Enhancing Brazilian Inflation Forecasts through Sentiment Analysis Using Large Language Models

Lucas Miranda Mendonça Rezende
Cézio Luiz Ferreira Junior
Mateus Tarcinalli Machado
Evandro Eduardo Seron Ruiz
Department of Computing and Mathematics
FFCLRP, University of São Paulo
Correspondence: evandro@usp.br

Abstract

Reliable inflation forecasts play a critical role in economic stability and policy decisions. Traditional econometric models perform well but often overlook qualitative signals that could improve predictive accuracy. Recent advances in AI-based Natural Language Processing enable the extraction of latent sentiment, offering a promising avenue for inflation forecasting. This study proposes a framework that combines Large Language Models (LLMs) to extract sentiment variables from the Brazilian Monetary Policy Committee (COPOM) minutes, optimize bias to match human-collected sentiment, and integrate them into ARIMA and LSTM models for one-step-ahead monthly IPCA prediction. Results show that LLM-generated sentiment trends are temporally coherent with historical inflation patterns and highly statistically significant ($p < 0.001$). Models whose sentiment evaluations aligned more closely with human assessments (*e.g.*, grok-4-fast and llama-4-maverick) achieved superior forecasting performance. ARIMA models consistently benefited from sentiment inclusion, while LSTM results were more variable.

1 Introduction

One of the central challenges in modern economics is generating reliable forecasts of macroeconomic variables that guide agents' decisions. Inflation, a key indicator of purchasing power and consumption trends, has therefore attracted extensive applied research. Time-series models have long been employed for this task in econometrics (*e.g.*, Katterbauer and Moschetta, 2022), and more recently, AI-based methods have been introduced and shown promising forecasting performance (*e.g.*, Ülke et al., 2018).

The analysis of public opinion on inflation-related matters has long been of critical importance, because specialized commentary shapes expectations, policy debates, and price dynamics (Istrefi and McMahon, 2025).

Forecasters often accompany numerical predictions with explanatory narratives. These accounts provide valuable context, clarify the underlying methodology, and describe justifying scenarios to ultimately enhance the forecast's credibility. Although qualitative, such narratives can be converted into quantitative data (Eugster and Uhl, 2024).

The general objective of this paper is to develop a proof of concept for a methodology that characterizes and extracts the sentiment bias of a given LLM on a given topic. As a concrete validation, this methodology is applied to Brazilian inflation forecasting, with three specific objectives: (i) to evaluate LLMs as sentiment detectors on COPOM (Central Bank of Brazil's Monetary Policy Committee) minutes and quantify the scorer bias each model introduces; (ii) to verify whether ARIMA and LSTM models achieve better predictive accuracy when fed bias-corrected, LLM-derived sentiment scores; and (iii) to determine whether these sentiment features improve inflation forecasting in a statistically significant way.

The results show that incorporating bias-corrected LLM sentiment scores can improve inflation forecasting accuracy over traditional models in the Brazilian context. The methodology is described in Section 3, results are presented in Section 4, and a brief literature survey follows in Section 2.

2 Related works

While machine learning (ML) offers high forecasting accuracy, its 'black-box' nature hinders the interpretability required for economic policymaking. To improve transparency, researchers are integrating ML with econometric methods using 'Sentometrics' – the sentiment analysis of textual data. By converting qualitative text from news and policy communications into quantitative sentiment in-

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 172–182

dices, Sentometrics captures subjective variables that influence inflation. These indices can be incorporated into traditional models, enhancing predictive power without sacrificing interpretability.

The foundational link between textual sentiment and economic forecasting was established in early literature (Tetlock, 2007). Tetlock constructs media-based sentiment measures from financial news and demonstrates their predictive power for cross-sectional stock returns, particularly during periods of extreme sentiment. The study shows that media tone conveys information beyond fundamentals and that sentiment's impact varies with news volume and investor attention.

Using similar methods, Ludvigson, 2004 examine whether consumer sentiment surveys add independent, actionable information about future spending beyond standard economic and financial indicators. Kräussl and Mirgorodskaya, 2017 study show how optimistic versus pessimistic news narratives affect investor perceptions, behavior, and long-term market performance. Casey and Owen, 2013 and Mehtab and Sen, 2019 examine expert market commentary, with Mehtab specifically analyzing technical texts from Twitter (now X).

Algaba et al., 2020 introduced a key methodological framework covering three dimensions: sentiment extraction, aggregation, and modeling, synthesizing approaches for converting qualitative data into sentiment indicators. Recent studies have shown that although LLMs show a basic sensitivity to sentiment, there are substantial variations in their accuracy and consistency (Liu et al., 2025). For this reason, we applied a variety of LLMs to assess sentiments.

COPOM minutes are a rich data source: they signal policymakers' uncertainty and confidence levels. Such *qualitative narratives* can enrich quantitative inflation forecasts by revealing tones of optimism, caution, or pessimism. Clements and Reade, 2020, analyzing the Bank of England's inflation reports, highlight three key insights:

1. Narrative texts frequently contain forward-looking information not captured by numerical forecasts alone.
2. Certain forecast errors can be more effectively interpreted by examining the tone of the accompanying narrative.
3. Narratives serve as a strategic communication tool for central banks, particularly useful for conveying uncertainty and policy rationale during periods of heightened volatility.

