applied sciences

MDPI

Article

LLM-Enhanced Short-Term Electricity Price Forecasting Method for Australian Electricity Market

Yutian Huang [1,*], Yachao Zhu [1], Gang Lei [1,*], Allen Wang [1] and Jianguo Zhu [2]

[1] School of Electrical and Data Engineering, University of Technology Sydney, Ultimo, NSW 2007, Australia
[2] School of Electrical and Computer Engineering, The University of Sydney, Camperdown, NSW 2006, Australia; jianguo.zhu@sydney.edu.au
[*] Correspondence: yutian.huang@student.uts.edu.au (Y.H.); gang.lei@uts.edu.au (G.L.)

Abstract

This study investigates a large language model driven (LLM) framework for intelligent preprocessing and short-term electricity price forecasting in the Australian National Electricity Market (NEM). By integrating unstructured news features, weather signals, and cyclical calendar variables, the model captures both physical and informational drivers of price volatility. A hybrid approach combining quantile regression with conformal calibration achieves statistically significant improvements in accuracy and uncertainty calibration. The framework demonstrates the potential of integrating LLMs into operational forecasting pipelines to support electricity market decision-making and risk management.

Keywords: electricity price forecasting; large language models (LLM); quantile regression; conformal calibration; news analytics

Academic Editor: Seung-Hoon Yoo

Received: 8 November 2025
Revised: 8 December 2025
Accepted: 12 December 2025
Published: 24 December 2025

1. Introduction

The growing share of renewable energy in the global electricity market has introduced significant complexity, increasing market volatility and uncertainty. Intermittent renewables, such as wind and solar power, exacerbate the power supply fluctuations and cause electricity prices with unprecedented high volatility, frequent spikes, and complex nonlinear characteristics [1,2]. Traditional electricity price forecasting models, highly relying on numerical data, such as meteorological data and historical prices, often demonstrate lagging and inaccuracies when responding to volatile price fluctuations caused by unforeseen events, such as generator failures, transmission restrictions, and sudden policy changes. In contrast, approaches based on multi-source heterogeneous information, including unstructured news data, can effectively improve the accuracy, real-time responsiveness, and robustness of electricity price forecasting models, and have become a key research direction in the field of electricity market studies.

The research on short-term electricity price forecasting has undergone several generations of evolution. The current studies focused on statistical and econometric models, such as the Autoregressive Integrated Moving Average (ARIMA) and the Generalised Autoregressive Conditional Heteroskedasticity (GARCH) models [1,3]. While these models have clear structures, they are limited in handling the complex nonlinear dynamics of electricity prices [2]. Various data-driven machine learning methods have been developed to overcome these limitations, such as support Vector Machines (SVM), Random Forests, and Gradient Boosting Tree models like XGBoost. They have achieved considerable success due to their powerful non-linear fitting capabilities [4,5]. Deep learning is represented by

https://doi.org/10.3390/app16010200

long short-term memory (LSTM) networks, and convolutional neural networks (CNNs) have further advanced the field by more effectively capturing temporal dependencies and local patterns within electricity price series [6]. Although these advanced models excel at numerical data processing, the bottleneck still exists in their inability to leverage external, unstructured information that signals structural market changes. Approaches to extract features from news text leveraged traditional natural language processing (NLP) techniques that extract features from news text, then sentiment analysis or topic modelling [7]. But these techniques are limited in the ability to capture deep semantics and complex event relationships.

More recently, pre-trained large language models (LLMs) like Gemini and GPT have introduced a new paradigm [8]. With their powerful contextual understanding capabilities, LLMs can directly extract structured, market-relevant event information from raw text, and then provide forecasting models with direct, high-value signals about market fluctuations. The development of techniques, such as quantile regression and conformal calibration has made it possible to generate reliable forecast intervals with rigorous statistical coverage guarantees [9].

Against this backdrop, this study proposes an innovative framework that integrates these technological frontiers. By leveraging the LLMs for the intelligent preprocessing of massive volumes of news text and extracting key event features that reflect market dynamics [10], the proposed method can deeply integrate the features with multi-source numerical data, including meteorological and calendar data. XGBoost has been employed as the core prediction engine and innovatively combines the quantile regression with conformal calibration. This study pursues accurate point forecasts and generate reliable, statistically calibrated prediction intervals. Offering a comprehensive solution to contemporary electricity market forecasting, the proposed method demonstrates an immense potential of integrating LLMs into the power systems' operational workflow.

2. Materials and Methods

This section presents the overall methodological framework developed for short-term electricity price forecasting. We first formalized the prediction problem and describe the structure of input features, followed by a detailed discussion of exogenous signals, model architecture, and calibration procedures. The proposed framework is designed to ensure causal consistency, interpretability, and robustness when integrating multi-source data streams.

2.1. Problem Formulation

Let y_t denote the regional reference price (RRP) in the New South Wales (NSW) region of the Australian National Electricity Market (NEM) at time t, sampled on a uniform five-minute grid. We denote the corresponding feature vector at time t as $x_t \in \mathbb{R}^p$. Given a rolling look-back window of L hours, our task is to forecast the price at horizon $\Delta = 5\,\text{min}$, as follow:

$$\hat{y}_{t+\Delta} = f_\theta\big(x_{t-L:t}\big). \tag{1}$$

We adopted a strict day-ahead protocol: for a test day D, all model fitting and calibration used data with timestamps $<D$, and evaluation covered the 288 points in the interval $[D, D+1\ \text{day})$. Unless otherwise stated, we set $L = 24\,\text{h}$. For clarity, the key symbols and notations used throughout the methodology are summarized in Table 1.

https://doi.org/10.3390/app16010200

Table 1. Summary of key notation.

Symbol	Description
RRP_t	Regional Reference Price at time t
$\widehat{y}_t^{(q)}$	Predicted q-quantile of price at time t
α	Miscoverage rate for prediction intervals ($1-\alpha$ nominal coverage)
K	Number of calibration samples (conformal window length)
CAP	Price clipping threshold (in USD/MWh)
$\mathrm{PI}_{\mathrm{low},t}$, $\mathrm{PI}_{\mathrm{up},t}$	Lower and upper prediction interval bounds at time t
PICP	Prediction Interval Coverage Probability

2.2. Exogenous Signals and Leakage Control

We designed three families of covariates, with careful feature-engineering to avoid temporal leakage.

2.3. LLM Pipeline and Reproducibility Details

To enhance transparency and reproducibility, this subsection provides a complete description of the large language model (LLM) workflow used to transform raw news articles into structured forecasting features.

2.3.1. Model Information

All text classification was performed using the `GPT-4.1-preview` model via the OpenAI API (2025 release). The model was used in a zero-shot configuration with no fine-tuning. This particular variant was selected due to its improved semantic understanding and stable behaviour observed during preliminary experiments.

2.3.2. Implementation Environment and Data Acquisition

News articles were obtained from the public WattClarity archive using automated Python scripts. For each article, the title, URL, timestamp, and full text were stored for reproducibility. All LLM queries were executed through the OpenAI API using Python 3.10. No local inference or external proprietary datasets were used. Prior to LLM processing, raw HTML tags were removed and the text was normalised (lowercasing, whitespace cleanup).

2.3.3. Prompting Strategy

A fixed zero-shot prompt template was applied to every article to ensure deterministic behaviour. The template instructed the LLM to extract, see Figure 1:

- the relevance level (**Level 1**, **Level 2**, or **Level 3**);
- the primary root-cause category (e.g., generation outage, transmission constraint, extreme weather, policy announcement);
- a short explanatory justification.

