GENEVA GRADUATE INSTITUTE
INSTITUT DE HAUTES ÉTUDES INTERNATIONALES ET DU DÉVELOPPEMENT
GRADUATE INSTITUTE OF INTERNATIONAL AND DEVELOPMENT STUDIES

Graduate Institute of International and Development Studies
International Economics Department
Working Paper Series

Working Paper No. HEIDWP23-2026

Experimenting with Large Language Models for Inflation Forecasting in Colombia

Aaron L. Garavito-Acosta
Central Bank of Colombia

Ugjkcp K_prglcx+Pgtcp_
Central Bank of Colombia

Hs_l H, Mqngl_+Rchcgpm
Central Bank of Colombia

Cbe_p A_gacbm+E_pag_
Central Bank of Colombia

Chemin Eugène-Rigot 2
P.O. Box 136
CH - 1211 Geneva 21
Switzerland

BCC Bilateral Assistance & Capacity Building for Central Banks

Experimenting with Large Language Models for Inflation Forecasting in Colombia[1]

Aaron L. Garavito-Acosta
(agaravac@banrep.gov.co)
Wilmer Martinez-Rivera
(wmartiri@banrep.gov.co)
Juan J. Ospina-Tejeiro
(jospinte@banrep.gov.co)
Edgar Caicedo-García
(ecaicega@banrep.gov.co)

August 28, 2026

Abstract

This paper develops and applies a standardized framework for forecasting year-on-year inflation in Colombia using large language models (LLMs). We conduct six sequential experiments in which the information set available to the models is progressively expanded by incorporating historical macroeconomic data, contextual indicators, explicit economic structure, and contemporaneous information retrieved through web search. Each configuration is executed daily and generates 24-month inflation forecasts in real time rather than retrospectively, together with qualitative explanations of the forecasts and, in the more advanced configurations, assessments of the shocks affecting inflation. This real-time design mitigates look-ahead bias and produces genuine forecast vintages. The framework is implemented as a programmatic pipeline in Python that queries the OpenAI and Google APIs, executes predefined experiment-specific prompts, and automatically processes and stores the model responses, while applying forecast validation and revision procedures in the more advanced configurations. The results show that richer information environments produce less monotonic inflation paths that remain above the 3% target over the forecast horizon and are more consistent with the contemporaneous domestic and external shocks affecting inflation. The qualitative analysis also shows that the models consistently identify relevant inflation drivers and their interactions. These richer forecast paths are broadly consistent with the pattern observed in survey-based inflation expectations. A formal evaluation of forecast accuracy will be conducted as additional real-time forecast vintages become available.

Keywords: Large language models, Inflation forecasting, Colombia, Prompt design.
JEL Codes: C53, E01, E31, E37, F31, E23.

[1]The authors thank Professor Francesco Audrino of the Swiss Finance Institute (SFI) for his academic supervision of this research. The authors also thank Cédric Tille and Nikhil Ray for their insightful comments and support. The authors gratefully acknowledge the excellent research assistance provided by Nicolás Quenguan, Isabella Polanco, Germán Vásquez, and María Camila Mejía. This document was edited with the supervised use of artificial intelligence tools. The opinions and ideas expressed are the authors' own and do not necessarily reflect those of the Central Bank of Colombia or its Board of Directors.

1 Introduction

Inflation forecasts are a key input to monetary policy design in inflation-targeting regimes, as the expected path of policy rates is determined in part by projected inflation, which also contributes to anchoring private-sector expectations and supports the design of central bank communication. Reflecting this policy relevance, the literature on inflation forecasting is extensive and methodologically diverse. Broadly, existing studies can be grouped into several methodological approaches, including traditional econometric and time series models (Atkeson and Ohanian, 2001; Stock and Watson, 2007; Bańbura et al., 2010), survey and text-based measures of expectations (Ang et al., 2007; Ramos-Veloza et al., 2024), machine learning approaches (Medeiros et al., 2021; Joseph et al., 2024; Martínez-Rivera et al., 2025), and semi-structural and DSGE models (González-Gómez et al., 2020; González et al., 2011).

Recent literature has increasingly explored an alternative methodological approach based on artificial intelligence, particularly Large Language Models (LLMs) (Ludwig et al., 2024; Faria-e Castro and Leibovici, 2024; Alam et al., 2026). This emerging technology has become an operational component of analytical pipelines, offering a promising addition to the forecaster's toolkit. However, despite their growing adoption, the use of LLMs in macroeconomic forecasting still lacks systematic evaluation protocols, standardized prompt design, and fully replicable experimental frameworks. Additionally, in Colombia the post-pandemic period has been marked by highly volatile price dynamics driven by global commodity shocks, domestic supply constraints, and shifting expectations. These conditions have strained traditional econometric models and increased the need for approaches, including AI tools, that can process large volumes of structured and unstructured data in real time.

In this context, the objective of this article is to develop and apply a controlled experimental framework for inflation forecasting with LLMs. We design a sequence of six experiments in which progressively richer contextual information is supplied to two models (GPT-4o and Gemini 2.5 Pro) under a common protocol that ensures comparability across steps. The framework is applied to Colombian headline CPI inflation (year-on-year), generating daily forecasts for each month over a 24-month horizon, together with qualitative explanations of the forecasts and, in the more advanced configurations, assessments of the shocks affecting inflation. Beyond producing the forecasts themselves, the sequential design allows us to examine how their dynamics and dispersion change as the information environment becomes progressively richer, assessing the role of historical data, contextual indicators, theory-based guidelines, and contemporaneous macroeconomic information. A formal evaluation of forecast accuracy against econometric and survey benchmarks is left for future work, because it requires a longer sample of forecast errors than is currently available.

The remainder of the paper is organized as follows. Section 2 reviews the literature on LLMs and economic forecasting. Section 3 provides the data description, and Section 4 presents the methodology, detailing the six experimental steps and the reproducibility, leakage, and consistency controls that govern the forecast generation process. Section 5 reports the results, comparing forecasts across experimental steps and model providers against relevant benchmarks. The final section concludes and discusses the implications for the use of LLMs in central bank forecasting practice.

2 Literature Review

The recent literature on LLMs and economic forecasting can be organized around several complementary uses. First, LLMs are used to generate macroeconomic forecasts directly (e.g., Faria-e Castro and Leibovici, 2024; Alam et al., 2026; Giannellis et al., 2026; Gonzalez, 2026; Bernasconi et al., 2025; André et al., 2025; Benny et al., 2025). Second, pretrained language models are adapted to process and forecast numerical time series (e.g., Gruver et al., 2023; Xue and Salim, 2023; Jin et al., 2024; Chang et al., 2025). Third, LLMs are used to extract predictive signals from textual sources that are subsequently incorporated into statistical or hybrid forecasting systems (e.g., Kwon et al., 2025; Jaworski, 2026; Allard et al., 2024; Fan et al., 2025). Fourth, a related body of work uses LLMs to simulate survey respondents or macroeconomic agents and generate economic expectations (e.g., Zarifhonarvar, 2026; Hansen et al., 2026; Lin et al., 2025; Wu et al., 2025).

Among these uses, direct forecasting is the one most closely related to our empirical approach. In these applications, the LLM processes the information provided in the prompt and directly generates a numerical macroeconomic forecast. With respect to inflation, Faria-e Castro and Leibovici (2024) use PaLM to generate retrospective, in-sample conditional forecasts of U.S. inflation and benchmark its performance against the Survey of Professional Forecasters (SPF), reporting lower mean-squared errors than the SPF overall, in most years, and at almost all horizons. Alam et al. (2026) employ ChatGPT-4 Turbo to forecast inflation for the current month and the subsequent twelve months. They find that ChatGPT nowcasts underperform the Cleveland Fed benchmark, while its longer-horizon out-of-sample forecasts are largely stale and unresponsive to incoming inflation data. On the other hand, they obtain stronger performance in pseudo out-of-sample exercises, raising concerns about information leakage and the validity of retrospective evaluations. In a similar vein, Giannellis et al. (2026) retrospectively evaluate GPT-5, GPT-4, and Gemini 2.5 Flash using one-month-ahead forecasts of year-over-year inflation in the United States and the euro area following the COVID-19 shock. Their results indicate that LLM forecasts generally outperform the random-walk benchmark as well as autoregressive and Kalman-filter models, with accuracy tending to improve when the forecasts are generated immediately after the latest inflation release rather than at the end of the previous month.

Other studies extend this direct forecasting approach to emerging economies and broader macroeconomic settings. Gonzalez (2026) retrospectively evaluates GPT-4o Mini by generating conditional short-term inflation forecasts for Argentina using forecast origins from January 2023 to August 2024 and compares them with the *Relevamiento de Expectativas de Mercado* (REM) of the Central Bank of Argentina. The results indicate that the LLM performs similarly to the REM when monthly inflation is at moderate or relatively low levels, whereas accuracy deteriorates during episodes of sharp inflationary acceleration: although the model correctly anticipates the direction of accelerating inflation, it systematically underestimates the magnitude of extreme spikes, a limitation also present, albeit to a lesser extent, in the REM forecasts. Likewise, Benny et al. (2025) evaluate LLMs for forecasting several macroeconomic variables in India for fiscal year 2025 to 2026, including inflation, and compare their predictions with traditional ARIMA models and forecasts from experts and leading institutions. Their results reveal substantial heterogeneity across models. They also show that combining LLM forecasts improves GDP forecasting accuracy relative to the best-performing individual model, although all evaluated approaches failed to anticipate the sharp decline in inflation observed in July 2025, highlighting the difficulty of forecasting structural breaks. Beyond inflation, LLMs have also been applied to GDP nowcasting and forecasting, with mixed results (Bernasconi et al., 2025; André et al., 2025).

A second line of work adapts language models specifically to numerical time series forecasting. Gruver et al. (2023) represent time series as strings of numerical digits and formulate forecasting as next-token prediction, while Xue and Salim (2023) express numerical inputs and outputs through natural-language prompts. More specialized approaches reprogram or fine-tune pretrained models to better capture temporal patterns, including Time-LLM (Jin et al., 2024) and time series alignment followed by forecasting fine-tuning (Chang et al., 2025). Related work focuses on macroeconomic applications. Koyuncu et al. (2026) develop BISTRO, a foundation model fine-tuned on BIS macroeconomic data, while Chib and Tan (2025) train a macroeconomic transformer using observed and DSGE-generated synthetic data. Carriero et al. (2025) compare Time Series Language Models with standard forecasting methods using FRED-MD data and find heterogeneous performance across models.

A third use of LLMs is to extract predictive information from textual sources. Kwon et al. (2025) use LLMs to construct macroeconomic sentiment indices that capture demand and supply drivers and improve the performance of simple forecasting models. In financial applications, Chen et al. (2025) extract predictive signals and Kirtac and Germano (2024) sentiment measures from news-based textual data to forecast stock-market outcomes. Text-derived signals have also been incorporated into hybrid macroeconomic forecasting systems. Jaworski (2026) uses GPT-4o to construct an inflation expectations indicator from social media and incorporates it into ARIMAX models, while Tawichsri et al. (2026), Gao et al. (2026), Allard et al. (2024), and Fan et al. (2025) use LLM- or AI-derived indicators from news or social media in inflation forecasting models. Similar approaches have been applied to GDP forecasting using textual sentiment measures derived from LLMs or related language models (de Bondt and Sun, 2025; Yin and Guo, 2026). Related work evaluates whether incorporating monetary-policy statements into GPT-4 prompts can complement traditional CPI forecasting models (Kaur et al., 2025), proposes hybrid frameworks that combine structured data, NLP, and LLMs for inflation prediction (Pandey et al., 2025), and reviews the broader use of LLMs across financial applications, including sentiment analysis, financial time series, forecasting, and reasoning (Nie et al., 2024).

