Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Event-Conditioned News Representations for Discovering Textual Information Regimes in Commodity Forecasting

Mihael Arcan

mihael.arcan@gmail.com

Home Lab

Research Article

Keywords: NLP, financial NLP, Textual information, Forecast horizons, Market forecasting

Posted Date: May 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9814312/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Event-Conditioned News Representations for Discovering Textual Information Regimes in Commodity Forecasting

Mihael Arcan[1]

[1]Home Lab, Galway, Ireland

May 21, 2026

Abstract

News is a central channel of information arrival in financial markets, but its predictive value is not uniform across assets, horizons, and regimes. This paper studies when NLP-derived news signals matter for commodity directional forecasting. We introduce an event-conditioned benchmark across eight commodity markets and five forecast horizons, combining price, macroeconomic, and headline-derived news features under global and asset-specific news provenance views. Event intensity is used to define the unconditional Full sample and progressively selective Top-q information buckets, enabling a direct test of whether textual information becomes more predictive during high-information periods.

Under a unified expanding-window walk-forward protocol, we evaluate ARIMA, LSTM, TCN, PatchTST, Ridge, XGBoost, and LightGBM across strict and max-data data-retention settings. Rather than reporting only winning configurations, we provide full-trajectory evidence: global family averages, Full-to-Top-q bucket transitions, best-per-asset–horizon frontiers, weak-configuration evidence, distribution tails, and robustness across data-retention protocols.

Results show that news-driven predictive edge is regime-local rather than uniform. Full-space directional accuracy remains close to chance for most model families, but targeted event-conditioned asset–horizon regions exhibit substantially stronger performance, especially for tabular machine-learning models that combine price, macro, news, and regime-conditioned features. At the same time, weak tails remain visible, showing that event filtering and news augmentation can degrade performance when misaligned with the asset, horizon, or feature regime. Strict filtering identifies sharper high-signal regions, while max-data improves coverage and average stability.

The central implication for financial NLP is that news text should not be evaluated only as a static predictive feature. In market forecasting tasks, textual information also functions as a regime-discovery signal: it helps identify when the relationship between language, market context, and future direction changes. The findings support a regime-specific view of NLP-based financial forecasting, in which headline-derived representations are most informative under particular asset, horizon, and event-intensity conditions rather than uniformly across the full prediction space.

1 Introduction

Commodity markets are driven by information. Energy, metals, and agricultural prices react to macroeconomic announcements, geopolitical shocks, inventory releases, weather events, supply disruptions, and changes in global risk sentiment. Much of this information reaches the market first through text, especially through news headlines. For forecasting, this makes news more than an

auxiliary feature source: it is a real-time record of information arrival. From a financial NLP perspective, this creates a representation problem and an evaluation problem. The representation problem is how to convert timestamped headlines into market-relevant textual signals. The evaluation problem is how to determine when those signals are actually informative. A headline representation may encode useful textual context, yet still be irrelevant to a given asset, horizon, or market state. This paper therefore treats news text not only as content to be represented, but also as evidence of changing information regimes.

The central insight of this paper is that the value of news is regime-specific. A headline does not have the same predictive meaning across all commodities, horizons, and market states. A geopolitical event may be highly informative for crude oil, while a monetary-policy signal may affect gold through risk sentiment and copper through growth expectations. Similarly, a supply-related headline may matter immediately for one commodity but only gradually for another. These differences suggest that NLP-derived news representations should be evaluated not only by whether they improve average prediction, but by whether they identify the market regimes in which directional predictability is strongest.

This paper studies event-conditioned textual news representations for commodity directional forecasting. The forecasting task provides a downstream testbed for evaluating when headline-derived NLP signals carry predictive information. The forecasting models are used as downstream probes of textual signal usefulness, not as the primary contribution of the paper. We formulate the task as binary prediction of forward price direction and evaluate whether textual news representations improve out-of-sample directional accuracy across assets and horizons. The goal is not to propose a trading strategy, but to study when textual information becomes predictive, which market settings benefit from it, and how event-conditioned news representations interact with model families and feature regimes.

We introduce an event-conditioned benchmark in which news is represented through two provenance views. `GlobalNews` captures broad market-relevant information, while `AssetNews` captures commodity-linked information. Event intensity is used to define the full sample and progressively selective Top-q buckets. This design makes it possible to test whether forecasting performance strengthens during periods of elevated textual information arrival.

The benchmark covers eight commodities and five forecast horizons, producing 40 asset–horizon cells. We combine price, macroeconomic, and news-derived features, and compare ARIMA, LSTM, TCN, PatchTST, Ridge, XGBoost, and LightGBM under a unified expanding-window walk-forward protocol. Two data-retention settings are evaluated: a strict protocol for high-precision signal discovery and a max-data protocol for broader coverage and robustness.

The study is organised around four research questions: **RQ1** whether NLP-derived news features improve commodity directional forecasting across the full configuration space; **RQ2** whether predictive gains are concentrated in event-conditioned asset–horizon regimes; **RQ3** which model families most effectively exploit price, macro, and news-derived information; and **RQ4** how observed gains change under strict filtering versus max-data retention.

From an NLP perspective, the contribution is an event-conditioned evaluation framework for financial text. Rather than treating news representations as static feature additions, we evaluate them as tools for regime discovery. Their value lies not only in encoding headline content, but also in identifying when information arrival is likely to matter for future market direction.

The paper makes four contributions. First, it introduces an event-conditioned evaluation framework for financial NLP, where headline-derived textual signals are assessed not only by average predictive accuracy but also by their ability to identify information regimes. Second, it operationalises textual information arrival through global and asset-specific news provenance views, allowing headline signals to be evaluated under both broad-market and commodity-linked interpretations. Third,

it provides a full-trajectory evaluation protocol for text-based financial forecasting, including global means, event-bucket transitions, best-per-asset–horizon frontiers, weak configurations, distribution tails, and strict-vs-max-data robustness. Fourth, it shows that the value of textual news representations is local and conditional: strong performance emerges only when text provenance, event intensity, asset, horizon, feature block, and model family are aligned.

The main finding is that NLP-derived news signals reveal local windows of directional predictability in commodity markets. Predictive performance is strongest when textual information intensity, asset context, forecast horizon, feature representation, model family, and data-retention protocol are aligned. This supports a regime-specific view of financial NLP: the question is not simply whether news predicts prices, but how news identifies the market conditions under which prediction becomes possible.

2 Related Work

This paper connects four areas: commodity forecasting, financial NLP, event-based prediction, and robust financial machine-learning evaluation. Commodity forecasting has traditionally used econometric and time-series models, with more recent work applying machine-learning and deep-learning methods to energy, metals, and agricultural markets [1, 19, 16]. These studies show that nonlinear models can be useful in selected settings, but performance is often asset-specific, horizon-specific, and sensitive to feature construction. This motivates evaluation designs that preserve asset–horizon heterogeneity rather than relying only on aggregate averages.

A related literature studies whether text helps explain or predict financial markets. Early work linked media tone to market movements and trading volume [18], while domain-specific financial dictionaries showed that general sentiment resources can misclassify financial language [15]. Subsequent studies have used news, filings, earnings calls, and social media to extract sentiment, topics, embeddings, uncertainty measures, and event indicators for return, volatility, or direction prediction [5]. The common premise is that text contains information not yet fully reflected in prices. However, textual signals are noisy, time-dependent, and context-sensitive; their value depends on source, timing, asset relevance, and market regime.

