
2024-01 to 2024-06	0.5552	25.61	27.55	**0.5664**	**27.90**	**29.36**
2024-04 to 2024-09	**0.5874**	25.83	**27.57**	0.5868	**25.96**	27.46
2024-07 to 2024-12	0.5638	**25.76**	26.40	**0.5645**	25.69	**26.53**
Full Period	0.5694	26.05	25.94	**0.5710**	**26.21**	**26.46**

Notes: R^2 is out-of-sample. QLIKE and SFE entries are improvements relative to a baseline HAR model; higher values indicate better performance. The evaluation sample is restricted to days with at least one extended-hours article; days with only intraday news are excluded from both specifications.

Table 4 shows that the extended-hours specification outperforms the full-day variant in five of seven periods. Over the full sample, the extended-hours model attains slightly higher R^2 (0.5710 vs. 0.5694), slightly higher QLIKE improvement (26.21 vs. 26.05), and a larger SFE improvement (26.46 vs. 25.94) relative to the HAR baseline. These results support the hypothesis that the most informative disclosures arrive outside regular trading, when prices cannot immediately incorporate news. Adding intraday articles contributes little incremental signal and can introduce noise, as releases are rapidly impounded into prices during continuous trading. Focusing on extended-hours news is therefore an efficient design choice for forecasting next-day volatility.

4.3 Performance by Sample Size

To verify that improvements are not mechanically driven by firms with more observations, we partition the cross-section into deciles by each ticker's number of training observations and recompute out-of-sample performance within each bin. This assesses whether the embedding-based gains are uniform across coverage levels. Across all deciles, the fused AW-HAR model consistently improves on the HAR baseline in R^2 and delivers positive percentage reductions in QLIKE and SFE, with the largest error reductions in the sparsest bins.

Table 5: Out-of-sample performance by sample-size deciles (full period)

Decile	**Sample count**	**HAR $\mathbf{R^2_{\text{oos}}}$**	**AW-HAR $\mathbf{R^2_{\text{oos}}}$**	**QLIKE**	**SFE**
1	8,741	0.2267	0.5466	39.35	41.37
2	11,776	0.3638	0.6008	38.13	37.25
3	13,687	0.2698	0.5455	38.11	37.76
4	16,036	0.3884	0.5915	35.38	33.21
5	17,702	0.3011	0.5169	31.95	30.87
6	18,994	0.3939	0.5764	30.79	30.10
7	21,414	0.3651	0.5210	25.38	24.56
8	24,045	0.4975	0.5938	21.02	19.18
9	30,806	0.5067	0.5910	17.49	17.08
10	35,379	0.5161	0.5541	8.97	7.85

Notes: Tickers are sorted into deciles by the number of per-ticker training observations in the expanding window. Metrics are evaluated out of sample within each decile. QLIKE and SFE report percentage error reductions relative to the HAR baseline (positive values indicate improvement).

The persistence of gains from the sparsest to the densest deciles indicates that the attention-weighted news channel generalizes across heterogeneous history lengths, while shared parameters transfer learning from richly covered tickers to thin-history tickers. Contrary to intuition, tickers with the most data exhibit smaller incremental gains because their baseline R^2 is already high—these firms are typically larger with more stable volatility. Even so, patterns learned from data-rich names propagate globally, as evidenced by the pronounced improvements in the lower deciles.

Table 6: Architecture ablation by period: out-of-sample R^2 and SFE improvements vs. HAR