A fourth application involves simulating expectations, surveys, or macroeconomic agents. Zarifhonarvar (2026), Hansen et al. (2026), Lin et al. (2025), and Wu et al. (2025) use LLMs to generate synthetic survey responses, simulate professional and household expectations, and construct economic agents with demographic or professional characteristics. These applications are appealing because expectations constitute a central channel in inflation dynamics and because simulated surveys can be conducted retrospectively or under alternative information treatments, at a fraction of the cost of human surveys. At the same time, the degree of heterogeneity and correspondence with human responses depends on model architecture, profile design, prior beliefs, and information conditioning. Retrospective applications also require careful control of survey dates and potential ex post knowledge.

Across these uses, several cross-cutting methodological issues emerge. The first dimension concerns prompt design and task formulation. Iadisernia and Camassa (2025) examine whether detailed role descriptions improve macroeconomic forecasts and find no measurable forecasting advantage relative to prompts without them, while the resulting LLM forecasts remain considerably more homogeneous than those of the human panel of the professional forecasters' survey. Osborn et al. (2025) show, in the context of labor-market forecasting, that the way forecasting tasks are formulated, including the choice of prompting strategy, affects forecast stability and performance. André et al. (2025) find that the language of the prompt affects forecasting performance, whereas narrative formats and requests for explanations have little effect on accuracy but can unlock the model's ability to produce a forecast. In turn, Liu et al. (2024) propose a prompting strategy that decomposes time-series forecasting into short- and long-term subtasks with distinct forecasting mechanisms. However, Schoenegger et al. (2025) find that most prompt-engineering modifications yield limited gains relative to a minimal baseline and that some strategies, including explicit Bayesian-reasoning instructions, can even reduce forecasting accuracy. Taken together, these studies suggest that prompt design matters for LLM forecasts, but that more elaborate instructions do not guarantee systematic improvements. These findings reinforce the case for transparent documentation of prompts, cutoff dates, generation parameters, output formats, and the number of model runs.

The second dimension concerns the methodological validity of LLM-based forecasting exercises. Crane et al. (2025) document errors in the recall of historical macroeconomic information, including the mixing of initial data releases with subsequent revisions and errors in recalling when the information became available. Karger et al. (2025) address potential outcome leakage by constructing a dynamic benchmark using only questions about events that are unresolved at the time of forecast submission, while Paleka et al. (2025) identify multiple forms of temporal leakage together with the difficulty of extrapolating evaluation results to real-world forecasting performance as central pitfalls in the assessment of language-model forecasters. Lopez-Lira et al. (2025) show that LLMs have memorized economic and financial data from their training period, so that genuine forecasting skill cannot be separated from recall within pre-cutoff samples, and Alam et al. (2026) show that pseudo out-of-sample exercises can substantially overstate real-time forecasting performance. More generally, Ludwig et al. (2024) formalize the no-training-leakage condition required for valid prediction exercises and emphasize the role of validation data when LLM outputs are used in downstream estimation. These studies justify strict controls against look-ahead bias, temporal leakage, and benchmark contamination. Beyond these threats to temporal validity, the second dimension also encompasses broader limitations of LLM-based forecasting. Su et al. (2024), Nie et al. (2024), and Abdullahi et al. (2025) discuss challenges related to hallucinations, data quality, generalization, explainability, and computational costs. Meanwhile, Tan et al. (2024) question whether the LLM component necessarily adds predictive value to time series models, showing that, in some settings, it can be removed or replaced by simpler components without reducing forecasting performance.

The third dimension concerns the evaluation of LLM-based forecasts relative to external benchmarks. Kim et al. (2024) compare GPT-4 with human analysts and conventional machine learning models in predicting the direction of earnings changes from financial statements, finding that GPT-4 with chain-of-thought prompting outperforms analysts and performs on par with a specialized machine-learning model. Halawi et al. (2024) develop a forecasting system that combines information retrieval, reasoning, and forecast aggregation and approaches, and in some settings surpasses, the crowd aggregate of competitive human forecasters. In macroeconomic applications, André et al. (2025), discussed above, benchmark LLM nowcasts against the Banque de France's suite of econometric models, while Osborn et al. (2025) benchmark LLM predictions against a moving average in labor-market forecasting tasks, where this simple baselines often remain competitive. Finally, Karger et al. (2025) compare LLMs with expert forecasters and the general public, finding that the median forecast of superforecasters outperforms the strongest LLMs. These results indicate that the evaluation of LLM-based forecasting methods should include simple statistical benchmarks, robust econometric or machine-learning models, and institutional or expert forecasts rather than relying solely on qualitative assessments.

This paper contributes to this literature by developing a standardized framework for generating real-time inflation forecasts. Unlike most existing studies, which rely on retrospective or pseudo out-of-sample exercises, our design produces genuine forecast vintages through six sequential experiments that progressively enrich the information available to the models, from macroeconomic data and theoretical guidance to contemporaneous web-based information. In doing so, the paper extends the evidence on LLM-based forecasting to an emerging economy and provides a protocol that can be adapted by central banks.

3 Inflation Dynamics and Data

3.1 Colombian Inflation Dynamics

The empirical analysis focuses on Colombia's headline inflation rate, measured at a monthly frequency as the year-on-year percentage change in the Consumer Price Index (CPI). Figure 1 illustrates its evolution over the 1994–2026 period. Following a record high of 32.4% in 1990, inflation declined steadily and reached single-digit levels by mid-1999. This prolonged disinflation reflected a combination of institutional and macroeconomic developments, notably the strengthening of the Banco de la República following the constitutional recognition of its independence in 1991 (Article 371 of the Political Constitution of Colombia), the gradual transition toward an inflation-targeting regime, the reduced prevalence of indexation mechanisms, and the pronounced contraction in aggregate demand associated with the financial crisis and recession of the late 1990s.

Inflation remained comparatively contained throughout most of the following decade, albeit with occasional episodes of renewed price pressure. In 2007–2008, annual inflation rose to nearly 8%, largely as a result of higher international food and commodity prices and robust domestic demand. These pressures subsequently receded as the global financial crisis of 2008–2009 weakened economic activity and dampened inflationary pressures. Inflation accelerated again in 2015–2016, when the sharp depreciation of the Colombian peso, an exceptionally severe El Niño episode, and a nationwide truck drivers' strike generated substantial supply-side pressures. Taken together, these shocks pushed annual inflation to almost 9% by mid-2016. Price pressures remained elevated in the following year and were further reinforced by the increase in the value-added tax rate from 16% to 19%. The monetary policy tightening that followed gradually steered inflation back toward the central bank's target.

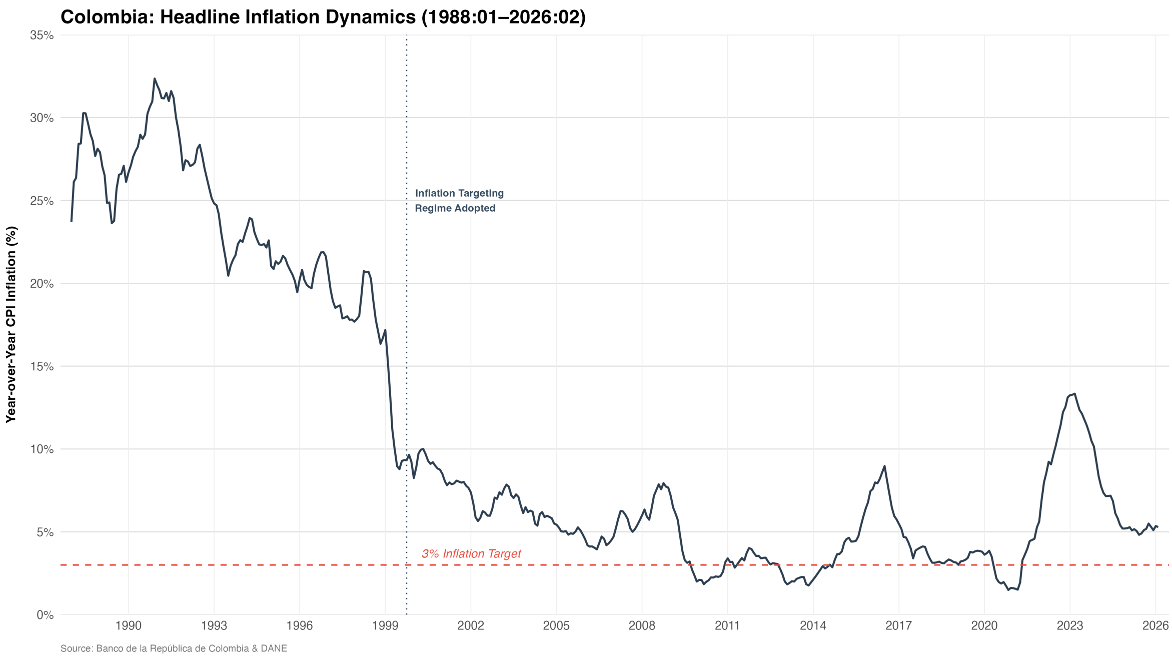

Figure 1: Colombian headline inflation

This adjustment was subsequently disrupted by the economic shock associated with the COVID-19 pandemic in 2020. The initial collapse in aggregate demand produced a temporary decline in inflation. However, this effect was reversed as economic activity recovered. The reopening of the economy coincided with persistent disruptions to global supply chains, rising transportation and commodity costs, exchange-rate depreciation, and, subsequently, additional pressures arising from Russia's invasion of Ukraine. The interaction of these domestic and external shocks led to a pronounced and persistent increase in inflation, which reached approximately 13% in 2023, its highest level since the late 1990s. Inflation has declined since then but has remained above the Banco de la República's 3% target. More recently, the disinflationary trend has been interrupted, with inflation resuming an upward trajectory in 2026. Annual headline inflation increased from 5.6% in March to 6.1% in June. This renewed acceleration reflected several factors such as pass-through of higher labor costs,robust private consumption and excess demand, supply disruptions affecting food prices, and rising international transportation and input costs. Taken together, these pressures outweighed the disinflationary effects of the Colombian peso's appreciation.

3.2 Data and Sequential Information Sets

For the development of the forecasting exercises, a macroeconomic dataset for Colombia is constructed at mixed frequencies (monthly, quarterly, and annual), according to the availability of each official source, as summarized in Table 1. Unlike traditional time series exercises, where the information set is typically fixed, this study adopts a sequential and cumulative design. Each new block of variables is incorporated only at the step at which it becomes relevant, allowing the information content of the prompt to expand progressively across experiments. Section 4 describes the six experimental steps in detail, including the information set supplied to the models at each stage.

Table 1 summarizes the variables employed in the forecasting exercises, while the discussion below explains their economic rationale and their role across the different experimental steps. The target variable is annual inflation, measured as the twelve-month change in the Consumer Price Index (CPI) reported by DANE for the period 1955:07–2026:05. This series is the only variable present in all experiments (Steps 0–5), ensuring direct comparability across specifications and between the GPT and Gemini models. At Step 2, the framework evolves from a univariate to a multivariate setting through the incorporation of three additional forecast variables. The exchange rate (TRM) is aggregated from daily to monthly averages; GDP is maintained at its official quarterly frequency to avoid artificial interpolation; and the monetary policy rate (TPM) is included to capture the policy stance of the Banco de la República and its implications for inflation expectations. Together with CPI inflation, these variables form the core set jointly projected in Step 2: inflation, the exchange rate, and the policy rate over a 24-month horizon, and GDP growth over eight quarters.

Starting in Step 3, the information set is expanded with contextual variables organized into four thematic categories motivated by the small open economy Phillips-curve literature: inflation expectations, economic slack, domestic cost pressures, and external pressures.