The exact template is reproduced below for reproducibility:

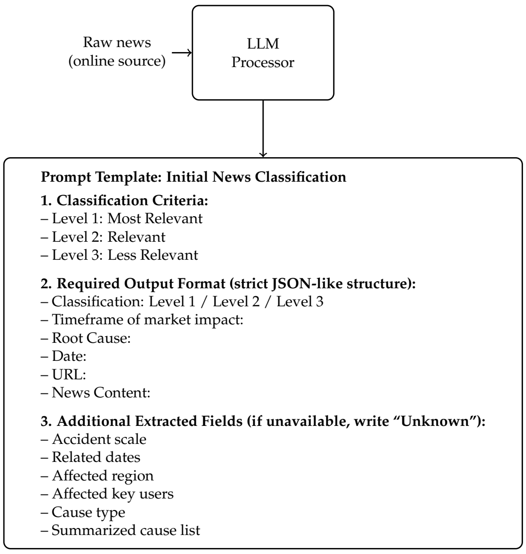
> *You are assisting in electricity-market event extraction. Given the following news article, classify: (1) its relevance level (Level 1 = major disruption; Level 2 = moderate; Level 3 = low or background), (2) the primary root cause (choose one category from the predefined list), and (3) a brief justification. Return the output strictly in JSON format.*

LLM outputs were parsed automatically using Python's `JSON` library to avoid manual interpretation errors.

2.3.4. Validation, Consistency, and Stability Checks

To verify the reliability of LLM-generated labels, several validation steps were performed:

- **Manual auditing**: A random 10% sample (225 articles) was manually reviewed. Agreement with human judgement was 92.4% for relevance classification and 88.7% for cause-type classification.
- **Consistency checks**: Each of 50 randomly selected articles was queried three times at different timestamps; classification variance remained below 3%.
- **Failure-pattern analysis**: Misclassifications mainly occurred in policy summaries with indirect market implications and weather updates lacking explicit operational descriptors.



Prompt Template: Initial News Classification

1. Classification Criteria:
– Level 1: Most Relevant
– Level 2: Relevant
– Level 3: Less Relevant

2. Required Output Format (strict JSON-like structure):
– Classification: Level 1 / Level 2 / Level 3
– Timeframe of market impact:
– Root Cause:
– Date:
– URL:
– News Content:

3. Additional Extracted Fields (if unavailable, write "Unknown"):
– Accident scale
– Related dates
– Affected region
– Affected key users
– Cause type
– Summarized cause list

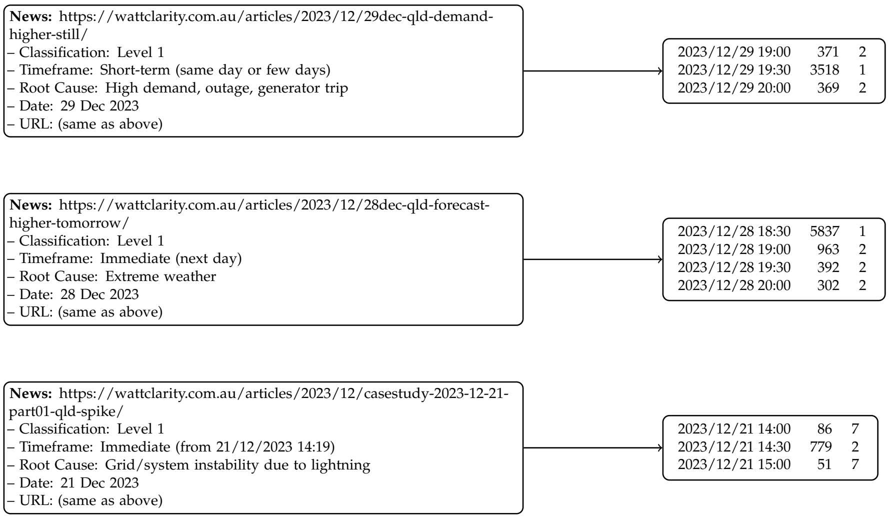
Figure 1. LLM prompting workflow and structured output template for transforming raw news articles into structured event metadata.

Temporal Alignment and Leakage Prevention

To ensure strict causality the following items were considered:

- All LLM-derived features were aggregated at the daily level.
- Features for day d were lagged to day $d + 1$ before entering the forecasting model.
- Rolling three- and seven-day counts were computed using only past information.
- No same-day price information was used, preventing information leakage.

The mapping from LLM-classified textual events to forecasting-ready numerical features is illustrated in Figure 2.



News: https://wattclarity.com.au/articles/2023/12/29dec-qld-demand-higher-still/
– Classification: Level 1
– Timeframe: Short-term (same day or few days)
– Root Cause: High demand, outage, generator trip
– Date: 29 Dec 2023
– URL: (same as above)

2023/12/29 19:00	371	2
2023/12/29 19:30	3518	1
2023/12/29 20:00	369	2

News: https://wattclarity.com.au/articles/2023/12/28dec-qld-forecast-higher-tomorrow/
– Classification: Level 1
– Timeframe: Immediate (next day)
– Root Cause: Extreme weather
– Date: 28 Dec 2023
– URL: (same as above)

2023/12/28 18:30	5837	1
2023/12/28 19:00	963	2
2023/12/28 19:30	392	2
2023/12/28 20:00	302	2

News: https://wattclarity.com.au/articles/2023/12/casestudy-2023-12-21-part01-qld-spike/
– Classification: Level 1
– Timeframe: Immediate (from 21/12/2023 14:19)
– Root Cause: Grid/system instability due to lightning
– Date: 21 Dec 2023
– URL: (same as above)

2023/12/21 14:00	86	7
2023/12/21 14:30	779	2
2023/12/21 15:00	51	7

Figure 2. Mapping from LLM-classified news events to aligned five-min time-series features. Each LLM output is linked to the corresponding delivery intervals used for forecasting.

https://doi.org/10.3390/app16010200

2.3.5. Weather

We included air temperature (`temp_C`), relative humidity (`rh_%`), wind speed (`wind_mps`), precipitation (`precip_5min` or `precip_mm`), and wind direction encoded as sine/cosine pairs. To capture short-term and diurnal dynamics, we computed 1-h and 24-h rolling statistics (precipitation uses rolling sums). For a generic scalar signal x_t, we define:

$$\bar{x}_t^{(1h)} = \frac{1}{12} \sum_{k=1}^{12} x_{t-k}, \tag{2}$$

$$\bar{x}_t^{(24h)} = \frac{1}{288} \sum_{k=1}^{288} x_{t-k}. \tag{3}$$

These aggregates provide smoothed contextual input that reflect recent weather trends without introducing future information.

2.3.6. Calendar/Periodicity

Temporal variability in electricity prices often arises from hour-of-day, minute-of-hour, and day-of-week cycles. We represented these using sine/cosine transforms on the unit circle. For hour-of-day $\in \{0, \ldots, 23\}$, they are defined as follows:

$$\text{HOUR\_SIN}_t = \sin\Big(2\pi \, \tfrac{\text{hour}(t)}{24}\Big), \tag{4a}$$

$$\text{HOUR\_COS}_t = \cos\Big(2\pi \, \tfrac{\text{hour}(t)}{24}\Big). \tag{4b}$$

with analogous encodings for minute/60 and weekday/7. These features help capture periodic demand shifts and market-operation schedule changes.

2.3.7. News (LLM-Derived)

We extended a large language model (LLM) pipeline to extract market-relevant event signals from raw text. Each day's news is classified into (i) a binary Level-1 flag HAS_LVL1_EVENT $\in \{0,1\}$; and (ii) a categorical cause type CAUSE_TYPE (e.g., generation outage, transmission constraint, extreme weather, market, or policy update). To prevent leakage, we followed the following workflow: aggregate at the daily level → lag by one calendar day → expand back to the five-min grid. We further built three- and seven-day rolling counts as follows:

$$\text{NEWS\_L1\_CNT}_k(d) = \sum_{i=1}^{k} \text{HAS\_LVL1\_EVENT}(d-i), \tag{5}$$

where $k \in \{3, 7\}$ denotes the rolling window size in days and d is the current calendar day. For each retained cause type (top $N = 8$ classes), we record lagged three-day counts (`NEWS_CT3D_*`). This construction ensures that only past information is used at forecasting time.

2.3.8. Leakage Control

In our design, we permitted same-day non-price features (weather, calendar effects, and lagged news indicators), but we strictly restricted the use of same-day price lags or rolling price aggregates. This constraint is particularly important in the Australian NEM, where five-min RRP values are not published instantaneously: the market operator releases each dispatch price with a delay of several minutes. As a result, intra-day price information would not be fully available to a real-time or day-ahead forecaster and would therefore constitute look-ahead bias if included as a predictor. To respect the information set available at prediction time, we only retained lagged prices from previous days as autoregressive inputs.