Architecture	$HAR \times w(\cdot)$			$HAR + c(\cdot)$			$HAR \times w(\cdot) + c(\cdot)$		
Period	$\mathrm{R}^2_{\mathrm{oos}}$	QLIKE	SFE	$\mathrm{R}^2_{\mathrm{oos}}$	QLIKE	SFE	$\mathrm{R}^2_{\mathrm{oos}}$	QLIKE	SFE
2023-01 to 2023-06	0.5084	23.18	18.43	**0.5121**	**24.51**	**19.04**	0.5022	23.04	17.40
2023-04 to 2023-09	0.5679	26.99	25.57	**0.5839**	**30.07**	**28.65**	0.5805	29.09	28.07
2023-07 to 2023-12	0.5715	23.89	24.42	0.5928	**26.68**	28.17	**0.5929**	26.24	**28.20**
2023-10 to 2024-03	0.5705	23.07	24.55	**0.5862**	**24.86**	**27.30**	0.5808	24.67	26.35
2024-01 to 2024-06	0.5458	25.03	26.02	**0.5664**	**27.90**	**29.36**	0.5658	27.75	29.27
2024-04 to 2024-09	0.5744	23.01	25.27	**0.5868**	**25.96**	**27.46**	0.5854	24.80	27.21
2024-07 to 2024-12	0.5548	24.07	24.88	0.5645	25.69	26.52	**0.5683**	**26.28**	**27.16**
Full Period	0.5575	24.14	23.89	**0.5710**	**26.21**	**26.46**	0.5685	25.95	25.78

Notes: This table compares the out-of-sample performance of three different model architectures for fusing news signals with the HAR baseline. R^2 values are reported directly, while QLIKE and SFE values represent the percentage improvement over the standard HAR model (higher is better). The evaluation period spans from January 2023 to December 2024. The best results are in bold for each period.

4.4 Model Architecture Ablation

To determine the most effective method for integrating the news-based signal with the HAR baseline, we conducted an architecture ablation study. We compared our proposed additive model, where the news-derived component $c(\cdot)$ is added to the HAR forecast, against two alternatives: (i) a multiplicative model where the HAR forecast is scaled by a news-driven weight $w(\cdot)$, and (ii) a hybrid model that incorporates both components.

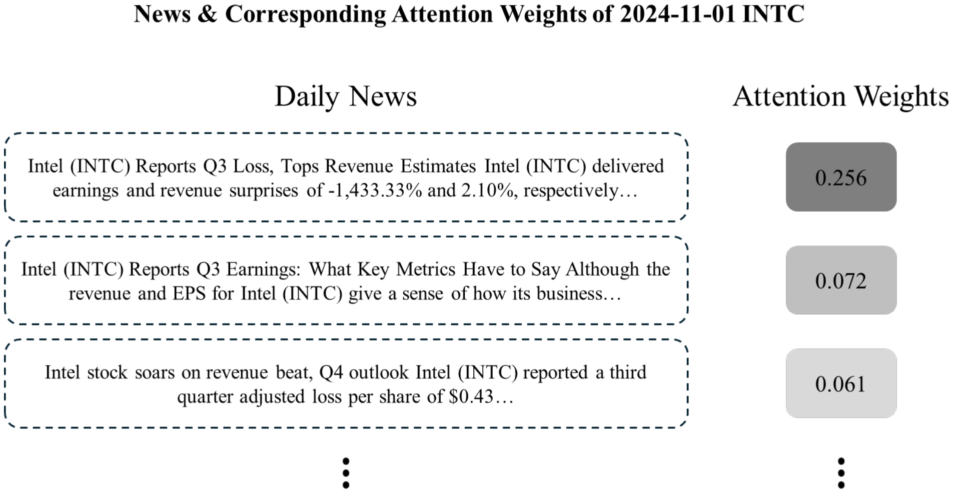
The out-of-sample results, presented in Table 6, show that the purely additive model, $HAR+c(\cdot)$, is the top-performing architecture. For the full period, it achieves the highest out-of-sample R^2 (0.5710) and the largest improvements in both QLIKE (26.21%) and SFE (26.46%) relative to the baseline HAR. Interestingly, the hybrid model, $HAR \times w(\cdot) + c(\cdot)$, fails to improve upon the additive version. This suggests that the multiplicative weight $w(\cdot)$ introduces noise; because the neural network learns this weight from news embeddings alone, it has no visibility into the HAR baseline value it is scaling. Consequently, it may amplify inherent estimation errors in the HAR term, degrading the final forecast. The superiority of the additive approach indicates that news is best modeled as an independent shock that provides an absolute correction to the baseline, validating our choice of this architecture for the AW-HAR framework.