Expectation-related variables. This block includes core inflation (CPI excluding food and regulated items), survey-based inflation expectations, and Break-Even Inflation (BEI), computed as the spread between the one-year COP and UVR zero-coupon yields. Together, these variables capture both survey and market measures of inflation expectations and the underlying persistence of inflation.

Economic slack variables. This category includes employment and unemployment rates, the Economic Monitoring Index (ISE), the output gap, and household consumption growth. The output gap is the direct measure of slack, while the remaining indicators track the cyclical position of demand at higher frequency.

Domestic cost-pressure variables. This block consists of eight Producer Price Index (PPI) measures that capture changes in production costs before they are transmitted to consumer prices.

External pressure variables. These include the Real Exchange Rate Index (RERI), Brent oil prices in Colombian pesos and U.S. dollars, Federal Reserve policy-rate projections, and sovereign credit default swaps (CDS). Collectively, they capture exchange-rate pass-through, imported inflation, and country-risk conditions.

Table 1: Summary of Variables and Data Sources

Variable Group	Series Included	Frequency	Source	Unit roots
Primary Forecast Variable				
Headline CPI Inflation	Consumer Price Index, 12-month change	Monthly	DANE	1
Variables for Multivariate Forecasting or Contextual Use				
Exchange Rate	Representative Market Exchange Rate	Monthly (Avg.)	BanRep	1
Economic Activity	Real GDP Growth	Quarterly	DANE	0
Monetary Policy	Monetary Policy Rate	Monthly	BanRep	2
Contextual Variables				
Expectation-Related Variables				
Underlying Inflation	CPI excluding food and regulated items	Monthly	DANE	1
Survey-Based Inflation Expectations (8 series)	Monthly inflation expectation; annual inflation expectation; end-of-current-year inflation expectation; 12-month-ahead inflation expectation; 12-month-ahead inflation expectation. Total; industry; financial sector; academics and consultants	Monthly/Quarterly	BanRep	0;1 1;1 1;1 1;1
Business Price and Cost Expectations (2 series)	12-month-ahead input-cost expectations; 12-month-ahead selling-price expectations	Quarterly	ODE	0;0
Market-Based Inflation Expectations (3 series)	One-year COP zero-coupon yield; one-year UVR zero-coupon yield; one-year Break-Even Inflation	Daily/Monthly	BanRep	1;1;1
Economic Activity and Slack Variables				
Labor Market Indicators (2 series)	Seasonally adjusted unemployment rate; seasonally adjusted employment rate	Monthly	DANE	1;1
Economic Activity and Demand Indicators (2 series)	Economic Monitoring Indicator (ISE), annual change; household consumption, annual change	Monthly/Quarterly	DANE	0;0
Output Gap	GDP output gap	Quarterly	BanRep	0
Domestic Cost-Pressure Variables				
PPI by Economic Use (4 series)	PPI Final Consumption Goods; PPI Intermediate Consumption Goods; PPI Capital Goods; PPI Construction Materials	Monthly	BanRep	1;0 1;1
Sectoral PPI Measures (4 series)	PPI Manufacturing Industries; PPI Manufacture of Food Products; PPI Agricultural Production; PPI Fuels and Gas	Monthly	BanRep	1;0 1;1
External Pressure Variables				
Real Exchange Rate	CPI-based Real Exchange Rate Index	Monthly	BanRep	1
Imported-Price Indicators (4 series)	Imported Goods PPI; Food Products PPI; Imported Goods Deflator; Imported Services Deflator	Monthly/Quarterly	BanRep/DANE	1;0 1;1
Brent Crude Oil Prices (2 series)	Brent crude oil price in U.S. dollars; Brent crude oil price in Colombian pesos	Monthly	U.S. EIA/BanRep	0;1
Global Commodity Price Indices (8 series)	Food Price Index; Agricultural Raw Materials Price Index; Metals Price Index; Fertilizer Price Index; Food Price Index in COP; Agricultural Raw Materials Price Index in COP; Metals Price Index in COP; Fertilizer Price Index in COP	Monthly	IMF/BanRep	0;0 0;0 1;1 0;0

Notes: Variables are grouped to keep the table concise, with the individual series included in each group reported in the second column. For the contextual dataset supplied to the LLMs, variables originally available at a quarterly frequency are transformed to monthly frequency using linear interpolation. For zero-coupon yields, the last daily observation of each month is used, whereas the daily TRM is aggregated using monthly averages. Break-Even Inflation (BEI) is calculated as the difference between the one-year COP and UVR zero-coupon yields. Commodity price indices expressed in Colombian pesos are constructed using the TRM.

4 Methodology

This section presents the empirical strategy used to generate monthly year-on-year CPI inflation forecasts with LLMs. The framework consists of six sequential experiments designed to examine how LLM-based forecasts respond to changes in the information available to the models, the macroeconomic context, and the structure of the forecasting instructions, with progressively different information sets and degrees of economic guidance across experiments. The experiments are implemented using two LLM families, OpenAI's GPT and Google's Gemini. Specifically, GPT-4o is used in Steps 0 to 4 and GPT-4.1 in Step 5, while Gemini 2.5 Pro is used throughout the six experiments.[2] Each experimental configuration is run daily, generating a new set of forecasts each day and allowing us to track high-frequency revisions in the models'

[2]Both models were queried at a temperature of 0 with a nominal maximum output budget of 32,000 tokens, both for the initial forecast and for the subsequent correction attempts, while auxiliary calls for web-based context retrieval, summarization, and consistency validation used a temperature of 0.1. Beyond temperature, GPT was also given a Top-p of 1.0 and a fixed random seed of 42; these two parameters were not specified for Gemini, so temperature is the only randomness control shared by both models. These settings improve consistency across executions rather than ensuring deterministic replication, consistent with Anesti et al. (2025), who show that lower temperature values tend to produce more consistent outputs.

projections over time[3]. The forecasts are generated in real time rather than reconstructed retrospectively, so that each daily run reflects the information environment prevailing when the forecast is produced. This real-time design mitigates look-ahead bias and produces a sequence of genuine forecast vintages that can be compared over time. In addition to numerical forecast paths, the models are asked to provide textual explanations of the economic reasoning underlying their projections, with the depth and structure of these explanations varying across experimental configurations. A detailed description of each experiment is provided in Table 2 and Appendix A.

An important distinction across the experiments is that their purpose evolves as the framework becomes more sophisticated. While Steps 0 to 3 focus on how forecasts respond to progressively richer information sets, Steps 4 and 5 additionally incorporate economic structure and consistency checks to guide the forecasting process, while Step 5 further integrates contemporaneous information retrieved through web search to identify and assess the factors affecting the inflation outlook. The forecasting exercises are implemented through a programmatic pipeline in Python using Jupyter notebooks, rather than through manual interaction with a conversational interface. In each daily run, the corresponding notebook loads the required historical and contextual data, the forecast date, and the model configuration. It then constructs the experiment-specific prompt and submits it to the OpenAI or Google Gemini API using the corresponding API credentials. Depending on the experimental configuration, the models may also access web-search capabilities, while the remaining information is supplied directly from the local data files. The responses are returned in structured formats and are automatically processed to extract the numerical forecasts and textual explanations. When required by the experimental design, the resulting forecasts are also subjected to validation and revision procedures before being stored together with the model responses and relevant execution metadata. Overall, this architecture provides a systematic workflow covering data ingestion, prompt construction, model invocation, response processing, validation, and output storage for each forecasting run.

Prompts for all experiments were written in English, as prompt language can affect LLM performance. In particular, André et al. (2025) report lower forecasting errors for English-language prompts. The prompts also assign the models an explicit economic forecasting role tailored to the Colombian macroeconomic context, with the degree of specialization adjusted to the complexity and objectives of each experiment. This design is consistent with Osborn et al. (2025), who document that role assignment can improve the stability of LLM responses. In addition, the prompts are organized into structured sequences of tasks, with the degree of decomposition increasing across experiments. In the more advanced configurations, the models are explicitly guided through intermediate analytical stages before generating or revising their forecasts. This design is consistent with Kim et al. (2024), who find that prompting strategies that decompose complex tasks into intermediate reasoning steps can outperform simpler prompting approaches. Finally, the date of each daily execution is explicitly included in the prompt to provide the models with a clear temporal reference for the information environment in which the forecast is generated, following the dated-prompt approach used by Giannellis et al. (2026).

To illustrate how the forecasting framework operates in practice, we describe Step 5 in greater detail, as it represents the most comprehensive implementation of the proposed methodology. This configuration combines historical and contextual macroeconomic data, contemporaneous web-based information, explicit economic relationships, joint forecast generation, and post-generation validation within a single computational workflow. The forecasting process begins by combining the historical series of the four forecast variables (headline CPI inflation, nominal exchange rate, monetary policy rate, and real GDP growth) with the broader set of contextual indicators described in Table 1. These data provide information on inflation expectations, economic activity and slack, domestic cost pressures, and external conditions. In addition to these historical and contextual data, Step 5 constructs a contemporaneous macroeconomic context assembled from domestic and international information retrieved through web search[4]. The retrieval is performed by Gemini using Google Search grounding, with one query per information category. To ensure that both models face an identical information set, this context is retrieved once per query date and stored; the same category-level summaries are then supplied to GPT and to Gemini.

The use of web search in Step 5 is motivated by the potential value of incorporating contemporaneous information into

[3]Daily execution commenced in 2026 on the following start dates: Step 0 on March 4 (Gemini) and February 19 (GPT); Step 1 on March 3 (Gemini) and February 13 (GPT); Step 2 on March 4 (Gemini) and February 16 (GPT); Step 3 on March 5 (Gemini) and February 19 (GPT); Step 4 on March 4 (Gemini) and March 6 (GPT); and Step 5 on April 28 (Gemini) and April 17 (GPT). Prior to these dates, runs were generated through backward in-sample execution using date restrictions embedded in the prompt. From these dates onward, all forecasts correspond to genuine real-time vintages. Because the exercise remains in daily operation, the sample of real-time vintages continues to expand.

[4]The contemporaneous macroeconomic context is constructed through category-specific web searches covering relevant domestic and external developments. This context is used to identify and characterize the shocks and factors that are relevant at the forecast date, to inform the assumptions underlying the macroeconomic scenario, and ultimately to condition the joint forecast paths.