Incorporating sentiment analysis into econometric forecasting enhances both predictive accuracy and interpretability. Qualitative narratives from the Brazilian Central Bank communications complement traditional numerical data, supporting more informed policy decisions.

3 Methodology

3.1 Mathematical Overview

3.1.1 ARIMA Model

The ARIMA model that incorporates an exogenous variable is commonly referred to as ARIMAX. In this article, we will use the terms ARIMA and ARIMAX interchangeably[1].

Let B be the backshift operator ($By_t = y_{t-1}$) and $\Delta = (1 - B)$. The ARIMA(p, d, q) model incorporating sentiment as an exogenous regressor is given by:

$$\phi(B)\Delta^d y_t = c + \beta(B)s_{t-k} + \theta(B)\varepsilon_t, \quad (1)$$

where $\phi(B) = 1 - \sum_{i=1}^{p} \phi_i B^i$, $\theta(B) = 1 + \sum_{j=1}^{q} \theta_j B^j$, and $\beta(B) = \sum_{\ell=0}^{L} \beta_\ell B^\ell$. Residuals satisfy $\varepsilon_t \sim$ i.i.d. $(0, \sigma_\varepsilon^2)$. Parameters $\Theta_{\text{ARIMA}} = \{c, \phi_i, \theta_j, \beta_\ell\}$ are estimated by conditional least squares or maximum likelihood:

$$\mathcal{L}_{\text{ARIMA}} = \sum_t \varepsilon_t^2. \quad (2)$$

The one-step-ahead forecast is obtained as

$$\hat{y}_{t+1|t}^{\text{ARIMA}} = \phi(B)^{-1}\big(c + \beta(B)s_{t+1-k}\big). \quad (3)$$

3.1.2 LSTM Model

To capture nonlinear dynamics between inflation and sentiment, we also train a recurrent neural network with Long Short-Term Memory (LSTM) units. Given a lookback window of length L, the multivariate input sequence is

$$\mathbf{x}_u = \begin{bmatrix} y_u \\ s_u \end{bmatrix} \in \mathbb{R}^2. \quad (4)$$

At each internal time step u, the LSTM evolves as:

$$f_u = \sigma(W_f \mathbf{x}_u + U_f h_{u-1} + b_f), \quad (5)$$
$$i_u = \sigma(W_i \mathbf{x}_u + U_i h_{u-1} + b_i), \quad (6)$$
$$\tilde{c}_u = \tanh(W_c \mathbf{x}_u + U_c h_{u-1} + b_c), \quad (7)$$
$$c_u = f_u \odot c_{u-1} + i_u \odot \tilde{c}_u, \quad (8)$$
$$o_u = \sigma(W_o \mathbf{x}_u + U_o h_{u-1} + b_o), \quad (9)$$
$$h_u = o_u \odot \tanh(c_u). \quad (10)$$

[1]The order of the ARIMA(1,1,1) model was obtained through Partial Autocorrelation Function (PACF) analysis.

The one-step prediction is produced from the last hidden state:

$$\hat{y}^{\text{LSTM}}_{t+1|t} = w_{\text{out}}^{\top} h_t + b_{\text{out}}. \tag{11}$$

Parameters Θ_{LSTM} are optimized via backpropagation through time, minimizing the mean squared error:

$$\mathcal{J}_{\text{LSTM}} = \frac{1}{N}\sum_{t=1}^{N}(y_{t+1} - \hat{y}_{t+1|t})^2 + \lambda\|\Theta_{\text{LSTM}}\|^2. \tag{12}$$

3.1.3 Forecast Comparison

Forecasts from both models are compared under an identical rolling-origin evaluation scheme. Predictive performance is assessed using the Root Mean Square Error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(y_{t+1} - \hat{y}_{t+1|t})^2}. \tag{13}$$

The comparison of $\text{RMSE}_{\text{ARIMAX}}$ and $\text{RMSE}_{\text{LSTM}}$ reveals whether sentiment extracted from monetary policy communications enhances inflation forecasting in linear versus nonlinear frameworks.

3.2 Creating the Phrase Dataset

Scraping

We collected COPOM minutes from the official listing (Banco Central do Brasil, 2025a), downloading both HTML and PDF versions when available.

The dataset C contains 251 COPOM minutes from January 1996 to July 2025. Each minute c has an associated date d_m and may be in HTML and/or PDF format.

Parsing

For each COPOM minute c in C:

1. Type-Specific Pre-Processing

   HTML: extract body content, remove formatting tags (*e.g.,*`strong`, `i`, `br`) while preserving inner content. Phrases are then extracted using periods as sentence boundaries.

   PDF: extract phrases using SpaCyLayout (Explosion AI, 2024) with `pt_core_news_lg` model.

   We create phrase lists P_c^{html} and P_c^{pdf}, each containing phrases from respective versions.

2. General Pre-Processing

   For each phrase in P_c^{html} and P_c^{pdf}: (i) Remove newlines and tabs; (ii) Remove tag entities (*e.g.*, ` `); (iii) Reduce consecutive spaces, commas, periods to single characters; (iv) Add period at end if missing. Pre-processing was mainly needed for text in HTML format. Stopwords were not removed, nor was stemming or even lemmatization applied, as such transformations are known to negatively affect the performance of pre-trained language models (Haviana et al., 2023).

