RGe
ISSN: 2177-3246

SOYBEAN PRICE FORECASTING VIA HYBRID LSTM-LLM ARCHITECTURE: STATISTICAL AND ECONOMIC EVALUATION OF BRAZILIAN AGRIBUSINESS NEWS SENTIMENT

PREVISÃO DE PREÇOS DA SOJA VIA ARQUITETURA HÍBRIDA LSTM-LLM: AVALIAÇÃO ESTATÍSTICA E ECONÔMICA DO SENTIMENTO DE NOTÍCIAS DO AGRONEGÓCIO BRASILEIRO

PREDICCIÓN DE PRECIOS DE LA SOJA MEDIANTE ARQUITECTURA HÍBRIDA LSTM-LLM: EVALUACIÓN ESTADÍSTICA Y ECONÓMICA DEL SENTIMIENTO DE NOTICIAS DEL AGRONEGOCIO BRASILEÑO

10.56238/revgeov17n6-094

Marco Antonio França Benjamim
Graduando em Ciência da Computação
Instituição: Universidade Tecnológica Federal do Paraná (UTFPR)
Endereço: Paraná, Brasil
E-mail: benjamimmarcoaf@gmail.com

Samuel Bellido Rodrigues
Doutor em Métodos Numéricos em Engenharia
Instituição: Universidade Tecnológica Federal do Paraná (UTFPR)
Endereço: Paraná, Brasil
E-mail: samuelb@utfpr.edu.br

Lucas da Silva Ribeiro
Doutor em Métodos Numéricos em Engenharia
Instituição: Universidade Tecnológica Federal do Paraná (UTFPR)
Endereço: Paraná, Brasil
E-mail: lribeiro@utfpr.edu.br

Levi Lopes Teixeira
Doutor em Métodos Numéricos em Engenharia
Instituição: Universidade Tecnológica Federal do Paraná (UTFPR)
Endereço: Medianeira, Paraná, Brasil
E-mail: levilopes@utfpr.edu.br

Tasia Hickmann
Doutora em Métodos Numéricos em Engenharia
Instituição: Universidade Tecnológica Federal do Paraná (UTFPR)
Endereço: Paraná, Brasil
E-mail: hickmann@utfpr.edu.br

RGe
ISSN: 2177-3246

Jairo Marlon Correa
Doutor em Métodos Numéricos em Engenharia
Instituição: Universidade Tecnológica Federal do Paraná (UTFPR)
Endereço: Paraná, Brasil
E-mail: jairocorrea@utfpr.edu.br

ABSTRACT
Soybean is Brazil's leading agricultural commodity, and its price volatility poses significant challenges to producers, traders, and policymakers, given the market's nonlinear dependence on exogenous factors such as climate conditions, trade policies, and the informational flow of news. This study investigates to what extent incorporating textual sentiment extracted by agribusiness-specialized LLMs improves the predictive accuracy and economic value of LSTM models for soybean futures price forecasting (SJCc1). To this end, six architectures were empirically compared — a naïve benchmark, pure LSTM, LSTM with frozen LLM in scalar and probabilistic outputs, and end-to-end versions of both — using 3,261 price records and a corpus of 27,024 Brazilian agribusiness news articles, with fine-tuning on 1,000 labeled news items and Bayesian hyperparameter optimization via TPE. Statistical comparison employed the Model Confidence Set (MCS) procedure at 90% confidence, complemented by a paired block bootstrap test for cumulative returns. It is observed that only the LSTM+LLM architecture with probabilistic output joined the MCS alongside the naïve benchmark — being the only model to generate statistically significant cumulative excess return over buy-and-hold (58.27%; $p \approx 0.003$; Sharpe ratio: 1.74) —, with its advantage amplifying during high-volatility periods. It is concluded that the predictive gain stems from the specific combination of a specialized LLM and probabilistic sentiment encoding, rather than from textual integration per se.

Keywords: Agricultural Commodities. Recurrent Neural Networks. Sentiment Analysis. Bayesian Optimization. Model Confidence Set.

RESUMO
A soja é a principal commodity agrícola brasileira e a volatilidade de seus preços impõe desafios significativos a produtores, traders e formuladores de políticas públicas, dada a dependência não linear do mercado a fatores exógenos como condições climáticas, políticas comerciais e fluxo informacional de notícias. Objetiva-se investigar em que medida a incorporação de sentimento textual extraído por LLMs especializados no agronegócio aprimora a acurácia preditiva e o valor econômico de modelos LSTM para previsão do contrato futuro de soja (SJCc1). Para tanto, seis arquiteturas foram comparadas empiricamente — benchmark naïve, LSTM pura, LSTM com LLM congelada em saídas escalar e probabilística, e versões end-to-end de ambas — utilizando 3.261 registros de preço e um corpus de 27.024 notícias brasileiras do agronegócio, com fine-tuning sobre 1.000 notícias rotuladas e otimização bayesiana de hiperparâmetros via TPE. A comparação estatística utilizou o procedimento Model Confidence Set (MCS) a 90% de confiança, complementada por teste bootstrap emparelhado em blocos para retorno acumulado. Observa-se que apenas a arquitetura LSTM+LLM com saída probabilística integrou o MCS ao lado do benchmark naïve — sendo a única a gerar retorno acumulado estatisticamente significativo sobre o buy-and-hold (58,27%; $p \approx 0{,}003$; Sharpe ratio: 1,74) —, com vantagem ampliada em períodos de alta volatilidade. Conclui-se que o ganho preditivo decorre da combinação específica entre LLM especializada e codificação probabilística do sentimento, não da integração textual per se.

Palavras-chave: Commodities Agrícolas. Redes Neurais Recorrentes. Análise de Sentimento. Otimização Bayesiana. Model Confidence Set.

