Reddit's Pulse on US Inflation: Forecasting with Large Language Models*

Andrea Del Monaco[1], Luigi Longo[2], Juri Marcucci[1], and Irene Tafani[3]

[1]Bank of Italy
[2]JRC - European Commission
[3]IMT School Lucca

2025-03-20 (19:59:48)

Abstract

This paper employs Large Language Models (LLMs) to extract meaningful signals from Reddit's social media data and forecast U.S. inflation. We focus on popular economics-related subreddits, constructing indexes that capture users' interests and perceptions of inflation and price dynamics. To derive relevant information from social media conversations, we fine-tune pre-trained LLMs (including BERT, Qwen, LLaMA or Gemma) using labels generated by human annotators and ChatGPT. Our results demonstrate that indicators derived from fine-tuned LLMs significantly improve the point and density forecasting performance of time series models for U.S. inflation.

JEL: E31, C32, C53, C55

Keywords: Economic Forecasting, Social Media, Reddit, Inflation, Text Mining, Text-as-data, Text Analysis, Natural Language Processing, Sentiment Analysis, Big Data, Large Language Models, ChatGPT, Generative Artificial Intelligence.

*We gratefully acknowledge comments from the participants of the 6^{th} Conference on "Nontraditional Data, Machine Learning, and Natural Language Processing in Macroeconomics - ECONDAT 2024 Fall Meeting", and the 2025 ECB FIAT and AI in Economics workshop on "Harnessing Artificial Intelligence for inflation assessment".
The views and ideas expressed herein are those of the authors and do not necessarily reflect those of the Bank of Italy, the Eurosystem, or the European Commission. All remaining errors are our own. Corresponding author: Luigi Longo, email: luigi.longo@ec.europa.eu.

1 Introduction

Accurate forecasting of macroeconomic variables, such as inflation, is essential for policymakers, investors, and economic analysts. Traditional economic forecasting methods primarily rely on historical data and established economic indicators derived from surveys run by government agencies. However, the increasing prevalence and influence of social media platforms present a real-time and rich source of data that can complement and enhance conventional forecasting models (see for example Angelico et al. (2022).

This paper introduces a novel approach to economic forecasting by leveraging the vast and dynamic textual data available on Reddit, a widely used social media platform. Our study highlights the value of integrating modern computational techniques, big data, and artificial intelligence with traditional economic and statistical and analysis.

specifically, we explore the potential of Reddit discussions to nowcast and forecast inflation in the United States. Reddit hosts numerous subreddits dedicated to economic discourse, where users frequently share their views and concerns regarding inflation and macroeconomic trends. By analyzing these discussions, we aim to extract text-based indicators that reflect public sentiment and interest in inflation, potentially serving as early signals of inflation dynamics.

To achieve this, we employ a combination of standard text-mining techniques (such as simple word counts and sentiment indicators) and advanced Large Language Models (LLMs). Our methodology integrates BERT-based models alongside LLaMA 3, Gemma 2, and Qwen 2.5, which are used to process and classify subreddit posts related to inflation and price dynamics. These models are fine-tuned using labels generated by conversational LLMs - ChatGPT and LLaMA 3 - through a zero-shot learning approach, later reviewed and refined by human annotators to enhance accuracy. In particular, leveraging open-source models enables us to train specialized LLMs, fine-tuning general-purpose models on Reddit discussions about inflation to improve their context awareness.

Our empirical results demonstrate that LLM-derived indicators significantly enhance the forecasting performance of traditional time series models. An autoregressive (AR) model augmented with a text-based indicator from Reddit outperforms a simple AR benchmark in out-of-sample forecasting accuracy across horizons ranging from one to 18 months ahead, over the period 2019m1-2023m12. Moreover, we show that LLM-derived text-based indicators improve both point and density forecasts, particularly for short-erm predictions. We further validate these results by assessing the marginal contribution of these indicators to the predictive model's likelihood function. Additionally, we find that fine-tuning LLMs to lavel domain-specific text enhances accuracy compared to using only pre-trained models, while also reducing the uncertainty of the extracted educated signal.

Our findings indicate that textual indicators from certain Reddit communities can an-

ticipate inflation fluctuations between six and twelve months in advance. Furthermore, we document heterogeneity across Reddit communities, suggesting that different subreddits capture distinct dimensions of inflation-related discourse.

The remainder of this paper is structured as follows. Section 2 reviews recent research on machine learning approaches and text-as-data applications in economic forecasting. Section 3 introduces the Reddit dataset used in this study and provides a descriptive analysis. Section 4 presents the large language models employed and details the fine-tuning methodology. Section 5 outlines the strategy used to extract an educated inflation signal from both pretrained and fine-tuned LLMs. In Section 6, we present the forecasting results and assess the predicitve performance of Reddit-based indicators. Finally, Section 7 concludes the paper and discusses directions for future research, particularly on the role of social media and LLMs in economic forecasting, social science research, and policy analysis.

2 Previous studies in forecasting with textual and social network data

The literature on extracting insights from textual data can broadly be divided into Natural Language Processing (NLP) and Generative AI/LLMs. NLP techniques serve as the foundation for more advanced generative models, such as LLMs, evolving from basic sentiment analysis to sophisticated language interpretation in diverse contexts.

In empirical macroeconomics and in macroeconomic forecasting, NLP methods have been extensively applied to extract structured indicators from unstructured data sources, such as policy speeches, articles, and social media platforms (see Marcucci (2024) for more details on forecasting macroeconomic variables with text-as-data). Over the past two decades, a growing body of research has utilized such data to construct macroeconomic indicators for forecasting and structural analysis.

One of the most influential contributions to this literature is by Baker et al. (2016), who developed the Economic Policy Uncertainty (EPU) index using word counts related to uncertainty in ten major U.S. newspapers. This index laid the groundwork for numerous applications of NLP in macroeconomic research. More recently, Angelico et al. (2022) leveraged Twitter/X data to construct a high-frequency measure of consumer inflation expectations in Italy. By combining Latent Dirichlet Allocation (LDA) with a dictionary-based approach (using bi-grams and tri-grams), they developed a measure closely aligned with survey expectations and professional forecasts.

Aruoba and Drechsel (2024) applied sentiment analysis using the Loughran and McDonald (2011) dictionary to measure Central Bank sentiment from policy documents, producing uncertainty indicators inspired by Baker et al. (2016). Similarly, Azqueta-Gavaldón et al.

(2023) applied LDA to European newspaper articles, achieving comparable results in detecting uncertainty indicators.

Beyond social media, textual resources like newspapers have proven valuable in economic forecasting, as highlighted by Marcucci (2024). For instance, Aprigliano et al. (2023) introduced a high-frequency Text-based Economic Sentiment Index (TESI) and a Text-based Economic Policy Uncertainty (TEPU) for Italy. These monthly indicators significantly reduced uncertainty in short-term forecasts of key macroeconomic aggregates and improved forecasting accuracy when combined with GDP data. In the UK context, Kalamara et al. (2022) used text from three major newspapers to improve macroeconomic forecasts for variables like GDP, inflation, and unemployment. Their findings underscore the effectiveness of combining term counts with supervised machine learning, especially during periods of economic stress. Similarly, Barbaglia et al. (2022) employed fine-grained, aspect-based sentiment analysis on nearly seven million U.S. newspaper articles. Their results demonstrated that economic sentiment tracks business cycle fluctuations and improves forecasts, particularly in capturing the tails of probability distributions.

Applications extend beyond macroeconomic forecasting. For example, Renault (2017) showed that sentiment analysis could predict intraday stock returns, outperforming both standard dictionary-based methods and advanced machine learning techniques.

Recent advancements in LLMs have opened new frontiers in economic forecasting, providing innovative tools and methodologies. Researchers increasingly use social media data to forecast trends, as illustrated by Gueta et al. (2024), who predicted financial outcomes from narratives extracted from tweets. Similarly, Carriero et al. (2024) demonstrated that Time Series Language Models (TSLMs) could rival established econometric models like Bayesian VARs and factor models, even without fine-tuning on economic data. Faria-e Castro and Leibovici (2024) highlighted the potential of LLMs like PaLM to produce inflation forecasts as accurate as traditional sources like the Survey of Professional Forecasters (SPF), offering cost-effective alternatives where survey data are lacking. Additionally, Bybee (2023) showcased how LLMs can be used to create proxies for expectations and beliefs, opening new avenues for research into human-like belief formation.

LLMs have also been applied in behavioral studies. For example, Horton (2023) demonstrated that GPT-3 could replicate human-based experimental findings at significantly lower costs. LLMs enable testing numerous variations, such as changes in wording or prompt order, with virtually unlimited sample sizes, all while eliminating ethical concerns associated with human subjects.

The importance of fine-tuning LLMs for domain-specific tasks has gained increasing attention. Fine-tuning involves manually labeling a subset of textual entries and training the model to optimize its predictive accuracy. For example, Antweiler and Frank (2004) analyzed over 1.5 million messages from Yahoo! Finance and Raging Bull to predict stock

market volatility. Using a naive Bayes classifier trained on a subset of 1,000 manually labeled messages, they found that disagreement in stock messages correlates with increased trading activity. Similarly, Shapiro et al. (2022) employed human-labeled articles to evaluate machine learning classifiers, demonstrating that BERT outperformed simpler models such as Bag-of-Words and GloVe.

Recent work by Audrino et al. (2024) utilized fine-tuning to construct an uncertainty index similar to the EPU. As noted by Ash and Hansen (2023), pre-trained LLMs can achieve high performance on specific tasks using relatively small labeled datasets, highlighting the importance of fine-tuning for adapting general models to specialized applications.