3. Length Filtering

   Iterating P_c^{html} and P_c^{pdf}: Discard single-word phrases and phrases with character count below μ, the mean character count from the respective source P_c^{x}. This empirically determined threshold removes very short or poorly-formed fragments that carry insufficient context for reliable sentiment assessment.

4. Blacklist Filtering

   Iterating P_c^{html} and P_c^{pdf}: Remove phrases containing: (v) JavaScript; (vi) *cookies*; (vii) *expand_less*; (viii) *content_copy*; or (ix) *Garantir a estabilidade do poder de compra da moeda*.

   While terms (v) to (viii) are related to web page elements and scripts, term (ix) is the Brazilian Central Bank's motto, which often appears in the minutes and is irrelevant for sentiment analysis.

Finally, we select the version with more phrases (PDF if equal, given its superior quality), producing F_{d_m} for each date d_m. F is the union of all F_{d_m}.

Phrase Selection

F is flattened into a list L of tuples (phrase, date).

Dense passage retrieval is performed using semantic similarity filtering. Embeddings are computed with Qwen3-Embedding-0.6B (Zhang et al., 2025), retaining phrases with cosine similarity above 0.6 to "inflation".

The final dataset F^{infl} contains 9,378 inflation-related phrases across 251 dates (≈37.4 phrases per date).

3.3 Creating the Sentiment Datasets

We evaluated phrase sentiment using nine LLMs from six different companies, listed in Table 1.

Model	Token Limit
gpt-5	1,024
gpt-oss-120b	512
gemini-2.5-pro	128
gemma-3-27b-it	8
deepseek-chat-v3.1	4
claude-sonnet-4	1
grok-4-fast	1
llama-4-maverick	1
phi-4	1

Table 1: Selected LLMs and their token limits.

Token limits are parameters passed to the LLM's API that determine the maximum number of tokens the response can contain, including reasoning capabilities when available. We determine token limits by testing on the first date's phrases; if any receives -2, we double the limit and repeat testing until all responses are successful. Table 1 also shows the final token limits used.

For each model, we submitted one independent, context-free request per phrase in F^{infl}.

The prompt, originally formulated in Brazilian Portuguese, explained the task and appended the phrase:

> **Optimism definition:** It occurs when projections indicate that inflation will remain below the target or comfortably within the tolerance range. An optimistic phrase may signal that the Central Bank sees room to lower interest rates or maintain a more accommodative monetary policy.
>
> **Pessimism definition:** It occurs when projections indicate inflation above the target or near the upper limit of the tolerance range. A pessimistic phrase suggests concern about inflationary pressures and may justify a more restrictive monetary policy.
>
> **Evaluate the sentence as:** Letter 'O' for optimistic, 'N' for neutral, or 'P' for pessimistic. Your answer must be only one letter, without any other additional text.

Models classify phrases as optimistic, neutral, or pessimistic, converting responses (O, N, P) to 1, 0, and -1, and assign -2 to unparseable responses, which rarely occur.

Evaluations not equal to 1 or -1 were discarded, as neutral sentiment is not expected in contexts of clear inflationary trends, is difficult to characterize consistently, and often reflects contradictory signals rather than a genuinely neutral stance. The results were concatenated into sets E_M for each model, containing tuples (phrase, date, sentiment). The set E_{Models} contains all E_M.

3.4 Human Evaluation Dataset

We created three human evaluation datasets:

1. *Open evaluation*

   We created a website for users to mark O/N/P for a group of randomly selected phrases from F^{infl}, limited to 10 phrases per browser for every 24h period. Distributed to economics graduate students at USP and Unicamp. Result: E_{open} with 278 tuples.

2. *Specialist evaluation*

   A subset $F^{\text{infl-350}}$ of 350 random phrases from F^{infl}, with date labels Base64-encoded to prevent bias. They were labeled as: 1 (optimistic), 0 (neutral), -1 (pessimistic). Result: $E_{\text{specialist}}$ with 350 tuples.

3. *Consolidated evaluation*

   $F^{\text{infl-350}}$ re-analyzed by the specialist and two additional professors together, discussing each phrase to reach consensus. Result: $E_{consolidated}$ with 220 tuples.

Evaluations not equal to 1 or -1 were discarded for the same reasons stated above. Set E_{Humans} contains all E_h.

3.5 Testing Inflation Prediction Performance

We test two inflation prediction models: ARIMAX and LSTM. See Equations 3 and 12, respectively.

The goal is to determine whether LLM-derived sentiment reduces RMSE over a data-only baseline, and whether human-based bias correction yields further gains.

Creating the Input Datasets

For each set of the power set of E_{Models}, except for the empty one, we will concatenate the tuples of the selected E_M sets into a single set named U_i.

For each U_i created, we will create j more tuples in the form (U_i, V_j), where V_j is one of the three human evaluation datasets in E_{Humans}.

For each tuple (U_i, V_j) created, we will create k more tuples in the form (U_i, V_j, eq_k), where eq_k is one of the equations to be used for bias correction later.

The tuple (U_i, V_j, eq_k) represents the sentiment evaluations from the selected LLM models combined with the human evaluation dataset V_j for bias correction using equation eq_k.

The possible equation forms for eq_k are: the linear form ($x + a$), the affine form ($bx + a$), a quadratic form ($cx^2 + bx + a$), and a cubic spline form ($dx^3 + cx^2 + bx + a$).