RESUMEN
La soja es la principal commodity agrícola brasileña y la volatilidad de sus precios impone desafíos significativos a productores, traders y formuladores de políticas públicas, dada la dependencia no lineal del mercado a factores exógenos como las condiciones climáticas, las políticas comerciales y el flujo informacional de noticias. Se investiga en qué medida la incorporación de sentimiento textual extraído por LLMs especializados en el agronegocio mejora la precisión predictiva y el valor económico de modelos LSTM para la predicción del contrato de futuros de soja (SJCc1). Para ello, seis arquitecturas fueron comparadas empíricamente — benchmark naïve, LSTM pura, LSTM con LLM congelada en salidas escalar y probabilística, y versiones end-to-end de ambas — utilizando 3.261 registros de precios y un corpus de 27.024 noticias brasileñas del agronegocio, con fine-tuning sobre 1.000 noticias etiquetadas y optimización bayesiana de hiperparámetros mediante TPE. La comparación estadística empleó el procedimiento Model Confidence Set (MCS) con 90% de confianza, complementada por una prueba bootstrap emparejada en bloques para el retorno acumulado. Se observa que únicamente la arquitectura LSTM+LLM con salida probabilística integró el MCS junto al benchmark naïve — siendo el único modelo en generar retorno acumulado estadísticamente significativo sobre el buy-and-hold (58,27%; $p \approx 0{,}003$; Sharpe ratio: 1,74) —, con ventaja ampliada en períodos de alta volatilidad. Se concluye que la ganancia predictiva proviene de la combinación específica entre un LLM especializado y la codificación probabilística del sentimiento, y no de la integración textual per se.

Palabras clave: Commodities Agrícolas. Redes Neuronales Recurrentes. Análisis de Sentimiento. Optimización Bayesiana. Model Confidence Set.

1 INTRODUCTION

Soybean is Brazil's most produced agricultural crop, with output projected to grow 1.1% (1.8 million tonnes) in the 2026 harvest (IBGE, 2025). The country is also the world's largest producer, with 45.9 million hectares harvested in 2024, ahead of the United States at 34.8 million hectares (FAO, 2024). As a strategic input across multiple sectors — from human and animal nutrition to bioenergy — soybean price dynamics are a central concern for producers, agribusinesses, traders, and policymakers.

Accurately forecasting soybean prices remains a considerable challenge: the market is highly efficient and subject to nonlinear dependencies, abrupt shocks, and exogenous factors such as climate conditions, trade policies, and macroeconomic conditions. Neural networks, particularly Long Short-Term Memory (LSTM) models, have consistently outperformed linear approaches and naïve benchmarks in commodity price forecasting (Zhang et al., 2018; Puchalsky et al., 2018; Ray et al., 2023; Song et al., 2024; Ali et al., 2025).

However, historical price series do not explicitly capture supply or demand shocks frequently conveyed through news. Integrating textual sentiment extracted by Large Language Models (LLMs) has shown consistent predictive gains in financial markets (Farimani et al., 2022; Zhang; Xia, 2022; Zhang; Dong; Xu, 2026). Despite these advances, four gaps persist in the agricultural commodities domain: (i) predominance of general-purpose LLMs lacking agribusiness vocabulary specialization; (ii) absence of systematic comparison across text-numeric fusion architectures; (iii) limited evidence on model robustness during high-volatility periods; and (iv) lack of rigorous statistical tests for cumulative return.

This study addresses these gaps through a comprehensive methodological and economic evaluation. Six next-day forecasting architectures were developed and empirically compared, combining historical soybean futures prices (SJCc1) with textual sentiment features extracted from a corpus of 27,024 Brazilian agribusiness news articles: (i) naïve benchmark; (ii) pure LSTM; (iii–iv) LSTM with frozen LLM in scalar and probabilistic outputs; and (v–vi) end-to-end trainable versions of both. Two LLMs were fine-tuned on 1,000 labeled news items with Bayesian hyperparameter optimization via TPE. Statistical comparison used the Model Confidence Set (MCS) procedure at 90% confidence, complemented by a paired block bootstrap test for cumulative excess returns over a buy-and-hold benchmark.

Specifically, this study investigates to what extent incorporating textual sentiment from specialized agribusiness news can enhance the predictive accuracy and economic value of neural network-based soybean price forecasting models.

2 MATERIALS AND METHODS

2.1 COMPUTATIONAL RESOURCES AND TECHNOLOGIES

Experiments were conducted on Google Colab using an NVIDIA Tesla T4 GPU (15 GB VRAM) and 12.7 GB of system RAM, with sessions limited to 5 hours. The implementation used Python 3.12.12 with the following libraries: Pandas 2.2.2, NumPy 2.0.2, torch 2.10.0, matplotlib 3.10.0, arch 8.0.0, Optuna 4.7.0, SciPy 1.16.3, transformers 5.0.0, and scikit-learn 1.6.1, BeautifulSoup 4.14.3, requests 2.32.4.

2.2 DATA ACQUISITION AND PREPROCESSING

This section describes the collection and preparation of the numerical and textual data underlying the study.

2.2.1 Historical Price Data

Daily soybean futures prices (SJCc1) were obtained from Investing.com, covering June 12, 2012 to July 19, 2025 — a total of 3,261 daily records. Data were split chronologically into three subsets, as shown in Table 1, with the training set comprising approximately 70% of all records:

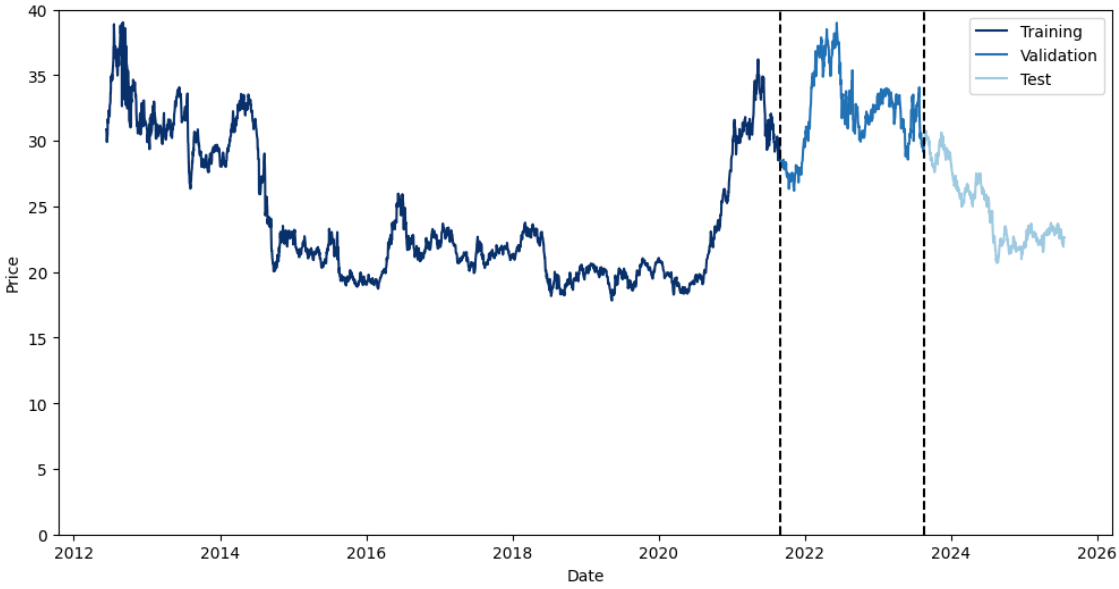
Table 1. Dataset partition.