Table 2: Main Characteristics of the Six Experimental Forecasting Configurations

Step	Forecast Variables and Horizon	Information Supplied to the LLM	Methodological Approach
0	Headline CPI inflation (monthly, 24 months)	The model retrieves the latest available headline inflation information through web search, prioritizing official sources such as DANE.	Baseline forecasting exercise. The LLM examines the available inflation information and its recent dynamics and generates a forecast for headline CPI inflation.
1	Headline CPI inflation (monthly, 24 months)	Historical headline inflation series supplied directly to the model. No external web information is used.	The univariate forecasting exercise of Step 0 is retained, but the information source is controlled. Rather than retrieving inflation data from the web, the historical series is supplied directly to the LLM, which uses the observed dynamics of the series to generate the forecast.
2	Headline CPI inflation, nominal exchange rate (TRM), and monetary policy rate (monthly, 24 months); real GDP growth (quarterly, 8 quarters)	Historical series of the four forecast variables supplied directly to the model. No external web information is used.	The historical-data-based forecasting approach of Step 1 is extended from inflation alone to four macroeconomic variables. The LLM analyzes their historical dynamics and generates the four forecasts jointly, requiring the projected paths to be mutually consistent within a common macroeconomic scenario.
3	Headline CPI inflation, nominal exchange rate (TRM), and monetary policy rate (monthly, 24 months); real GDP growth (quarterly, 8 quarters)	Historical series of the four forecast variables, complemented with the contextual variables reported in Table 1. No external web information is used.	The joint multivariate forecasting framework of Step 2 is retained, while the information set is expanded with contextual variables. The LLM assesses the relevance of this additional information and uses it to condition the four jointly determined forecast paths.
4	Headline CPI inflation, nominal exchange rate (TRM), and monetary policy rate (monthly, 24 months); real GDP growth (quarterly, 8 quarters)	Historical series of the four forecast variables and the contextual variables reported in Table 1, complemented with forward-looking macroeconomic assumptions. No external web information is used.	The multivariate framework of Step 3 is retained and further disciplined by an explicit macroeconomic structure. The LLM generates the four forecast paths jointly using relationships linking inflation, economic activity, monetary policy, and the exchange rate, including Phillips-curve, Taylor-rule, UIP, aggregate-demand mechanisms, among others. Forecasts are evaluated for economic and quantitative consistency. A limited number of revisions may be attempted, and the best-scoring candidate is retained.
5	Headline CPI inflation, nominal exchange rate (TRM), and monetary policy rate (monthly, 24 months); real GDP growth (quarterly, 8 quarters)	Historical series of the four forecast variables and the contextual variables reported in Table 1, combined with contemporaneous domestic and external macroeconomic information obtained through web search.	The theoretically disciplined multivariate framework of Step 4 is retained and extended with contemporaneous macroeconomic information. The LLM identifies and quantifies relevant shocks, evaluates their transmission to the four forecast variables using macroeconomic relationships and empirical evidence for Colombia, and incorporates these effects into the joint forecast paths. Forecasts are evaluated for theoretical and shock consistency. Up to four revisions may be attempted, and the best-scoring candidate, including the original forecast, is retained.

Notes: In Step 5, the GPT-4.1 implementation reuses the contemporaneous web context previously generated by Gemini when available. This design was adopted to reduce API and web-search costs while keeping the information environment more closely aligned across models; GPT nevertheless performs its own shock assessment and generates its forecasts independently.

LLM-based forecasting. Existing evidence is mixed. Halawi et al. (2024) show that the retrieved news can improve probabilistic prediction, while Bernasconi et al. (2025) report competitive nowcasting performance when Gemini is combined with web search. Step 5 therefore provides an empirical setting in which the contribution of contemporaneous web-based information can be evaluated within a theoretically disciplined forecasting framework. This web-based context is intended to capture recent economic developments, policy decisions, shocks, and other relevant information that may not yet be fully reflected in the historical dataset. Together with visual representations of the historical and conditioning variables, these inputs are incorporated into a structured prompt from which the LLM constructs a common macroeconomic scenario. This scenario includes forward-looking assumptions about relevant domestic and external conditions inferred from the available data, contemporaneous information, and identified shocks.[5]

The prompt further introduces economic discipline by requiring the LLM to evaluate the interactions among inflation, economic activity, monetary policy, and the exchange rate using a set of literature-based consistency rules covering Phillips-curve dynamics, monetary policy reaction functions, uncovered interest parity, aggregate-demand mechanisms, expectations, fiscal conditions, risk premia, and other relevant macroeconomic channels. Once the historical data, contextual indicators, contemporaneous macroeconomic context, and economic guidance have been incorporated into the prompt, the LLM uses this information to construct the macroeconomic scenario and shock assessment and, within the same generation step, produces the four forecast paths jointly. Inflation, GDP growth, the exchange rate, and the policy rate are therefore treated as co-determined outcomes of a common macroeconomic scenario rather than as independently generated projections. The forecast is produced directly by the language model as the output of this single generation step. No statistical or econometric model is estimated at any point. The forecast paths are therefore judgmental outputs of the model, disciplined by the theoretical framework embedded in the prompt. The prompt is designed to prevent the model's judgment from anchoring on two obvious references. The first is third-party projections, which the model may not adopt as its own forecasts, and the second is the tendency to smooth through the shocks identified at that date, so that a path which does not visibly reflect them is treated as inconsistent unless the forces that offset them are quantified and sourced. Under these conditions, the model returns twenty-four monthly paths for inflation, the exchange rate and the policy rate, and eight quarterly paths for GDP growth.

After the initial forecasts are generated, a coverage diagnostic checks whether the different components of the contemporaneous macroeconomic context have been represented in the shock analysis. The resulting forecast paths are then subjected to a unified validation procedure that combines shock-path and theoretical consistency checks. Specifically, the procedure assesses whether the direction and magnitude of the forecast movements are consistent with the quantified residual effects of the identified shocks and whether the joint dynamics of the four variables are economically plausible under the macroeconomic regime implied by the forecast. If material shock-path or theoretical inconsistencies are detected, the validation feedback is returned to the LLM through a correction prompt. Up to four revised forecasts may be generated, each one built on the most recent revision and re-evaluated under the same validation procedure. Every candidate receives a validation score that penalizes failed checks and critical violations and rewards the breadth of checks that could be evaluated, with lower scores indicating greater consistency. The original forecast is retained as the initial candidate and remains eligible for final selection, and the procedure stops as soon as a candidate satisfies all checks. The final output therefore corresponds to the best-scoring candidate among the original forecast and the revised alternatives, allowing the procedure to strengthen economic consistency without mechanically imposing a predetermined forecast path.

5 Results

This section presents the empirical findings from the sequential forecasting experiments described in Section 4. It first examines the main factors identified by Gemini in Step 5 as influencing the inflation outlook. Text mining techniques are then used to evaluate the relative importance of these factors and the relationships among them. The second part compares the 24-month-ahead inflation forecasts generated by GPT and Gemini across the six experimental steps, assessing how progressively richer information sets, explicit economic structure, and model-based identification of inflation drivers shape forecast dynamics and dispersion. Finally, the section analyzes the monthly updates to the December 2026 and

[5]The LLM translates the contemporaneous macroeconomic context into a structured shock analysis. Shock magnitudes are determined by the model but disciplined by literature-based empirical ranges, prioritizing evidence for Colombia and using midpoint values as default references. Prospective shocks are probability-weighted, while effects already reflected in the observed data are discounted so that only the remaining forward impact is considered. The resulting shock assessment is used both to inform the jointly generated forecast paths and to evaluate their consistency.

December 2027 inflation forecasts and compares the LLM-based projections with survey expectations from the Banco de la República. Since the models are instructed not to adopt these expectations as their own forecasts, any divergence from consensus is itself an object of interest.

5.1 Qualitative Analysis

The evidence presented here suggests that LLMs may enhance conventional forecasting frameworks by providing a structured assessment of the key factors influencing inflation dynamics and projections. Moreover, because LLM-based forecasts can be updated at high frequency, they may facilitate a more timely assessment of the evolving drivers of inflation over the forecast horizon, thereby enhancing risk analysis and informing policy discussions. The monthly evolution of the macroeconomic factors identified by Gemini as the main determinants of the inflation forecast is summarized in Figure 2. The figure is constructed from a text analysis of the inflation-related factors and shocks identified by Gemini in Step 5 as part of the forecast generation process. These factors are classified into macroeconomic categories. Specifically, the texts were processed using text-mining techniques that identify trigrams, defined as sequences of three consecutive words, which were subsequently classified into thirteen predefined categories of macroeconomic factors, including oil and fertilizer prices, global supply chains, international food prices, monetary policy, inflation expectations, the minimum wage, and international financial conditions, among others. For each month, the relative frequency of occurrence of these categories within the set of identified trigrams was computed, allowing the relative importance of each factor in explaining inflationary pressures to be quantified.

A first result is the high degree of stability exhibited by the dominant factors throughout the period. In all months analyzed, factors associated with international oil prices, global supply chains and transportation costs, as well as international food and fertilizer prices, appear among the most relevant categories. Another relevant finding is the recurrent presence of domestic factors related to wage indexation, increases in the minimum wage, inflation expectations, and domestic demand conditions. This finding suggests that the model consistently identifies the central role of external costs and terms-of-trade conditions as determinants of the expected behavior of Colombian inflation. A particularly noteworthy aspect is the model's ability to capture changes in the relative importance of certain factors across months. While some categories maintain structural relevance throughout the entire period, others display temporary variations associated with specific events. This feature suggests that the LLM does not operate solely on fixed historical relationships. Rather, it incorporates recent contextual information and dynamically adapts its interpretation of prevailing inflationary risks.

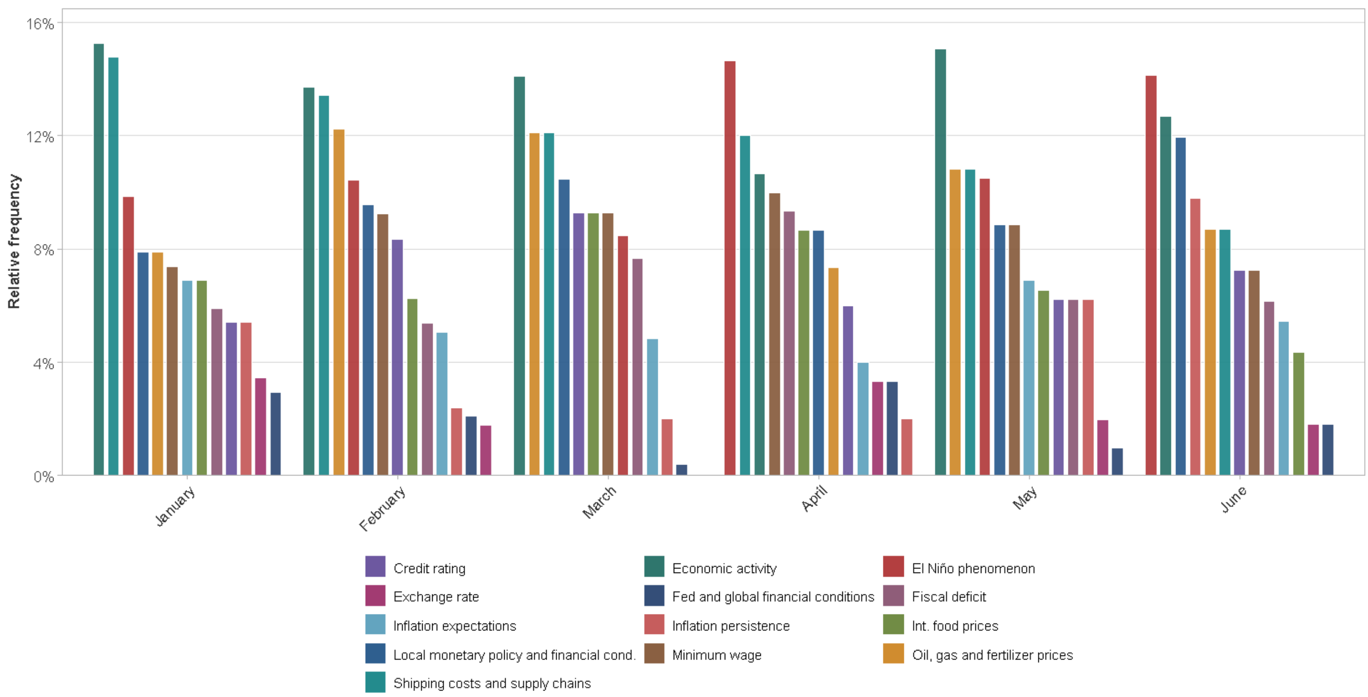

Figure 2: Relevant factors for inflation dynamics and forecast

Figure 3 complements the previous analysis by representing the internal structure of the factors identified by Gemini through a word network constructed from the most frequent trigrams observed in June. The network was constructed using the same texts employed to generate the forecasts. Initially, the content was processed using text-mining techniques to identify relevant trigrams, which were subsequently classified into the thirteen macroeconomic categories mentioned above. Each trigram was decomposed into its constituent words and transformed into a directed network in which links represent the sequence of word occurrence within each expression. Repeated connections were then aggregated and weighted according to their frequency of occurrence, while each word was assigned to the category that accumulated the greatest relative weight among the trigrams in which it appeared.