For each tuple (U_i, V_j, eq_k), we create three input datasets, each providing tuples of the form (Inflation, Sentiment):

1. *Inflation data only* (*Baseline*)

   IPCA monthly (Series 433) provided by the Banco Central do Brasil, 2025b. The exogenous sentiment variable is set to 0 for associated inflation values.

2. *Inflation + Sentiment (Uncorrected)*

   Combine the IPCA monthly (Series 433) with the sentiment variable, which is created as an average grade per date of the evaluations in U_i, interpolated by cubic spline and fitted to the available IPCA dates.

3. *Inflation + Sentiment (Corrected)*

   Combine the IPCA monthly (Series 433) with the sentiment variable, created as an average grade per date of the evaluations in U_i, which is interpolated by cubic spline, fitted to the available IPCA dates, and corrected based on the bias measured from V_j.

   The correction process works as follows:

   First, we average the LLM sentiment scores from U_i and the human evaluations from V_j by date, then interpolate them using a cubic spline to create a continuous daily time series.

   Next, we identify a single set of parameters for the transformation equation eq_k that minimizes the mean squared error (MSE) when applied individually to each date.

   We apply the equation for each date, using the variable x to represent the average LLM sentiment score on that date, and the resulting value represents the bias-corrected sentiment score.

   We optimize using gradient descent with the Adam optimizer (Kingma and Ba, 2014) for 1,000 epochs and a learning rate of 0.01, implemented in PyTorch.

   Finally, we apply these optimized parameters to the equation to transform the LLM sentiment score for each date in U_i, producing bias-corrected values that align with human judgment from V_j.

Finally, for each tuple (U_i, V_j, eq_k) created, we have three new associated lists of tuples in the form of (Inflation, Sentiment). Each list is called IN_{ijkl} where i is the LLM model combination used; j is the human evaluation dataset used for bias correction; k is the equation type used for bias correction; and $l \in \{$*Baseline*, *Uncorrected*, or *Corrected.*$\}$

The set IN contains all sets IN_{ijkl}.

Running the Tests

Looking at the IN set, we see that this approach involves repetition of IN_{ijkl} datasets since, for example, *Baseline* is the same for all tuples (U_i, V_j, eq_k).

While computationally inefficient, this provides a control for every experiment: *Baseline* for both *Uncorrected* and *Corrected*, and *Uncorrected* for *Corrected*.

For each IN_{ijkl} in IN, we run both ARIMAX and LSTM (Hochreiter and Schmidhuber, 1997) inflation prediction models on the respective dataset with a 70/30 train/test split.

ARIMAX incorporates sentiment as an exogenous variable (Moslemi et al., 2024) and is fitted under a walk-forward validation scheme. The LSTM comprises 5,000 neurons and is trained using the NAdam optimizer (Dozat, 2016) with a learning rate of 0.001, for up to 10,000 epochs and early stopping at patience 10. Predictive accuracy for both models is measured via Root Mean Squared Error (RMSE); further implementation details are available in the repository referenced in the Code Availability section. The choice of a heavily parameterized LSTM is motivated by recent findings on double descent (Schaeffer et al., 2023), which indicate that generalization can improve rather than degrade in the overparameterized regime.

In total, we conducted 36,792 tests, here explained: $(2^9 - 1)$ LLM combinations $\times$ 3 human datasets $\times$ 4 equation types $\times$ 3 dataset types $\times$ 2 models.

Testing the Statistical Significance: Student's T-test and Diebold-Mariano test

To assess statistical significance, we perform a one-sample T-test on the percentage RMSE improvements relative to Baseline, testing whether the mean differs significantly from zero for each model and correction type.

For the ARIMA and LSTM results produced by the IN_{ijkl} instances, we grouped by the l index value (*Uncorrected* or *Corrected*), excluding *Baseline*, since this set serves as the reference point. For each group, we calculated the percentage improvement by comparing the RMSE of each IN_{ijkl} against its corresponding baseline:

$$\text{Improvement}_{\%} = \frac{\text{RMSE}_{\text{baseline}} - \text{RMSE}_{ijkl}}{\text{RMSE}_{\text{baseline}}} \times 100$$

.

This yields four groups: LSTM-Uncorrected, LSTM-Corrected, ARIMA-Uncorrected, and ARIMA-Corrected. Each group aggregates results across all LLM combinations (i), human evaluation datasets (j), and equation types (k), resulting in $n = 6{,}132$ observations per group (*i.e.*, 511 LLM combinations $\times$ 3 human datasets $\times$ 4 equation types).

The null hypothesis $H_0 : \mu = 0$ states that sentiment inclusion provides no average improvement, tested against $\text{H}_1 : \mu \neq 0$. The T-statistic is:

$$t = \frac{\bar{x}}{s/\sqrt{n}};$$

where $\bar{x}$ is the mean improvement percentage across all experimental runs, s is the sample standard deviation, and n is the sample size. The p-value indicates the probability of observing such improvements if sentiment truly had no effect. We consider $p < 0.001$ as highly significant, $p < 0.01$ as very significant, and $p < 0.05$ as significant.