Split	Period	Records
Training	12/06/2012 – 01/09/2021	2,281
Validation	02/09/2021– 18/08/2023	490
Test	21/08/2023– 19/07/2025	490

Source: Elaborated by the authors (2026).

Figure 1 shows the complete series with partition boundaries indicated by vertical lines and color coding.

Figure 1. Price series partitioned into training/validation/test sets.



Source: Elaborated by the authors (2026).

2.2.2 Price Series Preprocessing

The historical price series contained missing values on trading days, concentrated mainly between 2012 and 2021. A two-level imputation strategy was applied. At the first level, simple linear regression was used with the CBOT soybean series as predictor. The high correlation between the two series (r = 0.9948) justified this choice, and fit quality was assessed via random train-test split, yielding $R^2 = 0.9924$ on the test set — indicating excellent imputation capacity. This method filled 1,265 values, all between 07/20/2012 and 03/29/2021. At the second level, for missing days without a corresponding CBOT value, linear interpolation between adjacent observations was applied, covering 24 additional cases. Missing days were validated against the Ibovespa trading calendar, a gap-free reference series. Weekends and holidays carry no records, as no trading occurs on those days. Of the 1,289 total imputations, only one fell within the test set, making the impact on model evaluation negligible.

For model training, the series was standardized via z-score normalization (zero mean and unit variance), as shown in Equation (1).

$$y*_i = \frac{y_i - \mu}{\sigma} \quad (1)$$

Where y_i is the price at time i, μ is the mean, σ the standard deviation, and $y*_i$ the resulting standardized value at instant i.

The parameters μ and σ were estimated exclusively on the training set and applied to all other sets, preventing data leakage.

2.2.3 Textual Dataset

A comprehensive news corpus was built from three Brazilian agribusiness portals to capture factors influencing soybean supply and demand. Sources, periods, and volumes are shown in Table 2.

Table 2. News sources, collection periods, and volumes.

Source	Collection Period	News Articles
Canal Rural	29/06/2012 – 01/08/2025	11,182
Notícias Agrícolas	02/01/2015 – 31/07/2025	13,521
Globo Rural	03/10/2022 – 08/08/2025	2,321
Total		27,024

Source: Elaborated by the authors (2026).

To ensure predictive rigor and avoid look-ahead bias, the integrated dataset was structured under a strict causal constraint: any price forecast for instant t+1used only closing prices (y_t) and sentiment from news published (n_t) up to the end of day t.

2.2.4 Labeled Dataset for Fine-Tuning

To specialize the LLM for the agribusiness domain, a labeled sentiment dataset was constructed. A random subsample of 1,000 news articles (without replacement) was drawn from the training period (12/06/2012 – 01/09/2021), ensuring temporal alignment between learned language patterns and the relevant market context.

Labels were assigned by Gemini 2.5 Pro (Google) using the following prompt: “Atue como um analista de commodities agrícolas e classifique as notícias de soja que eu enviar utilizando a seguinte lógica de pontuação: responda -1 para notícias baixistas que indiquem aumento de oferta ou queda na demanda, sugerindo que o preço irá cair; responda 0 para notícias neutras que não alterem o equilíbrio do mercado; e responda 1 para notícias altistas que indiquem quebras de safra, problemas logísticos ou aumento de demanda, sugerindo escassez e alta nos preços. Para cada notícia enviada, forneça primeiro a classificação numérica e, em seguida, uma breve justificativa técnica sobre o impacto esperado nos preços.” The use of an LLM as annotator is grounded in Gilardi et al. (2023), who showed that such models outperform non-specialist crowd workers in text classification tasks at significantly better cost-effectiveness. Gemini 2.5 Pro was selected for its state-of-the-art performance on public comprehension and reasoning benchmarks, surpassing the model evaluated in that study.

To validate label quality, a stratified sample of 99 articles (33 per class) was manually annotated by the authors. Agreement was measured by quadratically weighted Cohen's Kappa (κ_w), appropriate for ordinal scales, yielding $\kappa_w = 0.92$ — classified as almost perfect agreement by Landis and Koch (1977) — with 89.5% accuracy. The main disagreements occurred at the boundary between neutral and bullish classes.

2.5 BAYESIAN HYPERPARAMETER OPTIMIZATION

Optimal hyperparameter selection is critical for machine learning model performance (Hutter; Kotthoff; Vanschoren, 2019). Traditional methods such as Grid Search become computationally prohibitive as the search space grows, while Random Search does not leverage previous evaluations to guide future trials (Zhang et al., 2021). Bayesian optimization overcomes these limitations by iteratively building a probabilistic surrogate model of the hyperparameter-performance relationship, directing each new evaluation toward the most promising regions of the search space (Zhou et al., 2024).

This work adopted the Tree-structured Parzen Estimator (TPE) variant (Bergstra et al., 2011), particularly suited for mixed search spaces with continuous and categorical variables. In benchmarks, TPE finds robust optima with substantially fewer evaluations than grid or random methods (Bergstra et al., 2011; Liao et al., 2024).

2.6 LSTM ARCHITECTURE AND HYPERPARAMETER OPTIMIZATION

Long Short-Term Memory (LSTM), introduced by Hochreiter and Schmidhuber (1997), was developed to mitigate vanishing and exploding gradient problems in standard Recurrent Neural Networks (RNNs). Its gating mechanism — input, forget, and output gates — enables selective control of information flow through a memory cell, retaining long-term dependencies across the temporal sequence (Bengio; Simard; Frasconi, 1994; Fan et al., 2019). LSTM has been validated for agricultural commodity price forecasting, including cotton (Chandan; Kumari, 2025), supporting its adoption as the predictive core of this study.

Beyond the standard unidirectional LSTM, the bidirectional variant (Bi-LSTM) was included as a hyperparameter option. Bi-LSTM processes the sequence simultaneously in chronological and reverse directions, combining hidden states from both at each time step (Schuster; Paliwal, 1997). This ability to capture temporal context from two perspectives has shown performance gains in time series forecasting (Wang et al., 2023), and was indeed selected by Bayesian optimization in the hybrid LSTM+LLM architectures.