5 Case Study: Dynamic Model Behavior for Intel

To demonstrate the interpretability and state-dependent nature of the AW-HAR framework, we examine its predictions for Intel Corporation (`INTC`) during two distinct volatility regimes: a news-heavy shock event and a routine trading day.

5.1 Response to High-Volatility News: 2024-11-01

We first analyze November 1, 2024, a session characterized by significant market turbulence. Relying exclusively on historical price data, the baseline HAR model generated a modest volatility forecast of 0.0194. However, the after-hours news flow contained critical information regarding quarterly earnings, specifically identifying a quarterly loss alongside a revenue surprise.


News & Corresponding Attention Weights of 2024-11-01 INTC

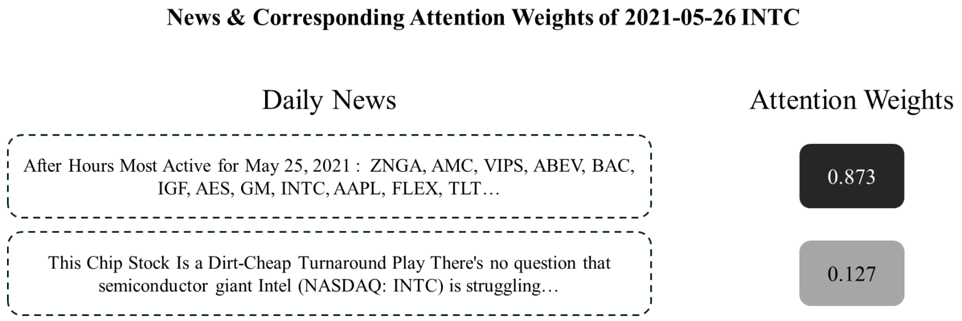
Daily News	Attention Weights
Intel (INTC) Reports Q3 Loss, Tops Revenue Estimates Intel (INTC) delivered earnings and revenue surprises of -1,433.33% and 2.10%, respectively…	0.256
Intel (INTC) Reports Q3 Earnings: What Key Metrics Have to Say Although the revenue and EPS for Intel (INTC) give a sense of how its business…	0.072
Intel stock soars on revenue beat, Q4 outlook Intel (INTC) reported a third quarter adjusted loss per share of $0.43…	0.061
⋮	⋮

Figure 10: Attention mechanism for Intel (`INTC`) on 2024-11-01. The model assigns high saliency weights to earnings-related disclosures, identifying them as drivers of next-day volatility.

The AW-HAR attention mechanism successfully identified the high economic salience of these disclosures. Consequently, the model produced a positive additive correction of +0.0071, raising the fused forecast to 0.0265. Although this adjustment did not fully capture the extreme realized volatility of 0.4793, the model correctly anticipated the direction and presence of a significant shock, offering a clear improvement over the static baseline.This example highlights the architecture's transparency, allowing analysts to pinpoint exactly which news items triggered the volatility premium.

5.2 Noise Filtering on Low-Volatility Days: 2021-05-26

In contrast, we examine May 26, 2021, a representative low-volatility session where news flow consisted of generic, non-material updates. For this date, the baseline HAR model predicted a volatility of 0.0105.



After Hours Most Active for May 25, 2021 : ZNGA, AMC, VIPS, ABEV, BAC, IGF, AES, GM, INTC, AAPL, FLEX, TLT…

This Chip Stock Is a Dirt-Cheap Turnaround Play There's no question that semiconductor giant Intel (NASDAQ: INTC) is struggling…

Figure 11: Attention mechanism for Intel (`INTC`) during a low-volatility regime. The model assigns low, dispersed weights to generic news, effectively neutralizing the news correction term.

