Title

Anonymous ACL submission

Abstract

TO DO

1 Introduction

2 Related Works

3 Methodology

3.1 Creating the Phrase Dataset

3.1.1 Scraping

We collected COPOM (Central Bank of Brazil's Monetary Policy Committee) minutes using Python and Selenium. We accessed `https://www.bcb.gov.br/publicacoes/atascopom/cronologicos`, which contains the listing of all of them. For each minute, we downloaded both the HTML and PDF content when available.

We ended up with a dataset C containing 251 COPOM minutes from January 1996 to July 2025. Each minute c in C has an associated date d_i and may have one or both HTML and PDF versions of the content.

3.1.2 Parsing

For each COPOM minute c in C:

1. Type-Specific Pre-Processing

   HTML file: if it exists, we extracted only the content inside the body tag. Tags such as `strong`, `i`, and `br` were removed while preserving their inner content. Other tags were removed along with their content.

   PDF file: if it exists, we used SpaCyLayout with the `pt_core_news_lg` model to extract individual phrases from PDF documents.

   After that, we created two separate phrase lists: one from the HTML source P_c^{html} and another from the PDF source P_c^{pdf}.

2. General Pre-Processing

   For each phrase in both P_c^{html} and P_c^{pdf}, we applied the following steps in that order: (1) Removed newlines and tabs; (2) Removed remaining tag entities (e.g., ` `); (3) Reduced multiple consecutive spaces, commas, and periods to single characters; (4) Added a period at the end if it did not exist.

3. Length Filtering

   For both P_c^{html} and P_c^{pdf} sets, we applied the following steps in that order: (1) Discarded single-word phrases; (2) Discarded phrases with character count below μ, the mean character count of phrases from the respective source P_c^{x}.

4. Blacklist Filtering

   We removed phrases containing at least one of the words from the following list: (1) *javascript*; (2) *cookies*; (3) *expand_less*; (4) *content_copy*; (5) *Garantir a estabilidade do poder de compra da moeda*.

   While terms (1) to (4) are related to web page elements and scripts, term (5) is the Brazilian Central Bank's motto, which often appears in the minutes and is not relevant for sentiment analysis.

Finally, we compared the number of phrases between sets P_c^{html} and P_c^{pdf} for each minute c. We selected the set with the most phrases; if both sets had equal size, we chose the PDF version as it appeared to have an overall superior phrase quality. When either source was unavailable or contained insufficient information, this step ensured we obtained the most reliable set for each minute.

At the end we obtained a set F made of smaller sets F_{d_i} for each date d_i, where d_i is the associated date of minute c. Each F_{d_i} contained 20 to 70 phrases.

3.1.3 Phrase Selection

We concatenated all phrases from the previous step while preserving their date labels. We performed vector search using cosine similarity for the word "inflation" in the dataset using the following Python libraries:

```
import os
import torch
import pandas as pd
from sklearn.metrics.pairwise import cosine_similarity
import numpy as np
from tqdm import tqdm
from langchain_huggingface import HuggingFaceEmbeddings
```

We retained only phrases with a similarity score greater than 0.6. Using the date labels, we reconstructed the minutes for each date with the selected phrases.

3.2 Model-Based Evaluation

3.2.1 Large Language Model Evaluation

We evaluated each phrase of the entire dataset using nine different large language models: openai/gpt-5, anthropic/claude-sonnet-4, google/gemini-2.5-pro, x-ai/grok-4-fast, openai/gpt-oss-120b, meta-llama/llama-4-maverick, google/gemma-3-27b-it, microsoft/phi-4, and deepseek/deepseek-chat-v3.1.

For each model, we used the same dataset obtained from the data collection phase. For each phrase, we prompted the model without providing previous context; each request was independent.

The evaluation prompt was formulated in Portuguese and asked the model to classify each phrase as optimistic, neutral, or pessimistic based on the following definitions provided by our specialist economist Cezio:

> **DEFINIÇÃO DE OTIMISMO:** Ocorre quando as projeções indicam que a inflação ficará abaixo da meta ou dentro do intervalo de tolerância com folga. Isso pode sinalizar que o Banco Central vê espaço para reduzir juros ou manter uma política monetária mais acomodatícia.
>
> **DEFINIÇÃO DE PESSIMISMO:** Ocorre quando as projeções apontam para inflação acima da meta ou próxima do teto do intervalo de tolerância. Isso sugere preocupação com pressões inflacionárias e pode justificar uma política monetária mais restritiva.
>
> **AVALIE A FRASE COMO:** O para OTIMISTA, N para NEUTRA, P para PESSIMISTA. SUA RESPOSTA DEVE SER APENAS UMA LETRA, SEM QUALQUER OUTRO TEXTO.