2.6.1 Search Space and Optimization

Table 3. Hyperparameters and candidate values for LSTM optimization.

Hyperparameter	Candidate Values
Input window size	5, 10, 15, 20, 30
Number of LSTM blocks	1, 2, 3, 4, 5
Hidden layer size	4, 8, 16, 32, 64, 128, 256, 512, 1024
Dropout rate	0.1 to 0.5
Bidirectionality	True, False
Optimizer	AdamW, SGD, RMSprop
Batch size	1, 4, 8, 16, 32
Learning rate	$1 \times e^{-5}$ to $1 \times e^{-2}$ (log scale)
Weight decay	$1 \times e^{-6}$ to $1 \times e^{-4}$ (log scale)
LR scheduler	False, ReduceLROnPlateau, CosineAnnealing, StepLR

Source: Elaborated by the authors (2026).

Optimization targeted minimization of Mean Square Error (MSE) on the validation set. The search space comprised the hyperparameters and value ranges shown in Table 3.

To ensure computational efficiency, each candidate configuration was trained for at most 20 epochs, with early stopping after 3 consecutive epochs without validation error reduction. The search was implemented using the Optuna library, whose native pruning mechanism automatically halts trials performing below the average of previous evaluations, concentrating the computational budget on the most promising regions of the search space.

2.7 LANGUAGE MODEL (LLM) AND HYPERPARAMETER OPTIMIZATION

The effectiveness of an LLM in specialized domains critically depends on its adaptation to domain-specific vocabulary and contextual relationships (Mu et al., 2026; Raiaan et al., 2024). In agricultural commodity markets, terms such as "rainfall" or "drought" carry direct economic implications for soybean supply that general-purpose models — or even corporate finance LLMs — tend to misinterpret. For this reason, a pretrained LLM was fine-tuned on the 1,000 labeled news articles (Section 2.2.4) with Bayesian hyperparameter optimization, aiming to specialize the model for sentiment classification in the Brazilian agribusiness context.

2.7.1 Fine-Tuning Process and Model Selection

The search space jointly considered base model selection and architecture and training hyperparameters, as shown in Table 4.

For each output format — scalar (Tanh) or probabilistic (Softmax) — 300 configurations were evaluated via TPE with Optuna pruning, with training limited to 20 epochs and early stopping after 3 consecutive epochs without validation loss reduction. The loss function was adapted to the output type: MSE for Tanh output and Cross-Entropy for probabilistic output.

In both variants, two dense layers were added on top of the base LLM: an intermediate layer with dimension and activation function defined by optimization, followed by an output layer with Tanh activation (scalar in [−1, 1]) or Softmax (probability vector for three classes: bearish, neutral, and bullish).

The selected hyperparameters for both specialized LLMs are shown in the last two columns of Table 4.

Table 4. Hyperparameters, candidate values, and selected configuration for LLM optimization.

Hyperparameter	Candidate Values	LLM-Prob	LLM-Tanh
Base Model	lucas-leme/FinBERT-PT-BR, ProsusAI/finbert, neuralmind/bert-base-portuguese-cased, distilbert-base-uncased, openai-community/gpt2	lucas-leme/FinBERT-PT-BR	neuralmind/bert-base-portuguese-cased
Freezing Strategy	No layers frozen, up to, only the last two layers unfrozen	No layers frozen	No layers frozen
Classification Layer Size	128, 256, 512, 1024	128	512
Activation Function	ReLU, GELU, Tanh, LeakyReLU, ELU, SiLU	Tanh	GELU
Dropout Rate	0.1 to 0.4	0.356	0.118
Batch Size	8, 16, 32	32	16
Learning Rate	$1 \times e^{-6}$ to $3 \times e^{-2}$ (log scale)	2.43×10^{-6}	3.42×10^{-4}
Weight Decay	$1 \times e^{-5}$ to $1 \times e^{-2}$ (log scale)	$6.76 \times e^{-5}$	$2.37 \times e^{-4}$
Optimizer	AdamW, SGD, RMSprop	RMSprop	AdamW
LR Scheduler	False, ReduceLROnPlateau, CosineAnnealing, StepLR	CosineAnnealing	ReduceLROnPlateau

Source: Elaborated by the authors (2026).

2.8 LSTM + LLM INTEGRATION STRATEGIES

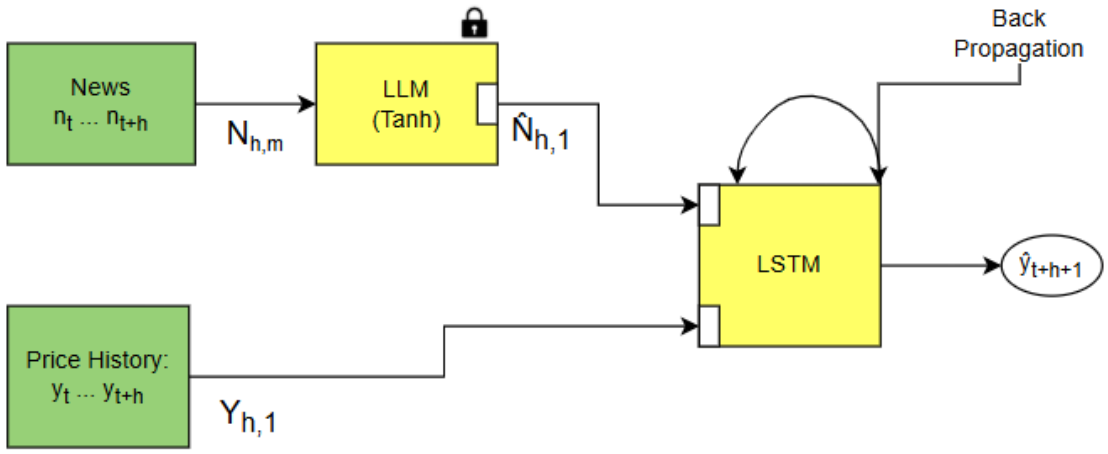
Table 5. LLM variant, output type, and optimization strategy for each of the 6 models.