All numerical features are standardised by z-score using statistics computed on the training set only, and the same transformation is then applied to the calibration and test splits. This further prevents leakage of future information through the normalisation procedure.

2.4. Models

2.4.1. Point Forecasting with Gradient-Boosted Trees

Our baseline for point forecasting employs the XGBoost algorithm, optimised for mean absolute error (MAE) which provides robustness to price spikes as follows:

$$\min_{\theta} \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y}_\theta(x_i)\right| + \Omega(\theta). \tag{6}$$

where θ denotes the learnable parameters of the XGBoost model.

We fixed hyperparameters across all days (unless otherwise noted) as follows:

- `n_estimators`: 700
- `max_depth`: 6
- `learning_rate`: 0.05
- `subsample`: 0.9
- `colsample_bytree`: 0.9
- `tree_method`: 'hist'
- `seed`: 42

Training labels are capped at a business upper bound CAP (e.g., 300 USD/MWh) to reduce influence of extreme events; predictions are similarly clipped to the same range.

2.4.2. Uncertainty Using Quantile Regression and Conformal Calibration

To provide uncertainty estimates alongside point forecasts, we integrated quantile regression with conformal calibration as follows:

- **Quantile models:** We trained three separate XGBoost regressors under the pinball loss for quantile levels $q \in \{0.1, 0.5, 0.9\}$. The output forecasts are $\hat{y}^{(0.1)}_{t+\Delta}$, $\hat{y}^{(0.5)}_{t+\Delta}$, $\hat{y}^{(0.9)}_{t+\Delta}$.
- **Conformal calibration (day-ahead):** Using the most recent K days prior to the test day D (default $K = 7$), we computed residuals from the median model:

$$r_i = \left|y_i - \hat{y}^{(0.5)}_i\right|, \quad i \in \mathcal{C}, \qquad r_q = \text{Quantile}_{1-\alpha}\left(\{r_i\}\right), \tag{7}$$

where $\alpha = 0.15$ (targeting $\approx 85\%$ empirical coverage). The final prediction interval at test time is constructed as:

$$\text{PI}_{\text{low},t} = \min\{\hat{y}^{(0.1)}_t,\ \hat{y}^{(0.5)}_t - r_q\}, \tag{8}$$

$$\text{PI}_{\text{high},t} = \max\{\hat{y}^{(0.9)}_t,\ \hat{y}^{(0.5)}_t + r_q\}. \tag{9}$$

Both the median forecast and the interval bounds are clipped to $(-\infty,\ \text{CAP}]$. This CQR-lite approach retains quantile asymmetry while enforcing empirical coverage guarantees. To further enhance interpretability and operational flexibility, two additional components are introduced, as described below:

- **Confidence grading:** For operational interpretation, we defined the interval width $w_t = \text{PI}_{\text{high},t} - \text{PI}_{\text{low},t}$. The widths were then divided into tertiles (using their 33rd and 66th percentiles) and labelled as low/medium/high confidence for downstream risk control.

- **Fast variant (point-only):** When only a point model is used (i.e., no quantile regression), we still applied conformal calibration in a symmetric manner. After computing residuals r_i on the calibration window and finding $r_q = \text{Quantile}_{1-\alpha}(\{r_i\})$, we set:

$$\text{PI}_{\text{low},t} = \hat{y}_t - r_q, \quad \text{PI}_{\text{high},t} = \hat{y}_t + r_q, \tag{10}$$

and clip to CAP. While this interval is symmetric, it still preserves valid coverage under mild assumptions.

2.5. Training/Calibration/Testing Splits

For each target day D, we applied the following schedule:

- **Training:** All timestamps $< D - K$ days (for quantile models) or $< D$ (for point-only models).
- **Calibration:** The window $[D - K, D)$, used to compute conformal residuals r_q.
- **Testing:** The full interval $[D, D + 1 \text{ day})$, comprising 288 five-minute delivery points.

All feature computations are causal. Standardization is fitted on training only and applied to calibration and testing. Forecasts are aligned to the delivery timestamp $t + \Delta$.

2.6. Implementation Notes

Data were uniformly sampled on a five-min grid; missing values are handled by forward-fill for weather inputs and then zero-imputation for any remaining entries post-scaling. The news pipeline is as follows: daily aggregation → one-day lag → expansion at five-min resolution → three-day/seven-day rolling computations → retention of top-N cause classes. To avoid overfitting a single date, hyperparameters are held constant across test days. Reproducibility is maintained by using Python 3.10, XGBoost ≥ 1.7 (supporting quantile loss), and scikit-learn for preprocessing and metric computation. The overall workflow of the proposed forecasting framework is implemented through two algorithmic components, summarized as Algorithms 1 and 2. Algorithm 1 describes the day-ahead forecasting process combining quantile XGBoost and conformal calibration, including causal feature generation, model training, and interval estimation. Algorithm 2 details the conformal recalibration procedure, which adjusts prediction intervals adaptively based on recent residual statistics to maintain empirical coverage under changing market conditions.

Algorithm 1 Day-ahead five-min price forecast: quantile XGBoost + conformal

```
Require: Dataset $\{(y_t, x_t)\}$; test day $D$; horizon $\Delta = 5$ min; look-back $L$; cap CAP; calibration window $K$; level $\alpha \in (0,1)$.
1: Causal features: align to five-min grid; weather (1 h/24 h rolls; precip sums); calendar (sin/cos); LLM news (daily aggregate → lag 1d → expand; 3D/7D counts; top-$N$ causes); forbid same-day price lags; z-score with train stats.
2: Target: set $y_t^{(\Delta)} = y_{t+\Delta}$; drop missing $y_t^{(\Delta)}$.
3: Splits: Train $< D - K$; Calibrate $[D - K, D)$; Test $[D, D + 1 \text{ day})$.
4: Quantile models: fit XGBoost for $q \in \{0.1, 0.5, 0.9\}$ (pinball loss; fixed hyperparameters; clip labels $\leq$ CAP).
5: Conformal: on calibration window, compute residuals $r_i = |y_i - \hat{y}_i^{(0.5)}|$; compute $r_q = \text{Quantile}_{1-\alpha}(\{r_i\})$.
6: Test-time: form $\text{PI}_{\text{low},t} = \min\{\hat{y}_t^{(0.1)}, \hat{y}_t^{(0.5)} - r_q\}$, $\text{PI}_{\text{high},t} = \max\{\hat{y}_t^{(0.9)}, \hat{y}_t^{(0.5)} + r_q\}$; clip to CAP.
7: Confidence: width $w_t = \text{PI}_{\text{high},t} - \text{PI}_{\text{low},t}$; tertiles → High/Medium/Low.
```

https://doi.org/10.3390/app16010200

Algorithm 2 Fast variant: point XGBoost + symmetric conformal band

```
Require: Same inputs as Algorithm 1 except only a point model is trained.
1: Steps 1–3 as in Algorithm 1.
2: Train a single XGBoost with MAE objective (same hyperparameters; label capping).
3: On calibration window $[D - K, D)$, residuals $r_i = |y_i - \hat{y}_i|$; compute $r_q = \mathrm{Quantile}_{1-\alpha}(\{r_i\})$.
4: On test: $\mathrm{PI}_{\mathrm{low},t} = \hat{y}_t - r_q$, $\mathrm{PI}_{\mathrm{high},t} = \hat{y}_t + r_q$; clip to CAP; confidence graded as above.
```