To complement the T-test analysis at the run level, we applied the Diebold-Mariano (DM) test (Diebold and Mariano, 1995) to compare per-timestep forecast accuracy between the baseline and sentiment-enhanced models. For each experimental run, the loss differential at time t is defined as

$$d_t = e^2_{\text{baseline},t} - e^2_{\text{model},t},$$

where e_t^2 denotes the squared forecast error at step t. A positive d_t indicates that the sentiment model achieves a lower squared error at that step.

The DM statistic tests whether the mean loss differential $\bar{d}$ is significantly different from zero:

$$\text{DM} = \frac{\bar{d}}{\widehat{\sigma}_{\bar{d}}},$$

where $\widehat{\sigma}_{\bar{d}}$ is a heteroskedasticity-and-autocorrelation-consistent (HAC) estimate of the standard error of $\bar{d}$. Under the null hypothesis $H_0 : \mathbb{E}[d_t] = 0$ (equal predictive accuracy), the statistic is asymptotically standard normal. A positive DM statistic indicates the sentiment-enhanced model is more accurate on average at the per-timestep level.

We apply the DM test to the same four groups (ARIMA-Uncorrected, ARIMA-Corrected, LSTM-Uncorrected, LSTM-Corrected) across all $n = 6{,}132$ experimental runs, reporting the mean DM statistic, median p-value, and the percentage of runs where the test is individually significant ($p < 0.05$) and where DM>0.[2]

4 Results

Figure 1 shows that, despite some variability, all LLMs follow a similar sentiment trend with peaks and valleys at the same dates. Even **deepseek-chat-v3.1**, which averages substantially lower, follows the same trend. Language models can capture sentiment dynamics, but their outputs should be treated with caution given potential biases; bias correction may improve downstream forecasting performance.

While inflation appears stable, its correlation with sentiment is inconsistent: it is inverse in 2002, direct in 2022, and absent in 2008. This decoupling suggests the Central Bank is successfully using monetary policy to maintain its inflation targets.

Figures 2, 3, and Table 2 show average grades and confidence intervals per dataset.

Significant variation exists across models, with all averages slightly negative (including the human-evaluated ones), suggesting a general pessimistic bias in the COPOM minutes.

Grok-4-fast and **llama-4-maverick** had bias closest to the human averages, while **deepseek-chat-v3.1** was the furthest and most pessimistic by a large margin.

The human confidence intervals are wider due to the smaller sample; overall, human evaluations

[2]Residual adequacy checks, (Ljung and Box, 1978) and (Jarque and Bera, 1980), were also performed to assess the statistical properties of the forecasting residuals. The numerical results are available in the repository referenced in the Code Availability section.

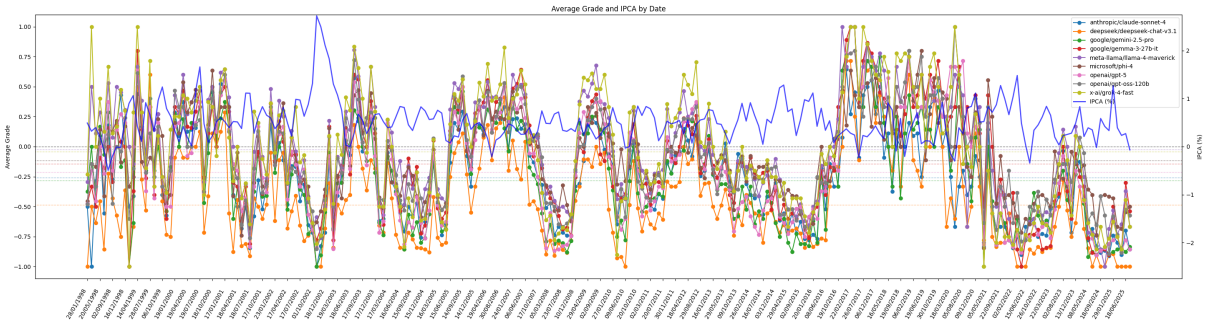

Figure 1: Average LLM sentiment grade by date and model (with IPCA inflation).

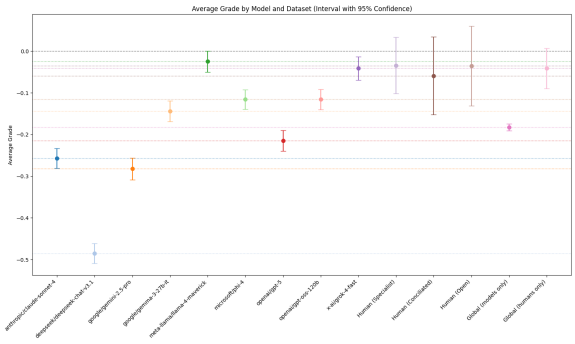

Figure 2: Average grade and confidence intervals by dataset at 95% confidence level.

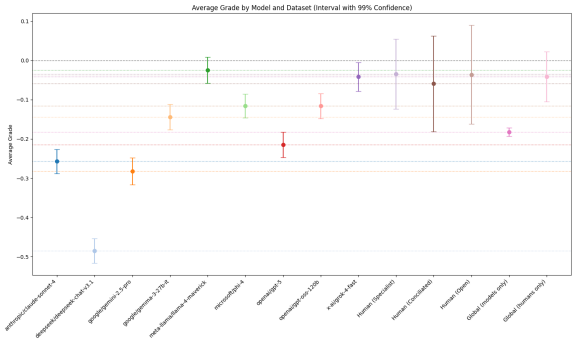

Figure 3: Average grade and confidence intervals by dataset at 99% confidence level.