3 Reddit Data

We utilize textual data from Reddit, a social media platform and discussion website where registered users (known as '*redditors*') can submit and vote on content.[1] Reddit comprises thousands of communities called subreddits, each focused on a specific topic or theme, denoted by '*r/Subreddit_Name*'. Users can subscribe to subreddits of interest to view submissions, analogous to tweets or posts in Twitter/X terminology, related to those topics.

Submissions on Reddit vary in format, including text posts, links to external websites, images, and videos. Other users can comment on these submissions, and comments can, in turn, receive replies, creating threaded discussions or comment trees (Choi et al., 2015). Submissions and comments can also be upvoted or downvoted, influencing their visibility on the platform. The net difference between upvotes and downvotes determines the 'score' of a submission or comment. Additionally, Reddit assigns '*karma points*' to users based on the popularity of their submissions and comments, serving as a form of social capital and reflecting their contributions to the community.

To identify relevant data for our study, we focused on the most populated economic subreddits, particularly those discussing inflation and price dynamics. These include r/Economics (4.3M members as of July 2024), r/wallstreetbets (14M members as of July 2024) r/Economy (1M members), r/AskEconomics (1M members), r/EconomicHistory (949k members), and r/badeconomics (768k members). For our analysis, we selected the largest subreddits, r/Economics, and r/wallstreetbets, and included r/Economy for robustness.[2]

We collected data using PRAW (Python Reddit API Wrapper), a versatile library for

[1] According to `https://foundationinc.co/lab/reddit-statistics/` (accessed December 31, 2024), 26.4 million Americans use Reddit monthly. Reddit boasts 330 million monthly active users, with 42% of U.S. internet users aged 18 to 24 actively engaging on the platform. It is also the 6th most popular website in the U.S.

[2] We excluded smaller subreddits such as r/EconMonitor (39k members) and r/EconInflation (94 members) due to their limited user bases.

interfacing with the Reddit API. PRAW simplifies data retrieval by offering user-friendly tools to access posts, comments, and user information from various subreddits. Its features include searching for specific topics, tracking subreddit activity, and collecting data based on customized criteria. Built-in pagination and rate limit management facilitate the extraction of large datasets, making it an essential tool for analyzing Reddit discussions.

In our study, we differentiate between submission data and comment data due to their distinct attributes. Submissions include metadata such as title, author, date, number of comments, score (net likes), and body text. Comments, while lacking certain submission-specific attributes (e.g., title, number of comments), still include scores reflecting their popularity.[3]

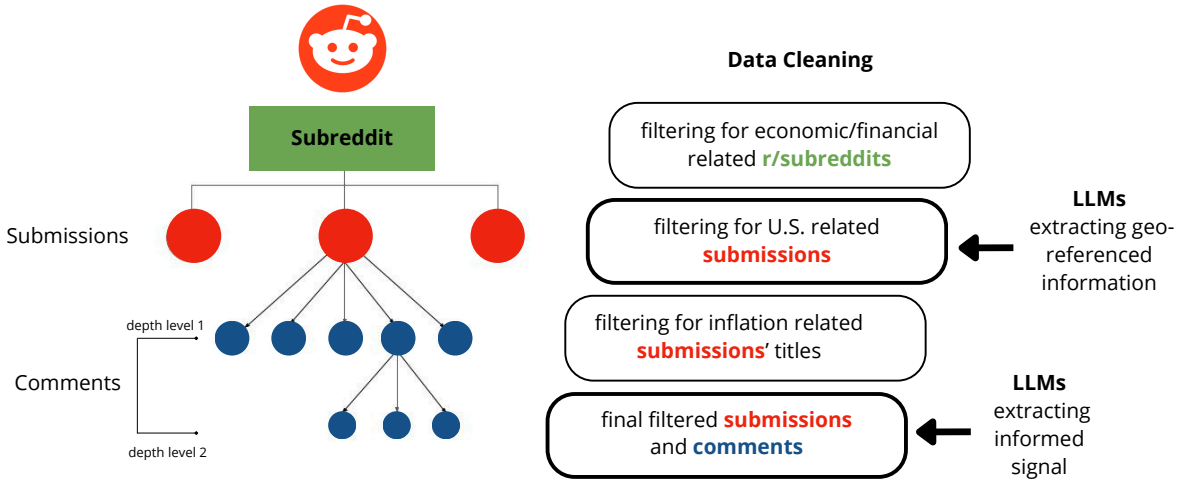


Figure 1: Left: The hierarchical structure of Reddit discussions. Subreddits (green) contain submissions (red), which generate comment threads (blue). The tree structure illustrates how discussions evolve within a subreddit, with submissions as the root and comments branching out as responses. **Right**: An overview of the LLM-based approach, depicting the multi-step filtering process applied to Reddit data.

Figure 1 illustrates the structure of Reddit and how our methodology integrates with social media data. Reddit follows a hierarchical structure: at the top level, it is divided into distinct subreddits, each dedicated to a specific topic or interest. Within each subreddit, users create submissions that serve as the core content.

Submissions receive comments from other users, including the original author, initiating discussions. Each comment adds a new layer of interaction, forming a hierarchical comment tree. Users can respond to both the original submission and individual comments, creating

[3]For a detailed overview of submission and comment features, refer to the Reddit PRAW documentation at `https://praw.readthedocs.io/en/stable/`.

branching discussions of varying depth. This structure facilitates complex, multi-level conversations, where each level represents a reply to the previous one.

To extract relevant signals from Reddit data, our approach integrates LLMs with a keyword-based filtering strategy to efficiently manage and reduce the dataset's dimensionality while preserving the relevance of the extracted information. The analysis begins by selecting content from subreddits focused on economics and finance, particularly discussions related to inflation. To further refine the dataset, an LLM-based classification model assigns country labels to each post, filtering for content specifically related to the United States. A keyword-based approach is then applied to post titles to retain only inflation-related submissions. Finally, fine-tuned proprietary LLMs are used to extract meaningful signals from the data. Throughout this process, both submissions and comments are analyzed to capture the social dimension of user interactions. The following sections provide a detailed description of each subreddit, the data-cleaning methodology, and descriptive statistics.

3.1 The subreddits r/Economics and r/economy

In this paper, we primarily use textual data from the subreddit r/Economics, the largest economics-focused community on Reddit. Created in January 2008, r/Economics serves as a central forum for discussing and learning about economic topics. As of July 10, 2024, the subreddit had 4,337,991 members. Its description reads:[4]

"The Dismal Science. Reddit's largest economics community. Serving as a central forum for users to read, discuss, and learn more about topics related to the economic discipline. We have written rules to support this aim and welcome those who want to learn and those who want to contribute. We aim to foster an environment where everybody feels safe and welcomed and where people feel encouraged to have healthy and productive discussions. Created Jan 25, 2008."

Table 1 provides examples of submission titles from r/Economics. These examples highlight the potential to extract valuable insights about price dynamics and economic trends directly from the titles, as curated by the community moderators and contributors. As a robustness check, we also incorporated data from r/Economy, another prominent subreddit focused on economic discussions, which had approximately 1 million members as of July 10, 2024.[5] Its description states:

"Forum for economy, business, politics, stocks, bonds, product releases, IPOs, advice, news, investment, videos, predictions, government, money, politics, debate, capitalism, current trends, and more."

[4]Retrieved from `https://www.reddit.com/r/Economics/`

[5]Retrieved from `https://www.reddit.com/r/economy/`

For both subreddits, we collected submissions and associated comments spanning the period from January 2008 to December 2023.

Table 1: Examples of Titles from r/Economics subreddit

Author	Title	Score	Date
Mighty_L_LORT	Crop prices soar to 8-year high, renewing food inflation fears	1.0	2021-07-14 17:16:27
scolfin	The mystery of rising prices. Are greedy corporations to blame for inflation?	1.0	2022-06-12 12:05:34
PreppingKangaroo	Why Is Inflation So High?	1.0	2022-05-30 13:01:54
reggie_morris	Food Prices Are New Inflation Threat for Governments and Central Banks	1.0	2023-07-12 09:25:14
nosotros_road_sodium	Inflation has cooled a lot. So why do things still feel so expensive?	1.0	2024-01-21 15:10:01
leopoldnick	Inflation Accelerated in March Due to Strengthening Economy, Rising Energy Prices	1.0	2021-06-05 22:05:46
NewHights1	Why Are Prices So High? Blame Supply Chain – And That's The Reason Inflation Is Here To Stay	1.0	2021-06-23 18:02:47
magenta_placenta	U.S. inflation rate rises to 13-year high of 5.4%	1.0	2021-05-04 19:08:58
Theauntgate	Jerome Powell believes inflation will be high until 2022 and sees rates hike premature	1.0	2021-03-24 16:01:05

3.2 The subreddit r/wallstreetbets

In addition to r/Economics and r/economy, we also include textual data from the subreddit r/wallstreetbets as part of our analysis. This subreddit was created in January 2012 and represents a unique community focused on financial markets, individual stock trading, and high-risk investment strategies. As of July 10, 2024, r/wallstreetbets has 14,045,881 members.

r/wallstreetbets gained significant attention during the GameStop short squeeze in early 2021, where the collective actions of its members demonstrated the power of retail investors to influence stock prices. As noted in Long et al. (2023), the subreddit has a distinct linguistic style characterized by community-specific slang, abbreviations, and memes, which makes traditional dictionary-based text analysis challenging (this additionally justifies the usage of LLMs). However, this linguistic uniqueness also provides an opportunity to capture nuanced signals about financial market sentiment and expectations.