Model responses (O, N, P) were converted to numerical values: 1 for optimistic, 0 for neutral, and -1 for pessimistic. Phrases that could not be parsed were labeled as -2; such cases were rare.

Inference was performed using the OpenRouter API to unify model access. Each model was assigned a maximum token limit determined through initial testing. Models were tested on the phrases from the first minute with an initial limit of 1 token. If any phrase received a -2 score, the limit was doubled and the test was repeated until the model could process all phrases successfully.

The resulting maximum token limits were: openai/gpt-5 (1024), google/gemini-2.5-pro (128), openai/gpt-oss-120b (512), google/gemma-3-27b-it (8), deepseek/deepseek-chat-v3.1 (4), and others (1).

3.3 Human-Based Evaluation

We performed human evaluation in three stages: specialist, consolidated, and open evaluation.

3.3.1 Specialist Evaluation

We concatenated all phrases from all minutes while encoding their date labels in Base64 to prevent human bias. We randomly selected 350 phrases for manual evaluation by our specialist economist. The specialist labeled each phrase as: 1 (optimistic), 0 (neutral), -1 (pessimistic), -2 (non-related), or -3 (did not understand). Using the date labels, we reconstructed the minutes for each date with the remaining phrases.

3.3.2 Consolidated Evaluation

The dataset from the specialist evaluation was reanalyzed by the specialist and two additional professors. They discussed each phrase and attempted to reach consensus. This consolidated evaluation resulted in a dataset of 220 phrases. Using the date labels, we reconstructed the minutes for each date with the remaining phrases.

3.3.3 Open Evaluation

We created a website (https://inflation-form.luvas.io) featuring the same evaluation prompt used for the LLM models. Users were presented with each phrase and could select from three options: optimistic, neutral, or pessimistic. Responses were later converted to numerical values (1, 0, -1, respectively) and stored. Each browser was limited to evaluating 10 phrases per 24-hour period.

We requested collaborating universities (USP and Unicamp) to share the website with their economics-related graduate students. The open evaluation phase resulted in 278 evaluated phrases. Using the date labels, we reconstructed the minutes for each date with the remaining phrases.

3.4 Model Comparison and Analysis

3.4.1 Baseline and Sentiment-Enhanced Datasets

We created three comparison datasets:

1. **Only Inflation (Baseline):** We used IPCA monthly inflation data (Series 433) from the Brazilian Central Bank API: https://api.bcb.gov.br/dados/serie/bcdata.sgs.433/dados?formato=json

2. **Inflation + Sentiment (without Correction):** We combined the baseline IPCA data with sentiment variables. For each LLM model combination, we computed sentiment as the average grade per date from the model evaluations. Sentiment values were interpolated using cubic spline fitting and aligned with available IPCA dates.

3. **Inflation + Sentiment (with Correction):** We combined the baseline IPCA data with bias-corrected sentiment variables. For each LLM model combination and each human evaluation dataset, we computed sentiment as described above, then applied a correction factor based on the selected human evaluation bias.

3.4.2 Evaluation Framework

We compared model performance for each LLM model combination against each of the three human evaluation datasets: specialist, consolidated, and open. This multi-faceted comparison allowed us to assess the quality of LLM-based sentiment classification relative to human expert judgment across different evaluation methodologies.

4 Results

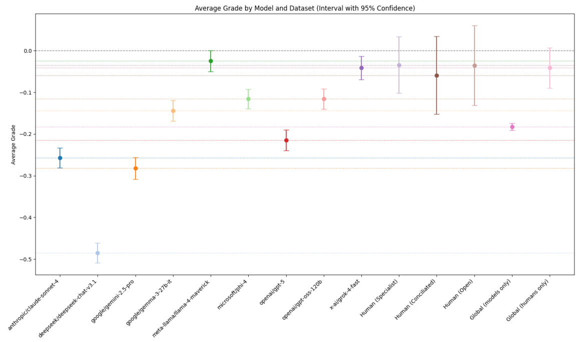

Figure 2: Average confidence intervals by dataset at 95% confidence level.