Event-based forecasting treats news not as a generic feature stream, but as information that may change the forecasting environment. Ding et al. [9] show that structured event representations extracted from news can support stock prediction. Regime-specific forecasting is closely related: the mapping from predictors to outcomes may change across volatility states, macro regimes, or information environments. This perspective is especially relevant for commodities, where energy, metals, and agricultural markets respond to different information channels. The present paper extends this idea by using NLP-derived event intensity to define information regimes for commodity directional forecasting.

NLP representations for financial markets have evolved from dictionary-based sentiment and bag-of-words features to contextual and domain-adapted language models. General-purpose models such as BERT [8] and financial-domain variants such as FinBERT [2] demonstrate the value of richer textual representations. Yet accurate text representation does not guarantee forecasting usefulness: a headline may be linguistically informative but irrelevant to the asset, horizon, or regime being predicted. For this reason, we evaluate headline-derived representations through downstream event-conditioned forecasting performance.

The modelling literature also motivates comparing heterogeneous forecasting families. Classical statistical models provide transparent baselines; LSTM, TCN, and PatchTST represent modern sequence-learning approaches [3, 17]; and tabular models such as Ridge, XGBoost, and LightGBM

can capture structured interactions among price, macro, news, and regime features. In event-conditioned settings, such interactions may be central because the value of news can depend jointly on asset state, macro context, event intensity, provenance, and horizon.

Finally, financial machine-learning studies are vulnerable to look-ahead bias, overfitting, data snooping, and winner-only reporting. Prior work emphasises leakage-aware validation, backtest-overfitting risk, and the dangers of selecting from large model spaces [14, 4, 20, 10]. This paper addresses these concerns through expanding-window walk-forward validation, strict and max-data protocols, full configuration retention, weak-tail reporting, and complete asset–horizon matrices. The contribution is therefore both empirical and methodological: we ask not simply whether news improves prediction on average, but when textual information becomes useful, for which commodities, at which horizons, and under which modelling conditions.

3 Methodology

This section describes the forecasting procedure used in the study. The methodology follows a reproducible recipe: define the directional forecasting task, construct time-aligned inputs, represent news events, build feature blocks, create event-conditioned samples, evaluate models with walk-forward validation, test performance against chance, and aggregate configuration-level outputs into the result tables.

The main design principle is temporal discipline. At every stage, models only use information that would have been available at the time of prediction. This rule applies to price features, macroeconomic variables, headline-derived news signals, feature scaling, event-bucket construction, model fitting, and evaluation.

3.1 Forecasting Task

The task is binary directional forecasting. For each commodity asset and forecast horizon, the model predicts whether the future price movement is positive or non-positive.

For asset a at time t, let $P_{a,t}$ denote the observed price. The forward return over horizon h is:

$$r_{a,t,h} = \frac{P_{a,t+h} - P_{a,t}}{P_{a,t}}.$$

The target label is:

$$y_{a,t,h} = \begin{cases} 1, & r_{a,t,h} > 0, \\ 0, & r_{a,t,h} \leq 0. \end{cases}$$

A value of 1 represents an upward future move, while a value of 0 represents a flat or downward future move. The forecasting problem is therefore directional rather than magnitude-based.

The primary evaluation metric is directional accuracy (DA), defined as the share of correct out-of-sample directional predictions:

$$\text{DA} = \frac{\text{number of correct directional predictions}}{\text{number of out-of-sample predictions}}.$$

Directional accuracy is used as a classification metric for evaluating predictive signal, not as a proxy for economic profitability. The benchmark does not simulate trading, portfolio allocation, transaction costs, slippage, liquidity constraints, leverage, or risk-adjusted returns. The empirical question is whether event-conditioned textual information improves directional classification under temporally disciplined validation.

Date	Category	Provenance	Matched asset	Headline (excerpt)
2022-09-21	WORLD NEWS	GlobalNews	–	Biden At UN To Call Russian War An Affront To Body's Charter
2022-09-20	WORLD NEWS	GlobalNews	–	4 Russian-Controlled Ukrainian Regions Schedule Votes This Week To Join Russia
2022-09-19	POLITICS	GlobalNews + AssetNews	Copper	Biden Says U.S. Forces Would Defend Taiwan If China Invaded
2022-09-13	U.S. NEWS	GlobalNews + AssetNews	Gold, Silver	U.S. Inflation Falls For 2nd Straight Month On Lower Gas Costs
2022-09-11	WORLD NEWS	GlobalNews	–	Last Reactor At Ukraine's Zaporizhzhia Nuclear Plant Stopped

Table 1: Illustrative strict-pipeline examples from the headline dataset. Provenance assignment follows implemented category filtering, macro-keyword filtering, and asset-keyword matching rules.

3.2 Time-Aligned Information Sets

For every prediction, the model input is constructed from information available at or before the forecast origin. The input information comes from three sources: price history, macroeconomic context, and news-derived event features.

Price features describe recent market behaviour. Macro features describe the broader economic environment. News features are derived from timestamped headlines and represent textual information arrival. The same temporal rule is applied to all feature types: no future price, macro, or news information is used when making a prediction.

3.3 News Event Representation

The news component converts timestamped headlines into daily event signals aligned with commodity price observations. The objective is to identify information-rich periods in which directional predictability may improve, rather than to assume that all headlines are equally informative.

Each news item contains a timestamp, headline text, and metadata. The strict pipeline applies category filtering and macro-keyword filtering to identify broad market-relevant headlines. Surviving headlines enter the `GlobalNews` view. In parallel, asset-keyword matching assigns headlines to commodity-specific `AssetNews` views when an asset rule is satisfied. The max-data protocol relaxes these filtering rules to retain a broader set of observations.

Table 1 shows representative examples from the headline dataset after strict-pipeline filtering. The examples illustrate how a headline may enter only `GlobalNews`, or both `GlobalNews` and `AssetNews` when commodity-linked keyword rules match.

After filtering and provenance assignment, headlines are aggregated at daily frequency. The resulting daily news signals are aligned with the commodity observation calendar and used both as news-derived features and as the basis for event-intensity scoring. Higher event-intensity values correspond to days with stronger headline-based information arrival.

For prediction at time t, only headlines published at or before t are used. No future headline enters feature construction, event-intensity scoring, bucket assignment, model training, or evaluation.

Concretely, the implemented news representation combines daily aggregated sentence-transformer embeddings, headline-volume and category-count signals, and rule-based provenance assignment

(GlobalNews/AssetNews).

3.4 Headline-Derived Textual Features

The textual component produces structured headline-derived features rather than feeding raw text directly into forecasting models. The implemented pipeline uses a sentence-transformer encoder (all-MiniLM-L6-v2) to obtain headline embeddings, then aggregates them to daily mean vectors. These daily embedding features are stored as emb_* columns and used in the News feature block.

In parallel, the pipeline constructs interpretable daily textual signals from metadata: total headline volume (article_count) and category-count indicators (cat_*). Provenance views are then formed through rule-based assignment: GlobalNews via category + macro-keyword filtering, and AssetNews via commodity-specific keyword matching.

To reduce leakage risk and dimensional instability, embedding compression is applied fold-safely inside walk-forward training (PCA fit on training folds only, then applied to test folds). Event-intensity subsets (Top-q) are defined from train-anchored quantiles of daily headline volume, so bucketing never uses future information.

This representation is deliberately hybrid: semantically informed (contextual embeddings) and operationally interpretable (volume/category/provenance rules). The NLP contribution is therefore the event-conditioned textual representation and its leakage-aware downstream evaluation, rather than pretraining or fine-tuning a new language model.