Table 2 presents the number of submissions and comments for each subreddit before and after the data-cleaning process. For each subreddit, we define a relevant time range corresponding to the data used in our empirical evaluation. Initial observations with minimal social activity were excluded, as they provided limited value for signal extraction. As we are using different useful time range for some of the subreddit, the results in the forecasting application of 6 might slightly change for the same model because the in-sample estimation window differs.

3.3 Data preprocessing and descriptive analysis

This paper systematically analyzes textual data derived from both Reddit submissions and a substantial portion of their associated comments. Comments often provide richer informa-

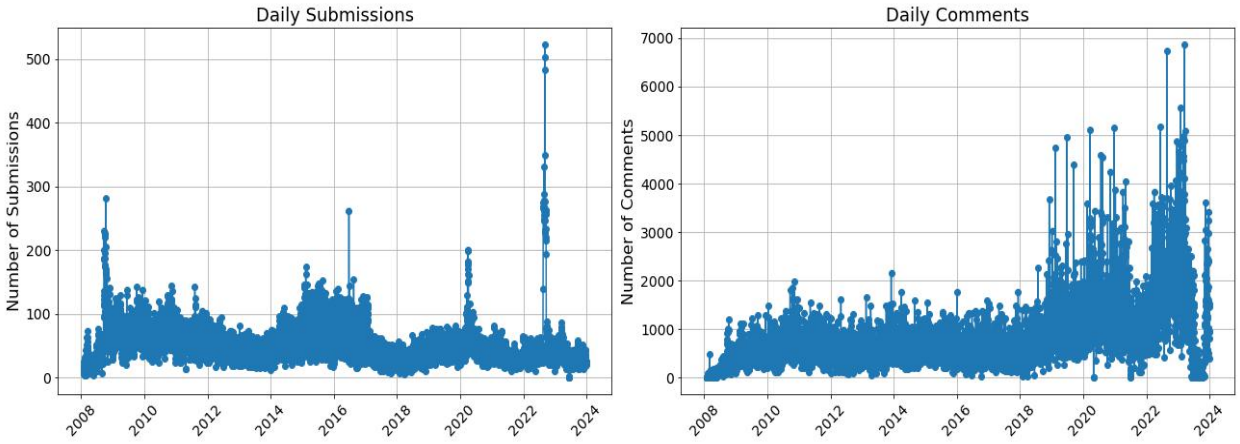
Table 2: Summary statistics for submissions and comments for each subreddit.

Subreddit	time range	useful time range	submissions	comments	submissions (inflation)	comments (inflation)
r/Economics	2008m1-2023m12	2009m1-2023m12	327,370	5,395,099	8,929	57,898
r/economy	2008m3-2023m12	2009m1-2023m12	238,884	2,244,968	9,986	35,079
r/wallstreetbets	2012m4-2023m12	2015m1-2023m12	2,349,125	77,590,749	4,708	55,610

tion due to their granularity and potential insights. This study focuses on submissions and comments from the r/Economics, r/wallstreetbets, and r/Economy subreddits.

The large volume of available data played a key role in shaping our methodology. As shown in Figure 2, the number of submissions and comments in the r/Economics subreddit, as well as across Reddit more broadly, has grown steadily over time. Notably, comments increased significantly after 2018, with sharp spikes in submissions between 2022 and 2023. During these periods, daily comment counts exceeded 7,000, while daily submissions surpassed 500.

Figure 2: Daily counts of submissions and comments in the subreddit r/Economics.



Given the complexity and size of the Reddit dataset, a rigorous data processing approach was essential to extract information relevant to our research objectives.

To refine our dataset, we first filtered submissions that explicitly mention or reference inflation-related topics. Next, to ensure the geographical relevance of the data, we used an LLM to geo-locate each submission's title, categorizing content based on both explicit and implicit country references. Submissions related to the United States or without a specific geographical reference were retained, while those explicitly linked to other countries were excluded.

Using the filtered dataset, we conducted a term analysis by generating a word cloud to

visually identify the most frequently mentioned words in the comments section. As shown in Figure 3, prominent terms such as "US," "money," "America," and "inflation" emerged, while common verbs like "find" or "ongoing" were excluded. This visualization offers a concise and informative overview of the central topics discussed within the subreddits.

Figure 3: Wordcloud subreddit r/Economics.



The final step of our data cleaning process was to retain only submissions and related comments containing specific inflation-related keywords:[6]

{*inflation, price, prices, deflation, disinflation, hyperinflation*}

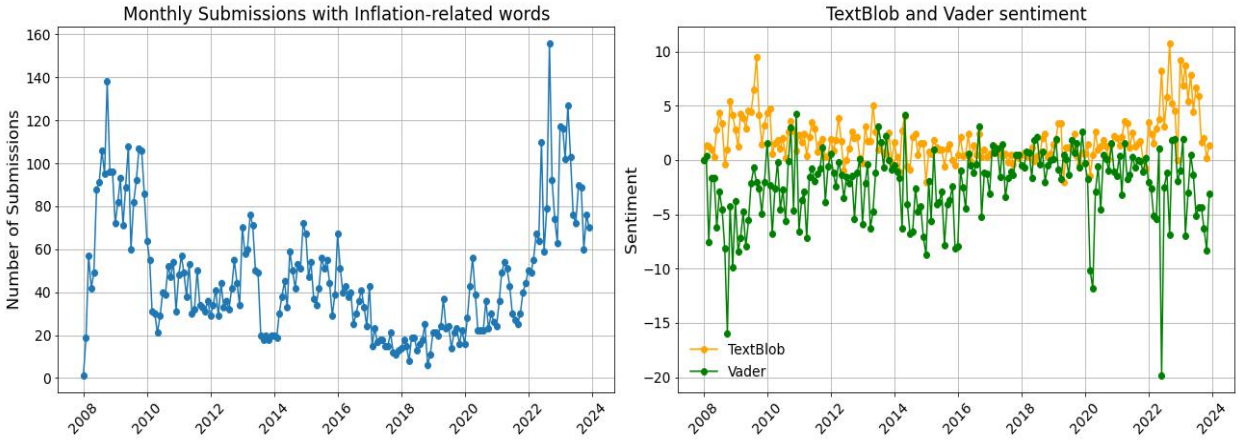
At this stage, to further analyze our data, we conducted a sentiment analysis (see A.1) using the TextBlob (see A.1.1) and VADER (see A.1.2) sentiment dictionaries. These tools are commonly used in economic forecasting studies for their simplicity and effectiveness in rule-based sentiment extraction. Both pre-trained libraries enable efficient large-scale text processing, assigning polarity scores to each sentence in our filtered dataset and providing a conventional measure of sentiment within the submissions. For these reasons, these techniques will also be included in our analysis as a baseline comparison to our new method.

Figure 4 presents the monthly counts of inflation-related submissions (left) and their corresponding sentiment scores (right). To enable comparison with official inflation data,

[6]To ensure robustness, we also included additional subsets with terms such as "wages" and "jobs."

which is released monthly, we aggregated both submission counts and sentiment scores at a monthly level. This aggregation offers a clearer visualization of trends over time and aligns with the release frequency of inflation statistics, making it suitable for our forecasting task.

Figure 4: Monthly counts of inflation-related submissions (**left**) and sentiment scores (**right**).



While TextBlob and VADER offer notable advantages, such as ease of use and adaptability to diverse text inputs, they also have limitations, as demonstrated in our analysis. These models are not specifically trained to capture the unique linguistic nuances of Reddit discussions, particularly those related to economic topics like inflation. As a result, they may struggle with community-specific slang, jargon, or subtle contextual variations common in subreddits. This underscores the importance of using advanced methods like LLMs to account for the linguistic idiosyncrasies of online communities when conducting sentiment analysis.

The following paragraph provides a detailed explanation of the selected LLMs and their implementation, with a focus on fine-tuning the models to extract refined signals for our prediction task.

4 A Primer on Large Language Models (LLMs)

Large Language Models (LLMs) are advanced artificial intelligence systems designed to understand and generate human language. These models are trained on vast amounts of text data, allowing them to predict and produce coherent text based on the input they receive. Examples of well-known LLMs include BERT, GPT, LLaMA, and Gemma. Their ability to recognize patterns and relationships in language makes them valuable tools for tasks such as text classification, summarization, and sentiment analysis.

Although LLMs are trained on general text from the internet, they may not always perform well on domain-specific tasks, such as analyzing discussions about inflation on Reddit or other social media.

Fine-tuning is the process of adapting a pretrained LLM to a specific downstream task by training it on a smaller dataset relevant to that domain. This process helps the model understand the specific language, terminology, and context of the target application, improving its accuracy and relevance. Fine-tuning large models can be computationally expensive and time-consuming. To reduce the training cost, parameter-efficient fine-tuning (PEFT) methods have been introduced to fine-tune pre-trained models with a minimal number of parameters. To address this, researchers have developed efficient methods that reduce the amount of data and computing power required without sacrificing performance. Two key techniques are `LoRA` (Low-Rank Adaptation) and `DoRA` (Weight-Decomposed Low-Rank Adaptation).