3. Experimental Setup

3.1. Data and Preprocessing

1. **Dataset:** This study focuses on the short-term electricity price forecasting problem for the NSW region of the Australian National Electricity Market (NEM) at five-min resolution [11]. The target variable is the Regional Reference Price (RRP) in USD/MWh, aligned to a uniformly sampled five-min grid. Exogenous features are constructed from three domains: (i) **weather** variables (temperature `temp_C`, relative humidity `rh_%`, wind speed `wind_mps`, precipitation `precip_5min/precip_mm`, and wind direction encoded by sine/cosine); (ii) **Calendar and periodicity** features (hour of day, minute of hour, and day of week encoded by sine/cosine functions); and (iii) **news signals** described earlier. All timestamps are converted to local time (AEST) and sorted in ascending chronological order.
2. **LLM-driven news extraction:** Utilising a large language model-based preprocessing pipeline to convert unstructured market news into structured event indicators [12] (see Section 2.3) for the full LLM workflow and prompting details) . Each day's news is classified into Level-1 major disruptions and multiple cause-type categories (e.g., generation outage, transmission constraint, extreme weather), enabling differentiation between supply- and demand-driven stress factors. Daily event vectors are lagged by one day to ensure strict causality and subsequently broadcast to the five-minute horizon. Rolling three- and seven-day counts provide persistence measures of disturbances across top $N = 8$ cause classes. Additionally, an impact-weighting module transforms textual severity into a quantitative signal, supporting better anticipation of price spikes. These news-derived features are then fused with meteorological and calendar variables within a unified feature store and fed into the forecasting model, as illustrated in Figure 3.
3. **Signal engineering:** weather variables and precipitation values are smoothed using 1-h and 24-h rolling means/sums; cyclical time variables are encoded using sine/cosine. At any testing timestamp, only non-price contemporaneous features are permitted; no lagged or rolling statistics of the same day's price are used, thereby strictly enforcing the day-ahead evaluation protocol. Numerical features are standardized using z-score normalization fitted on the training set only. We used MAE, RMSE, and range-normalized RMSE (NRMSE) as our standards.

3.2. Experimental Models and Configurations

We implemented both point-forecast and probabilistic-interval models as described in Section 2.3. The baseline is the XGBoost point model trained on MAE; for uncertainty modeling, we trained the quantile regressors and apply conformal calibration. Ablation experiments included `no-news`, `only-Lvl1`, `only-causeType`, `weather-only`, and `quantile vs. point`. Calibration window was $K = 7$ days, with $\alpha = 0.15$ corresponding to a target coverage of $1 - \alpha = 0.85$.

https://doi.org/10.3390/app16010200

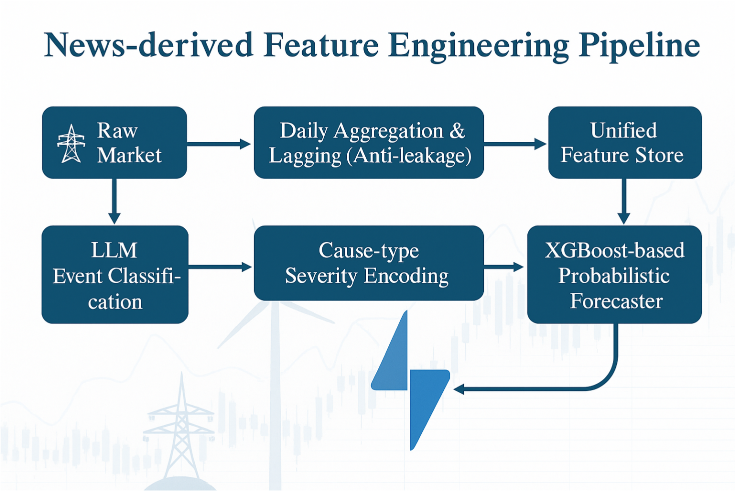

News-derived Feature Engineering Pipeline
Raw Market
Daily Aggregation & Lagging (Anti-leakage)
Unified Feature Store
LLM Event Classification
Cause-type Severity Encoding
XGBoost-based Probabilistic Forecaster

Figure 3. LLM-enhanced news feature engineering workflow for transforming raw market events into forecasting-ready signals.

Additional Baseline Models

To provide a comprehensive comparison against state-of-the-art forecasting techniques, we evaluated three representative baseline models widely used in recent electricity price and load forecasting literature:

(1) NGBoost (Uncertainty-Aware Gradient Boosting)

NGBoost produces probabilistic forecasts by modelling a full predictive distribution instead of point estimates. We adopted the normal distribution parameterization with tree-based base learners and 500 boosting iterations. NGBoost serves as a strong benchmark for distributional forecasting without conformal calibration. However, its predictions tend to be overdispersed under high-volatility conditions in five-min NEM data.

(2) Informer (Transformer Architecture)

Informer is an efficient transformer model optimized for long sequence time-series forecasting using ProbSparse attention and generative decoding. We used a 24-h look-back window, two encoder layers, two decoder layers, a hidden dimension of 64, and the Adam optimizer. Informer is capable of modelling long-range dependencies but often smooths extreme price spikes when trained on high-frequency electricity markets.

(3) CNN–LSTM Hybrid Network

The CNN–LSTM baseline combines one-dimensional convolutions for local temporal feature extraction with LSTM layers for sequential modelling. The network uses two convolutional layers, followed by two LSTM layers (hidden size 64), and a dense output head. Although commonly effective in load forecasting, CNN–LSTM architectures typically underfit abrupt spike patterns unless trained with volatility-aware objectives.

All baseline models use the same look-back window ($L = 24$ h) and follow the same day-ahead training–calibration–testing protocol described in Section 2.5.

3.3. Evaluation Metrics and Statistical Tests

We evaluated predictive accuracy using MAE and RMSE. We further computed range-normalized RMSE (NRMSE) defined as:

$$\mathrm{NRMSE} = \frac{\mathrm{RMSE}}{\max_t(y_t) - \min_t(y_t)} \times 100\%. \tag{11}$$

For probabilistic forecasts, we evaluated prediction interval coverage probability (PICP) and mean prediction interval width (MPIW). We also reported the absolute coverage error $|\text{PICP} - (1 - \alpha)|$. When comparing models with vs. without news features, we applied the Diebold–Mariano (DM) test with the Newey–West (HAC) variance estimation.

3.4. Experimental Settings

Unless otherwise stated, we adopted a look-back window $L = 24$ h, a horizon $\Delta = 5$ min, calibration window $K = 7$ days, and $\alpha = 0.15$. The random seed was fixed at 42 for reproducibility. To avoid date-specific overfitting, all hyperparameters remained fixed across test days. The choice of $\alpha = 0.15$ followed operational practice in energy-market forecasting, where 85% prediction intervals offered a practical balance between informativeness and reliability. Higher confidence levels (90–95%) were tested, but they produced considerably wider intervals that reduced the usefulness for day-ahead decision-making in trading applications. The empirical coverage (PICP) corresponding to the nominal 85% interval was reported in all case studies.

Ablation sets consider no-news, only-Lvl1, only-causeType, weather-only, quantile vs. point configurations. In the main text, we focused on the most informative contrast, namely the presence vs. absence of news features (no-news vs. with-news), while the other ablation variants are summarised more briefly in Section 4.3 due to space limitations.

4. Results and Analysis

We evaluated the predictive performance of the proposed short-term electricity price forecasting framework using data from the New South Wales (NSW) region in May 2024 at a five-minute resolution. All models shared identical architectures and hyperparameters. To prevent information leakage, all news-derived features were aggregated at the daily level and lagged by one day before being broadcast back to five-minute granularity.

Although we presented detailed case studies for selected volatile days (such as 18 May and 3 February), all models in this study were trained and evaluated on the complete NSW 12-month dataset in 2024. The case studies were included solely to illustrate behavioural patterns during extreme market conditions and did not represent the full training or evaluation window used in our experiments.

4.1. Case Study on 18 May 2024

Unless otherwise stated, the blue shaded region in the forecasting figures indicates the calibrated $(1 - \alpha)$ prediction interval, where $\alpha = 0.15$. The median forecast ($q = 0.5$) is shown as an orange curve, while blue markers indicate actual observations. Background colors encode forecast confidence based on the prediction interval width (green: high, yellow: medium, and red: low).

Figure 4 presents the baseline XGBoost model without news features. The model yields an NRMSE of 8.79%, which is acceptable for intra-day horizons but exhibits a delayed response during the evening price surge. The blue line represents the actual RRP, while the orange line denotes the predicted values.