3.5 Textual information regimes

We use the term textual information regime to describe a time-local market state in which headline-derived signals have elevated relevance for a forecasting task. A textual information regime is not defined only by the presence of news, but by the interaction between news intensity, provenance, asset context, forecast horizon, and feature representation. In this framing, NLP-derived signals serve two roles. First, they encode information from headline text into structured features. Second, they provide a conditioning signal that helps identify when the relationship between text and future market direction may differ from the unconditional sample.

This distinction is important because poor average performance of textual features does not necessarily imply that text is uninformative. Instead, text may be informative only under specific event-conditioned regimes. The benchmark is designed to evaluate this possibility directly by comparing unconditional and event-conditioned samples while retaining weak and negative configurations.

3.6 Feature Blocks

The study evaluates whether different combinations of information improve forecasting. The core feature blocks are:

$$\{\texttt{Price}, \texttt{Macro}, \texttt{News}, \texttt{Price+Macro}, \texttt{Price+News}, \texttt{Macro+News}, \texttt{All}\}.$$

The single-modality blocks test whether one information source is sufficient. The combined blocks test whether predictive value emerges when information sources are used together. For example, Price+News tests whether news adds value beyond recent market behaviour, while Macro+News tests whether news is more useful when interpreted in macroeconomic context.

For the tabular machine-learning models, additional regime-aware feature variants are evaluated:

$$\{\texttt{RegimeSpecific}, \texttt{All+RegimeInteraction}, \texttt{All+RegimeGated}\}.$$

These variants allow the benchmark to test whether predictive relationships change across market states.

3.7 Event Conditioning

The central methodological idea is event conditioning. Instead of assuming that news is equally useful at all times, the study tests whether news becomes more predictive during periods of elevated information arrival.

Each observation receives an event-intensity score from the news pipeline. This score is used to divide observations into event buckets:

$$\{\texttt{Full}, \texttt{Top50}, \texttt{Top60}, \texttt{Top70}, \texttt{Top80}\},$$

with `Top90` generated as a candidate bucket and retained only when minimum-row constraints are satisfied.

The `Full` bucket contains all eligible observations. The Top-q buckets contain increasingly event-intensive observations. For example, `Top80` keeps only the highest-intensity observations according to the event-intensity score.

Event conditioning is treated as an empirical question, not an assumption. High news intensity may identify informative regimes, but it may also remove useful observations, reduce sample size, or concentrate noisy information. The benchmark therefore evaluates both successful and weak event-conditioned configurations.

3.8 Walk-Forward Validation

All models are evaluated using expanding-window walk-forward validation. This validation design mirrors how a forecasting system would operate in real time.

For each configuration, the procedure is:

1. Arrange observations in chronological order.
2. Train the model on an initial historical window.
3. Predict the next out-of-sample test segment.
4. Record the predictions and realised labels.
5. Expand the training window forward.
6. Repeat the process until all valid test segments have been evaluated.

At every fold, the model is trained only on past observations and evaluated only on future observations. Test observations are never used to fit transformations, select features, construct event buckets, or train the model for that fold.

3.9 Model-Family Evaluation

The benchmark compares seven model families under the same forecasting target, validation design, and evaluation metric.

ARIMA is used as a classical time-series baseline. LSTM, TCN, and PatchTST are used as neural sequence baselines in a dedicated baseline path. Ridge, XGBoost, and LightGBM are used

as tabular machine-learning models over engineered price, macro, news, and regime-conditioned features.

The purpose of comparing these families is not only to find the best model. The comparison also tests which modelling assumptions are most compatible with event-conditioned commodity forecasting: autoregressive dynamics, neural sequence learning, linear tabular structure, or nonlinear feature interactions.

3.10 Strict and Max-Data Protocols

The full benchmark is evaluated under two data-retention protocols. The **strict protocol** applies conservative news filtering, with category and keyword filters enabled. It is designed to prioritise cleaner samples and sharper signal discovery. The **max-data protocol** retains a broader set of observations primarily by disabling category and keyword news filters. It is designed to increase coverage and test whether results remain stable under less restrictive data retention. Both protocols use the same target definition, feature blocks, event buckets, model families, validation design, metric, statistical tests, and aggregation logic. The only intended difference is the data-retention rule.

3.11 Statistical Testing

For every configuration, directional accuracy is tested against a no-skill benchmark of 0.5. A value of 0.5 corresponds to chance-level directional prediction.

The study uses a one-sided normal-approximation test of directional accuracy against the chance benchmark of 0.5. Wilson confidence intervals are also reported to describe uncertainty around the directional accuracy estimate [21].

The statistical tests are used descriptively. Because the benchmark contains many configurations, individual high-performing cases are not interpreted as universal proof of model superiority. Instead, the paper focuses on family-level summaries, event-bucket transitions, weak-tail behaviour, and strict-vs-max-data robustness. The reported p-values should therefore be read as nominal, configuration-level diagnostics rather than multiplicity-corrected confirmatory evidence.

3.12 Configuration-Level Outputs

Each experimental run produces one configuration-level record. A configuration is defined by the asset, horizon, news provenance view, feature block, event bucket, model family, and data-retention protocol.

For each configuration, the benchmark stores:

- mean directional accuracy across out-of-sample folds;
- number of valid test observations;
- number and share of correct predictions;
- one-sided normal-approximation p-value against 0.5;
- Wilson confidence interval bounds;
- identifiers for asset, horizon, feature block, news view, event bucket, model family, and protocol.

This configuration-level record is the basic unit of the empirical analysis. Every table in the Results section is generated from these records.

Component	Specification
Assets	8 commodities
Horizons	1, 3, 5, 10, 20
Target	Binary forward direction
Modalities	Price, macro, news
News provenance	GlobalNews, AssetNews
Event buckets	Full, Top50, Top60, Top70, Top80; Top90 if min-row constraint is met
Model families	ARIMA, LSTM, TCN, PatchTST, Ridge, XGBoost, LightGBM
Protocols	Strict, MaxData
Validation	Expanding-window walk-forward
Metric	Directional accuracy
Testing	One-sided normal-approximation test vs 0.5
Uncertainty	Wilson confidence intervals

Table 2: Summary of the experimental benchmark.

3.13 Aggregation Strategy

To avoid winner-only reporting, we aggregate the configuration-level outputs along five complementary views. First, *full-space family performance* reports the mean directional accuracy across all configurations within each model family and protocol. Second, *bucket-level performance* reports the mean, median, and nominal significance share within each event bucket. Third, the *best-per-asset–horizon frontier* selects the best configuration for each asset–horizon cell and averages these best-cell values by family. Fourth, *tail behaviour* reports the low-5% and high-5% mean directional accuracy to summarise downside and upside configuration risk. Fifth, *weak-configuration evidence* explicitly reports the weakest specifications so that negative evidence remains visible. Together, these summaries distinguish broad average performance, local regime discovery, event-bucket behaviour, and configuration risk.

4 Experimental Setup

This section records the empirical ingredients used to execute the benchmark. Whereas Section 3 defines the forecasting procedure, this section specifies the data sources, benchmark dimensions, model families, realised configuration counts, and stored outputs required for reproducibility. Table 2 summarises the full experimental design.

4.1 Data Sources and Alignment

The benchmark combines three aligned data sources: timestamped news headlines, daily commodity prices, and macroeconomic context variables. Table 3 summarises the empirical data components used to construct the forecasting panel.

The textual source is the HuffPost News Category Dataset.[1] The dataset is a JSON-lines corpus of timestamped headlines with metadata fields including `headline`, `category`, `date`, `authors`, and `link`. The news file used in this study spans 2012-01-28 to 2022-09-23. Headlines are processed into daily `GlobalNews` and `AssetNews` signals and used to construct news-derived features and event-intensity scores.