- `LoRA`: as shown by Hu et al. (2022), `LoRA` simplifies fine-tuning by adjusting only a small number of parameters rather than the entire model. It works by introducing small, trainable layers that modify the model's behavior without altering its core structure. This approach is like adding a custom filter to a camera lens—enhancing specific features without changing the camera itself. `LoRA` significantly reduces the computational cost of fine-tuning, making it more accessible. `LoRA` leverages the concept of low-rank decomposition to approximate the weight matrix W of the LLMs by introducing low-rank matrices A and B so that[7]

$$W \;=\; A \times B.$$

- `DoRA` decomposes the pre-trained weights into their magnitude and directional components and then fine-tunes both. `DoRA` was suggested by Liu et al. (2024) who build on `LoRA` by distributing fine-tuning adjustments across different parts of the model. Instead of applying all modifications in one place, `DoRA` spreads them out, improving the model's adaptability while maintaining efficiency. `DoRA` goes a step further by decomposing high-rank `LoRA` layers into structured single-rank components and introducing the concept of directional updates, which enhances both learning capacity and stability. This involves decomposing the weight matrix W into two components, magnitude and direction. While the direction is handled by normal `LoRA`, the magnitude is handled by a separate learnable parameter, so that

$$W' \;=\; W + U \times V^T$$

[7]It is similar to reduced rank regression in econometrics.

Another key technique used during fine-tuning is **quantization** (Q), which reduces the size of an LLM by representing its internal data using fewer bits. For example, instead of storing each number with 32 bits, quantization might use only 8 bits. This reduces the model's memory footprint and speeds up its processing, making it more practical for real-world applications. Despite using fewer bits, quantized models can still deliver high performance, especially when fine-tuned with methods like `LoRA` and `DoRA`.

By combining quantization with fine-tuning techniques like `LoRA` and `DoRA` (using `QLoRA` or `QDoRA`), researchers can create LLMs that are both accurate and efficient. This approach allows businesses, researchers, and developers to leverage advanced language models without requiring extensive computational resources, making AI tools more accessible and cost-effective.

In the context of our study, these methods enable us to adapt LLMs to analyze Reddit discussions, capturing nuanced language patterns and extracting signals relevant to inflation forecasting.

4.1 LLMs adopted

We used the following LLMs:

4.1.1 BERT (Base), FinBERT and InflaBERT

BERT (Bidirectional Encoder Representations from Transformers) is a transformer-based model developed by Devlin et al. (2018) at Google. It is pre-trained on the Toronto BookCorpus and English Wikipedia using Masked Language Modeling and Next Sentence Prediction tasks. BERT is available in two versions: the *base* version with 110 million parameters and the *large* version with 340 million parameters. In addition, we employed the classical FinBERT variant[8] (Araci, 2019) and the recently released InflaBERT variant (Allard et al., 2024), which is fine-tuned to predict inflation-related sentiment from news articles.

4.1.2 LLaMA 3 (1B, 3B, and 8B)

LLaMA (Large Language Model Meta AI) is a series of open-source LLMs released by Meta in early 2023. The LLaMA 3.x models have been available since April 2024. For example,

[8]FinBERT is a domain-specific variant of BERT designed for financial text analysis. It undergoes a two-step training process: pretraining on general language data followed by fine-tuning on financial datasets. This adaptation enhances its performance in tasks such as sentiment analysis, stock price prediction, and risk assessment. Evaluation metrics, including accuracy, cross-entropy loss, and macro F1 average, demonstrate its superiority over general-purpose and other domain-specific language models.

LLaMA 3 8B features a vocabulary of more than 128,000 tokens, and its training dataset is seven times larger than that used for LLaMA 2.[9]

4.1.3 Gemma 2 (2B, 9B, and 27B)

Gemma is a family of open-source LLMs developed by Google and released in early 2024. Version 2, released in late June 2024, includes architectural enhancements such as sliding window attention and logit soft-capping. It is pre-trained on twice the data of the original version and features a vocabulary of 256,128 tokens.[10] The model is available in three sizes: 2, 9, and 27 billion parameters.[11]

4.1.4 Qwen 2.5 (0.5B, 1.5B, and 7B)

The Qwen 2.5 family of open-source LLMs was developed by Alibaba Cloud's Qwen team. These models have been extensively pre-trained on a dataset of 18 trillion tokens, significantly enhancing their capabilities in common-sense reasoning, domain expertise, and overall language understanding. Designed to meet diverse needs, the Qwen 2.5 models deliver robust performance in both general and specialized tasks. The family includes models ranging from smaller versions with 0.5 billion parameters to larger ones with 7 billion parameters.[12]

5 Guided Inflation Signal Extraction Using LLMs

5.1 Methodology

We first used LLMs to determine the country associated with posts from the selected subreddits. A LLaMA 3.1 70B model was asked to zero-shot label[13] the country of a sample comprising approximately 20% of the available posts, stratified by year. We then fine-tuned a LLaMA 3.1 8B model on this machine-labeled training set to predict the country of the remaining posts. This approach improved the accuracy of country detection compared to

[9] `https://ai.meta.com/blog/meta-llama-3/`

[10] `https://huggingface.co/blog/gemma`

[11] `https://storage.googleapis.com/deepmind-media/gemma/gemma-2-report.pdf`

[12] `https://qwenlm.github.io/blog/qwen2.5/`

[13] Zero-shot labeling refers to the ability of a model to classify or label data it has never explicitly encountered during training. Unlike traditional models that require labeled examples for each category, zero-shot models can generalize their knowledge to new tasks or labels based on textual descriptions or contextual understanding. This approach is particularly useful when labeled data is limited or unavailable.

using a lexicon of country-related keywords and enabled the exclusion of posts unrelated to the US.[14]

After assigning countries to the posts, we selected those explicitly related to the US. Additionally, we included posts for which the LLMs could not assign a country and those with general content applicable to the US. The next step involved filtering these posts to retain only those discussing inflation.

To extract signals from the filtered posts, we first human-labeled a random sample comprising approximately 20% of the dataset. We then asked both ChatGPT-4 and LLaMA 3.1 70B to double-check the labels, using their feedback to refine the human annotations.[15] We subsequently fine-tuned a range of LLMs to predict the signals in the remaining posts. These models included BERT-based models (BERT base, FinBERT, and InflaBERT) and models from the LLaMA 3, Gemma 2, and Qwen 2.5 families.

The BERT-based models were fully fine-tuned in a supervised manner.For the LLMs, we applied a modified version of the `QLoRA` technique (see Dettmers et al. (2023)). First, the models were quantized to 4 bits. Next, we applied the 'weight decomposed' variant of Low-Rank Adaptation (`LoRA`), known as `DoRA` (see Liu et al. (2024)). The `DoRA` technique was used in two configurations: the standard version (`QDoRA+`) and 'extreme' version (`xQDoRA+`), where the rank was reduced by a factor of 2^3.

To enhance the efficiency and effectiveness of the fine-tuning, we employed the `LoRA+` optimizer (see Hayou et al. (2024)). This optimization technique assigns different learning rates to the approximation matrices that would otherwise be updated using the same learning rate. The learning rates were updated using the AdamW variant of the Adam optimizer, which is specifically designed to improve weight decay during training (see Loshchilov and Hutter (2017)). We will refer to these fine-tuning approaches as `QDoRA+` and `xQDoRA+`.

The objective function was the binary cross-entropy loss with logits. Since the dataset was slightly imbalanced toward negative sentiment, we addressed this issue by assigning weights to each class using the standard balanced weighting method:

$$weight_i = \frac{\text{total number of samples}}{\text{number of classes} \times \text{number of samples in class } i}$$

For performance evaluation, we used the F1 weighted score, which extends the F1 score to multi-class classification problems by accounting for the support (i.e., the number of true

[14]For example, this approach allowed us to exclude posts refferring to "Aussie," slang for the Australian dollar.

[15]Specifically, the authors 'discussed' with ChatGPT posts that had conflicting classifications between human annotators.

instances) of each class. The weighted F1 score is calculated as:

$$\text{F1}_{\text{weighted}} = \sum_{i=1}^{C} \frac{n_i}{N} \times \text{F1}_i$$

where n_i is the number of true instances for class i, N is the total number of instances across all classes, and F1_i is the F1 score for class i.

To ensure the statistical robustness of our results, each model was fine-tuned 20 times using a fixed set of pre-sampled random seeds. The final model selected to predict signals in the unlabeled posts was the one associated with the seed that achieved the median (high) F1 weighted score across all runs.

5.2 Manual labeling with the help of ChatGPT

Since traditional sentiment indicators may not fully capture the nuances of discourse on the social media platform Reddit, we began by manually annotating approximately 500 submissions. Specifically, the authors aimed to determine whether the author of each subreddit post expected: i) prices or inflation to increase (an 'UP' signal/label); ii) prices or inflation to decrease (a 'DOWN' signal/label); or iii) prices or inflation to remain the same (a 'NEUTRAL' signal/label).

These labeled posts were then used to initially fine-tune an LLM, such as BERT. We examined all submissions where there was a mismatch between the predicted and true labels. This analysis was conducted using ChatGPT, with the authors carefully 'discussing' each mismatch. We continued labeling additional posts and repeated this process to compile a substantial number of well-labeled posts.

5.3 Fine-Tuning Large Language Models to Extract an 'Educated' signal

To generate leading indicators for our nowcasting and forecasting exercise, we first labeled approximately 1,400 submissions containing the word *'inflation'*. The submissions' titles were classified as UP, DOWN, or NEUTRAL, indicating whether they conveyed a signal of increasing inflation/prices, decreasing inflation/prices, or no clear signal, respectively. Labeling was performed using both human judgement and ChatGPT. Specifically, we manually labeled the submissions' titles of r/Economics and then used these labeled examples to predict the remaining unlabeled ones using a BERT model. Any mismatches between the predicted and the true labels were carefully reviewed and discussed using ChatGPT.