No	Name	LLM	LLM Output	Optimization
1	Naïve	-	-	-
2	Pure LSTM	-	-	TPE 500 iter.
3	LSTM+LLM_Tanh	Frozen	Scalar	TPE 500 iter.
4	LSTM+LLM_Prob	Frozen	3-class vector	TPE 500 iter.
5	LSTM+LLM_Tanh(E2E)	Trainable	Scalar	Single run
6	LSTM+LLM_Prob(E2E)	Trainable	3-class vector	Single run

Source: Elaborated by the authors (2026).

Six next-day price forecasting architectures were designed and compared to empirically evaluate the added value of textual information and the impact of different integration schemes on predictive performance. Table 5 summarizes the variants.

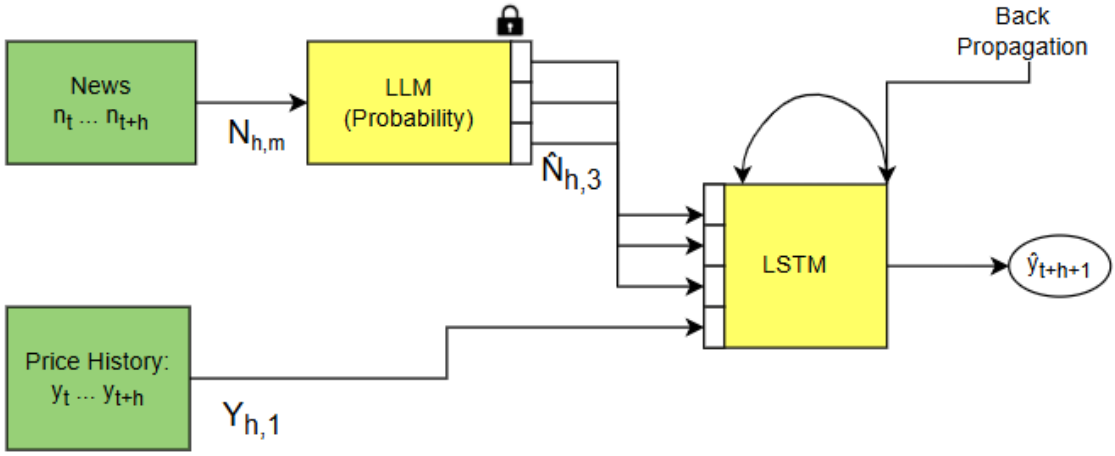
Figure 2. LSTM + LLM-Tanh model architecture; the LLM is not trained jointly with the LSTM.



Source: Elaborated by the authors (2026).

Figure 3 illustrates the joint LSTM+LLM architecture with probabilistic output via Softmax activation. In contrast to the previous model, the LLM provides the LSTM with a three-class vector [bearish, neutral, bullish] instead of a continuous scalar, enriching the sentiment representation fed into the recurrent network.

Figure 3. LSTM + LLM-Prob model architecture; the LLM is not trained jointly with the LSTM.



Source: Elaborated by the authors (2026).

The naïve model (model 1) serves as the basic benchmark, assuming the best predictor of the current day's price is the previous day's price: $\hat{y}_{(t)} = y_{(t-1)}$.

In frozen LLM architectures (models 3 and 4), the sentiment extractor is fixed after fine-tuning and produces a feature — scalar via Tanh or three-class vector via Softmax — concatenated to the price sequence as LSTM input. In end-to-end architectures (models 5 and 6), LLM parameters remain trainable during joint training, enabling fine-tuning guided directly by the price forecasting error.

For days with multiple publications, daily sentiment is aggregated as the arithmetic mean of up to 20 news items. For days without news, a neutral signal is imputed: an equiprobable vector [0.33; 0.34; 0.33] in probabilistic architectures and scalar value 0 in Tanh models. News published on non-trading days — weekends and holidays — is appended to the most recent preceding trading day, avoiding the discard of potentially relevant textual information.

Models 2, 3, and 4 underwent Bayesian optimization with 500 iterations; their optimal hyperparameters are shown in Table 6. End-to-end models (5 and 6) were not subjected to extensive optimization — given the prohibitive computational cost — and inherited the hyperparameters of their frozen LLM counterparts: model 5 from model 3, and model 6 from model 4.

Table 6. LSTM-only hyperparameters selected via Bayesian optimization for each model

Hyperparameter	Pure LSTM	LSTM Prob	LSTM Tanh
Window Size	15	10	5
Hidden Units	[512]	[1024]	[1024, 1024, 1024, 1024]
Layers per Block	1	1	1
Dropout Rate	0.164	0.406	0.265
Bidirectional	False	True	True
Optimizer	RMSprop	AdamW	AdamW
Learning Rate	4.49×10^{-4}	1.15×10^{-4}	1.06×10^{-4}
Weight Decay	1.53×10^{-4}	7.04×10^{-6}	3.89×10^{-4}

Warmup Strategy	Linear (0.1 ratio)	None	None
Batch Size	8	1	1

Source: Elaborated by the authors (2026).

2.9 DIRECTIONAL RETURN CONSTRUCTION

To assess economic implications, model signals were converted into daily returns of a simple directional strategy. The rule is: take a long position ($s_t = 1$) if the forecast for the next day's closing price exceeds the current day's closing, and a short position ($s_t = -1$) otherwise. The daily strategy return is therefore:

$$R_t^{\text{model}} = s_t * r_t \quad (2)$$

$$r_t = (y_t - y_{t-1})/y_{t-1} \quad (3)$$

Where y_t is the closing price on day t and r_t is the observed market return between t−1 and t. The benchmark corresponds to buy-and-hold:

$$R_t^{\text{benchmark}} = r_t \quad (4)$$

Cumulative returns are computed by compounding:

$$\text{cumret} = \prod_{t=1}^{T}(1 + R_t) - 1 \quad (5)$$

For the naïve model, the directional signal is defined as $s_t = 1$ if $y_{t-1} > y_{t-2}$ and $s_t = -1$ otherwise — a necessary adaptation so the model takes bidirectional positions, consistent with the other models.

The system uses no leverage and deliberately excludes transaction costs, as the objective is to assess the relative directional value of forecasts rather than present an operational trading strategy.

2.10 STATISTICAL EVALUATION FRAMEWORK

The six architectures were evaluated along three complementary dimensions: predictive accuracy (error metrics), statistical robustness (MCS procedure), and economic value (bootstrap for cumulative returns).