The results reveal several thematic clusters associated with the main sources of inflationary pressure identified by the model. Nodes related to international food prices, oil prices, transportation costs, monetary policy, the minimum wage, inflation expectations, and global financial conditions form clearly differentiated groupings, reflecting the way in which these concepts jointly appear within the economic context used by Gemini. The proximity among terms within each cluster suggests that the model recognizes coherent and recurrent economic relationships among the various factors considered in the forecasting process.

Moreover, the network is not dominated by a single category but instead exhibits multiple interconnected thematic cores, reflecting the multidimensional nature of the analysis performed by the model. This result is consistent with the evidence presented in Figure 2, where several categories simultaneously contribute to explaining the expected path of inflation. Taken together, the structure of the network suggests that Gemini does not base its inferences on isolated words, but rather on sets of concepts that systematically appear within broader economic narratives. This capability allows the model to identify not only which factors are relevant, but also how they relate to one another within the macroeconomic environment considered in the forecasting exercise.

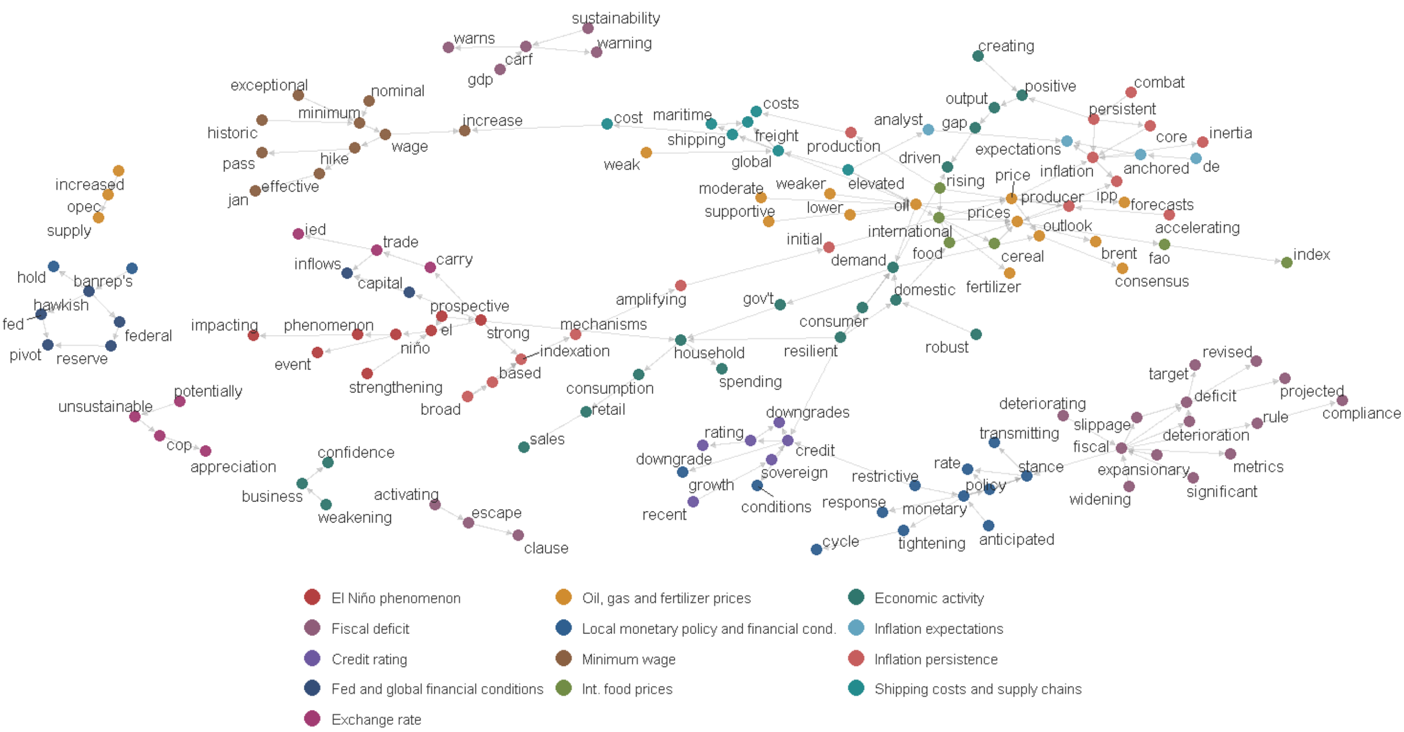

Figure 3: Word Network for June Based on the Most Frequent Trigrams

5.2 Inflation Forecast

Building on the descriptive analysis of inflation dynamics in Colombia, we now turn to the forecast results. Figure 4, together with the first panel of Figure 5, presents the inflation forecasts obtained in Steps 0, 1, 2, and 3. Despite differences across models and experimental configurations, a salient feature of the early experiments is the tendency of the forecast paths to move relatively quickly toward the 3% inflation target, particularly in Steps 0 to 3. This behavior is consistent with the information structure used in these configurations. The inflation target forms part of the information set supplied to the LLMs and these early configurations do not provide the models with systematic access to contemporaneous web-based information on the broader macroeconomic environment: Step 0 uses web search only to retrieve and verify the latest inflation information, while Steps 1 to 3 restrict the models to the data supplied in the prompt. Consequently, although Steps 2 and 3 expand the information set through joint forecasting and additional conditioning variables, they still lack a systematic assessment of contemporaneous shocks and their persistence.

Figure 5 reports the inflation forecasts obtained in Step 4. Relative to the earlier configurations, the forecast dynamics change substantially. Rather than displaying an almost monotonic convergence toward the 3% target, both LLMs allow inflationary pressures to remain elevated over the short horizon before a gradual disinflation process emerges. Gemini produces a higher and more dispersed near-term profile followed by a comparatively pronounced decline, whereas GPT exhibits a more gradual adjustment and remains at higher inflation rates toward the end of the forecast horizon. The expectations reported by the Monthly Survey of Economic Analysts' Expectations[6] display a broadly similar qualitative pattern, with an initial increase followed by gradual disinflation, although the magnitude and speed of convergence differ across forecasts. This change in behavior is consistent with the richer economic structure introduced in Step 4, in which the historical and conditioning information is complemented with explicit macroeconomic relationships and forward-looking factors that may sustain inflation away from the target. An important limitation of this configuration, however, concerns the updating of its forward-looking information set because Step 4 does not retrieve contemporaneous macroeconomic information through web search. Instead, the factors expected to exert upward or downward pressure on inflation are supplied through an expert-curated input that must be updated manually, typically on a quarterly basis. Consequently, the information set may incorporate new shocks or relevant developments only with a delay, while its manual maintenance makes the procedure relatively cumbersome. These limitations motivate Step 5, which systematically updates the contemporaneous macroeconomic context through web search.

Results for Step 5 show a further change in the inflation forecast dynamics. Rather than converging rapidly and monotonically toward the 3% target, forecasts from both LLMs initially increase and subsequently enter a gradual disinflationary phase, while remaining above the target at the end of the forecast horizon. For Gemini, the average forecast rises from approximately 6.1% in mid-2026 to around 7.1% toward the end of the year, before declining to about 4.6% in December 2027. This non-monotonic profile is consistent with the information structure introduced in Step 5. In contrast to Step 4, contemporaneous domestic and external macroeconomic information is systematically retrieved through web search and incorporated into the forecasting process. The LLM uses this information to identify relevant shocks, assess their persistence, quantify their remaining forward effects, and incorporate their transmission into the jointly determined forecast paths. Notably, forecasts from Step 5 tend to be slightly higher than survey expectations during the second half of 2026. With respect to the underlying trend, the results consistently exhibit a broadly similar pattern across all specifications. The projected increase during 2026 is consistent with the coexistence of domestic and external inflationary pressures. Domestic factors include persistent core and services inflation, the increase in the minimum wage and its pass-through to labor costs through indexation mechanisms, the de-anchoring of inflation expectations, and strong domestic demand supported by consumption, a positive output gap, and the fiscal impulse. These pressures are compounded by higher international food and fertilizer prices, the potential effects of El Niño on agricultural supply and food prices, rising production and transportation costs, and disruptions to global supply chains.

The stabilization of inflation at around 7.1% toward the end of 2026 reflects the interaction between these pressures and several offsetting forces. The appreciation of the Colombian peso and carry-trade inflows partially reduce imported inflation, while the restrictive monetary policy stance and weak investment gradually moderate aggregate demand. However, monetary policy transmission operates with a lag, allowing indexation, the second-round effects of the wage shock, and persistently elevated expectations to prolong inflationary persistence. The subsequent decline during 2027 and early 2028 reflects the cumulative effects of high interest rates, tighter credit conditions, and greater economic slack, although convergence remains incomplete because of persistent rigidity in core and services inflation. Forecast dispersion increases with the horizon, reflecting greater uncertainty regarding the persistence of indexation, the strength of the monetary policy response, fiscal developments, and the transmission of external shocks.

[6]The survey, conducted by the Banco de la República, collects inflation expectations at several horizons, including the following month, December of the current year, one year ahead, and December of the subsequent year. These expectations are depicted in Figure 5 as purple dots, allowing for a direct comparison between survey-based expectations and the forecasts obtained in Steps 4 and 5.