The dataset of submission titles from r/Economics exceeded 6,000 entries, as did the dataset from r/Economy. To gain a more detailed understanding of Reddit users' views on

inflation and price dynamics, we fine-tuned various LLMs using the human-labeled titles. First, we used BERT base uncased to predict the labels of unlabeled titles. Additionally, we employed LLMs with a higher number of parameters, including LLaMA 3 7B v03, Gemma 2 9B, and Qwen 2.5 7B. The same training pipeline was applied to each LLM.[16] The best-performing model was selected using the F1 score. To enable training on a single GPU (Nvidia A100), we applied 4-bit quantization and the 'weight-decomposed' variant of Low-Rank Adaptation (LoRA), known as DoRA (see Liu et al. (2024)). Quantization compresses the model by reducing the precision of its weights, with minimal impact on performance. LoRA, on the other hand, introduces low-rank matrices into each layer of the LLM architecture, making fine-tuning more efficient and less resource-intensive.

The fine-tuned models were also used to predict the sentiment of comments referring to submissions containing the keywords specified in Section 3.3. To reduce dimensionality and filter uninformative comments, we retained only those posted within two weeks of the submission date. The process of combining submissions and comments in explained in detail in Section 5.4.

It is important to note that reducing the number of trainable parameters did not degrade the performance of the fine-tuned models, as shown in Table 3 and Table 4.

From Table 4, it is evident that LLMs with a higher number of parameters yielded the best F1 scores. BERT base uncased achieved an F1 score of approximately 71% on the test set, while the domain-specific InflaBERT reached around 73%. Larger LLMs with 7B or 9B parameters exceeded an F1 score of 78%, with Gemma 2 9B model achieving the highest F1 score of 86%. Notably, Gemma 2 27B, the largest LLM we fine-tuned, achieved an F1 score of around 83%, comparable to smaller LLMs with only 25% of the parameters.

The efficiency of our fine-tuning process is demonstrated in Table 3, which shows that using `QDoRA+`, we trained at most 0.9% of the LLM parameters, while using `xQDoRA+`, we trained only 0.1%. This combination of quantization and fine-tuning enabled a high degree of efficiency, significantly reducing the number of retrained parameters.[17]

Moreover, Figure 5 illustrates that fine-tuning LLMs substantially increases the median F1 score, improving from an average of approximately 30% in raw models (Figure 5a) to over 75% in fine-tuned models (Figure 5b). Interestingly, using off-the-shelf models exhibit higher volatility in accuracy measures such as the F1 score. In contrast, fine-tuned LLMs using `QDoRA+` or `xQDoRA+` yield significantly more stable F1 scores. Additionally, from panel a) of

[16]For training the LLM models, we used 6,052 titles from r/Economics spanning from 2008 to 2022. The dataset was split into three subsets: 10% as the test set, 90% as the training set, with the training set further divided into 80% for training and 20% for validation. All datasets were one-hot encoded, expanding horizontally across the three labels 'UP', 'Neutral', and 'DOWN'. Predicted labels for the 4,840 titles from r/Economy were computed after model training.

[17]We are currently fine-tuning even larger LLMs to evaluate whether they can extract signals with higher forecasting performance.

Table 3: Percentage of trained parameters for each employed model

Model	**size**	**percentage**
BERT (base)	0.109 B	100.0%
FinBERT	0.109 B	100.0%
InflaBERT	0.082 B	100.0%
Qwen 2.5 0.5B	0.498 B	0.8782% (QDoRA+) 0.1198% (xQDoRA+)
Qwen 2.5 1.5B	1.553 B	0.5682% (QDoRA+) 0.0773% (xQDoRA+)
Qwen 2.5 7B	7.091 B	0.2880% (QDoRA+) 0.0391% (xQDoRA+)
LLaMA 3.2 1B	1.243 B	0.5555% (QDoRA+) 0.0760% (xQDoRA+)
LLaMA 3.2 3B	3.231 B	0.5753% (QDoRA+) 0.0788% (xQDoRA+)
LLaMA 3.1 8B	7.533 B	0.3664% (QDoRA+) 0.0499% (xQDoRA+)
Gemma 2 2B	2.627 B	0.4930% (QDoRA+) 0.0677% (xQDoRA+)
Gemma 2 9B	9.278 B	0.3911% (QDoRA+) 0.0538% (xQDoRA+)
Gemma 2 27B	27.273 B	0.1680% (QDoRA+) 0.0230% (xQDoRA+)

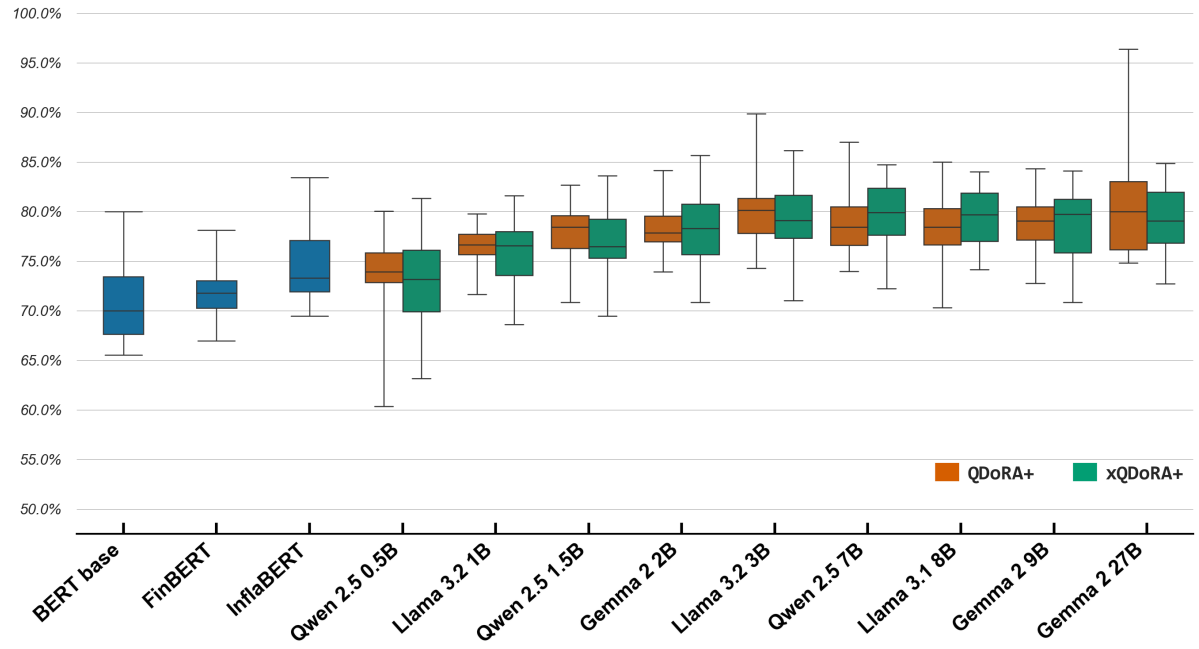
Figure 5, we observe that smaller models such as InflaBERT, which are specifically trained on inflation-related news, can achieve higher F1 score than much larger LLMs, including Gemma 2 27B.

Table 4: Performance metrics on the test sample for different LLMs classifying subreddit titles as UP, DOWN, or NEUTRAL