Dataset	Average
Global	
Models only	−0.1826
Humans only	−0.0413
By LLM	
claude-sonnet-4	−0.2572
deepseek-chat-v3.1	−0.4851
gemini-2.5-pro	−0.2823
gemma-3-27b-it	−0.1442
llama-4-maverick	−0.0248
phi-4	−0.1158
gpt-5	−0.2146
gpt-oss-120b	−0.1160
grok-4-fast	−0.0415
By Humans	
Specialist	−0.0343
Conciliated	−0.0591
Open	−0.0360

Table 2: Average sentiment grades.

Model	Uncorrected	Corrected
LSTM	0.1581%	0.2534%
ARIMA	1.2209%	0.7403%

Table 3: RMSE reduction by model.

are more optimistic than LLM ones.

Figure 4 compares the best configurations across the six setups. Sentiment grades generally improve baseline predictions, consistently enhancing ARIMA models, though LSTM results remain mixed. Notably, **grok-4-fast** and **llama-4-maverick**, the models most aligned with human sentiment, frequently perform best, suggesting that human-like sentiment bias improves inflation forecasting.

Table 3 shows RMSE reductions from adding sentiment (*Corrected* and *Uncorrected*) against the *Baseline* across all 36,792 tests. All configurations improve, with ARIMA benefiting most.

While ARIMA models showed a reduction in predictive performance when using corrected sentiment grades, LSTM models showed an improvement.

Table 4 shows that all four configurations yield $p < 0.001$, but the strength of evidence varies: ARIMA models produce T-statistics that dwarf the LSTM counterparts (268.0 and 76.3 versus 8.4 and 14.1). The effect of correction is also

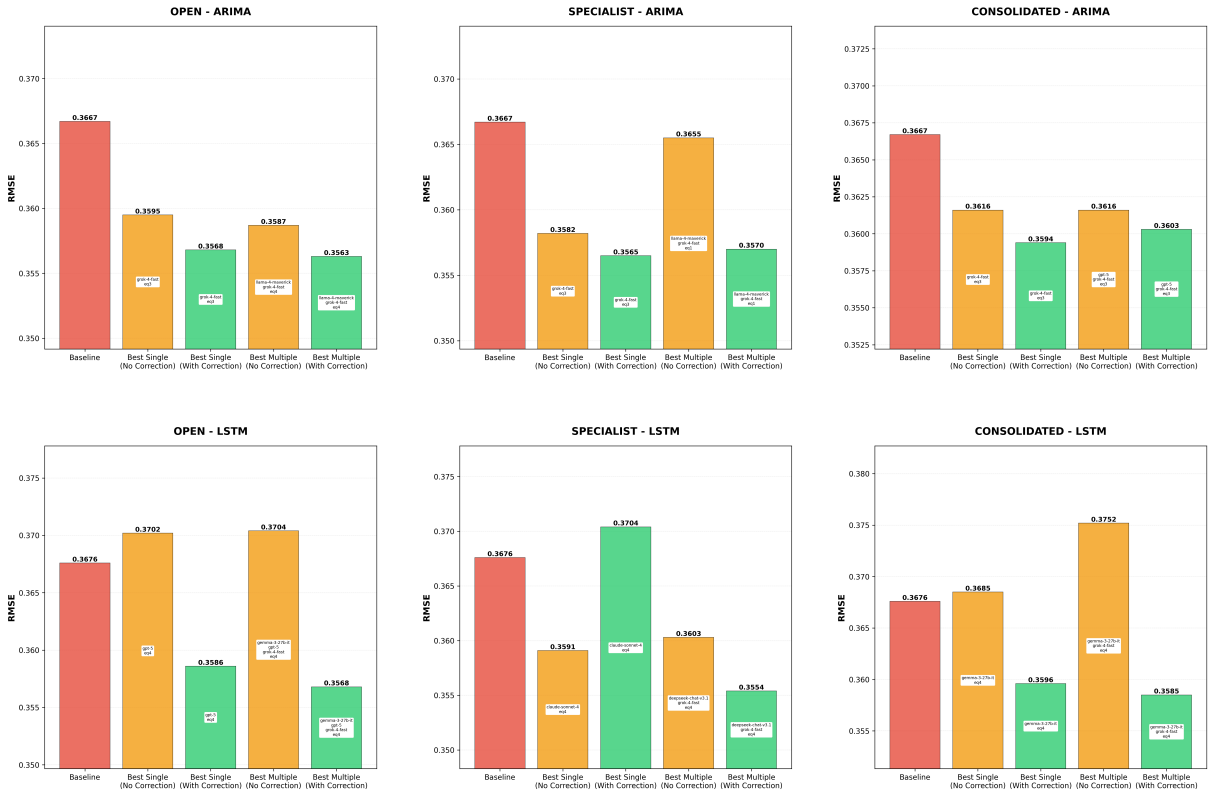

Figure 4: Best models: RMSE comparison across six different configurations.