With the inclusion of event-driven news features (Figure 5), the NRMSE improves to 8.40%. The model anticipates the upturn between 18:00 and 20:00 earlier and produces a smoother trajectory, reducing both overshooting and oscillation during the high-price regime.

https://doi.org/10.3390/app16010200

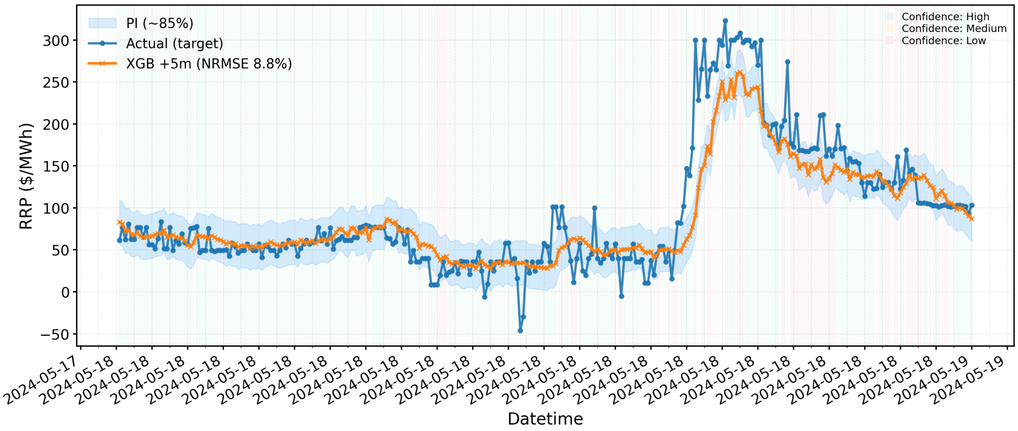


Figure 4. Forecasting results on 18 May 2024 using the baseline point XGBoost model without news features (NRMSE = 8.79% and PICP = 77.8%).

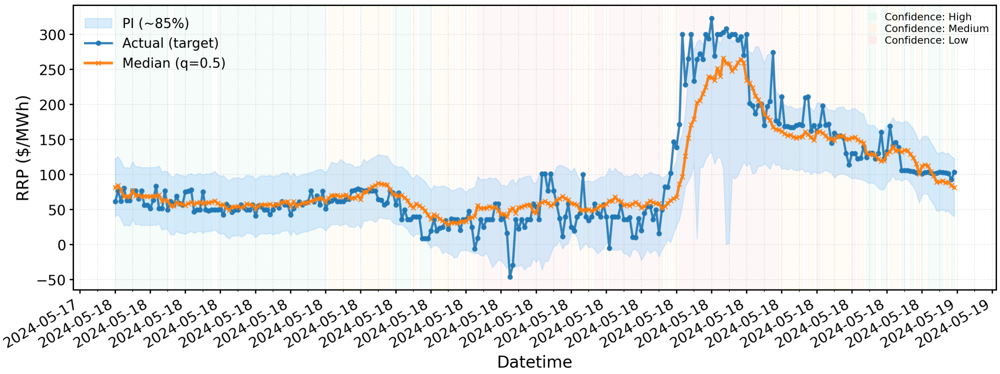


Figure 5. Forecasting results on 18 May 2024 using the proposed quantile XGBoost model with news features (NRMSE = 8.40 % and PICP = 90.6%).

For the typical high-load day of 18 May 2024, the point XGBoost model with conformal calibration achieves an NRMSE of 8.79% and an empirical coverage (PICP) of 77.8% for the nominal 85% prediction intervals (Figure 4). When using the quantile based XGBoost with conformal recalibration, the NRMSE is further reduced to 8.40%, and the empirical coverage increases to 90.6%, which is slightly above the nominal 85% level (Figure 5). These results indicate that the quantile-based specification yields both more accurate point forecasts and more conservative, yet better calibrated, uncertainty bands on this day.

4.2. Case Study: Prolonged High-Price Event on 04 February 2024

We further evaluated a prolonged high-price event on 4 February 2024, during which NSW prices exceeded USD 300/MWh for over 100 min (18:10–19:45). As illustrated in Figure 6, the baseline point XGBoost model without news features fails to anticipate the rapid price surge, resulting in substantial underestimation during the high-price regime.

In this extreme-price case study, we applied an upper clipping level of CAP $= USD300/\mathrm{MWh}$ to the training labels to stabilise model fitting in the presence of rare and highly volatile spikes. This threshold is used purely as a practical upper bound for learning and does not correspond to the regulatory Market Price Cap (MPC). Only upper clipping is applied; lower prices, including negative values, remain unmodified so that the characteristic asymmetry of NEM price distributions is preserved.

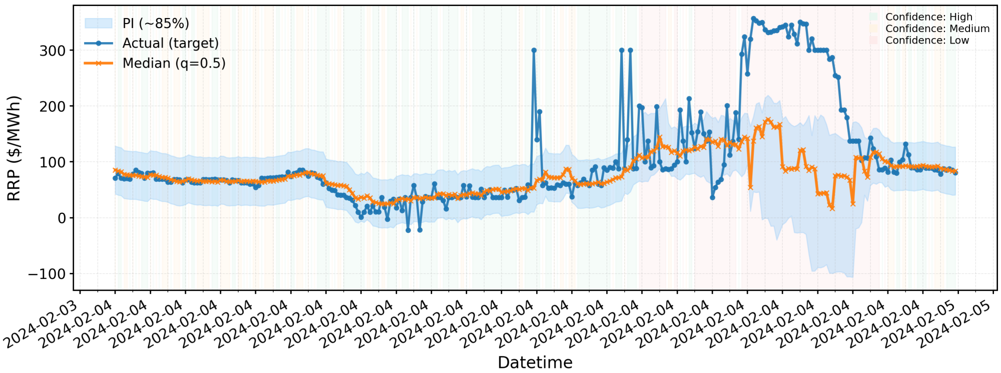


Figure 6. Forecasting results on 4 February 2024 using the baseline point XGBoost model without news features (NRMSE = 21.86% and PICP = 80.2%).

In contrast, Figure 7 shows the corresponding prediction results with news features, where the model adjusts upwards earlier and reduces spike-related errors, demonstrating improved responsiveness to stress events driven by market disruptions. For this prolonged high-price event on 4 February 2024, the point XGBoost baseline in Figure 7 attains an NRMSE of 16.97% and an empirical coverage (PICP) of 73.96% for the nominal 85% prediction intervals, whereas the quantile-based model in Figure 6 yields an NRMSE of 21.86% and a PICP of 80.2%. This reflects a trade-off between point accuracy and interval calibration. The quantile-based specification provides better calibrated uncertainty bands under extreme price spikes, at the cost of slightly reduced point-wise accuracy. The comparison between Figures 6 and 7 clearly indicates that news-derived signals enhance early recognition of upward price deviations, improving situational awareness for decision-making under sustained volatility.

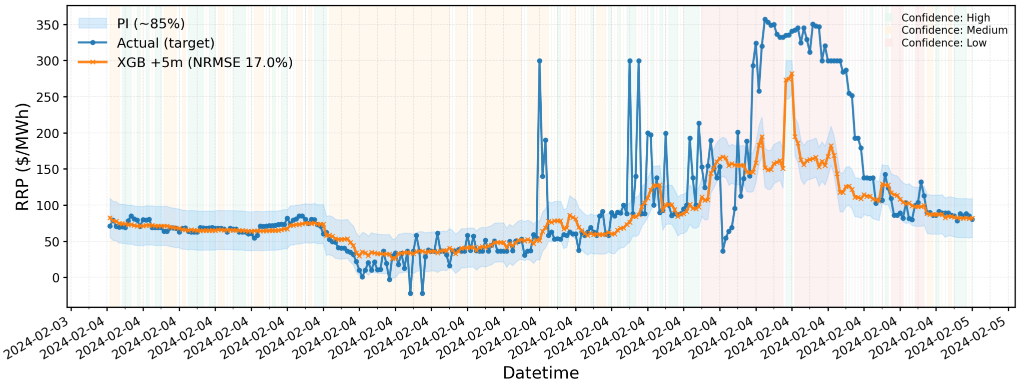


Figure 7. Forecasting results on 4 February 2024 using the proposed quantile XGBoost model with news features (NRMSE = 16.97%) and PICP = 73.96%.