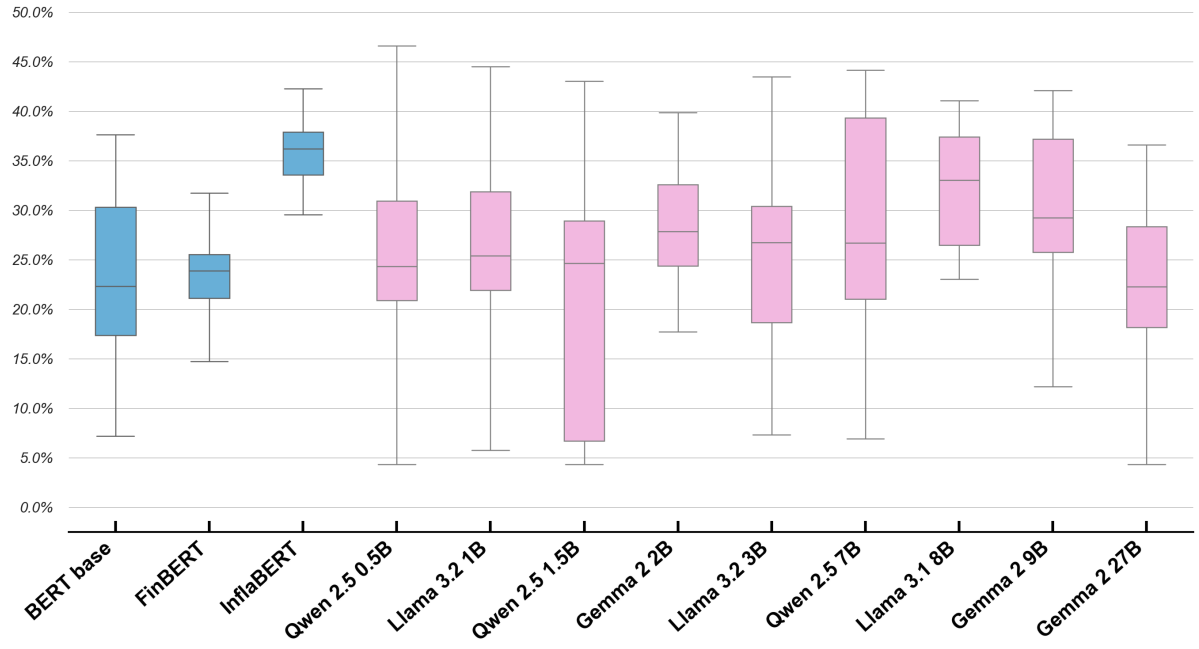
Model	Recall	Precision	F1	F1 weighted	Accuracy
BERT (base)	70.504%	70.504%	70.504%	70.561%	70.504%
FinBERT	71.223%	72.263%	71.739%	71.723%	70.504%
InflaBERT	71.942%	74.627%	73.260%	73.448%	71.942%
Qwen 2.5 0.5B	71.942% (QDoRA+) 71.223% (xQDoRA+)	77.519% (QDoRA+) 75.000% (xQDoRA+)	74.627% (QDoRA+) 73.063% (xQDoRA+)	74.509% (QDoRA+) 73.167% (xQDoRA+)	68.345% (QDoRA+) 69.784% (xQDoRA+)
Qwen 2.5 1.5B	76.259% (QDoRA+) 73.381% (xQDoRA+)	78.519% (QDoRA+) 79.688% (xQDoRA+)	77.372% (QDoRA+) 76.404% (xQDoRA+)	77.379% (QDoRA+) 76.116% (xQDoRA+)	73.381% (QDoRA+) 70.504% (xQDoRA+)
Qwen 2.5 7B	79.137% (QDoRA+) 76.978% (xQDoRA+)	78.014% (QDoRA+) 83.594% (xQDoRA+)	78.571% (QDoRA+) 80.150% (xQDoRA+)	78.353% (QDoRA+) 80.201% (xQDoRA+)	73.381% (QDoRA+) 74.820% (xQDoRA+)
LLaMA 3.2 1B	74.820% (QDoRA+) 76.259% (xQDoRA+)	78.195% (QDoRA+) 75.177% (xQDoRA+)	76.471% (QDoRA+) 75.714% (xQDoRA+)	76.785% (QDoRA+) 75.765% (xQDoRA+)	71.942% (QDoRA+) 73.381% (xQDoRA+)
LLaMA 3.2 3B	78.417% (QDoRA+) 76.978% (xQDoRA+)	80.741% (QDoRA+) 82.946% (xQDoRA+)	79.562% (QDoRA+) 79.851% (xQDoRA+)	79.914% (QDoRA+) 79.784% (xQDoRA+)	76.978% (QDoRA+) 76.978% (xQDoRA+)
LLaMA 3.1 8B	77.698% (QDoRA+) 78.417% (xQDoRA+)	79.412% (QDoRA+) 79.562% (xQDoRA+)	78.545% (QDoRA+) 78.986% (xQDoRA+)	78.453% (QDoRA+) 78.878% (xQDoRA+)	74.820% (QDoRA+) 76.978% (xQDoRA+)
Gemma 2 2B	76.978% (QDoRA+) 77.698% (xQDoRA+)	79.259% (QDoRA+) 78.261% (xQDoRA+)	78.102% (QDoRA+) 77.978% (xQDoRA+)	78.096% (QDoRA+) 78.042% (xQDoRA+)	74.820% (QDoRA+) 76.978% (xQDoRA+)
Gemma 2 9B	79.137% (QDoRA+) 77.698% (xQDoRA+)	79.710% (QDoRA+) 80.000% (xQDoRA+)	79.422% (QDoRA+) 78.832% (xQDoRA+)	79.360% (QDoRA+) 78.762% (xQDoRA+)	75.540% (QDoRA+) 70.504% (xQDoRA+)
Gemma 2 27B	79.856% (QDoRA+) 77.698% (xQDoRA+)	79.286% (QDoRA+) 80.000% (xQDoRA+)	79.570% (QDoRA+) 78.832% (xQDoRA+)	79.486% (QDoRA+) 78.816% (xQDoRA+)	76.978% (QDoRA+) 76.978% (xQDoRA+)

Figure 5: Uncertainty of the weighted F1 score of the test set

(a) fine-tuned models



(b) raw models



5.4 The Time Series Indicators

Using the classification obtained by the LLM in Table 4, we construct different leading indicators. In particular, we aggregate the submissions for each date labeled as UP and DOWN and compute the daily average difference ($UP - DOWN$), with NEUTRAL considered as 0. The text-based indicators from LLMs are constructed as follows:

$$\tilde{X}_{LLM,t} = \sum_{n}^{N} (UP - DOWN); \tag{1}$$

where N is the total number of submissions/comments on day t. The indicator aggregated at monthly frequency is denoted as $\tilde{X}_{LLM,m} = \tilde{X}_{LLM,1:30}$ and is used in the forecast regression. If no submissions are available on a given day, the indicator takes a value of zero.

To reduce short-term volatility, we smooth the indicators using a moving average (MA) window, assessing performance with different window lengths.

We consider submissions and comments jointly in the final indicator. Reddit data allows for constructing a network structure of submissions and comments. In particular, after identifying inflation-related submissions, we link comments to the parent ID of the original submission. This allows us to average the signal from submissions and their comments, capturing the discussion thread generated by each submission. To mitigate noise from extensive threads, we exclude comments that do not mention inflation-related words. The process is illustrated in Figure 6.

Figure 6 preasents an example of the network identification process. The central title represents an inflation-related submission, with linked comments selected based on the presence of inflation-related words. Comments posted more than two weeks after the submission are excluded to reduce dimensionality and focus on timely discussions, enhancing the signal-to-noise ratio.[18]

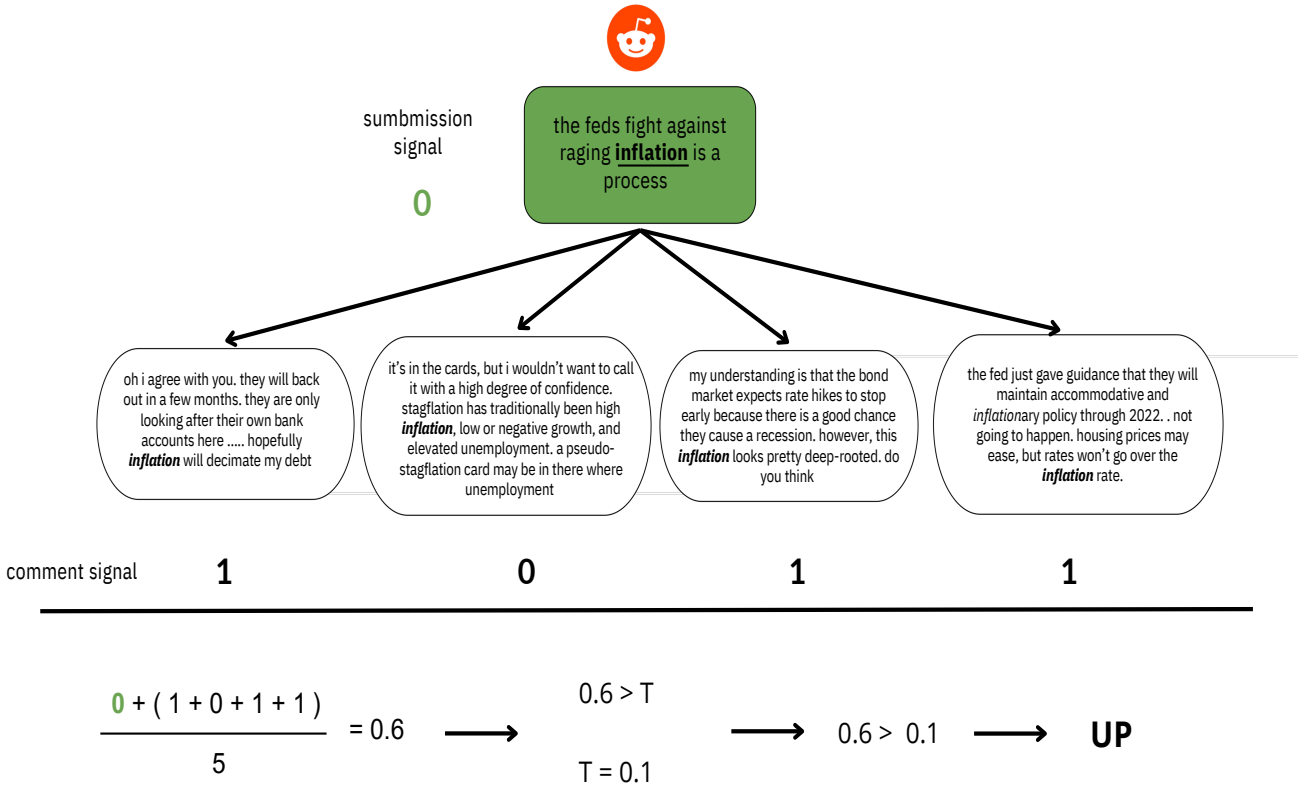
Once the network is identified, the submission signal is re-computed by averaging the submission and comment signals. Specifically, for each thread, we compute an equally weighted average between the submission signal (IS_i) and the comment signals ($IC_{i,j}$):

$$\bar{S}_i = \frac{IS_i + \sum_{j=1}^{J} IC_{i,j}}{J+1}. \tag{2}$$

We then refine the UP/DOWN/NEUTRAL classification for each submission. If the average falls within the threshold, say $(-thr, thr)$, the NEUTRAL label (0) is assigned. The UP label is assigned if the average exceeds thr, and the DOWN label if it falls below $-thr$.

[18] We experimented with re-weighting comment signals using their score (upvotes - downvotes). However, this approach did not improve signal extraction because the score reflects overall comment appreciation rather than its inflation-related content. In fact, using the score increased the signal-to-noise ratio.

Figure 6: Example of a submission-comments network threads



Notes: The initial submission is labelled as **NEUTRAL** (0). The submission's signal is re-weighted by using the comments signals. In particular, an average is computed and evaluated against a threshold: $\bar{S}_i > thr$: **UP**, $\bar{S}_i < thr$: **DOWN**.