Commodity price series are taken from Macrotrends.[2] The price panel contains synchronized

[1] `https://www.kaggle.com/datasets/rmisra/news-category-dataset`
[2] `https://www.macrotrends.net/`

Dataset component	Description
News text	Timestamped HuffPost headlines with category, date, author, and link metadata
Commodity prices	Daily price series for Gold, Silver, Platinum, Copper, Crude Oil, Heating Oil, Corn, and Coffee
Macroeconomic features	Date-aligned macro-regime and cross-market context variables
News unit	Headline-level text
News aggregation	Daily headline aggregation into global and asset-specific news views
Prediction targets	Forward returns at 1, 3, 5, 10, and 20 steps, evaluated by directional sign
Forecasting panel	Daily asset-time observations formed by aligning price, macro, and news features
Validation design	Expanding-window walk-forward evaluation

Table 3: Data components used to construct the forecasting panel.

daily series for the eight benchmark commodities: `Gold`, `Silver`, `Platinum`, `Copper`, `Crude Oil`, `Heating Oil`, `Corn`, and `Coffee`. These series are used to construct forward-horizon return targets and price-derived predictors.

Macroeconomic context variables are merged by date with the commodity-price panel. These variables provide cross-market context, including rates, foreign-exchange, and related macro indicators, for regime-aware forecasting.

All data sources are aligned to a common daily observation calendar. For a forecast made at date t, predictors are restricted to information available at or before t. Observations with insufficient price history, missing forward targets, or unavailable required feature blocks are excluded according to the corresponding data-retention protocol.

4.2 Benchmark Dimensions

The asset universe contains eight commodities: `Coffee`, `Copper`, `Corn`, `Crude`, `Gold`, `HeatingOil`, `Platinum`, and `Silver`. These assets cover agricultural commodities, industrial metals, precious metals, and energy markets. Forecast horizons are fixed at 1, 3, 5, 10, and 20 steps, yielding 40 asset–horizon cells. The target is binary forward direction: 1 for a positive future return and 0 for a non-positive future return.

The benchmark uses three input modalities: **Price**, containing price-derived market-history features; **Macro**, containing macroeconomic and regime-context features; and **News**, containing headline-derived NLP features and event-intensity signals. News is represented under two provenance views: `GlobalNews`, for broad market-relevant news, and `AssetNews`, for commodity-linked news. This design permits comparison between broad information conditions and asset-specific textual signals.

4.3 Feature Settings and Event Buckets

The core tabular feature blocks are `Price`, `Macro`, `News`, `Price+Macro`, `Price+News`, `Macro+News`, and `All`. These blocks evaluate single-modality signals, pairwise modality combinations, and the combined price–macro–news representation. The regime-conditioned variants are `RegimeSpecific`, `All+RegimeInteraction`, and `All+RegimeGated`; these variants test whether feature usefulness changes across market states.

Model family	Strict configs	MaxData configs
ARIMA	40	40
LSTM	40	–
TCN	40	40
PatchTST	40	40
Ridge	1950	3000
XGBoost	1950	3000
LightGBM	1950	3000

Table 4: Realised configuration counts by model family and data-retention protocol.

Event conditioning uses `Full`, `Top50`, `Top60`, `Top70`, and `Top80`. `Top90` is generated as a candidate bucket and retained only when the minimum-row constraint is satisfied. The `Full` bucket contains all eligible observations, while Top-q buckets retain observations with higher event-intensity scores.

4.4 Model Families

Seven model families are evaluated: `ARIMA`, `LSTM`, `TCN`, `PatchTST`, `Ridge`, `XGBoost`, and `LightGBM`. They are selected to compare classical time-series modelling, neural sequence learning, regularised linear learning, and nonlinear tabular learning under a common forecasting protocol.

`ARIMA` is included as a classical statistical baseline following the Box–Jenkins time-series modelling framework [6]. It tests whether autoregressive temporal structure alone provides useful directional signal. `LSTM` is a recurrent neural sequence model designed to capture temporal dependence through gated memory [11]. `TCN` is a causal convolutional sequence model that captures temporal patterns through dilated convolutions and has been proposed as a strong alternative to recurrent sequence models [3]. `PatchTST` is a transformer-based time-series model that represents modern patch-based neural forecasting by segmenting time series into patch tokens [17].

`Ridge` is a regularised linear tabular model used to test whether stable linear combinations of engineered features are predictive [12]. `XGBoost` and `LightGBM` are gradient-boosted tree models used to capture nonlinear interactions among price, macro, news, event-intensity, and regime-conditioned features [7, 13]. Together, these families provide a deliberately heterogeneous benchmark: a classical autoregressive baseline, neural sequence baselines, a regularised linear tabular baseline, and high-capacity nonlinear tabular learners.

4.5 Protocol Variants and Configuration Counts

Two data-retention protocols are used. `Strict` applies conservative filtering and tighter data-alignment requirements, while `MaxData` uses broader retention and larger effective sample coverage. Both protocols share the same target definition, validation design, metric, statistical test, and aggregation logic. Their difference is the retained sample produced by the filtering and alignment rules. Table 4 reports the realised configuration counts by model family and data-retention protocol.

The 40 configurations for ARIMA and the neural sequence baselines correspond to the 40 asset–horizon cells. The larger counts for Ridge, XGBoost, and LightGBM reflect the expanded tabular grid over feature blocks, news provenance views, event buckets, and regime-conditioned variants. LSTM is reported only under the strict protocol because the realised max-data benchmark contains ARIMA, TCN, PatchTST, Ridge, XGBoost, and LightGBM outputs.

Protocol	Family	n_{cfg}	DA_{all}	DA_{Full}	$\text{DA}_{\text{Top}q}$	DA_{best}	DA_{max}	DA_{low5}	DA_{high5}	Sig. share
Strict	ARIMA	40	0.4972	0.4972	–	0.4972	0.6038	0.4146	0.5875	0.175
	LSTM	40	**0.5113**	0.5113	–	0.5113	0.5745	0.4540	0.5698	**0.225**
	TCN	40	0.4930	0.4930	–	0.4930	0.5569	0.4406	0.5473	0.100
	PatchTST	40	0.4889	0.4889	–	0.4889	0.5564	0.4122	0.5554	0.075
	Ridge	1950	0.4915	0.4959	0.4900	0.5968	0.7867	0.3655	0.6359	0.082
	XGBoost	1950	0.4929	**0.5026**	0.4895	**0.6071**	**0.8545**	0.3607	0.6426	0.070
	LightGBM	1950	0.4942	0.5007	**0.4920**	0.6069	0.8000	0.3644	**0.6429**	0.070
MaxData	ARIMA	40	0.4933	0.4933	–	0.4933	0.5622	0.4372	0.5568	**0.175**
	TCN	40	0.4949	0.4949	–	0.4949	0.5465	0.4387	0.5423	0.075
	PatchTST	40	0.4908	0.4908	–	0.4908	0.5484	0.3736	0.5466	0.100
	Ridge	3000	0.4976	0.5010	0.4963	0.5680	0.7263	0.3601	0.6446	0.131
	XGBoost	3000	0.5068	0.5087	**0.5061**	**0.5885**	**0.8154**	0.3820	0.6463	0.139
	LightGBM	3000	**0.5070**	**0.5099**	0.5060	0.5862	0.8000	0.3774	**0.6528**	0.140

Table 5: Baseline-to-best progression by model family and data-retention protocol. DA_{all} is the mean directional accuracy across all configurations; DA_{Full} is the mean for the unconditional Full bucket; $\text{DA}_{\text{Top}q}$ averages event-conditioned Top-q buckets; DA_{best} is the mean best-per-asset–horizon value; DA_{max} is the best observed configuration; DA_{low5} and DA_{high5} are the low- and high-5% tail means. Bold values highlight the strongest headline results within each protocol.