4.3. Ablation Study

In this section, we report the main ablation contrast between the base model without news features (no-news) and the base model with news features (with-news). The intermediate variants (only-Lvl1, only-causeType, weather-only, and point vs. quantile) achieve performances between these two extremes and are therefore discussed only briefly. The ablation study in Table 2 indicates that removing news features degrades forecasting quality for 18 May 2024: (i) NRMSE increases of 0.39 percentage points, (ii) peak deviations become more pronounced, (iii) Coverage of prediction intervals degrades during volatility spikes.

Table 2. Ablation study on the impact of news features on forecasting performance.

Model	Features	MAE (USD/MWh)	RMSE (USD/MWh)	NRMSE (%)
Base Model	Price + Weather + Calendar	20.76	32.45	8.79
Base Model with news feature	Price + Weather + Calendar + News	20.25	31.00	8.40

4.4. Comparison with State-of-the-Art Baseline Models

To assess performance relative to contemporary forecasting approaches, we evaluated NGBoost, informer, and CNN–LSTM models under the same day-ahead protocol used for our XGBoost-based framework. Figures 8–10 illustrate their predictions for 18 May 2024.

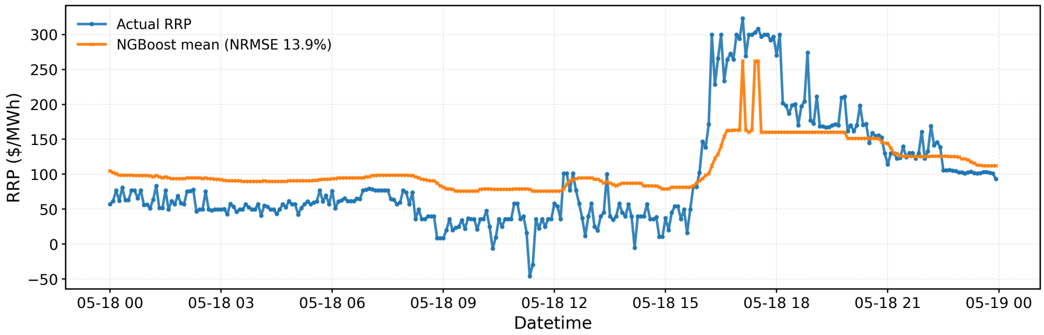


Figure 8. NGBoost predictive distribution on 18 May 2024. The mean forecast (orange line) systematically underestimates the evening surge in actual RRP (blue line), resulting in an NRMSE of 13.9%.

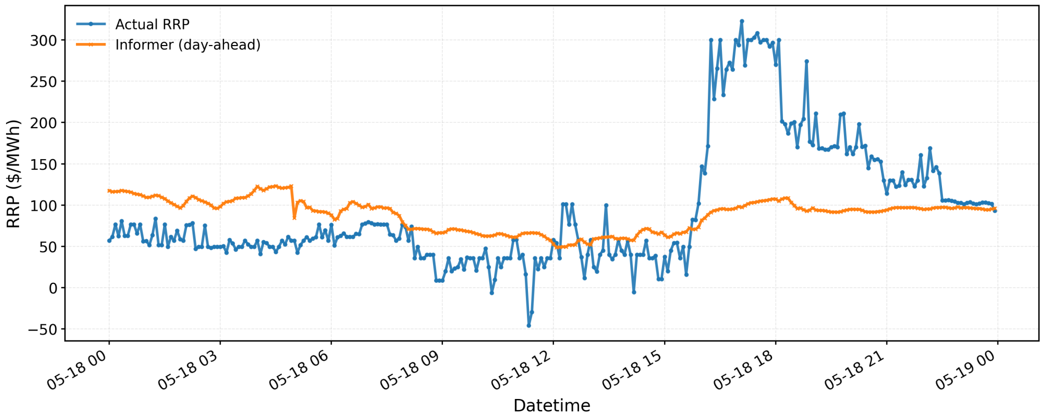


Figure 9. Comparison between the informer forecast (orange line) and actual RRP (blue line) on 18 May 2024. The model captures general diurnal trends but smooths high-frequency fluctuations, failing to anticipate the magnitude of the evening spike.

https://doi.org/10.3390/app16010200

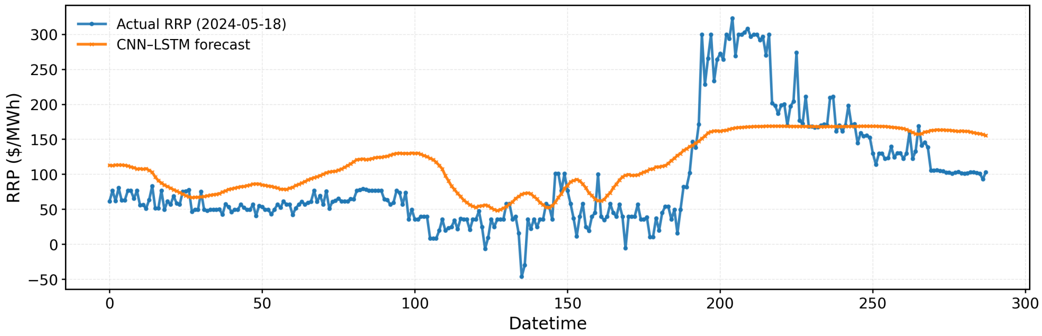


Figure 10. CNN-LSTM forecasting performance on 18 May 2024. The predicted trajectory (orange line) captures the coarse temporal shape of the actual RRP (blue line) but heavily smooths price dynamics, failing to anticipate the evening surge.

NGBoost produces a wide predictive interval (80% PI) spanning nearly USD ±600/MWh throughout the day. Although uncertainty aware, NGBoost significantly underestimates the evening surge and yields an NRMSE of 13.9%, indicating poor responsiveness to structural breaks.

The informer transformer model captures broad diurnal dynamics and identifies the onset of the 18:00–20:00 surge. However, it systematically smooths high-frequency fluctuations and underestimates spike magnitudes—a known limitation of transformer architectures trained on volatile energy price data.

The CNN–LSTM hybrid captures low-frequency patterns but exhibits strong smoothing and fails to respond to the evening spike altogether, leading to higher overall error. This behaviour aligns with previous studies where recurrent architectures underfit spike-driven price series without explicit volatility conditioning.

Table 3 summarizes the quantitative performance of all models. The proposed news-enhanced XGBoost model achieves the best accuracy among evaluated methods, outperforming deep learning and probabilistic baselines, particularly during high-volatility intervals.

Table 3. Performance comparison with state-of-the-art baseline models on 18 May 2024. Bold values indicate the best performance among the compared models.

Model	MAE (USD/MWh)	RMSE	NRMSE (%)
XGBoost (No-News)	20.76	32.45	8.79
XGBoost (News)	**20.25**	**31.00**	**8.40**
NGBoost (Normal)	40.95	51.36	13.91
Informer (Transformer)	54.47	67.47	18.28
CNN–LSTM	51.89	62.86	17.54

4.5. Statistical Significance Across Multiple Days

To ensure robustness beyond a single test day, we evaluated 22 days from 10–31 May 2024. A paired Diebold–Mariano (DM) test was used to compare absolute forecast errors at each five-min timestamp.

Across all hours:

$$DM_{\text{all}} = -0.35, \quad p = 0.728$$

indicates no statistically significant overall improvement.

However, restricting the comparison to spike periods (USD 100/MWh threshold):

$$DM_{\text{spike}} = -4.32, \quad p < 0.001$$

confirms a significant reduction in forecast error when volatility is high.

Figure 11 further illustrates that improvements accumulate specifically during structural price shocks.