We set $thr = 0.10$, noting that variations in this threshold do not significantly alter the results. This process can be interpreted as regularization, adjusting each labeled submission based on the extent to which comments agree or disagree with the original statement posted by the user with the submission.

The example in Figure 6 illustrates a case where a Gemma 2 27B signal is re-weighted from NEUTRAL to UP based on the comments' discussion.

Combining submissions and comments allows us to capture both content generation (submissions) and user engagement (comments), providing a more comprehensive measure of subreddit activity that may enhance the model's predictive power. Additionally, this approach accounts for the social dimension of user-generated content, reflecting how interactions between users in Reddit influence the signal coming from the single submissions. Since many submissions are reposted news headlines, comments often provide individual perspectives,

discussions, and debates, enriching the overall signal captured within each thread.

All the results presented in the empirical application are based on the signal averaged within each network thread.[19]

These indicators are used in the forecasting application in the next Section, where we evaluate their performance in predicting US inflation, specifically **CPI** (CPIAUCSL) and **PCE** (PCEPLIFE) inflation.

6 The predictive ability of the Reddit signals

In this section, we empirically evaluate the predictive ability of the educated signals constructed from Reddit data in combination with the LLM-generated insights. In particular, we assess whether our approach provides a significant signal for forecasting the year-over-year U.S. inflation rate. To validate our measures, we conduct multiple applications, including forecast evaluations and predictive regressions, analyzing both the conditional mean (point estimation) and the distribution (density estimation). The forecasting exercise assesses the predictive power of our signal relative to other indices, while the predictive regression framework examines whether our signal offers statistically significant explanatory power when included alongside other predictors. Our results are conditional on several factors, such as forecast horizons, selected subreddits, and the moving average smoothing windows applied to the Reddit indicators. To enhance readability and facilitate interpretation, Table 5 provides a comprehensive summary of the validation results.

Table 5: Summary of the results.

	Forecast	**Predictive regression**
Point estimation	Better at short-term	Reddit signals exhibit positive and statistically significant coefficients
Density estimation	Better at all horizons	Reddit signals exhibit positive and statistically significant coefficients and their partial correlation is higher during the hyperinflation period

Overall, the objective of the empirical application is to validate our indicators and assess whether they can be integrated into traditional macroeconometric analysis. In addition to a standard pseudo-out-of-sample forecast evaluation, we employ an inference approach based on regression to avoid constraints imposed by a limited out-of-sample.

In our analysis, we also include a larger LLM model - LLaMA 3.1 70B-instruct - which we use via prompting. Specifically, we query the model to determine whether a Reddit

[19] Results using only submission signals can be provided upon request.

submission or comment signals a direction for inflation (UP, DOWN, or NEUTRAL). The inclusion of this additional LLM serves as a benchmark, providing a high-parameter model that may enhance the accuracy of signal prediction. Ultimately, we evaluate the effectiveness of our proprietary fine-tuned smaller models against a third party, larger LLM to assess their relative performance.

6.1 A Point-Forecast Application

In this forecasting application, we compare an $AR(1)$ benchmark model with an $AR-X(1)$ model, which is an autoregressive model augmented with a leading indicator computed using a text-based index from Reddit.[20] The text-based indicators from LLMs are constructed as in (1) on a daily basis and subsequently averaged at a monthly frequency to align with inflation release dates.

We forecast U.S. inflation as measured by the Consumer Price Index (CPI) and the Personal Consumption Expenditures Price Index (PCE) for horizons of 1, 2, 3, 4, 6, 9, 12, and 18 months ahead.

As highlighted by Bybee (2023), LLM models can be interpreted as belief generators. The Reddit-based indicators we construct can be considered as reflecting the beliefs of a particular subgroup (i.e., Reddit users) and are thus comparable to alternative measures of inflation expectations. To benchmark our indicators, we include the University of Michigan Consumer Expectations survey as a key predictor of inflation.[21] Additionally, we incorporate the 1-year inflation Swap rate as a real-time proxy for market expectations.

Table 6: Variables used in the forecasting application

Indicator Name	**Transformation**	**Release date**	**Month considered**
CPI (CPIAUCSL)	y-o-y difference	+0.5 month delay	
PCE (PCEPLIFE)	y-o-y difference	+0.5 month delay	
Michigan Expectations	level	+0.5 month delay	next month
1-Y Inflation Swap	level	available in real-time	current month
Reddit Signals	smoothed (MA)	available in real-time	current month

Notes: This table presents the variables used in the forecasting application, along with their transformations and release dates. The last column specifies how each variable is incorporated into the regression model. CPI and PCE serve as the dependent variables and are included for the current month.

[20]The model includes the lagged dependent variable and the text-based indicator.

[21]SPF indicators are publicly available only on a quarterly basis and are therefore excluded from our analysis.

Table 6 provides a description of the variables used, including data transformations, release dates, and how they enter the model (third column). CPI and PCE are included in the current month, without adjusting for their release delays. Since they are the dependent variables, lagging them does not affect the identification of the best-performing indicators in the predictive regression.

The Michigan survey expectations are typically released at the end of the following month, whereas inflation Swap rates and Reddit signals have the advantage of being available in real time. To account for real-time information, we include Michigan expectations lagged by one month in the regression. The forecasting analysis is conducted using data from two primary subreddits: r/wallstreetbets and r/Economics. We leave the analysis for r/economy as a robustness check in the appendix. Additionally, we apply traditional sentiment analysis tools - VADER and TextBlob - to Reddit data.[22]

The following regression model is estimated:

$$\hat{\pi}_{t+h} = \alpha + \beta_1 \pi_t + \beta_2 X_{\tilde{t}} \tag{3}$$

where $X_{\tilde{t}}$ represents each of the additional variables (inflation expectations, swaps, sentiment indicators, Reddit signals) and $\tilde{t}$ is the most recent available date.

We present our results using histograms and tables. The histograms are shown in Figure 7 and provide an intuitive visualization of the results for $h = 1$ forecast horizon. All the forecast horizons, as well as the significance of the difference in forecasts, are presented in Table format in the following paragraphs. The forecast results are presented for different MA window smoothing of the Reddit indicators. Additional MA windows and robustness check for the subreddit r/economy are provided in section A.4 of the technical appendix.

[22]See the appendix for more details on these more standard techniques.

Figure 7: RMSE for different MA windows. Target = PCE, 1-step-ahead forecast. Results shown for different subreddits.

(a) r/Economics

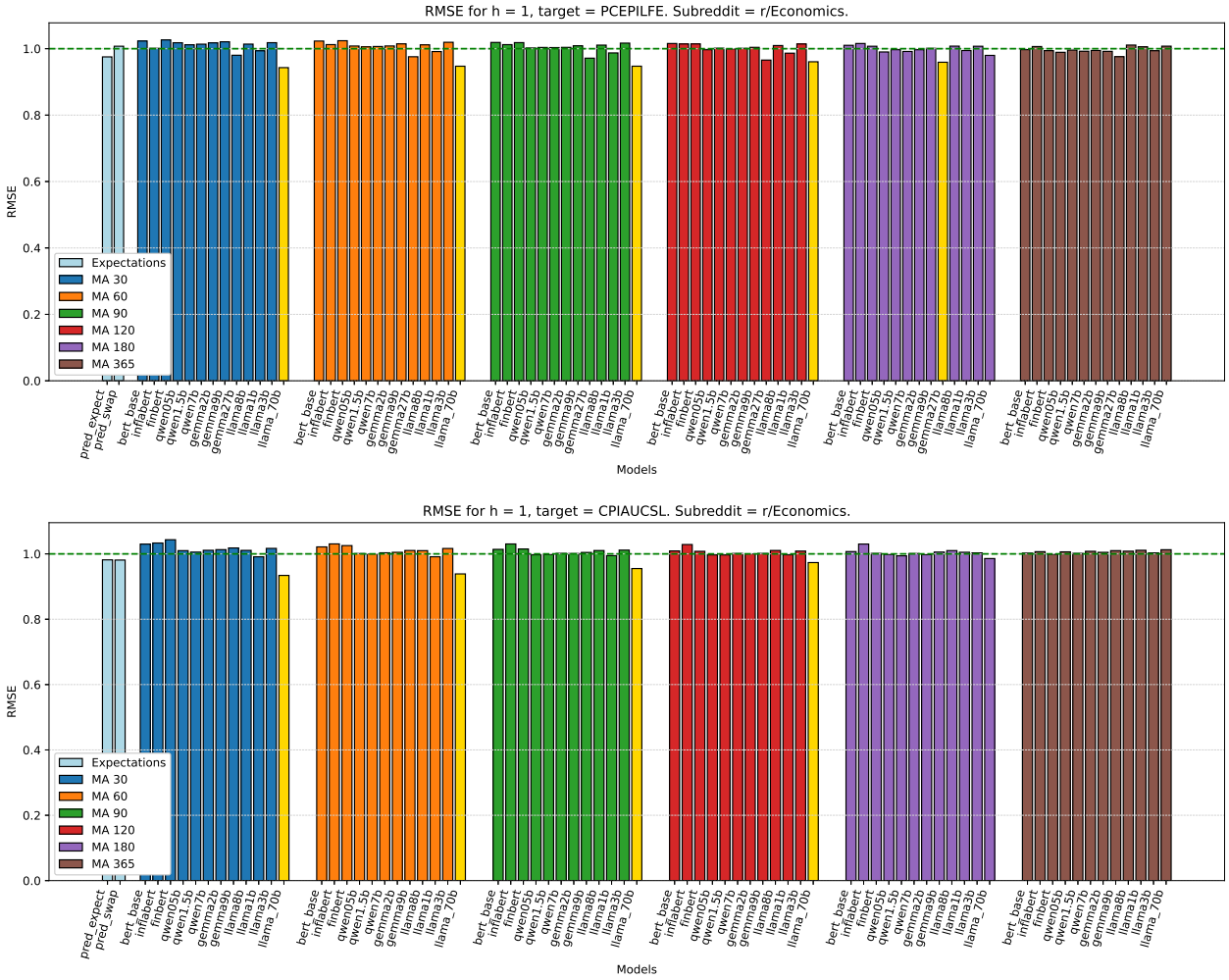


Figure 7: RMSE for different MA windows. Target = PCE, 1-step-ahead forecast. Results shown for different subreddits. (cont.)