2.10.1 Performance Metrics

Forecast accuracy was quantified using three error metrics widely adopted in the time series forecasting literature, computed on the hold-out test set:

Mean Squared Error (MSE):

$$\mathrm{MSE} = 1/\mathrm{n} \sum_{i=1}^{n} (\mathrm{y_i} - \hat{y}_\mathrm{i})^2 \quad (6)$$

Mean Absolute Error (MAE):

$$\mathrm{MAE} = 1/\mathrm{n} \sum_{i=1}^{n} |y_i - \hat{y}_\mathrm{i}| \quad (7)$$

Mean Absolute Percentage Error (MAPE):

$$\mathrm{MAPE} = 100\%/\mathrm{n} \sum_{i=1}^{n} \left| \frac{(y_i - \hat{y}_\mathrm{i})}{y_i} \right| \quad (8)$$

where y_i is the observed value, $\hat{y}_\mathrm{i}$ is the model forecast, and n is the total number of observations in the test set.

2.10.2 Model Confidence Set (MCS)

Different models may exhibit statistically equivalent performance. To formally identify a set of models that cannot be distinguished from one another at a given confidence level, the Model Confidence Set (MCS) procedure was applied (Hansen; Lunde; Nason, 2011). This method has been widely adopted in time series model comparison studies, including research evaluating the statistical performance of neural networks for price forecasting (Zhu et al., 2024; Mendoza; Kristjanpoller; Minutolo, 2023), providing greater rigor than simple pointwise metric comparisons.

The MCS procedure starts with the full set of models and iteratively eliminates the worst performer until the Equal Predictive Ability (EPA) hypothesis can no longer be rejected for the remaining models (Liang et al., 2022). Elimination is based on a sequential hypothesis test using a loss function — MSE in this study. The procedure was conducted at 90% confidence ($\alpha = 0.10$), consistent with prior work (Wang; Wang, 2020). The final output is a set of models containing the best performer(s) with controlled probability, providing robust statistical evidence for performance comparison.

2.10.3 Paired Bootstrap Test for Cumulative Returns

To assess whether each model's cumulative return exceeds the buy-and-hold benchmark, a paired block bootstrap was applied — a method suited for series with temporal dependence and heteroskedasticity (Politis & Romano, 1994). This analysis complements the MSE/MCS evaluation, as the two approaches capture distinct performance dimensions and together provide a more complete picture of practical model utility.

Alongside the bootstrap test, each strategy reports the compounded cumulative return, the approximate Sharpe ratio — the ratio of mean daily return to its standard deviation — and the maximum drawdown, defined as the largest peak-to-trough decline, serving as a measure of strategy risk.

3 RESULTS AND ANALYSIS

This section presents the empirical results of the six-architecture evaluation.

3.1 MODEL PERFORMANCE

Table 7 reports error metrics for the full test period. Note that end-to-end models (5 and 6) were trained without extensive hyperparameter optimization, as discussed in Section 2.6, and their results are exploratory in nature.

Table 7. Performance of the 6 models across evaluation metrics and MCS p-values.

Model	MSE	MAE	MAPE	p-value
LSTM+LLM_Prob	0.0664	0.1948	0.86%	1.000
Naïve	0.0667	0.1922	0.85%	0.793
Pure LSTM	0.0737	0.2097	0.92%	0.080
LSTM+LLM_Tanh(E2E)	0.0773	0.2093	0.92%	0.080
LSTM+LLM_Tanh	0.0778	0.2142	0.95%	0.068
LSTM+LLM_Prob(E2E)	0.0858	0.2262	1.00%	0.000

Source: Elaborated by the authors (2026).

The LSTM+LLM_Prob architecture achieved the lowest MSE among all models, closely followed by the naïve benchmark. The statistical significance of these differences is assessed in the next section via the MCS procedure.

3.2 STATISTICAL SIGNIFICANCE

MCS p-values are reported in the last column of Table 7. The set of models with statistically equivalent predictive ability contains only two: the Naïve benchmark and LSTM+LLM_Prob. All others were eliminated, indicating statistically inferior performance.

Notably, the Pure LSTM — subjected to the same Bayesian optimization as LSTM+LLM_Prob — was excluded from the MCS, while the hybrid model was retained.

For context, the validation period exhibited higher average volatility than the test period (0.97% vs. 0.85%), indicating a market regime shift between the two sets. During validation, Pure LSTM and LSTM+LLM_Prob performed similarly; in the test set, Pure LSTM recorded an MSE of 0.0737 — substantially higher than both LSTM+LLM_Prob (0.0664) and the Naïve model (0.0667).

3.3 PERFORMANCE UNDER VOLATILE CONDITIONS

To evaluate model behavior under market stress, two analyses were conducted on test period subsets. The first selected the 208 days with absolute price variation above the mean (0.85%), representing higher-turbulence conditions. The second — designed to avoid bias favoring the leading model — selected the 208 days on which LSTM+LLM_Prob recorded its largest squared errors, i.e., its most adverse performance.

Table 8. MSE and p-values for the six models under two conditions, A and B.

	(A) High Volatility		(B)LSTM+LLM_Prob Worst Days	
Model	MSE	p-value	MSE	p-value
LSTM+LLM_Prob	0.1489	1.000	0.1501	1.000
Naïve	0.1513	0.617	0.1502	0.935
Pure LSTM	0.1557	0.416	0.1543	0.669
LSTM+LLM_Tanh	0.1604	0.325	0.1612	0.537
LSTM+LLM_Tanh (E2E)	0.1674	0.228	0.1673	0.215
LSTM+LLM_Prob(E2E)	0.1760	0.030	0.1762	0.035

Source: Elaborated by the authors (2026).

MCS results for the high-volatility subset are shown in Table 8, column group A. LSTM+LLM_Prob maintained a p-value of 1.000, while the Naïve model's p-value dropped to 0.617 — suggesting that the hybrid model's relative advantage amplifies precisely under the conditions where reliable forecasts are most critical for decision-making.

In the adverse subset — the leading model's own worst days — results are shown in Table 8, column group B. Even in this deliberately unfavorable scenario, LSTM+LLM_Prob maintained the lowest mean MSE, with the Naïve model presenting a p-value of 0.935. The remaining models

exhibited even larger errors, suggesting that on days when the hybrid model underperforms, competitors fail more markedly.

Although neither subset reaches robust statistical significance at 90%, results consistently point in the same direction: the hybrid model's advantage tends to widen under stress conditions, and its underperformance on worst days is proportionally smaller than that of competing models.

3.4 ECONOMIC EVALUATION AND DIRECTIONAL PERFORMANCE

Table 9. Economic metrics and p-values for the six models.