The neural network assessed the available text as lacking informational value, generating a negligible correction of +0.0002The resulting fused prediction of 0.0107 remained virtually identical to the baseline, avoiding the introduction of noise when the actual realized volatility was low (0.0066). This behavior demonstrates the model's capacity to act as a noise filter: it effectively reverts to the HAR baseline during quiet periods while preserving sensitivity to genuine information shocks.

6 Conclusion

This paper examines whether attention-weighted embeddings of extended-hours news can systematically and interpretably improve next-day realized volatility forecasts relative to HAR and sentiment-augmented HAR benchmarks. We propose AW–HAR, an interpretable fusion architecture that preserves a per-stock HAR backbone and adds an additive news correction c_t derived from attention pooled transformer embeddings of after-hours articles. In an expanding-window backtest over 2023–2024, AW–HAR lifts out-of-sample R^2 to 0.5710 and reduces QLIKE and SFE by roughly 26% relative to the HAR baseline, with gains that are statistically significant and concentrated in high-volatility tails where history-only models perform worst.

A series of ablations and diagnostics clarify why the model works and how it uses news. Architecture experiments show that the additive design $HAR+c(\cdot)$ dominates multiplicative and hybrid alternatives, indicating that news is best modeled as an absolute shock rather than a pure scale factor on historical volatility. Z-score-sorted performance and the distribution of the learned correction c_t confirm that the news channel remains near-dormant on typical days and becomes active in the upper tail of the volatility distribution. Text-level analyses TF–IDF rankings, attention weights, and word clouds—show that the model consistently concentrates mass on earnings-related and uncertainty-laden phrases, while case studies for Intel demonstrate large upward corrections on event-driven days and negligible adjustments on quiet days. These properties jointly support an economic interpretation in which the news module identifies and prices discrete information shocks rather than amplifying noise.

Design choices around timing and regularization further reinforce both performance and interpretability. An ablation on news windows finds that extended-hours disclosures (16:00–09:30 ET) are slightly more informative for next-day volatility than full-day aggregation, validating the focus on information released when prices cannot immediately adjust. Methodologically, the framework combines attention pooling over transformer embeddings with per-stock HAR re-estimation and a cross-sectionally shared news module, pooling statistical strength across firms while retaining a transparent, stock-specific baseline. Training practices—log transformation of the target, Huber loss on the natural volatility scale, and gradient clipping—stabilize learning in the presence of volatility spikes without sacrificing sensitivity to tail events.

Taken together, these results suggest that integrating attention-weighted news semantics into transparent volatility models is both practically useful and compatible with deployment in risk management and regulated environments. AW–HAR offers a template for fusing deep text representations with econometric backbones in a way that preserves clear economic meaning: historical volatility dynamics remain visible and auditable, while a news module contributes an interpretable, state-dependent correction. Future work could extend this framework to multi-step horizons, option-implied volatility, intraday risk control, or cross-asset settings, as well as explore alternative encoders and summarization schemes that further enrich the news signal while maintaining the same level of transparency.

References

F. Corsi. A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2):174–196, November 2008. ISSN 1479-8417. doi: 10.1093/jjfinec/nbp001. URL `http://dx.doi.org/10.1093/jjfinec/nbp001`.

Robert F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50(4):987, July 1982. ISSN 0012-9682. doi: 10.2307/1912773. URL `http://dx.doi.org/10.2307/1912773`.

Tim Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327, April 1986. ISSN 0304-4076. doi: 10.1016/0304-4076(86)90063-1. URL `http://dx.doi.org/10.1016/0304-4076(86)90063-1`.

Ole E. Barndorff-Nielsen and Neil Shephard. Econometric analysis of realized covariation: High frequency based covariance, regression, and correlation in financial economics. *Econometrica*, 72(3):885–925, May 2004. ISSN 1468-0262. doi: 10.1111/j.1468-0262.2004.00515.x. URL `http://dx.doi.org/10.1111/j.1468-0262.2004.00515.x`.