4.6 Outputs and Reproducibility

For every configuration, the benchmark stores directional accuracy, the number of out-of-sample test observations, the number of correct predictions, the one-sided p-value against 0.5, Wilson confidence interval bounds, and configuration identifiers. These configuration-level outputs are the source for all reported result tables. The code framework used to implement the benchmark is publicly available on GitHub.[3]

5 Results

This section reports the full empirical trajectory from global average behaviour to local best-case performance. The emphasis is on distributional evidence rather than winner-only reporting: we retain weak configurations, non-winning model families, bucket-level transitions, low and high tails, and strict-vs-max-data robustness. This structure allows us to distinguish broad average predictability from sparse regime-specific predictability.

Because the benchmark contains many configurations, the results are interpreted as regime-discovery evidence rather than as confirmatory proof of universal predictability. High-performing local configurations identify candidate asset–horizon–event regimes where textual information may be useful, while weak configurations indicate where the same textual signals fail or become misleading.

5.1 Baseline-to-Best Trajectory Across All Families

Table 5 summarises the full performance trajectory by model family and data-retention protocol. The table separates average behaviour over the full configuration space, unconditional Full-bucket performance, event-conditioned Top-q performance, best-per-asset–horizon performance, extreme

[3] `https://github.com/miharcan/event-driven-alpha`

observed performance, tail behaviour, and the share of configurations exceeding the nominal significance threshold.

The first pattern is that full-space directional accuracy remains close to chance for most families. Under the strict protocol, ARIMA, TCN, and PatchTST cluster around 0.49–0.50, while LSTM is modestly higher at 0.5113. The tabular families also have near-chance full-space means: Ridge reaches 0.4915, XGBoost 0.4929, and LightGBM 0.4942. This shows that event-driven predictive signal is not uniformly distributed across the full configuration space.

The second pattern is that local performance is much stronger than global average performance. Under the strict protocol, the tabular families achieve substantially higher best-per-asset–horizon means: XGBoost reaches 0.6071, LightGBM 0.6069, and Ridge 0.5968. These values are well above the corresponding full-space means and show that the main predictive gains are concentrated in selected asset–horizon regimes rather than expressed uniformly across all specifications.

The third pattern is that strong local performance coexists with weak tails. For the strict tabular families, the low-5% tail means are approximately 0.36–0.37, while the high-5% means are above 0.63. This wide spread confirms that configuration choice is central: the same model family can contain both high-signal and materially weak specifications.

The strict and max-data protocols provide complementary evidence. MaxData improves broad average behaviour for the tabular families: XGBoost increases from 0.4929 to 0.5068, and LightGBM from 0.4942 to 0.5070. It also raises the share of nominally significant configurations. By contrast, strict filtering produces stronger best-per-asset–horizon performance, with XGBoost declining from 0.6071 under strict filtering to 0.5885 under MaxData. This indicates a precision–coverage trade-off: strict filtering is better at isolating sharp local regimes, while MaxData improves average stability and coverage.

5.2 From Full Baseline to Event-Conditioned Buckets

Table 6 reports the transition from the unconditional `Full` sample to increasingly event-conditioned Top-q buckets. This table is central to the paper because it tests whether news-derived event intensity changes average forecasting behaviour, rather than only producing isolated best-case configurations.

Under the strict protocol, event conditioning is selective rather than monotonic. For the tabular families, the highest mean DA is obtained in the `Full` bucket: Ridge reaches 0.4959, XGBoost reaches 0.5026, and LightGBM reaches 0.5007. More aggressive filtering does not automatically improve performance. In particular, the `Top80` bucket is weak under strict filtering, with XGBoost falling to 0.4693 and LightGBM to 0.4697. This indicates that high event intensity alone is not sufficient: the event bucket must also preserve enough informative observations and align with the relevant asset, horizon, feature block, and model family.

The max-data protocol reveals a different pattern. With broader sample retention, event-conditioned buckets become more productive for the tabular families. Ridge reaches its strongest bucket-level mean at `Top70` with a DA of 0.5237. XGBoost improves from 0.5087 in the `Full` bucket to 0.5212 in `Top70` and 0.5225 in `Top80`. LightGBM follows a similar pattern, increasing from 0.5099 in `Full` to 0.5174 and 0.5176 in `Top70` and `Top80`. These results show that event conditioning can improve average directional performance when the retained sample is sufficiently broad.

The contrast between strict and max-data is informative. Strict filtering identifies sharper local regimes elsewhere in the benchmark, but at the bucket-average level aggressive event filtering can become too selective. Max-data retains more observations and therefore allows high-intensity news regimes to remain statistically useful. The result is a precision–coverage trade-off: event

Setting	Family	Bucket	Configs	Mean DA	Median DA	Share($p \leq 0.05$)
Strict	ARIMA	Full	40	0.4972	0.4985	0.175
Strict	LSTM	Full	40	0.5113	0.5080	0.225
Strict	TCN	Full	40	0.4930	0.4945	0.100
Strict	PatchTST	Full	40	0.4889	0.4901	0.075
Strict	Ridge	Full	500	**0.4959**	0.4922	0.106
Strict	Ridge	Top50	400	0.4901	0.4848	0.062
Strict	Ridge	Top60	400	0.4948	0.4957	0.085
Strict	Ridge	Top70	400	0.4912	0.4870	0.077
Strict	Ridge	Top80	250	0.4805	0.4800	0.068
Strict	XGBoost	Full	500	**0.5026**	0.5010	0.088
Strict	XGBoost	Top50	400	0.4934	0.4933	0.070
Strict	XGBoost	Top60	400	0.4929	0.4957	0.075
Strict	XGBoost	Top70	400	0.4949	0.4957	0.065
Strict	XGBoost	Top80	250	0.4693	0.4667	0.036
Strict	LightGBM	Full	500	**0.5007**	0.5000	0.086
Strict	LightGBM	Top50	400	0.4935	0.4848	0.068
Strict	LightGBM	Top60	400	0.4971	0.4957	0.068
Strict	LightGBM	Top70	400	0.4993	0.4957	0.072
Strict	LightGBM	Top80	250	0.4697	0.4667	0.040
MaxData	ARIMA	Full	40	0.4933	0.4841	0.175
MaxData	TCN	Full	40	0.4949	0.4961	0.075
MaxData	PatchTST	Full	40	0.4908	0.4964	0.100
MaxData	Ridge	Full	800	0.5010	0.5027	0.136
MaxData	Ridge	Top50	800	0.4905	0.4844	0.090
MaxData	Ridge	Top60	800	0.4909	0.4885	0.126
MaxData	Ridge	Top70	300	**0.5237**	0.5250	0.293
MaxData	Ridge	Top80	300	0.4990	0.5167	0.073
MaxData	XGBoost	Full	800	0.5087	0.5100	0.207
MaxData	XGBoost	Top50	800	0.4990	0.4969	0.092
MaxData	XGBoost	Top60	800	0.5016	0.5021	0.111
MaxData	XGBoost	Top70	300	0.5212	0.5100	0.180
MaxData	XGBoost	Top80	300	**0.5225**	0.5167	0.113
MaxData	LightGBM	Full	800	0.5099	0.5097	0.207
MaxData	LightGBM	Top50	800	0.5014	0.5000	0.092
MaxData	LightGBM	Top60	800	0.5019	0.5000	0.113
MaxData	LightGBM	Top70	300	0.5174	0.5100	0.187
MaxData	LightGBM	Top80	300	**0.5176**	0.5167	0.113

Table 6: Path from the unconditional Full sample to event-conditioned buckets for every family. Bold values indicate the highest Mean DA within each tabular model family and data-retention protocol.