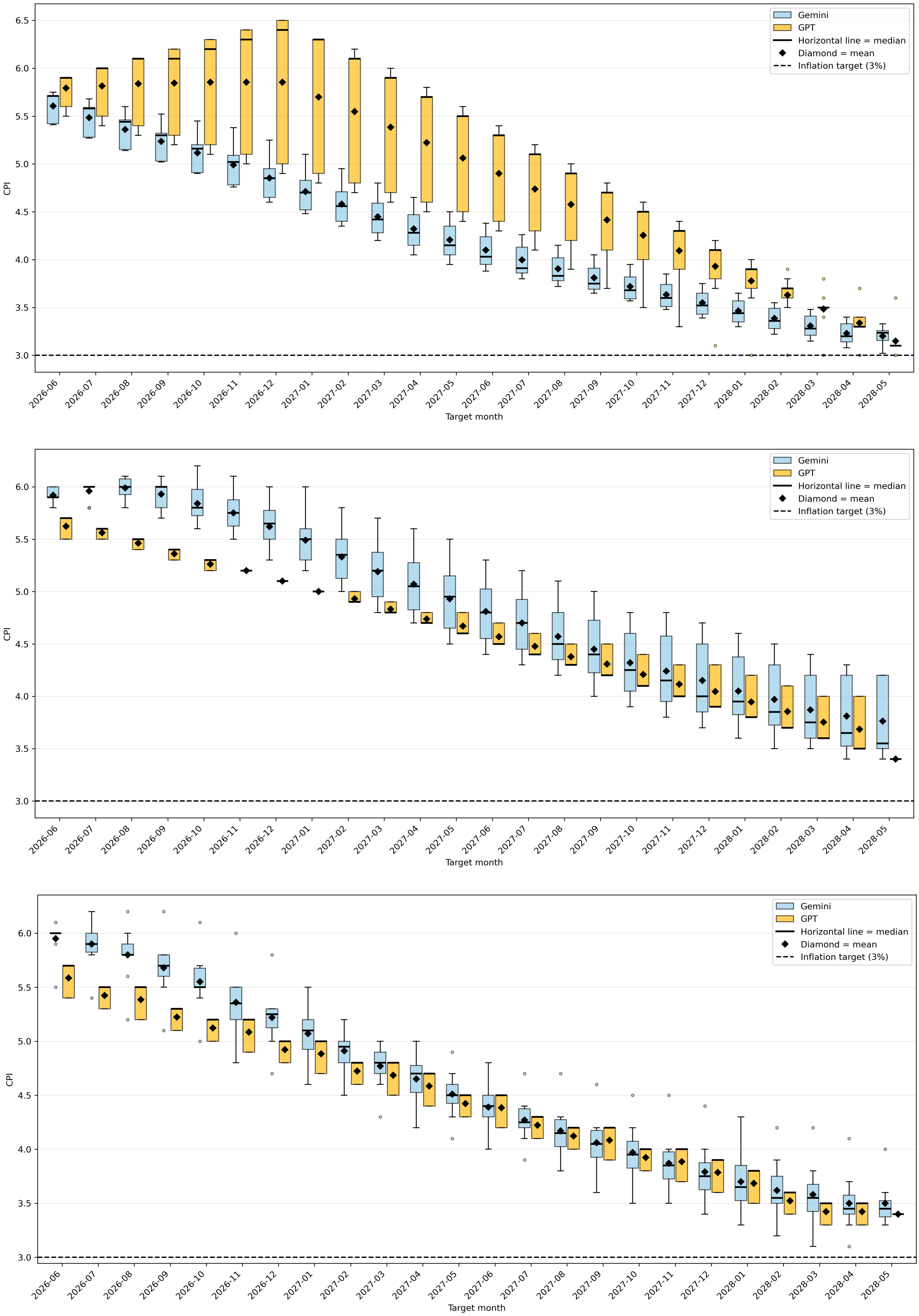

Figure 4: Inflation forecasts for Steps 0 to 2. Boxplots summarize daily forecasts generated from June 1 to June 20, 2026. Panels from top to bottom correspond to Steps 0, 1, and 2, respectively.

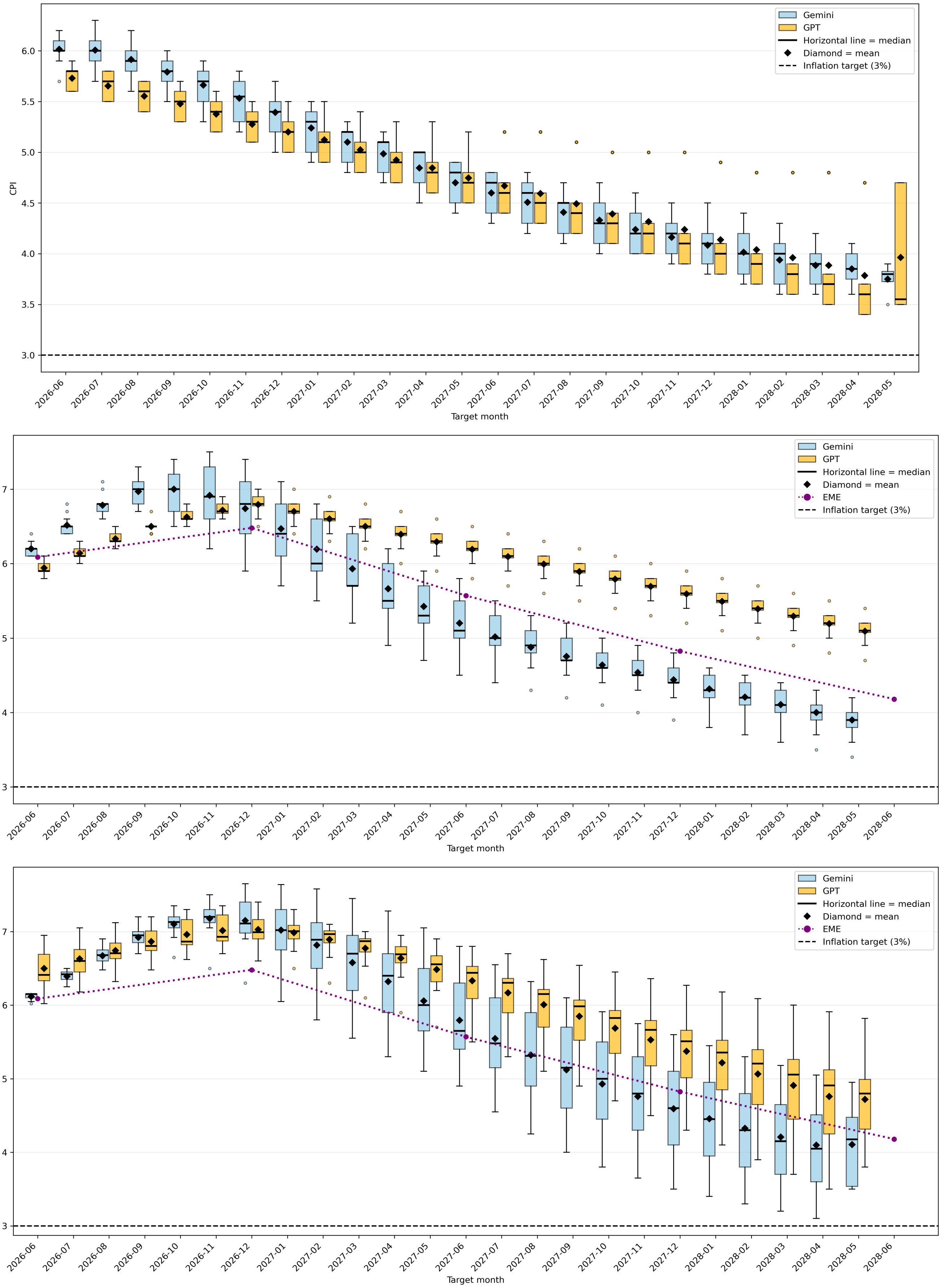

Figure 5: Inflation forecasts for Steps 3 to 5. Boxplots summarize daily forecasts generated from June 1 to June 20, 2026, to facilitate comparability with the Monthly Survey of Economic Analysts' Expectations (EME, Spanish acronym), shown by the purple dots.

5.3 Monthly Updates to Year-End Inflation Forecasts

We focus on the Step 5 forecasts for December 2026 and December 2027. Figure 6 reports the forecasts for December 2026 generated between January and June 2026 by GPT and Gemini[7], together with the corresponding EME expectations. Step 5 forecasts remain generally above the 3% inflation target throughout the period and, particularly in the later part of the sample, tend to exceed survey expectations. For Gemini, the average forecast increases from approximately 6.5% at the beginning of the exercise to around 7.1% in June. GPT drifts upward at a similar pace and converges on the same figure, so the revision characterizes the exercise rather than a single model.

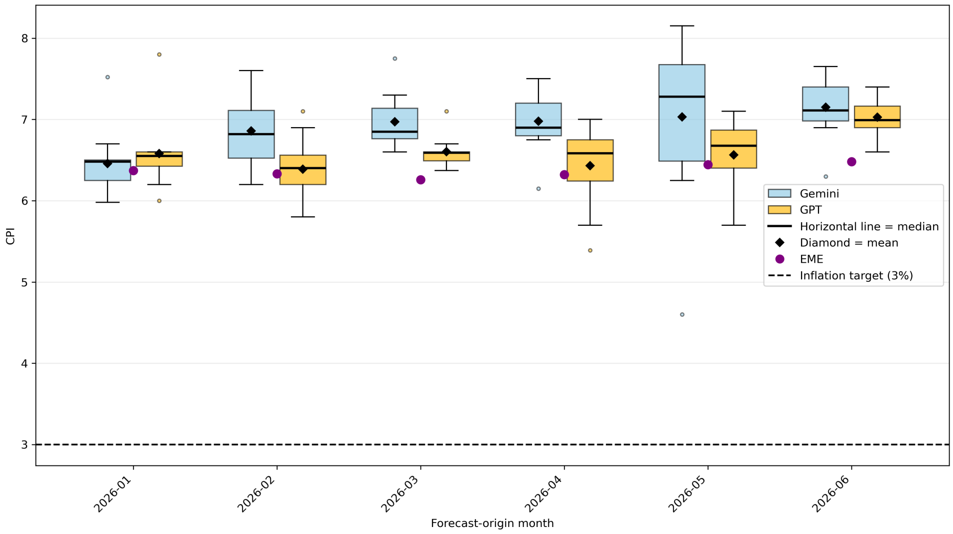

Figure 6: Step 5 forecasts for December 2026 generated between January and June 2026. Purple dots represent EME inflation expectations.[8]

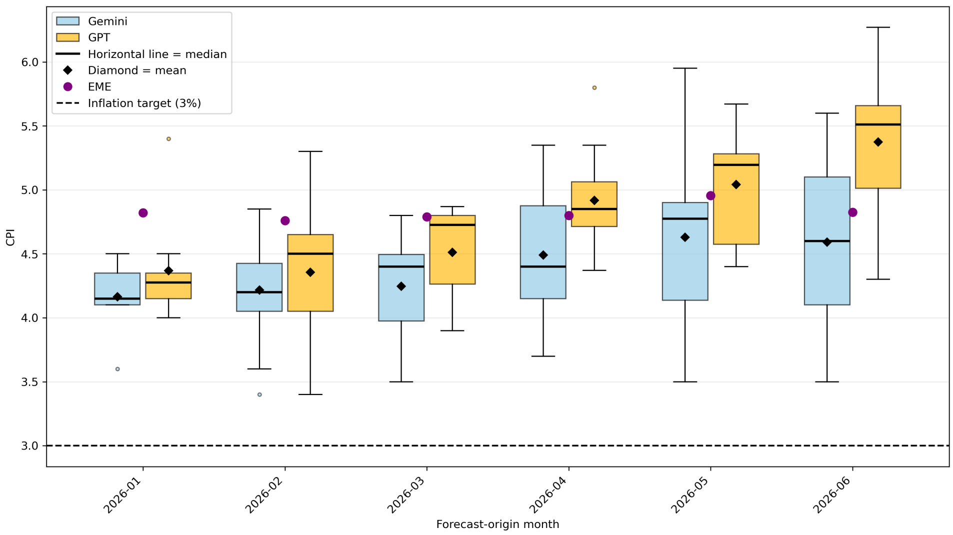

Figure 7: Step 5 forecasts for December 2027 generated between January and June 2026. Purple dots represent EME inflation expectations.[8]

The upward revision in Gemini's and GPT's December 2026 forecasts is consistent with the accumulation and persistence of the domestic and external inflationary pressures identified by Step 5, and with the incoming data, since observed

[7]Each point is a forecast produced on that date using only information available up to that date, so the information set expands as the sample advances.

[8] Experiments were conducted in real time, except for forecasts corresponding to dates before mid-April, which were generated through backward in-sample runs using date restrictions in the prompt.

inflation rose from 5.10% in December 2025 to 5.84% in May 2026 and each run starts from the latest observation. On the domestic side, these include the minimum-wage increase and its pass-through through indexation mechanisms, robust consumption, a positive output gap, fiscal pressures, and elevated inflation expectations. External and cost pressures include higher international food and fertilizer prices, climate-related risks, production and transportation costs, and disruptions to global supply chains. These forces are partially offset by the appreciation of the Colombian peso and restrictive monetary conditions, but the contemporaneous information incorporated by Step 5 suggests that their combined effects remain sufficiently persistent to keep projected inflation elevated through the end of 2026.