Torben G Andersen, Tim Bollerslev, Francis X Diebold, and Paul Labys. Modeling and forecasting realized volatility. *Econometrica*, 71(2):579–625, March 2003.

O. E. Barndorff-Nielsen. Power and bipower variation with stochastic volatility and jumps. *Journal of Financial Econometrics*, 2(1):1–37, December 2004. ISSN 1479-8417. doi: 10.1093/jjfinec/nbh001. URL `http://dx.doi.org/10.1093/jjfinec/nbh001`.

Peter Reinhard Hansen and Asger Lunde. Consistent ranking of volatility models. *J. Econom.*, 131(1-2):97–121, March 2006.

Francis X Diebold and Roberto S Mariano. Comparing predictive accuracy. *J. Bus. Econ. Stat.*, 13(3):253, July 1995.

Andrew J. Patton. Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160(1):246–256, January 2011. ISSN 0304-4076. doi: 10.1016/j.jeconom.2010.03.034. URL `http://dx.doi.org/10.1016/j.jeconom.2010.03.034`.

Stavros Degiannakis and George Filis. Forecasting oil price realized volatility using information channels from other asset classes. *Journal of International Money and Finance*, 76:28–49, September 2017. ISSN 0261-5606. doi: 10.1016/j.jimonfin.2017.05.006. URL `http://dx.doi.org/10.1016/j.jimonfin.2017.05.006`.

Dimos S. Kambouroudis, David G. McMillan, and Katerina Tsakou. Forecasting realized volatility: The role of implied volatility, leverage effect, overnight returns, and volatility of realized volatility. *Journal of Futures Markets*, 41(10):1618–1639, July 2021. ISSN 1096-9934. doi: 10.1002/fut.22241. URL `http://dx.doi.org/10.1002/fut.22241`.

B.J. Christensen and N.R. Prabhala. The relation between implied and realized volatility. *Journal of Financial Economics*, 50(2):125–150, November 1998. ISSN 0304-405X. doi: 10.1016/s0304-405x(98)00034-8. URL `http://dx.doi.org/10.1016/S0304-405X(98)00034-8`.

Ser-Huang Poon and Clive W. J Granger. Forecasting volatility in financial markets: A review. *Journal of Economic Literature*, 41(2):478–539, June 2003. ISSN 0022-0515. doi: 10.1257/jel.41.2.478. URL `http://dx.doi.org/10.1257/jel.41.2.478`.

Peter K. Clark. A subordinated stochastic process model with finite variance for speculative prices. *Econometrica*, 41(1):135, January 1973. ISSN 0012-9682. doi: 10.2307/1913889. URL `http://dx.doi.org/10.2307/1913889`.

Thomas W. Epps and Mary Lee Epps. The stochastic dependence of security price changes and transaction volumes: Implications for the mixture-of-distributions hypothesis. *Econometrica*, 44(2):305, March 1976. ISSN 0012-9682. doi: 10.2307/1912726. URL `http://dx.doi.org/10.2307/1912726`.

George E. Tauchen and Mark Pitts. The price variability-volume relationship on speculative markets. *Econometrica*, 51(2):485, March 1983. ISSN 0012-9682. doi: 10.2307/1912002. URL `http://dx.doi.org/10.2307/1912002`.

James M. Patell and Mark A. Wolfson. The intraday speed of adjustment of stock prices to earnings and dividend announcements. *Journal of Financial Economics*, 13 (2):223–252, June 1984. ISSN 0304-405X. doi: 10.1016/0304-405x(84)90024-2. URL `http://dx.doi.org/10.1016/0304-405X(84)90024-2`.

Torben G Anderson, Tim Bollerslev, Francis X Diebold, and Clara Vega. Micro effects of macro announcements: Real-time price discovery in foreign exchange. *American Economic Review*, 93(1):38–62, February 2003. ISSN 0002-8282. doi: 10.1257/000282803321455151. URL `http://dx.doi.org/10.1257/000282803321455151`.