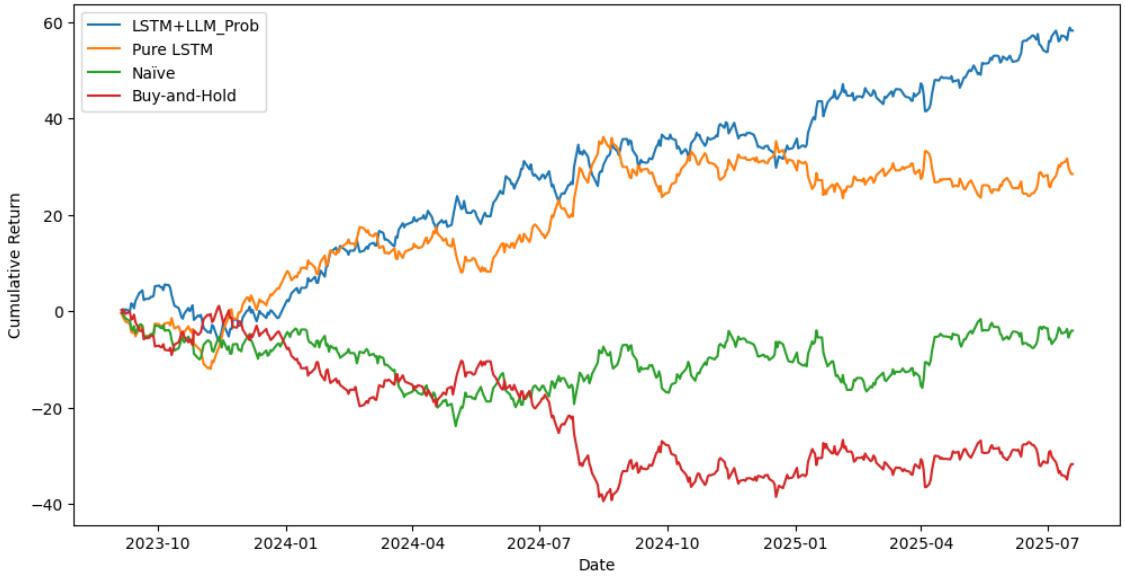
Model	Return	Sharpe	Max Drawdown	p-value	90% Conf. Int.
LSTM+LLM_Prob	58.27%	1.7368	-11.68%	0.003	(0.3819, 1.7032)
Pure LSTM	28.53%	0.5768	-20.35%	0.141	(-0.2794, 1.3102)
Naïve	-3.98%	-0.1839	-12.40%	0.2775	(-0.2140, 0.6831)
LSTM+LLM_Tanh	-2.25%	-0.1627	-12.86%	0.272	(-0.1826, 0.2163)
LSTM+LLM_Tanh(E2E)	-7.15%	-0.2756	-19.47%	0.3337	(-0.3648, 0.1838)
LSTM+LLM_Prob(E2E)	-4.13%	-0.2051	-14.87%	0.2996	(-0.2432, 0.1750)

Source: Elaborated by the authors (2026).

Table 9 summarizes, for each model, the compounded cumulative return, approximate Sharpe ratio, maximum drawdown, and empirical p-value from the paired bootstrap test against the buy-and-hold benchmark.

Over the 490-day test period, only LSTM+LLM_Prob generated a statistically significant positive cumulative return over the buy-and-hold benchmark (cumret ≈ 58.27%; $p \approx 0.003$ for block = 10; 90% bootstrap interval: [0.3819, 1.7032]). Pure LSTM achieved a positive return of 28.53%, but without statistical evidence of benchmark outperformance ($p > 0.14$). The cumulative returns of the two positively performing strategies, the naïve model, and the buy-and-hold are illustrated in Figure 4.

Figure 4. Cumulative returns of the positively performing strategies, the naïve model, and the market return.



Source: Elaborated by the authors (2026).

4 CONCLUSION

This study demonstrated that specializing an LLM in Brazilian agribusiness vocabulary, combined with probabilistic sentiment encoding, can bring an LSTM to the predictive level of the naïve benchmark in highly efficient markets. Among the six architectures compared, LSTM+LLM_Prob was the only one to join the MCS — achieving equivalence with the naïve model while Pure LSTM, subjected to the same optimization process, was excluded — and the only one to generate statistically significant cumulative excess return over buy-and-hold (cumret ≈ 58.27%; p ≈ 0.003). These results indicate that the predictive gain stems from the specific combination of a specialized LLM and probabilistic sentiment encoding, not from textual integration per se: scalar-output architectures and end-to-end models without extensive optimization did not replicate this performance.

This pattern was evident in the validation-to-test transition: although the two periods exhibited distinct average volatility (0.97% vs. 0.85%), with Pure LSTM and LSTM+LLM_Prob performing similarly in validation, only the hybrid model sustained that level in the test set — recording MSE of 0.0664 against 0.0737 for Pure LSTM. This suggests that specialized textual information acts as a contextualization mechanism, capturing signals associated with market condition changes not reflected in historical prices. This effect amplifies under critical conditions: on high-volatility days, the hybrid model's relative advantage over the naïve benchmark is larger; and even on its worst days, competing models show proportionally greater errors. For producers, traders, and policymakers, this translates into more reliable signals precisely when uncertainty is highest.

The main limitation is the inability to exhaustively optimize end-to-end models, whose prohibitive computational cost prevented Bayesian search at the same scale applied to other models.

This follows directly from treating all architectures under the same processing constraints: the per-iteration cost of end-to-end models precludes equivalent search. Their full potential remains unexplored and represents a natural direction for future work with greater computational resources or more efficient techniques such as LoRA or quantization. Future research may also explore: (i) applying the framework to other commodities such as corn, coffee, and cotton; (ii) replacing the LSTM with architectures such as Transformers or Mamba; and (iii) using the sentiment extractor as a diagnostic tool to quantify the price impact of specific textual events.

In summary, the hybrid pipeline developed and validated here establishes a solid methodological foundation for agricultural price forecasting systems grounded not only in historical series, but in contextual market understanding.

ACKNOWLEDGMENTS

This work was supported by UTFPR through the Institutional Scientific Initiation Scholarship Program (PIBIC).

REFERENCES

ALI, Z. et al. CMGM: A novel cross-market assets and multi-market modeling graph neural networks for financial market forecasting leveraging market states dependencies. **Alexandria Engineering Journal**, [S.l.], v. 128, p. 1101-1124, 2025.