Asset	Horizon	ARIMA	LSTM	TCN	PatchTST	Ridge	XGBoost	LightGBM
Coffee	1	0.5365	0.5301	0.4902	0.5020	0.5135	**0.6000**	0.5652
Coffee	3	0.5404	0.5650	0.5020	0.5176	0.5600	0.5697	**0.5818**
Coffee	5	0.5712	0.5631	0.5373	0.5059	**0.5826**	0.5467	0.5333
Coffee	10	0.5212	0.5456	0.5569	0.4647	**0.5879**	0.5867	0.5867
Coffee	20	0.5192	0.5631	0.5235	0.5275	0.5846	0.6000	**0.6061**
Copper	1	0.4667	0.4971	0.4745	0.5039	0.6133	0.5733	**0.6267**
Copper	3	0.5333	0.4913	0.4824	0.5431	0.6000	**0.6400**	0.6133
Copper	5	0.4990	0.5010	0.4471	0.5000	0.5600	**0.6400**	0.6200
Copper	10	0.5048	0.4854	0.4588	0.4686	**0.6400**	**0.6400**	**0.6400**
Copper	20	0.5524	0.5184	0.4784	0.5353	**0.7043**	0.6824	0.6824
Corn	1	0.4865	0.5353	0.5426	0.4634	0.5600	0.5826	**0.6000**
Corn	3	0.4654	0.5098	0.5149	0.4673	0.5273	0.5565	**0.5733**
Corn	5	0.4865	0.5039	0.5109	0.4772	0.5565	0.5577	**0.5826**
Corn	10	0.4558	0.5333	0.5267	0.4970	0.5913	**0.6000**	0.5913
Corn	20	0.4654	0.5333	0.4990	0.5564	**0.6522**	0.6435	0.6435
Crude	1	0.4560	0.5333	0.4894	0.4979	0.6182	0.6727	**0.7000**
Crude	3	0.4440	0.5167	0.4809	0.4766	**0.7000**	0.6750	0.6545
Crude	5	0.4720	0.5583	0.4681	0.4681	0.6250	**0.6800**	0.6533
Crude	10	0.4760	0.4500	0.5021	0.4426	**0.6800**	0.6533	0.6533
Crude	20	0.5440	0.5542	0.4340	0.4766	0.7867	**0.8545**	0.8000
Gold	1	0.5412	0.4760	0.5224	0.4490	0.5647	**0.6000**	0.5875
Gold	3	0.5569	0.4640	0.4898	0.4449	**0.5882**	0.5750	0.5686
Gold	5	0.5098	0.4840	0.4612	0.3878	0.6000	**0.6250**	**0.6250**
Gold	10	0.5255	0.5080	0.4531	0.4612	0.6500	0.6500	**0.7000**
Gold	20	0.5176	0.5080	0.4612	0.4490	0.5625	0.5804	**0.6250**
HeatingOil	1	0.4621	0.5294	0.5050	0.4752	0.4913	0.5600	**0.5867**
HeatingOil	3	0.4175	0.5451	0.5327	0.4871	0.5733	**0.5867**	0.5467
HeatingOil	5	0.4233	0.5745	0.5426	0.4931	**0.6000**	**0.6000**	**0.6000**
HeatingOil	10	0.4117	0.5235	0.5069	0.5149	**0.5867**	**0.5867**	**0.5867**
HeatingOil	20	0.4505	0.4863	0.4653	0.5545	**0.5739**	0.5304	0.5217
Platinum	1	0.5010	0.4923	0.4951	0.5146	**0.6000**	0.5733	0.5600
Platinum	3	0.5238	0.4942	0.4544	0.4699	0.5647	0.5913	**0.5941**
Platinum	5	0.5486	0.5058	0.4718	0.4680	0.5391	**0.5882**	0.5733
Platinum	10	0.5181	0.4692	0.4621	0.5126	**0.5314**	0.5176	0.5235
Platinum	20	**0.6038**	0.5404	0.5010	0.5340	0.5676	0.5143	0.5124
Silver	1	0.4980	0.4760	0.4980	0.4939	0.5833	0.6000	**0.6167**
Silver	3	0.4980	0.4560	0.4980	0.4367	0.6000	**0.6250**	**0.6250**
Silver	5	0.4824	0.4760	0.4776	0.4776	0.5250	**0.5750**	**0.5750**
Silver	10	0.4235	0.4560	0.4939	0.4939	**0.7000**	**0.7000**	**0.7000**
Silver	20	0.4784	0.5000	0.5102	0.5469	**0.6250**	0.5500	0.5412

Table 7: Complete strict-protocol matrix: best DA per family for every asset–horizon pair. Bold values indicate the highest DA within each asset–horizon row; ties are bolded jointly.

conditioning is valuable, but its effectiveness depends on how much data support remains after filtering.

The classical and neural sequence baselines appear only in the `Full` bucket and remain close to chance on average. Their role in this table is therefore calibrational. They establish the baseline difficulty of the forecasting task, while the tabular machine-learning families show where event-conditioned news, price, and macro features begin to produce stronger average performance.

5.3 Complete Asset–Horizon Matrix

Table 7 reports the strict-protocol best DA achieved by each model family for every asset–horizon pair. Unlike winner-only summaries, this matrix retains the full cross-family structure. It therefore shows not only which family performs best in each cell, but also how close the competing families are and where baseline models remain competitive.

The strongest pattern is that predictive performance is highly local. Tabular machine-learning families dominate many asset–horizon cells, especially where price, macro, news, and regime-conditioned features appear to interact strongly. The clearest example is `Crude` at horizon 20, where Ridge reaches 0.7867, XGBoost reaches 0.8545, and LightGBM reaches 0.8000. This is the strongest row in the matrix and indicates a pronounced event-conditioned signal for longer-horizon crude forecasting. This result should be interpreted as a candidate local regime rather than as a general claim about crude-oil predictability. Its value is primarily diagnostic: it identifies a setting in which headline-derived event information, horizon choice, and tabular feature interactions appear unusually aligned.

At the same time, the matrix shows that no model family dominates uniformly. Classical and neural baselines remain competitive in several cells: ARIMA is strongest for `Platinum` at horizon 20; LSTM leads in some short- and medium-horizon agricultural and energy cases; TCN leads for `Coffee` at horizon 10; and PatchTST is strongest for `Silver` at horizon 20. These cases are important because they show that the benchmark is not simply rewarding one modelling class across all settings.

Overall, the asset–horizon matrix supports the central empirical claim of the paper: commodity predictability is sparse, local, and regime-dependent. Strong directional accuracy emerges in targeted asset–horizon cells, especially for tabular models, but the full matrix shows that useful signal depends on the interaction between asset, horizon, model family, and feature regime.

5.4 Explicit Weak-Configuration Evidence

Table 8 reports representative weak-performing configurations. These rows are included deliberately: the objective is to show that event-conditioned forecasting contains not only local gains, but also configurations that perform materially below the no-skill benchmark.

The strict protocol shows the clearest weak-tail behaviour. The weakest tabular configurations are concentrated in news-heavy or event-conditioned specifications, especially `Top80` variants involving `GlobalNews` and combined macro–news or price–news blocks. The weakest strict XGBoost specification reaches a mean DA of only 0.4549, corresponding to a 4.51 percentage-point deficit relative to chance. Ridge and LightGBM show similar weak-tail behaviour, with representative configurations around 0.457–0.463. These results show that aggressive event filtering can degrade performance when the event bucket, feature block, and asset–horizon regime are misaligned.