Kenneth R. French and Richard Roll. Stock return variances. *Journal of Financial Economics*, 17(1):5–26, September 1986. ISSN 0304-405X. doi: 10.1016/0304-405x(86) 90004-8. URL `http://dx.doi.org/10.1016/0304-405X(86)90004-8`.

Linda Smith Bamber, Orie E. Barron, and Thomas L. Stober. Trading volume and different aspects of disagreement coincident with earnings announcements. *The Accounting Review*, 72(4):575–597, 1997. ISSN 00014826. URL `http://www.jstor.org/stable/248176`.

Werner Antweiler and Murray Z. Frank. Is all that talk just noise? the information content of internet stock message boards. *The Journal of Finance*, 59(3):1259–1294, June 2004. ISSN 1540-6261. doi: 10.1111/j.1540-6261.2004.00662.x. URL `http://dx.doi.org/10.1111/j.1540-6261.2004.00662.x`.

PAUL C. TETLOCK. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3):1139–1168, May 2007. ISSN 1540-6261. doi: 10.1111/j.1540-6261.2007.01232.x. URL `http://dx.doi.org/10.1111/j.1540-6261.2007.01232.x`.

PAUL C. TETLOCK, MAYTAL SAAR-TSECHANSKY, and SOFUS MACSKASSY. More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance*, 63(3):1437–1467, May 2008. ISSN 1540-6261. doi: 10.1111/j.1540-6261.2008.01362.x. URL `http://dx.doi.org/10.1111/j.1540-6261.2008.01362.x`.

TIM LOUGHRAN and BILL MCDONALD. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65, January 2011. ISSN 1540-6261. doi: 10.1111/j.1540-6261.2010.01625.x. URL http://dx.doi.org/10.1111/j.1540-6261.2010.01625.x.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019. URL https://api.semanticscholar.org/CorpusID:52967399.

Yi Yang, Mark Christopher Siy UY, and Allen Huang. Finbert: A pretrained language model for financial communications, 2020. URL https://arxiv.org/abs/2006.08097.

Tian Guo and Emmanuel Hauptmann. Fine-tuning large language models for stock return prediction using newsflow, 2024. URL https://arxiv.org/abs/2407.18103.

Francesco Audrino, Fabio Sigrist, and Daniele Ballinari. The impact of sentiment and attention measures on stock market volatility. *International Journal of Forecasting*, 36 (2):334–357, April 2020. ISSN 0169-2070. doi: 10.1016/j.ijforecast.2019.05.010. URL http://dx.doi.org/10.1016/j.ijforecast.2019.05.010.

Fernando Moreno-Pino and Stefan Zohren. Deepvol: volatility forecasting from high-frequency data with dilated causal convolutions. *Quantitative Finance*, 24(8): 1105–1127, August 2024. ISSN 1469-7696. doi: 10.1080/14697688.2024.2387222. URL http://dx.doi.org/10.1080/14697688.2024.2387222.

Sophia Zhengzi Li and Yushan Tang. Automated volatility forecasting. *Management Science*, 71(7):6248–6274, July 2025. ISSN 1526-5501. doi: 10.1287/mnsc.2023.01520. URL http://dx.doi.org/10.1287/mnsc.2023.01520.

John Bridle. Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. In D. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 2. Morgan-Kaufmann, 1989. URL https://proceedings.neurips.cc/paper_files/paper/1989/file/0336dcbab05b9d5ad24f4333c7658a0e-Paper.pdf.

Yuxin Shi, Lu Wang, and Chao Liang. Uncertain har-rv models and their extensions: A new perspective on forecasting the volatility of china's crude oil futures. *Journal of Futures Markets*, October 2025. ISSN 1096-9934. doi: 10.1002/fut.70049. URL http://dx.doi.org/10.1002/fut.70049.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference, 2024. URL https://arxiv.org/abs/2412.13663.

Marc Brinner and Sina Zarriess. Semcse: Semantic contrastive sentence embeddings using llm-generated summaries for scientific abstracts, 2025. URL https://arxiv.org/abs/2507.13105.