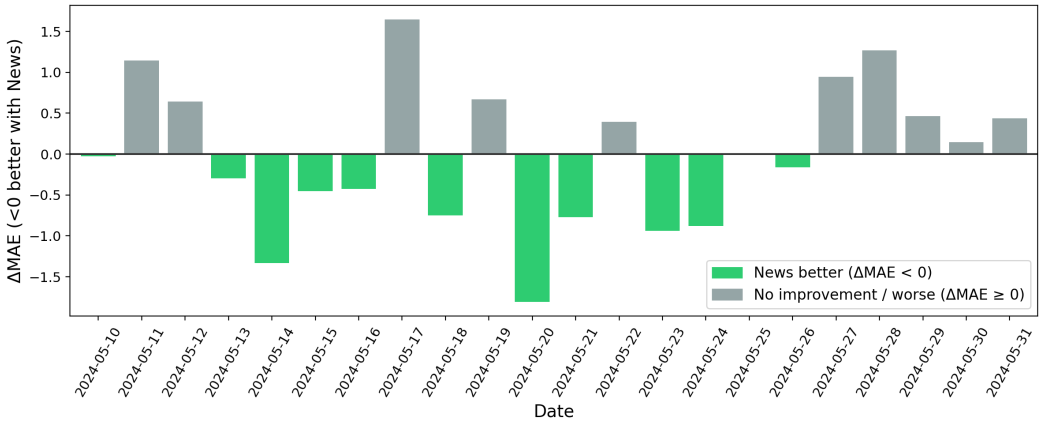


Figure 11. Daily MAE comparison between models with and without news features (10–31 May 2024). Green bars denote days where the news-enhanced model outperformed the baseline (ΔMAE < 0), demonstrating consistent gains during volatility.

To further verify whether these forecasting gains persist over consecutive days, we examined the cumulative difference in absolute errors between the two models.

As shown in Figure 12, improvements are not uniformly distributed but are concentrated around days with structural price shocks, supporting the conclusions from the spike-only DM test.

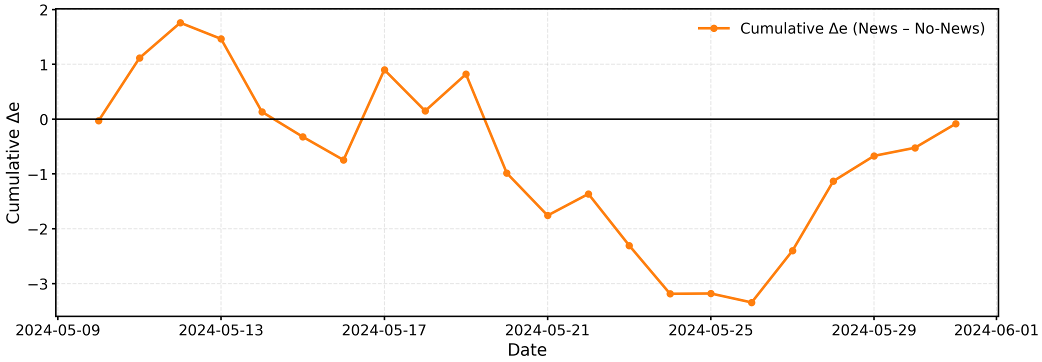


Figure 12. Cumulative difference in absolute prediction error. The orange trajectory trends downward, confirming that the inclusion of news features yields consistent cumulative accuracy improvements over the 22-day period.

These results demonstrate that news features offer marginal gains under normal market behaviour but provide substantial utility during rapid regime shifts where numerical features alone fail to capture emerging stress signals.

4.6. Interpretation of LLM-Driven Features and Economic Significance

To better understand why LLM-derived news features improve forecasting accuracy particularly during extreme volatility, we conducted several complementary analyses.

4.6.1. SHAP-Based Interpretation

Figure 13a presents the global SHAP importance ranking for 18 May 2024, showing that meteorological, demand, and calendar variables drive baseline predictability under normal operating conditions. However, when the analysis was restricted to high-price intervals

(RRP > 100), the feature ranking changes substantially, as shown in Figure 13b. In this regime, shock-related variables—particularly RRP_ROLL_MEAN_1H—dominate, indicating that recent volatility becomes the primary driver of extreme price fluctuations.

Figure 13c isolates SHAP contributions from LLM-derived news features. The most influential components include the following:

- **Weather-related Level-1 events**, consistent with storm-driven demand surges and generation deratings;
- **Mechanical and equipment failures**, which often precede forced outages and contribute to supply scarcity;
- **High Level-1 event density over the previous seven days**, capturing persistent system stress.

These patterns demonstrate that the LLM pipeline is not merely adding noise; rather, it consistently extracts semantically meaningful disruptions that align with real operational risks, thereby enhancing the model's responsiveness during structural market shocks.

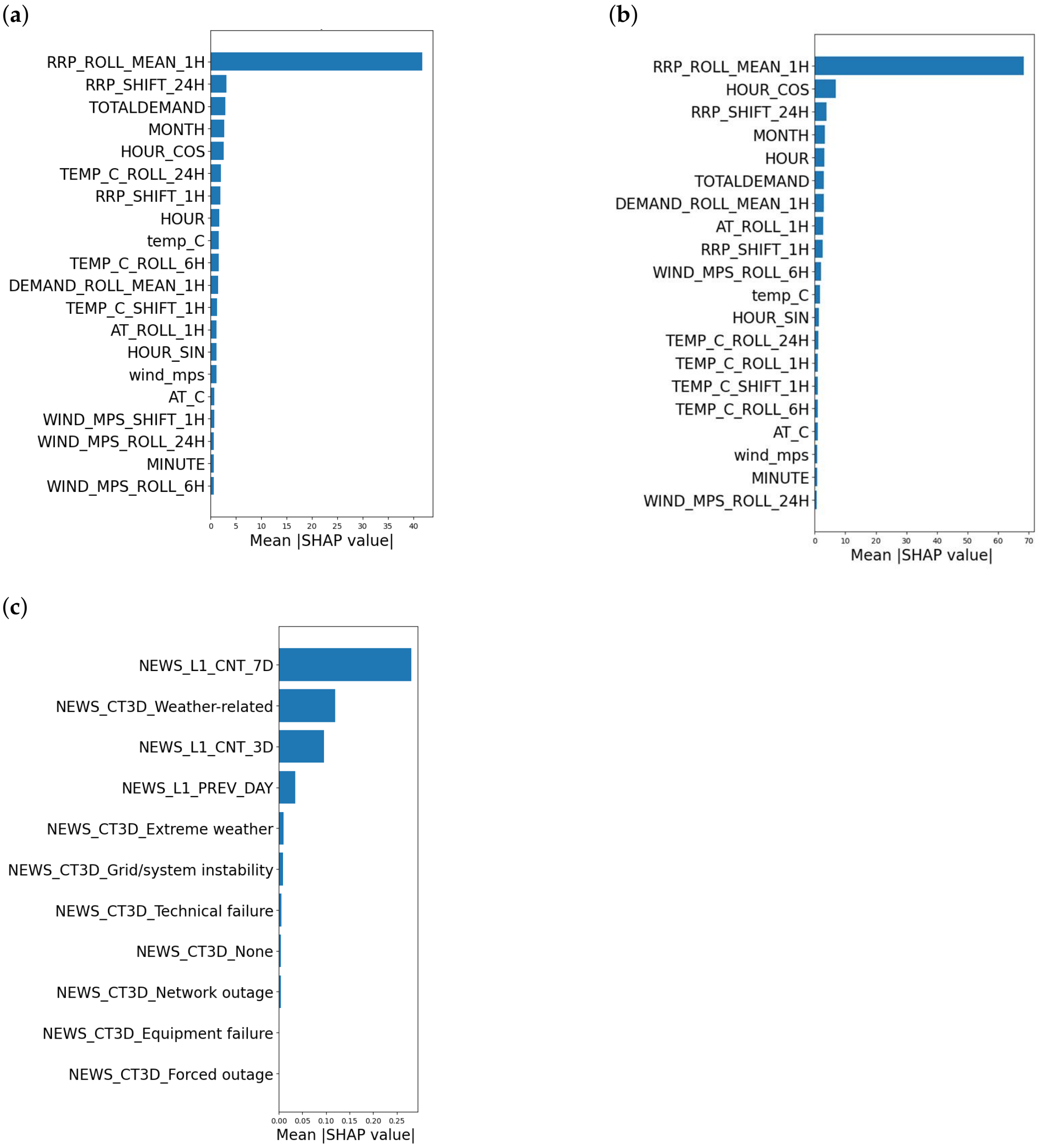


Figure 13. SHAP feature importance analysis: (**a**) global feature rankings vs. (**b**) rankings during price spikes ($RRP > 100$), illustrating regime-dependent driver shifts. (**c**) Importance of LLM-derived news features during spikes, identifying accumulated Level-1 events and weather signals as the primary contributors.

https://doi.org/10.3390/app16010200

4.6.2. Case Studies of LLM-Identified Events

Across multiple volatile days, we observed that LLM-classified events such as generator outage, extreme weather warning, and transmission constraint coincide with model improvements in peak-hour intervals. These events typically corresponded to physical system stress that was not fully captured by demand, weather, or autoregressive components alone.