Weak configurations also remain visible under the max-data protocol, although their deficits are generally smaller. For Ridge, the weakest reported news-only Top80 configurations reach 0.4588. For XGBoost and LightGBM, the weakest max-data examples are closer to chance, with gaps

Setting	Family	Model spec	Config count	Mean DA	Gap to 0.5
Strict	ARIMA	ARIMA_Full	40	**0.4972**	**-0.0028**
Strict	LSTM	LSTM_Full	40	**0.5113**	**+0.0113**
Strict	TCN	TCN_Full	40	**0.4930**	**-0.0070**
Strict	PatchTST	PATCHTST_Full	40	**0.4889**	**-0.0111**
Strict	Ridge	AssetNews__Full_Macro	10	**0.4570**	**-0.0430**
Strict	Ridge	GlobalNews__Top80_Macro+News	25	0.4629	-0.0371
Strict	XGBoost	GlobalNews__Top80_Macro+News	25	**0.4549**	**-0.0451**
Strict	XGBoost	GlobalNews__Top80_Price+News	25	0.4587	-0.0413
Strict	LightGBM	GlobalNews__Top80_All	25	**0.4576**	**-0.0424**
Strict	LightGBM	GlobalNews__Top80_Price	25	0.4592	-0.0408
MaxData	ARIMA	ARIMA_Full	40	**0.4933**	**-0.0067**
MaxData	TCN	TCN_Full	40	**0.4949**	**-0.0051**
MaxData	PatchTST	PATCHTST_Full	40	**0.4908**	**-0.0092**
MaxData	Ridge	AssetNews__Top80_News	15	**0.4588**	**-0.0412**
MaxData	Ridge	GlobalNews__Top80_News	15	**0.4588**	**-0.0412**
MaxData	XGBoost	AssetNews__Top50_News	40	**0.4794**	**-0.0206**
MaxData	XGBoost	GlobalNews__Top50_News	40	0.4812	-0.0188
MaxData	LightGBM	AssetNews__Top50_News	40	**0.4867**	**-0.0133**
MaxData	LightGBM	GlobalNews__Top50_News	40	0.4869	-0.0131

Table 8: Representative weak-performing configurations. Bold values indicate the weakest mean DA within each setting–family group; ties are bolded jointly.

ranging from approximately 1.3 to 2.1 percentage points below 0.5. This suggests that broader data retention reduces, but does not eliminate, weak-tail risk.

The table strengthens the paper's full-trajectory argument. Event-conditioned news representations can identify high-performing local regimes, but the same design space also contains fragile specifications. Reporting these weak configurations prevents winner-only interpretation and makes the configuration risk of NLP-based commodity forecasting explicit.

5.5 Robustness Under Strict vs Max-Data

Table 9 compares model-family robustness under the strict and max-data protocols. The table separates three complementary views of performance: full-space average DA, best-per-asset–horizon DA, and the share of configurations that exceed the nominal significance threshold.

The strict protocol identifies sharper local regimes. Under strict filtering, XGBoost achieves the strongest best-per-asset–horizon DA at 0.6071, closely followed by LightGBM at 0.6069 and Ridge at 0.5968. This indicates that the strongest event-conditioned signals are concentrated in selected asset–horizon cells rather than distributed uniformly across the full configuration space.

The max-data protocol improves broad average behaviour for the tabular machine-learning families. LightGBM reaches the highest full-space mean DA at 0.5070, while XGBoost follows closely at 0.5068. Max-data also increases the share of nominally significant configurations, with LightGBM reaching 0.140 and XGBoost 0.139. This suggests that broader data retention improves coverage and stabilises average effects, even though it slightly reduces the sharpest best-per-cell performance.

The classical and neural sequence baselines remain comparatively stable across protocols and mostly near chance. Their role is therefore primarily calibrational: they provide useful reference points for interpreting whether the tabular models are discovering regime-specific structure beyond

Setting	Family	Configs	Mean DA (all)	Mean DA (best A×H)	Share($p \leq 0.05$)
Strict	ARIMA	40	0.4972	0.4972	0.175
	LightGBM	1950	0.4942	0.6069	0.070
	LSTM	40	**0.5113**	0.5113	**0.225**
	PatchTST	40	0.4889	0.4889	0.075
	Ridge	1950	0.4915	0.5968	0.082
	TCN	40	0.4930	0.4930	0.100
	XGBoost	1950	0.4929	**0.6071**	0.070
Max-data	ARIMA	40	0.4933	0.4933	**0.175**
	LightGBM	3000	**0.5070**	0.5862	0.140
	PatchTST	40	0.4908	0.4908	0.100
	Ridge	3000	0.4976	0.5680	0.131
	TCN	40	0.4949	0.4949	0.075
	XGBoost	3000	0.5068	**0.5885**	0.139

Table 9: Robustness of model-family performance under strict and max-data specifications. Bold values indicate the strongest result within each protocol and metric column.

standard time-series and sequence baselines.

Overall, the robustness evidence shows a precision–coverage trade-off. Strict filtering is stronger for identifying high-signal local regimes, while max-data is stronger for broad average stability and significance density. This supports the paper's central claim: event-conditioned NLP signals are most useful when evaluated as regime-specific information rather than as uniformly predictive features.

6 Discussion and Limitations

The results support a regime-specific view of NLP-based commodity forecasting. The central finding is not that news improves every forecast, but that headline-derived event information helps identify local asset–horizon regimes in which directional predictability becomes materially stronger. This distinction is important: full-space averages are close to chance for most families, but best-per-asset–horizon results reveal substantially stronger local performance, especially for tabular machine-learning models.

6.1 Regime-Specific Predictability

The full-space results show that predictive signal is sparse. Across the complete configuration space, most model families remain near chance-level directional accuracy. This indicates that commodity direction is not broadly predictable through generic feature augmentation. However, the best-per-asset–horizon frontier shows that this average view hides important local structure. Under the strict protocol, XGBoost reaches a mean best-per-cell DA of 0.6071, LightGBM reaches 0.6069, and Ridge reaches 0.5968. These values show that event-conditioned signal exists, but it is concentrated in specific combinations of asset, horizon, feature block, event bucket, and model family.

The complete asset–horizon matrix strengthens this interpretation. The strongest example is Crude at horizon 20, where XGBoost reaches 0.8545, LightGBM reaches 0.8000, and Ridge reaches 0.7867. This suggests that longer-horizon crude forecasting contains a particularly strong event-conditioned structure. At the same time, the matrix shows that no single model family dominates uniformly. ARIMA, LSTM, TCN, and PatchTST remain competitive in selected cells, confirming

that the forecasting problem is heterogeneous rather than reducible to one universally superior architecture.

6.2 Event Conditioning and the Precision–Coverage Trade-off

The bucket-transition results show that event conditioning is useful, but not mechanically monotonic. Under the strict protocol, the strongest tabular bucket-level means occur in the `Full` bucket, and aggressive filtering to `Top80` weakens performance. This shows that high news intensity alone is not equivalent to directional relevance. A high-intensity period may contain conflicting, redundant, delayed, or already-priced information.