Figure 7 presents the corresponding Step 5 forecasts for December 2027. Although projected inflation is substantially lower than for December 2026, it remains above the 3% target. For Gemini, the average of the forecasts produced in January is approximately 4.2%, rising to 4.6% in June. GPT revises considerably more, from an average of 4.4% to 5.3%, so the agreement observed at the one-year horizon does not carry over to the second year. This gradual upward revision is consistent with the possibility that some of the inflationary pressures identified for 2026 have persistent effects that extend into 2027, particularly through wage indexation, second-round effects, and elevated inflation expectations. Measured as the gap between the two horizons, the implied disinflation is essentially unchanged for Gemini, at 2.4 and 2.6 percentage points at the start and the end of the sample, but narrows for GPT, from 2.1 to 1.8, so GPT revised not only the level of inflation but also the speed at which it is expected to recede. At the same time, restrictive monetary policy, weaker demand, and softer external cost pressures support a gradual disinflation process. The resulting path therefore points to a substantial decline in inflation relative to 2026, but not to full convergence to the target by the end of 2027. Although GPT and Gemini produce similar forecasts for December 2026, their projections diverge at longer horizons, suggesting that agreement between models decreases as forecast uncertainty increases over time.

6 Conclusions

As the use of Large Language Models (LLMs) in macroeconomic forecasting continues to grow, there is increasing interest in understanding how these models perform under different information environments. In this context, the main contribution of this paper is the development of a framework for forecasting inflation in Colombia using LLMs. We implement six experiments designed to evaluate how LLMs (GPT-4o, GPT-4.1 and Gemini 2.5 Pro) produce macroeconomic nowcasts and forecasts under different prompt specifications and information environments. The framework evolves from simple forecasting exercises based primarily on inflation dynamics to specifications that incorporate additional macroeconomic variables, contextual indicators, economic relationships, and contemporaneous information. This sequential design allows us to assess how forecasts change as the informational and economic content of the forecasting environment becomes richer.

The empirical results show that the information environment plays a central role in shaping LLM-based inflation forecasts. While Steps 0–3 generate paths that converge relatively quickly toward the 3% target, Steps 4 and 5 produce more persistent and non-monotonic inflation dynamics. Indeed, Step 5 combines economic structure, contemporaneous web-based information, shock identification, and consistency validation to generate jointly determined macroeconomic forecasts that reflect both current conditions and their transmission to inflation. Additionally, forecasts from Step 5 display a broadly similar directional pattern to the Banco de la República's Monthly Survey of Economic Analysts' Expectations (EME), while appearing to incorporate domestic and external inflationary pressures more explicitly than the earlier experimental configurations. These findings suggest that richer and more structured information sets lead LLMs to generate forecasts that better reflect the complexity and persistence of inflationary pressures.

In addition, we report month-by-month updates to the Step 5 forecasts for December 2026 and December 2027 over the January–June 2026 period. Throughout this period, forecasts remained above the 3% inflation target and were progressively revised upward as new information became available. Overall, these revisions are consistent with the persistence of the domestic and external inflationary pressures identified by the framework and point to a gradual disinflation process rather than a rapid return to the target. While the two models produced broadly similar projections for December 2026, their forecasts diverged at longer horizons, suggesting that model agreement weakens as forecast uncertainty increases.

Finally, while the results presented in this study provide encouraging evidence regarding the potential informational value of LLMs for inflation forecasting, they do not yet allow definitive conclusions about the relative forecasting accuracy of the proposed methodology. A key limitation is that the sample of genuine real-time forecasts remains relatively short, constraining the scope for rigorous statistical evaluation. To address this issue, we will continue running the forecasting exercise on a daily basis in order to accumulate a sufficiently large set of real-time forecast vintages and ex post forecast errors. Expanding the evaluation sample will enable a comprehensive assessment of predictive performance using formal statistical tests and a systematic comparison with relevant benchmark models proposed in the literature, including traditional econometric specifications and alternative forecasting approaches. Such comparisons are essential to determine whether the gains suggested by the current results translate into statistically and economically significant improvements in forecasting accuracy. More broadly, a longer real-time evaluation period will allow us to assess the robustness, stability, and reliability of the proposed Step 5 framework across different macroeconomic environments and inflation regimes, thereby providing a stronger basis for evaluating its practical usefulness as a forecasting and policy-support tool.

References

Abdullahi, S., Danyaro, K. U., Zakari, A., Aziz, I. A., Zawawi, N. A. W. A., and Adamu, S. (2025). Time-series large language models: A systematic review of state-of-the-art. *IEEE Access*, 13:30235–30261.

Alam, M. J., Boyle, S., Li, H., and Sekhposyan, T. (2026). Chatmacro: Evaluating inflation forecasts of generative ai. Federal Reserve Bank of San Francisco Working Paper 2026-04, Federal Reserve Bank of San Francisco.

Allard, M.-A., Teiletche, P., and Zinebi, A. (2024). Enhancing inflation nowcasting with llm: Sentiment analysis on news. *arXiv preprint arXiv:2410.20198*.

André, J., Bessec, M., and Goulby, Z. (2025). Llms vs. econometric models for nowcasting gdp growth: A practitioner's view. HAL preprint hal-05128225.

Anesti, N., Hill, E., and Joseph, A. (2025). Inflation attitudes of large language models. *arXiv preprint arXiv:2512.14306*. Version 1 submitted December 16, 2025.

Ang, A., Bekaert, G., and Wei, M. (2007). Do macro variables, asset markets, or surveys forecast inflation better? *Journal of Monetary Economics*, 54(4):1163–1212.

Atkeson, A. and Ohanian, L. E. (2001). Are Phillips curves useful for forecasting inflation? *Federal Reserve Bank of Minneapolis Quarterly Review*, 25(1):2–11.

Bańbura, M., Giannone, D., and Reichlin, L. (2010). Large Bayesian vector auto regressions. *Journal of Applied Econometrics*, 25(1):71–92.

Benny, B. J., James, C. K., and Gupta, H. (2025). Large language models in economic forecasting: A comprehensive analysis of predictive performance and benchmarking against traditional methods for India FY 2025–26. In *2025 5th International Conference on Mobile Networks and Wireless Communications (ICMNWC)*. IEEE.

Bernasconi, M. S., Gómez García, F., and Manzano Quiroga, J. (2025). Ask, think, predict: Llm-based nowcasting of argentina's gdp. Working paper, Universidad de San Andrés.

Carriero, A., Pettenuzzo, D., and Shekhar, S. (2025). Macroeconomic forecasting with large language models. *arXiv prepint arXiv:2407.00890*. Version 4 revised September 23, 2025.

Chang, C., Chen, T.-F., Wang, W.-Y., and Pen, W.-C. (2025). Llm4ts: Aligning pre-trained llms as data-efficient time-series forecasters. *ACM Transactions on Intelligent Systems and Technology*, 16(3).

Chen, J., Tang, G., Zhou, G., and Zhu, W. (2025). Chatgpt and deepseek: Can they predict the stock market and macroeconomy? *Olin Business School Center for Finance Accounting Research Paper No. 2023/18*.

Chib, S. and Tan, F. (2025). Learning the macroeconomic language. *arXiv preprint arXiv:2512.21031*. Version 2 revised December 27, 2025.

Crane, D. L., Karra, A., and Soto, P. E. (2025). Total recall? evaluating the macroeconomic knowledge of large language models. Finance and Economics Discussion Series 2025-044, Board of Governors of the Federal Reserve System, Washington.

de Bondt, G. J. and Sun, Y. (2025). Enhancing gdp nowcasts with chatgpt: A novel application of pmi news releases. ECB Working Paper 3063, European Central Bank.

Fan, Y., Lv, J., Sun, A., and Wang, Y. (2025). Llm-powered cpi prediction inference with online text time series. *arXiv preprint arXiv:2506.09516*.

Faria-e Castro, M. and Leibovici, F. (2024). Artificial intelligence and inflation forecasts. *Federal Reserve Bank of St. Louis Review*, 106(12):1–14. Fourth Quarter 2024.

Gao, T., Sun, A., Wang, Y., Liu, J., and Hsiao, C. (2026). How does llm help regional cpi forecast: An llm-powered deep panel modeling framework. *arXiv preprint arXiv:2604.06894*.

Giannellis, N., Hall, S. G., Kouretas, G. P., Tavlas, G. S., and Wang, Y. (2026). Macroeconomic and financial forecasting in the aftermath of the covid shock: Inflation forecasting with large language models. *Journal of Forecasting*, 45(5):2115–2121.

González, A., Mahadeva, L., Prada, J. D., and Rodríguez, D. (2011). Policy analysis tool applied to Colombian needs: PATACON model description. *Borradores de Economía*, (656).

Gonzalez, D. A. (2026). Forecasting inflation with artificial intelligence: A comparative analysis. *Revista de Métodos Cuantitativos para la Economía y la Empresa*, 41:1–27.

González-Gómez, A., Guarín-López, A., Rodríguez, D., and Vargas-Herrera, H. (2020). 4gm: A new model for the monetary policy analysis in Colombia. *Borradores de Economía*.

Gruver, N., Finzi, M., Qiu, S., and Wilson, A. G. (2023). Large language models are zero-shot time series forecasters. *arXiv preprint arXiv:2310.07820v3*.

Halawi, D., Zhang, F., Yueh-Han, C., and Steinhardt, J. (2024). Approaching human-level forecasting with language models. In *Advances in Neural Information Processing Systems 37*. Neural Information Processing Systems Foundation.

Hansen, A. L., Horton, J. J., Kazinnik, S., Puzzello, D., and Zarifhonarvar, A. (2026). Simulating the survey of professional forecasters. *SSRN Electronic Journal*. Working paper, revised January 22, 2026.

Iadisernia, G. and Camassa, C. (2025). Prompting for policy: Forecasting macroeconomic scenarios with synthetic llm personas. *arXiv preprint arXiv:2511.02458*. Accepted at ICAIF 2025.

Jaworski, K. (2026). Ai-driven inflation forecasting in the aftermath of covid-19. *Journal of Forecasting*, 45(5):2525–2548.

Jin, M., Wang, S., Ma, L., Chu, Z., Zhang, J. Y., Shi, X., Chen, P.-Y., Liang, Y., Li, Y.-F., Pan, S., and Wen, Q. (2024). Time-llm: Time series forecasting by reprogramming large language models. In *Proceedings of the 12th International Conference on Learning Representations*.

Joseph, A., Potjagailo, G., Chakraborty, C., and Kapetanios, G. (2024). Forecasting UK inflation bottom up. *International Journal of Forecasting*, 40(1):1521–1538.

Karger, E., Bastani, H., Yueh-Han, C., Jacobs, Z., Halawi, D., Zhang, F., and Tetlock, P. (2025). Forecastbench: A dynamic benchmark of AI forecasting capabilities. In *The Thirteenth International Conference on Learning Representations*.

Kaur, R., Liu, J., and Sim, C. (2025). Validating large language models (llms) for economic and inflation forecasting in aotearoa new zealand's public sector. In Varastehpour, S. and Shakiba, M., editors, *Proceedings: AIOT Global Summit 2025: Economic Growth, 15–16 July*, pages 93–99. ePress, Unitec.

Kim, A., Muhn, M., and Nikolaev, V. (2024). Financial statement analysis with large language models. *arXiv preprint arXiv:2407.17866*. Withdrawn working paper; version 3 revised February 20, 2025.

Kirtac, K. and Germano, G. (2024). Sentiment trading with large language models. *Finance Research Letters*, 62:105227.

Koyuncu, B., Kwon, B., Lombardi, M. J., Perez-Cruz, F., and Shin, H. S. (2026). Bistro: A general purpose oracle for macroeconomic time series. *BIS Quarterly Review*. March 2026.

Kwon, B., Park, T., Rungcharoenkitkul, P., and Smets, F. (2025). Parsing the pulse: Decomposing macroeconomic sentiment with llms. BIS Working Paper 1294, Bank for International Settlements. October 2025.

Lin, J., Sun, L., and Yan, Y. (2025). Simulating macroeconomic expectations in survey experiments with llm-based economic agents. *arXiv preprint arXiv:2505.17648*. Version 5 revised June 1, 2026.