4.6.3. Why Diebold–Mariano Significance Is Strong Only During Spikes

Although average-day improvements are modest, Diebold–Mariano tests applied to spike-only intervals show strong statistical significance. The reason is economic: the cost of forecast error is highly asymmetric. Errors during normal conditions have little operational impact, whereas errors during spikes correspond to mis-priced risk exposure, suboptimal dispatch decisions, and heightened balancing costs.

Thus, even small improvements in spike forecasts translate into large economic benefits. For example, improved anticipation of a USD 300/MWh event may outweigh hundreds of normal periods with small errors. This highlights the economic importance of LLM-derived signals for real-world decision-making.

5. Discussion

The findings show that while structured numerical variables (historical price dynamics, weather conditions, and calendar effects) remain the primary drivers of short-term electricity price forecasts, the inclusion of event-driven news features provides complementary benefits. Specific findings are as the following:

- *Interpretability*: News features allow a clearer linkage between forecast deviations and identifiable external events (e.g., outage announcements, extreme weather conditions, policy interventions). This enhances transparency for market operators and traders.
- *Resilience during shocks*: In periods of high volatility or unexpected disruptions, the news-augmented model exhibits improved responsiveness, capturing turning points and abnormal shifts better than purely numerical models.
- *Risk-aware forecasting*: The combined quantile and conformal framework produces practical uncertainty intervals that widen in volatile times and provide actionable confidence levels (High/Medium/Low) for operational decision-making.

In addition to the internal ablation, comparisons with NGBoost, CNN–LSTM, and informer demonstrate that the proposed framework remains competitive with state-of-the-art probabilistic and deep-learning approaches. While transformer and recurrent architectures capture smooth temporal profiles, they often under-represent high-frequency price spikes, whereas the news-enhanced XGBoost method provides more accurate responsiveness during event-driven volatility.

From an operational perspective, the approach also benefits from substantially lower computational cost compared to transformer-based models, making it more practical for frequent retraining and real-time deployment in market operations.

Nevertheless, important caveats remain. The improvement from textual features is modest in aggregate, reflecting the dominance of numerical signals under normal conditions. Moreover, because the model operates under a strict day-ahead setting, only daily aggregated and one-day-lagged news information is available at prediction time. Intra-day effects cannot be captured—not as a limitation of the method, but as a structural property of the forecasting problem itself. Future work might explore finer-grained event timestamps, richer semantic embeddings of news, and cross-regional propagation of disturbance signals.

In summary, integrating news-derived event signals into price forecasting augments model robustness and interpretability even if the incremental accuracy improvement is moderate in stable regimes.

6. Conclusions

In this study, we described a data-driven framework for short-term electricity price forecasting in the NSW region of the Australian National Electricity Market, combining structured numerical signals with exogenous event-driven features extracted using an LLM pipeline. By coupling quantile regression with conformal calibration, the model produces not only accurate point forecasts but also calibrated uncertainty intervals essential for risk-aware operations.

Empirical results on the five-minute NSW price data demonstrate that news-driven features lead to measurable, though moderate, improvement in accuracy (approximately 0.39 percentage-point reduction in NRMSE) and enhance performance during volatile market episodes. The key contribution lies not only in accuracy gains but in enabling interpretable, event-aware forecasting that supports decision-making in high-stakes electricity market contexts. Incorporating baseline models such as NGBoost, informer, and CNN-LSTM demonstrates that the proposed method achieves competitive or superior performance relative to modern deep learning architectures, while offering clearer interpretability and lower complexity.

For future work, several promising directions emerge. First, leveraging transformer-based LLMs and richer event embeddings could extract deeper semantic structure from news [13]. Second, extending this framework to multi-region forecasting and exploring cross-border propagation dynamics would enhance generality. Third, embedding the forecasting module into optimization engines, for example, bidding strategies or demand-response planning, could unlock operational value. Even incremental improvements in forecast accuracy can yield substantial economic benefits when paired with interpretability and uncertainty awareness.

Overall, this work affirms the value of integrating structured exogenous signals with unstructured textual insights for electricity price forecasting, offering both enhanced reliability and practical usability in dynamic market environments.

Author Contributions: Conceptualization, Y.H.; methodology, Y.H.; project administration, Y.H.; supervision, G.L., A.W. and J.Z.; validation, Y.H. writing—original draft, Y.H. and Y.Z.; writing—review and editing, Y.H. and Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is unavailable due to privacy or ethical restrictions.

Acknowledgments: The author would like to acknowledge the support of Xingxin Tang (University of New South Wales) and Ruikun Zhou (The University of Sydney) for significantly supporting this research.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Weron, R. *Modeling and Forecasting Electricity Loads and Prices: A Statistical Approach*; John Wiley & Sons: Chichester, UK, 2006.
2. Weron, R. Electricity price forecasting: A review of the state-of-the-art with a look into the future. *Int. J. Forecast.* **2014**, *30*, 1030–1081. [CrossRef]
3. Contreras, J.; Espinola, R.; Nogales, F.J.; Conejo, A.J. ARIMA models to predict next-day electricity prices. *IEEE Trans. Power Syst.* **2003**, *18*, 1014–1020. [CrossRef]
4. Lago, J.; De Ridder, F.; De Schutter, B. Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark. *Appl. Energy* **2018**, *225*, 936–951. [CrossRef]

https://doi.org/10.3390/app16010200

5. Lago, J.; Marcjasz, G.; De Schutter, B.; Weron, R. Forecasting electricity prices: Are deep learning models really superior? *Renew. Sustain. Energy Rev.* **2021**, *150*, 111441.
6. Bedi, J.; Toshniwal, D. Deep learning framework to forecast electricity price. *Appl. Energy* **2019**, *238*, 1312–1326. [CrossRef]
7. Kaur, P.; Edalati, M. Sentiment analysis on electricity Twitter posts. *arXiv* **2022**, arXiv:2206.05042.
8. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems 33*; Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2020; pp. 1877–1901.
9. Romano, Y.; Patterson, E.; Candes, E.J. Conformalized Conformal Calibration. In *Advances in Neural Information Processing Systems 32*; Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2019; pp. 3544–3554.
10. Lu, X.; Qiu, J.; Lei, G.; Zhu, J. Scenarios modelling for forecasting day-ahead electricity prices: Case studies in Australia. *Appl. Energy* **2022**, *308*, 118296. [CrossRef]
11. Australian Energy Market Operator. Aggregated Price and Demand Data. Australian Energy Market Operator, 2025. Available online: https://www.aemo.com.au/energy-systems/electricity/national-electricity-market-nem/data-nem/aggregated-data (accessed on 28 October 2025).
12. Huang, Y.; Huang, L.; Zhu, Y.; Lei, G.; Wang, A.; Zhu, J. Application of Large Language Models in Intelligent Preprocessing and Forecasting of Electricity Price. In Proceedings of the 2025 4th International Conference on Smart Grid and Green Energy (ICSGGE), Sydney, Australia, 28 February–2 March 2025; IEEE: Piscataway, NJ, USA, 2025; pp. 246–250.
13. Lu, X.; Qiu, J.; Yang, Y.; Zhang, C.; Lin, J.; An, S. Large Language Model-Based Bidding Behavior Agent and Market Sentiment Agent-Assisted Electricity Price Prediction. In *IEEE Transactions on Energy Markets, Policy and Regulation*; IEEE: Piscataway, NJ, USA, 2024.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.