The max-data protocol reveals the complementary side of the trade-off. With broader sample retention, event-conditioned buckets become more productive: Ridge peaks at `Top70`, while XGBoost and LightGBM improve in `Top70` and `Top80`. This suggests that event conditioning benefits from sufficient sample support. Strict filtering can isolate sharper local regimes, while max-data improves average stability and significance density. The two protocols therefore should not be interpreted as competing specifications, but as complementary lenses: strict is better for high-signal discovery, while max-data is better for coverage and robustness.

6.3 Model-Family Implications

The strongest local results are produced by tabular machine-learning models, particularly XGBoost, LightGBM, and Ridge. This indicates that the useful predictive structure is often feature-interactive rather than purely sequential. The relevant signal appears to arise from combinations of price state, macro context, news provenance, event intensity, regime conditioning, and forecast horizon. Tree-based models are well suited to this structure because they can capture nonlinear interactions among engineered features.

The neural sequence models and ARIMA remain important calibration baselines. Their near-chance average performance confirms the difficulty of the task and helps show that the strongest tabular results are not simply an artefact of easy asset–horizon cells. Instead, gains appear when the representation and model family are aligned with the appropriate event-conditioned regime.

6.4 Implications for Financial NLP

The main implication for financial NLP is that textual representations should be evaluated not only as predictive features, but also as regime-identification signals. In this benchmark, the value of news is not uniform across time. News becomes useful when it helps identify periods in which the forecasting relationship changes.

This shifts the evaluation question from "Does text improve prediction on average?" to "When does text identify a useful decision regime?" That framing is especially important in financial forecasting, where average effects can be weak while local conditional effects are strong. The paper therefore argues for full-trajectory evaluation in financial NLP: global averages, bucket transitions, local frontiers, weak configurations, distribution tails, and protocol robustness should all be reported.

6.5 Threats to Validity and Multiple Testing

The benchmark intentionally evaluates a large configuration space across assets, horizons, news provenance views, feature blocks, event buckets, model families, and data-retention protocols. This design is useful for studying heterogeneity, but it also creates a multiple-testing and model-selection

risk: some high-performing configurations may arise by chance when many alternatives are evaluated. For this reason, individual best-case results should not be interpreted as standalone confirmatory evidence.

The analysis mitigates this risk in three ways. First, it reports full-space family averages rather than only the strongest configurations, showing that average directional accuracy remains close to chance for many model families. Second, it reports weak configurations, low-tail performance, and non-winning model families, making negative evidence visible rather than selecting only favourable cases. Third, it compares strict and max-data protocols to test whether the observed patterns are sensitive to data-retention choices.

Even with these safeguards, the strongest local results should be interpreted as evidence of candidate event-conditioned regimes rather than definitive proof of general predictive superiority. A fully confirmatory follow-up would require pre-registering the selected asset–horizon regimes, feature blocks, event buckets, and model families, then evaluating them on later unseen data or on an independently collected news and price corpus.

6.6 Limitations

Several limitations should guide the interpretation of the results. First, the evaluation focuses on directional accuracy rather than realised trading performance. Directional accuracy measures out-of-sample classification skill, but it should not be interpreted as evidence of tradable profitability. The benchmark does not account for transaction costs, slippage, liquidity constraints, position sizing, drawdowns, market impact, portfolio construction, or risk-adjusted returns. The results therefore support claims about conditional predictive structure in textual information regimes, not claims about executable trading strategies.

Second, event conditioning is based on information intensity rather than causal event classification. High textual activity does not necessarily imply directional relevance. A high-intensity news period may contain information that is redundant, delayed, contradictory, or already reflected in prices. The results therefore show that textual information regimes can be predictive, but they do not prove that the measured news events causally move commodity prices.

Third, the benchmark uses headline-level representations. Headlines are timely and information-dense, but they omit details available in full articles. Future work could incorporate document-level representations, richer event taxonomies, source credibility, entity linking, and temporal decay of news relevance.

Fourth, the benchmark contains a large configuration space. The paper addresses this through full-trajectory reporting, weak-tail evidence, and strict-vs-max-data robustness, but individual high-performing configurations should be interpreted as evidence of local predictive regimes rather than standalone proof of universal model superiority.

Finally, the study evaluates a fixed commodity universe and fixed forecast horizons. The conclusions may differ for intraday data, alternative commodities, longer investment horizons, other asset classes, or different news sources.

7 Conclusion

This paper introduced an event-conditioned framework for evaluating when NLP-derived news signals improve directional forecasting across commodity markets. Rather than treating news as a universally useful feature source, we modelled headline-derived information as a conditional event signal whose value depends on asset context, forecast horizon, event intensity, feature representation, model family, and data-retention protocol.

The empirical evidence supports three conclusions. First, predictive edge is regime-local rather than uniform: average directional accuracy remains close to chance across much of the full configuration space. Second, strong local predictability emerges in targeted asset–horizon regimes, especially for tabular machine-learning models that combine price, macroeconomic, news, and regime-conditioned features. Third, event conditioning creates a precision–coverage trade-off: strict filtering identifies sharper high-signal regimes, while max-data retention improves average stability, coverage, and significance density.

The main contribution is therefore both methodological and empirical. Methodologically, the paper provides a full-trajectory evaluation design for financial NLP forecasting, combining strict temporal validation, event-bucket analysis, best-per-asset–horizon frontiers, weak-configuration evidence, distribution tails, and robustness across data-retention protocols. Empirically, it shows that headline-derived NLP signals can identify local windows of commodity directional predictability when textual information intensity, asset context, horizon, representation, and model family are aligned. For financial NLP, the central implication is that text should be treated not only as a representation problem, but also as a regime-discovery problem. News matters most when it helps identify the market states in which the forecasting relationship changes. This shifts the evaluation question from whether text improves prediction on average to when textual information defines a useful decision regime.

The strongest local configurations should therefore be viewed as hypothesis-generating evidence for event-conditioned textual regimes. They identify where textual information appears to align with market structure, but they do not establish universal predictability or tradable profitability. The appropriate interpretation is that event-conditioned NLP can help discover candidate regimes for further validation.

Future work should extend this framework with richer event taxonomies, document-level news representations, source-credibility measures, temporal decay features, and economically realistic trading simulations. A further direction is adaptive event conditioning, where models learn which textual events matter for each commodity and horizon rather than relying only on fixed intensity buckets.

References

[1] Hachmi Ben Ameur et al. Forecasting commodity prices: Empirical evidence using deep learning tools. *Annals of Operations Research*, 2023.

[2] Dogu Araci. FinBERT: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*, 2019.

[3] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

[4] David H. Bailey, Jonathan M. Borwein, Marcos López de Prado, and Qiji Jim Zhu. The probability of backtest overfitting. *Journal of Computational Finance*, 20(4):39–69, 2017.

[5] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[6] George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung. *Time Series Analysis: Forecasting and Control*. John Wiley & Sons, Hoboken, NJ, 5 edition, 2015.

[7] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

[9] Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. Deep learning for event-driven stock prediction. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, pages 2327–2333, 2015.

[10] Peter R. Hansen. A test for superior predictive ability. *Journal of Business & Economic Statistics*, 23(4):365–380, 2005.

[11] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[12] Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[13] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[14] Marcos López de Prado. *Advances in Financial Machine Learning*. Wiley, 2018.

[15] Tim Loughran and Bill McDonald. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *Journal of Finance*, 66(1):35–65, 2011.

[16] R. L. Manogna et al. Enhancing agricultural commodity price forecasting with deep learning. *Scientific Reports*, 2025.

[17] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

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

[19] A. Varshini et al. How good are different machine and deep learning models in forecasting metal futures in commodity markets? *Energy Policy*, 2024.

[20] Halbert White. A reality check for data snooping. *Econometrica*, 68(5):1097–1126, 2000.

[21] Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.