BENGIO, Y.; SIMARD, P.; FRASCONI, P. Learning long-term dependencies with gradient descent is difficult. **IEEE Transactions on Neural Networks**, [S.l.], v. 5, n. 2, p. 157-166, mar. 1994.

BERGSTRA, J. et al. Algorithms for Hyper-Parameter Optimization. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, 24., 2011. **Anais**[...] [S.l.]: Curran Associates, Inc., 2011.

BRASIL. Instituto Brasileiro de Geografia e Estatística – IBGE. **IBGE prevê safra de 332,7 milhões de toneladas para 2026, queda de 3,7% frente a 2025**. Brasília, DF: IBGE, 13 nov. 2025. Disponível em: https://agenciadenoticias.ibge.gov.br/agencia-sala-de-imprensa/2013-agencia-de-noticias/releases/45124-ibge-preve-safra-de-332-7-milhoes-de-toneladas-para-2026-queda-de-3-7-frente-a-2025. Accessed: Dec. 14, 2025.

CHANDAN, G. Y.; KUMARI, P. Exogenous variable driven cotton prices prediction: comparison of statistical model with sequence based deep learning models. **Big Data Research**, [S.l.], v. 42, p. 100569, 2025.

FAN, C. et al. Assessment of deep recurrent neural network-based strategies for short-term building energy predictions. **Applied Energy**, [S.l.], v. 236, p. 700-710, 2019.

FAO – FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS. **FAOSTAT: Crops and livestock products**. Rome: FAO, 2024. Disponível em: https://www.fao.org/faostat/en/#data/QCL. Accessed: Dec. 14, 2025.

FARIMANI, S. A. et al. Investigating the informativeness of technical indicators and news sentiment in financial market price prediction. **Knowledge-Based Systems**, [S.l.], v. 247, p. 108742, 2022.

GILARDI, F.; ALIZADEH, M.; KUBLI, M. ChatGPT outperforms crowd workers for text-annotation tasks. **Proceedings of the National Academy of Sciences**, [S.l.], v. 120, n. 30, p. e2305016120, 2023.

HANSEN, P. R.; LUNDE, A.; NASON, J. M. The Model Confidence Set. **Econometrica**, [S.l.], v. 79, n. 2, p. 453-497, 2011.

HOCHREITER, S.; SCHMIDHUBER, J. Long Short-Term Memory. **Neural Computation**, [S.l.], v. 9, n. 8, p. 1735-1780, nov. 1997.

HUTTER, F.; KOTTHOFF, L.; VANSCHOREN, J. (Ed.). **Automated Machine Learning: Methods, Systems, Challenges**. Cham: Springer, 2019.

LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. **Biometrics**, [S.l.], v. 33, n. 1, p. 159-174, 1977.

LIANG, C. et al. Climate policy uncertainty and world renewable energy index volatility forecasting. **Technological Forecasting and Social Change**, [S.l.], v. 182, p. 121810, 2022.

LIAO, M. et al. Improving the model robustness of flood hazard mapping based on hyperparameter optimization of random forest. **Expert Systems with Applications**, [S.l.], v. 241, p. 122682, 2024.

MENDOZA, C.; KRISTJANPOLLER, W.; MINUTOLO, M. C. Market index price prediction using Deep Neural Networks with a Self-Similarity approach. **Applied Soft Computing**, [S.l.], v. 146, p. 110700, 2023.

MU, Z. et al. Exploring financial sentiment analysis via fine-tuning large language model and attributed graph neural network. **Neural Networks**, [S.l.], v. 199, p. 108620, 2026.

POLITIS, D. N.; ROMANO, J. P. The stationary bootstrap. **Journal of the American Statistical Association**, [S.l.], v. 89, n. 428, p. 1303-1313, 1994.

PUCHALSKY, W. et al. Agribusiness time series forecasting using Wavelet neural networks and metaheuristic optimization: An analysis of the soybean sack price and perishable products demand. **International Journal of Production Economics**, [S.l.], v. 203, p. 174-189, 2018.

RAIAAN, M. A. K. et al. A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges. **IEEE Access**, [S.l.], v. 12, p. 26839-26874, 2024.

RAY, S. et al. An ARIMA-LSTM model for predicting volatile agricultural price series with random forest technique. **Applied Soft Computing**, [S.l.], v. 149, p. 110939, 2023.

SCHUSTER, M.; PALIWAL, K. K. Bidirectional recurrent neural networks. **IEEE Transactions on Signal Processing**, [S.l.], v. 45, n. 11, p. 2673-2681, nov. 1997.

SONG, Y. et al. Multi-decomposition in deep learning models for futures price prediction. **Expert Systems with Applications**, [S.l.], v. 246, p. 123171, 2024.

WANG, B.; WANG, J. Deep multi-hybrid forecasting system with random EWT extraction and variational learning rate algorithm for crude oil futures. **Expert Systems with Applications**, [S.l.], v. 161, p. 113686, 2020.

WANG, K. et al. Short-term electricity price forecasting based on similarity day screening, two-layer decomposition technique and Bi-LSTM neural network. **Applied Soft Computing**, [S.l.], v. 136, p. 110018, 2023.

ZHANG, D. et al. Prediction of soybean price in China using QR-RBF neural network model. **Computers and Electronics in Agriculture**, [S.l.], v. 154, p. 10-17, 2018.

ZHANG, F.; XIA, Y. Carbon price prediction models based on online news information analytics. **Finance Research Letters**, [S.l.], v. 46, p. 102809, 2022.

ZHANG, M. et al. Convolutional Neural Networks-Based Lung Nodule Classification: A Surrogate-Assisted Evolutionary Algorithm for Hyperparameter Optimization. **IEEE Transactions on Evolutionary Computation**, [S.l.], v. 25, n. 5, p. 869-882, 2021.

ZHANG, Y.; DONG, Z.; XU, W. Integrative stock price trend prediction via hierarchical LLM text processing and patch-based transformer with co-attention. **Expert Systems with Applications**, [S.l.], v. 302, p. 130441, 2026.

ZHOU, N. et al. Enhancing photovoltaic power prediction using a CNN-LSTM-attention hybrid model with Bayesian hyperparameter optimization. **Global Energy Interconnection**, [S.l.], v. 7, n. 5, p. 667-681, 2024.

ZHU, M. et al. Energy price prediction based on decomposed price dynamics: A parallel neural network approach. **Applied Soft Computing**, [S.l.], v. 164, p. 111972, 2024.