Liu, H., Zhao, Z., Wang, J., Kamarthi, H., and Prakash, B. (2024). Lstprompt: Large language models as zero-shot time series forecasters by long-short-term prompting. *arXiv preprint arXiv:2402.16132v1*.

Lopez-Lira, A., Tang, Y., and Zhu, M. (2025). The memorization problem: Can we trust llms' economic forecasts? *arXiv preprint arXiv:2504.14765*. Version 2 revised December 15, 2025.

Ludwig, J., Mullainathan, S., and Rambachan, A. (2024). Large language models: An applied econometric framework. *arXiv preprint arXiv:2412.07031*. Version 4 revised December 9, 2024.

Martínez-Rivera, W., González-Molano, E., and Caicedo-Garcia, E. (2025). Forecasting inflation from disaggregated data. *Applied Stochastic Models in Business and Industry*, 41(4):e70023.

Medeiros, M. C., Vasconcelos, G. F. R., Veiga, Á., and Zilberman, E. (2021). Forecasting inflation in a data-rich environment: The benefits of machine learning methods. *Journal of Business & Economic Statistics*, 39(1):98–119.

Nie, Y., Kong, Y., Dong, X., Mulvey, J. M., Poor, H., Wen, Q., and Zohren, S. (2024). A survey of large language models for financial applications: Progress, prospects and challenges. *arXiv preprint arXiv.2310.07820*.

Osborn, S., Valecha, R., Rao, H. R., Sass, D., and Rios, A. (2025). How ai forecasts ai jobs: Benchmarking llm predictions of labor market changes. *arXiv preprint arXiv:2510.23358*.

Paleka, D., Goel, S., Geiping, J., and Tramèr, F. (2025). Pitfalls in evaluating language model forecasters. *arXiv preprint arXiv:2506.00723v1*.

Pandey, A. K., Al Murshidi, G., Ali, A., and Al Zaabi, A. (2025). When ai meets economics: A hybrid framework for inflation prediction using nlp and llms. *Data Plus*, 3(2):81–96.

Ramos-Veloza, M. A., Orozco Rios, D. M., and Muñoz Martínez, J. A. (2024). Tweeting inflation: Real-time measures of inflation perception in Colombia. *Latin American Economic Review*, 33:1–39.

Schoenegger, P., Jones, C. R., Tetlock, P. E., and Mellers, B. (2025). Prompt engineering large language models' forecasting capabilities. *arXiv preprint arXiv:2506.01578*.

Stock, J. H. and Watson, M. W. (2007). Why has U.S. inflation become harder to forecast? *Journal of Money, Credit and Banking*, 39(s1):3–33.

Su, J., Jiang, C., Jin, X., Xiao, T., Ma, H., Wei, R., Jing, Z., Xu, J., and Lin, J. (2024). Large language models for forecasting and anomaly detection: A systematic literature review. *arXiv*, abs/2402.10350.

Tan, M., Merrill, M. A., Gupta, V., Althoff, T., and Hartvigsen, T. (2024). Are language models actually useful for time series forecasting? *arXiv preprint arXiv:2406.16964v2*. Version 2 revised October 26, 2024.

Tawichsri, T., Tuarob, S., Nookhwun, N., and Sangasaeng, C. (2026). News-based inflation expectations: Llm-assisted measurement and forecasting. PIER Discussion Paper 252, Puey Ungphakorn Institute for Economic Research. May 2026.

Wu, J. C., Xi, J., and Xie, S. (2025). Llm survey framework: Coverage, reasoning, dynamics, identification. NBER Working Paper 34308, National Bureau of Economic Research.

Xue, H. and Salim, F. D. (2023). Promptcast: A new prompt-based learning paradigm for time series forecasting. *arXiv preprint arXiv:2210.08964v5*. Version 5 revised December 10, 2023.

Yin, M. and Guo, M. (2026). The emotional drive of economic forecasting: Gdp predictions based on large language models. *Computational Economics*, 67(5).

Zarifhonarvar, A. (2026). Generating inflation expectations with large language models. *Journal of Monetary Economics*, 157:103859.

A Detailed Description of the Experimental Forecasting Configurations

Step 0: This step defines the baseline univariate forecasting exercise. The LLM is instructed to retrieve the most recent headline CPI inflation information from a predefined set of official sources, with DANE as the primary source, and to verify that the latest observation is current and consistent with the official release. Once the starting information has been validated, the model generates a month-by-month inflation forecast over a 24-month horizon. The forecast is based primarily on recent inflation dynamics, while allowing monetary policy considerations and convergence toward the 3 percent inflation target to become increasingly relevant over longer horizons. This configuration therefore evaluates the model when it must both retrieve and interpret the information used to construct the forecast.

Step 1: This step retains the same univariate forecasting task as Step 0 but removes the information-retrieval stage. The historical headline inflation series is supplied directly to the LLM, and the use of external web information is explicitly prohibited. The model is instructed to identify relevant features of the historical series, including trends, persistence, cycles, seasonality, and potential structural changes, and to use these patterns to generate the 24-month forecast. Step 1 therefore provides a more controlled setting in which the information available to the LLM is restricted to the supplied inflation history.

Step 2: This step retains the historical-data-based approach and the restriction on external information introduced in Step 1, but extends the forecasting system from inflation alone to four macroeconomic variables. The LLM is instructed to examine the historical dynamics of each series and their cross-variable relationships before constructing a common macroeconomic scenario. Inflation, GDP growth, the exchange rate, and the policy rate are then forecast jointly rather than as separate exercises, with the prompt explicitly requiring the four projected paths to be mutually consistent. Thus, the main methodological extension is the transition from a univariate forecast to a jointly determined multivariate forecasting system.

Step 3: The joint multivariate framework of Step 2 is retained, but the information set is expanded with the contextual variables reported in Table 1. Before generating the forecasts, the LLM performs a relevance assessment in which it identifies the most informative indicators within each group of contextual variables for each of the four forecast variables and considers the corresponding transmission mechanisms. These additional variables are not themselves forecast; instead, the signals extracted from them are used to condition the common macroeconomic scenario. The four forecast paths are then generated jointly under the same mutual-consistency requirement as in Step 2. This configuration therefore isolates the contribution of a richer macroeconomic information set while preserving the underlying multivariate forecasting structure.

Table 3: Dictionary of Thematic Expressions

El Niño Phenomenon	**International Food Prices**	**Credit Rating**
probability El Niño	rising international food	sp sovereign credit
prospective El Niño	rising international fao	sovereign credit rating
El Niño event	international food fao	sovereign credit downgrades
El Niño probability	international food index	sp credit rating
moderate-to-strong El Niño phenomenon	international food price	credit rating downgrades
strong El Niño phenomenon	international food outlook	downgrades sp fitch
high El Niño phenomenon	international food hike	downgrades sp bb-
prospective EL Niño phenomenon	international food falling	downgrades sp moodys
probability El Niño phenomenon	rising international cereals	rating downgrade bb
	international fertilizer prices	
Fed & Global Financial Conditions	**Fiscal Deficit**	**Inflation Expectations**
fed persistent inflation	significant fiscal slippage	persistent inflation expectations
hawkish fed stance	fiscal slippage deficit	de-anchoring inflation expectations
leading fed stance	fiscal slippage risk	significant inflation expectations
fed stance maintaining	expansionary fiscal stance	rising inflation expectations
fed stance reaccelerating	expansionary fiscal suspended	inflation expectations analyst
fed stance sticky	expansionary fiscal policy	inflation expectations risk
persistent hawkish fed	expansionary fiscal uncertainty	significant de-anchoring inflation
hawkish fed hold	expansionary fiscal risk	de-anchoring inflation analyst
hawkish fed market	fiscal stance uncertainty	rate inflation expectations
hawkish fed dollar	persistent fiscal deficit	inflation expectations signaling
hawkish fed expectations	large fiscal deficit	inflation expectations banrep
hawkish fed policy	carf	inflation expectations forecasts
hawkish fed rates	escape clause	medium-term inflation expectations
Inflation Persistence	**Local Monetary Policy and Financial Conditions**	**Minimum Wage**
restrictive persistent inflation	inflation monetary policy	nominal minimum wage
strong persistent inflation	hawkish monetary policy	large minimum wage
remains persistent inflation	restrictive monetary policy	massive minimum wage
confirms persistent inflation	monetary policy tightening	historic minimum wage
persistent inflation core	monetary policy rate	extraordinary minimum wage
persistent inflation elevated	monetary policy transmission	massive wage hike
persistent inflation rate	policy stance rate	effects minimum wage
persistent inflation central	interest rate banrep	record minimum wage
indexation past inflation	cycle policy rate	wage hike inflation
price past inflation	conditions credit growth	wage hike jan
broad past inflation	dynamic credit growth	unprecedented minimum wage
inertia past inflation	credit growth consumption	shock wage hike
transport past inflation	credit growth demand	significant minimum wage
services past inflation	credit growth domestic	
past inflation minimum	resilient credit growth	
past inflation wage	stable credit growth	
broad-based indexation mechanisms	credit growth aggregate	

Note: The multiple n-grams reported in this table result from the combination of different expressions associated with each thematic category. First, all n-grams were extracted from the model-generated shocks. These regular expressions include alternative expressions, synonyms and lexical variations related to the main topic. For expositional purposes, the table reports only the most representative expressions within each category; the full set contains additional terms that are essentially lexical variations or close formulations of the expressions displayed here.

Then, to avoid double counting, each n-gram was classified using a boolean detection function (*str_detect* in R), which determines whether an n-gram matches at least one expression within a given category. Therefore, regardless of how many expressions in the regular expression match a particular n-gram, it is assigned to a category only once. This procedure ensures that each n-gram contributes a single observation to the thematic frequency counts while allowing different lexical formulations of the same context to be grouped under a common category.

Table 4: Dictionary of Thematic Expressions (continued)

Oil, Gas and Fertilizer Prices	**Economic Activity**
high oil prices	positive output gap
moderately oil prices	persistent output gap
volatile oil prices	demand output gap
oil prices brent	iseyoy output gap
oil prices brent opec+	resilient domestic demand
international fertilizer prices	strong domestic demand
index fertilizer prices	growth domestic demand
oil price brent	domestic demand consumption
domestic natural gas	risks weak investment
natural gas price	growth weak investment
imported natural gas	output gap household
natural gas electricity	robust domestic demand
high fertilizer prices	output gap identified
Shipping Costs and Supply Chains	**Exchange Rate**
global supply chain	strong carry trade
costs supply chain	cop carry trade
freight supply chain	capital carry trade
shipping supply chain	inflows carry trade
supply chain pressures	remittances carry trade
supply chain disruptions	carry trade significant
supply chain gscpi	carry trade disinflationary
freight costs geopolitical	interest rate differential
chain pressures maritime	high rate differential
chain pressures shipping	wide rate differential
global shipping costs	cop appreciation
shipping costs geopolitical	currency appreciation
global shipping geopolitical	currency depreciation
supply chain container	

Note: The multiple n-grams reported in this table result from the combination of different expressions associated with each thematic category. First, all n-grams were extracted from the model-generated shocks. These regular expressions include alternative expressions, synonyms and lexical variations related to the main topic. For expositional purposes, the table reports only the most representative expressions within each category; the full set contains additional terms that are essentially lexical variations or close formulations of the expressions displayed here.

Then, to avoid double counting, each n-gram was classified using a boolean detection function (*str_detect* in R), which determines whether an n-gram matches at least one expression within a given category. Therefore, regardless of how many expressions in the regular expression match a particular n-gram, it is assigned to a category only once. This procedure ensures that each n-gram contributes a single observation to the thematic frequency counts while allowing different lexical formulations of the same context to be grouped under a common category.