Model	Correction	Mean (%)	Std (%)	t-statistic	p-value	Significant
LSTM	Uncorrected	0.1581	1.47	8.4154	4.82×10^{-17}	Yes ($p < 0.001$)
LSTM	Corrected	0.2534	1.40	14.1432	1.03×10^{-44}	Yes ($p < 0.001$)
ARIMA	Uncorrected	1.2209	0.36	268.0091	≈ 0	Yes ($p < 0.001$)
ARIMA	Corrected	0.7403	0.76	76.3413	≈ 0	Yes ($p < 0.001$)

Table 4: Student's t-test results. Higher t-statistics indicate stronger evidence of improvement over the *Baseline*.

opposite between architectures: for ARIMA, the uncorrected variant yields the higher T-statistic, while for LSTM the corrected variant outperforms. ARIMA-Uncorrected has the tightest spread (0.36%), against LSTM variants both exceeding 1.40%.

Table 5 shows the Diebold-Mariano test results, which assess predictive superiority at the individual forecast level. All configurations yield a positive mean DM statistic, but the gap between architectures is large: ARIMA-Uncorrected achieves DM>0 in 99.9% of runs with a mean of 0.7582, while LSTM-Uncorrected barely exceeds chance at 50.9% with a mean of 0.1571, nearly 5× smaller. As with the T-test, correction has opposite effects on each architecture: it reduces both the mean DM statistic and the DM>0 rate for ARIMA, while increasing both for LSTM.

Individual run significance ($p < 0.05$) is effectively absent in three of the four configurations (0.0%), which is expected given that each DM test operates on a single time series with limited statistical power. ARIMA-Corrected is the sole exception at 0.9%, a counterintuitive result given its lower mean DM statistic and lower DM>0 rate relative to ARIMA uncorrected. A plausible explanation is that correction introduces higher per-run variance: rather than consistently producing small positive values, the corrected variant yields a wider spread of outcomes, pushing a small fraction of runs past the significance threshold while also increasing the share of runs where the baseline wins.

Limitations Two aspects of the sentiment feature construction involve an indirect look-ahead. The cubic spline is fit over all COPOM meeting dates simultaneously, so interpolated values depend on future meetings, including those in the test period.

Model	Correction	Mean DM-stat	Median p-val	% Sig. ($p<0.05$)	% DM > 0
ARIMA	Uncorrected	0.7582	0.4482	0.0%	99.9%
ARIMA	Corrected	0.3905	0.4768	0.9%	76.6%
LSTM	Uncorrected	0.1571	0.5156	0.0%	50.9%
LSTM	Corrected	0.2037	0.5751	0.0%	61.6%

Table 5: Diebold-Mariano test results. DM>0 indicates the sentiment model achieves lower per-timestep squared errors than the *Baseline*. % Sig. reports the fraction of runs where this advantage is individually significant.

Similarly, the bias-correction parameters are estimated over all available dates before the 70/30 split is applied. Both effects are likely subtle, since the sentiment documents themselves predate each IPCA release.

The use of LLMs as sentiment instruments also introduces several sources of variability. Outputs can differ across runs (Liu et al., 2025), across models with different pre-training corpora (Alfiansyah et al., 2025), and across small prompt changes (Zhang and Han, 2025; Zhan et al., 2024). Any of these factors can shift the estimated bias profile.

5 Conclusion

The inclusion of sentiment analysis in the selected inflation forecasting models has demonstrated a measurable, if modest, improvement in predictive accuracy. Averaging results across all LLM combinations, human datasets, and correction equations, the improvements persist, suggesting they can be attributed to the sentiment integration method itself rather than to any particular configuration choice.

The T-test results, presented in Table 4, confirm that the mean RMSE improvement over the *Baseline* is statistically significant across all four configurations ($p < 0.001$). The strength of this evidence differs substantially by architecture, however: ARIMA models yield T-statistics orders of magnitude larger than their LSTM counterparts, and the effect of correction runs in opposite directions (ARIMA is weaker when *Corrected*, while LSTM is stronger), suggesting the two architectures interact with the sentiment signal in different ways.

The Diebold-Mariano results (see Table 5) reinforce this at the individual forecast level. For ARIMA-Uncorrected, the sentiment model outperforms the *Baseline* in virtually every single run (99.9%, DM>0), confirming that the aggregate advantage is consistent rather than driven by a few large wins. For LSTM, however, the uncorrected variant barely exceeds chance (50.9%), meaning the per-forecast advantage is unreliable despite being real in aggregate. Individual run significance ($p < 0.05$) is effectively absent in three of the four configurations (0.0%), which is expected given the limited statistical power of a single time series; ARIMA-Corrected is the sole exception at 0.9%, likely due to the higher per-run variance introduced by correction, as discussed in Section 4.

The two tests are consistent: ARIMA's improvement is large, consistent, and robust, while LSTM's is small and detectable only in aggregate. The unusually small p-values in the T-test warrant caution in interpreting practical significance.

Models whose evaluations were closer to human assessments also tended to perform better, underscoring the value of human alignment.

Future Work Eliminating the indirect look-ahead in the sentiment pipeline is a natural next step, either by replacing the global cubic spline with a causal interpolation procedure or by estimating bias-correction parameters using only the training period. A second direction is to isolate the sources of LLM variability by measuring the separate effects of reruns, model pre-training differences, and prompt design on the resulting sentiment bias profile.

Code Availability The code used for these experiments is available at this repository.