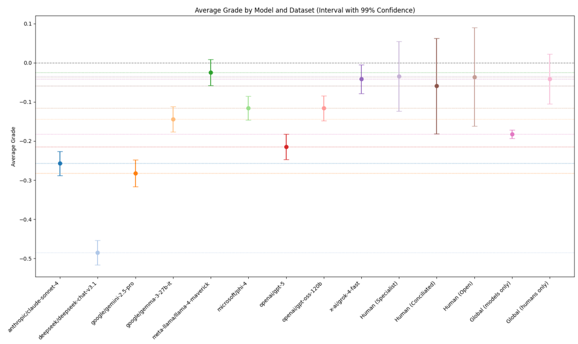

Figure 3: Average confidence intervals by dataset at 99% confidence level.

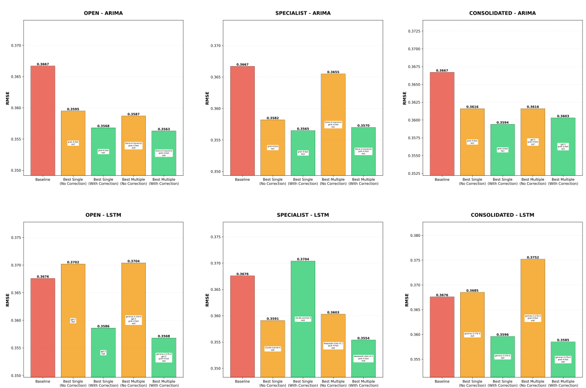

Figure 4: RMSE comparison across six different model configurations.

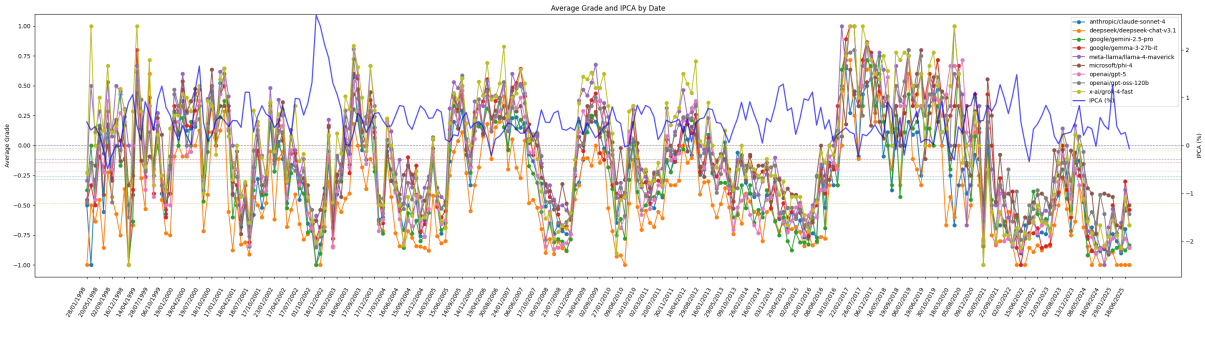

Figure 1: Average sentiment grade by date with IPCA inflation data.

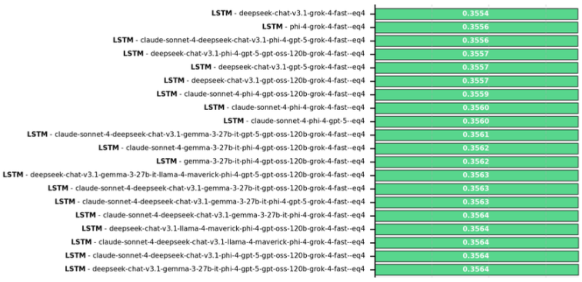

Figure 5: Model ranking based on specialist economist evaluation.

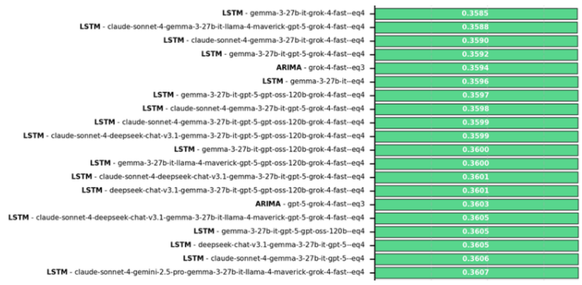

Figure 6: Model ranking based on consolidated evaluation.

Figure 7: Model ranking based on open evaluation.

5 Conclusion

References

A. Author and B. Author. 2025. Placeholder article title. *Journal Name*, 1:1–10.