Acknowledgement This work was carried out at the Center for Artificial Intelligence of the University of São Paulo (C4AI – http://c4ai.inova.usp.br/), with support by the São Paulo Research Foundation (FAPESP Grant #2019/07665-4) and by the IBM Corporation.

References

Fahri Alfiansyah, Mahendra Dwifebri Purbolaksono, and Alfian Akbar Gozali. 2025. Llm-based sentiment analysis in private social media: Case study of my tel-u app. In *2025 International Conference on Data Science and Its Applications (ICoDSA)*, pages 1100–1105. IEEE.

Andres Algaba, David Ardia, Keven Bluteau, Samuel Borms, and Kris Boudt. 2020. Econometrics meets sentiment: An overview of methodology and applications. *Journal of Economic Surveys*, 34(3):512–547.

Banco Central do Brasil. 2025a. Atas do Comitê de Política Monetária – Copom. https://www.bcb.gov.br/publicacoes/atascopom/cronologicos.

Banco Central do Brasil. 2025b. IPCA Monthly Series 433 – API. https://api.bcb.gov.br/dados/serie/bcdata.sgs.433/dados?formato=json.

Gregory P. Casey and Ann L. Owen. 2013. Good news, bad news, and consumer confidence. *Social Science Quarterly*, 94(1):292–315.

Michael P. Clements and J. James Reade. 2020. Forecasting and forecast narratives: The Bank of England inflation reports. *International Journal of Forecasting*, 36(4):1488–1500.

Francis X. Diebold and Roberto S. Mariano. 1995. Comparing Predictive Accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263.

Timothy Dozat. 2016. Incorporating Nesterov Momentum into Adam. Technical Report, Stanford University. ICLR 2016 Workshop track.

Patrick Eugster and Matthias W. Uhl. 2024. Forecasting inflation using sentiment. *Economics Letters*, 236:111575.

Explosion AI. 2024. spaCy-layout: Process PDFs, Word documents and other files with spaCy. GitHub repository. Accessed: 2026-03-10.

Sam Farisa Chaerul Haviana, Sri Mulyono, and Badie'Ah. 2023. The effects of stopwords, stemming, and lemmatization on pre-trained language models for text classification: A technical study. In *2023 10th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, pages 521–527. IEEE.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Klodiana Istrefi and Michael McMahon. 2025. Communication and inflation expectations. In *Research Handbook on Inflation*, pages 358–376. Edward Elgar Publishing.

Carlos M Jarque and Anil K Bera. 1980. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics letters*, 6(3):255–259.

Klemens Katterbauer and Philippe Moschetta. 2022. An innovative Artificial Intelligence and Natural Language Processing framework for asset price forecasting based on islamic finance: A case study of the Saudi stock market. *Econometric Research in Finance*, 6(2):183–196.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*. Published at ICLR 2015.

Roman Kräussl and Elizaveta Mirgorodskaya. 2017. Media, sentiment and market performance in the long run. *The European Journal of Finance*, 23(11):1059–1082.

Yang Liu, Xichou Zhu, Zhou Shen, Yi Liu, Min Li, Yujun Chen, Benzi John, Zhenzhen Ma, Tao Hu, Zhi Li, Zhiyang Xu, Wei Luo, and Junhui Wang. 2025. Do large language models possess sensitivity to sentiment? *Preprint*, arXiv:2409.02370.

Greta M Ljung and George EP Box. 1978. On a measure of lack of fit in time series models. *Biometrika*, 65(2):297–303.

Sydney C. Ludvigson. 2004. Consumer confidence and consumer spending. *Journal of Economic Perspectives*, 18(2):29–50.

Sidra Mehtab and Jaydip Sen. 2019. A robust predictive model for stock price prediction using deep learning and natural language processing. *arXiv preprint arXiv:1912.07700*.

Zahra Moslemi, Logan Clark, Sarah Kernal, Samantha Rehome, Scott Sprengel, Ahoora Tamizifar, Shawna Tuli, Vish Chokshi, Mo Nomeli, Ella Liang, Moury Bidgoli, Jeff Lu, Manish Dasaur, and Marty Hodgett. 2024. Comprehensive Forecasting of California's Energy Consumption: A Multi-Source and Sectoral Analysis Using ARIMA and ARIMAX Models. *arXiv preprint arXiv:2402.04432*.

Rylan Schaeffer, Mikail Khona, Zachary Robertson, Akhilan Boopathy, Kateryna Pistunova, Jason W. Rocks, Ila Rani Fiete, and Oluwasanmi Koyejo. 2023. Double Descent Demystified: Identifying, Interpreting & Ablating the Sources of a Deep Learning Puzzle. *arXiv preprint arXiv:2303.14151*.

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

Volkan Ülke, Afsin Sahin, and Abdulhamit Subasi. 2018. A comparison of time series and machine learning models for inflation forecasting: empirical evidence from the USA. *Neural Computing and Applications*, 30:1519–1527.

Tong Zhan, Chenxi Shi, Yadong Shi, Huixiang Li, and Yiyu Lin. 2024. Optimization techniques for sentiment analysis based on LLM (GPT-3). *Applied and Computational Engineering*, 77:251–257.

Kai Zhang and Yupeng Han. 2025. Harnessing commonsense: Llm-driven knowledge integration for fine-grained sentiment analysis. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 4106–4116.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. *arXiv preprint arXiv:2506.05176*.