(b) r/wallstreetbets

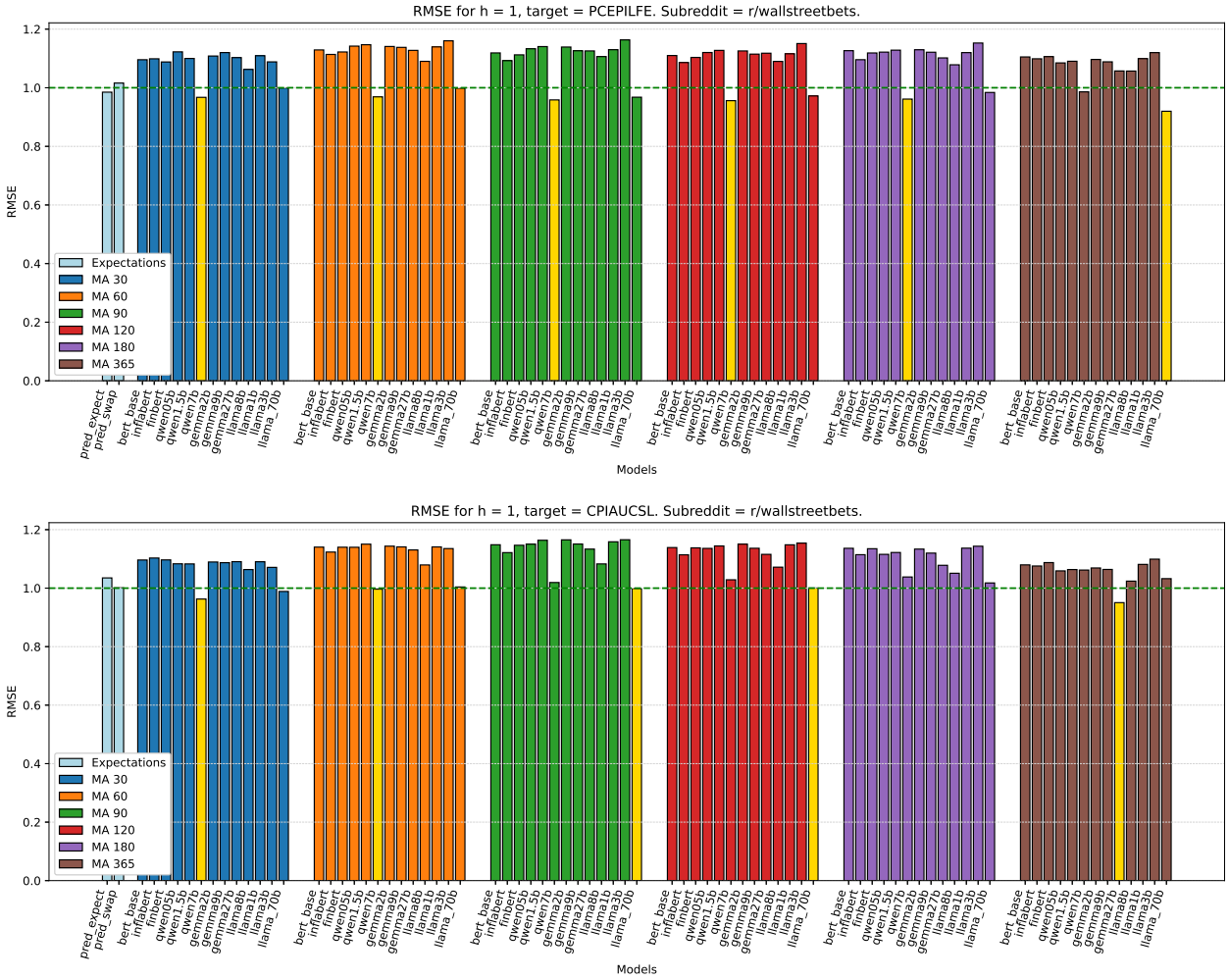


Figure 7 provides an overview of the shortest-term forecast horizon ($h = 1$) results for different MA windows. Each group in the histograms correspond to a MA window for all the LLM models processing Reddit indicators, while the first group considers the regression with expectations (Michigan/Swap). The best Reddit indicator (the one yielding the lowest RMSE in the OOS window) for each MA window is colored in yellow if provides an RMSE lower than the expectations. For the subreddit r/Economics, the histograms shows that, overall, the best model at $h = 1$ is the LLaMA 3.1 70B. However, there is a case of MA = 180 where the fine-tuned Gemma 2 27B provides the best result. For r/wallstreetbets, the results are in favour of the fine-tuned model, particularly for the Qwen 2.5 7B that yields that best forecast performance in most of the cases. In turn, the Gemma 2 27B and LLaMA 3.1 70B provide the best forecast for some MA windows.

Model	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=9$	$h=12$	$h=18$
Expectations									
Michigan Survey	0.975*	0.935***	0.883***	0.837***	0.810***	0.796***	0.782***	0.720***	0.629***
1-Y Inflation Swap	1.007	0.989	0.960	0.929*	0.904**	0.876***	0.843***	0.808***	0.759***
Sentiment									
VADER	1.007	0.996	0.995	0.991	0.993	0.995	0.997	1.001	1.006
TextBlob	1.021	1.011	1.012	1.012*	1.010**	1.008**	1.005	0.997	0.974*
LLM signals (30-days MA smooth)									
BERT-base	1.023	1.013	1.013	1.010	1.007	1.004	0.991	0.955**	0.895**
InflaBERT	1.002	0.964	0.947	0.948	0.945*	0.949**	0.945***	0.897***	0.810***
FinBERT	1.026	1.013	1.012	1.010	1.007	1.003	0.991	0.953**	0.899**
Qwen 2.5 0.5B	1.018	1.019	1.025	1.019	1.011	1.003	1.006	1.005	0.990
Qwen 2.5 1.5B	1.012	1.012	1.015	1.010	1.003	0.995	0.993	1.004	0.995
Qwen 2.5 7B	1.014	1.010	1.013	1.008	0.999	0.993	1.002	1.007	0.996
Gemma 2 2B	1.018	1.017	1.020	1.015	1.009	1.002	1.000	1.007	0.997
Gemma 2 9B	1.021	1.021	1.029*	1.027**	1.024**	1.020*	1.014*	1.007	0.995
Gemma 2 27B	0.980**	0.960***	0.953***	0.956***	0.968***	0.984**	1.007**	1.009***	1.001
LLaMA 3.2 1B	0.993	0.973	0.942*	0.915**	0.896***	0.876***	0.813***	0.783***	0.801***
LLaMA 3.2 3B	1.018	1.014	1.021*	1.022**	1.020***	1.016**	1.009**	0.993	0.959*
LLaMA 3.1 8B	1.014	1.011	1.011	0.999	0.992	0.986	0.979*	0.982**	1.003
LLaMA 3.1 70B-instruct	0.937	0.877*	0.862**	0.871**	0.874***	0.888***	0.893***	0.866**	0.769***
LLM signals (60-days MA smooth)									
BERT-base	1.023	1.017	1.017	1.012	1.008	1.006	0.994	0.959***	0.888**
InflaBERT	1.013	0.977	0.953	0.942	0.940*	0.944*	0.942***	0.891***	0.787***
FinBERT	1.024	1.016	1.016	1.012	1.008	1.006	0.995	0.957***	0.889**
Qwen 2.5 0.5B	1.008	1.013	1.023	1.018	1.011	1.004	1.004	1.008	0.992
Qwen 2.5 1.5B	1.006	1.006	1.012	1.008	1.000	0.992	0.990	1.004	0.996
Qwen 2.5 7B	1.006	1.007	1.013	1.005	0.997	0.991	0.999	1.009	0.998
Gemma 2 2B	1.008	1.011	1.018	1.012	1.006	1.001	0.997	1.008	0.999
Gemma 2 9B	1.015	1.019	1.030	1.030**	1.028**	1.024**	1.015*	1.011	0.997
Gemma 2 27B	0.971**	0.956***	0.947***	0.947***	0.957***	0.973***	1.003	1.009**	1.000
LLaMA 3.2 1B	0.991	0.965	0.939*	0.917**	0.899***	0.879***	0.807***	0.772***	0.788***
LLaMA 3.2 3B	1.019	1.015	1.022*	1.023**	1.021***	1.017**	1.010**	0.996	0.957*
LLaMA 3.1 8B	1.012	1.009	1.008	0.998	0.989	0.980*	0.974**	0.977**	1.002
LLaMA 3.1 70B-instruct	0.941	0.872*	0.837**	0.831***	0.843***	0.861***	0.874***	0.842***	0.725***
LLM signals (90-days MA smooth)									
BERT-base	1.019	1.017	1.021	1.015	1.011	1.008	0.996	0.965***	0.890**
InflaBERT	1.013	0.988	0.965	0.946	0.941*	0.943*	0.938***	0.887***	0.778***
FinBERT	1.018	1.015	1.018	1.014	1.010	1.008	0.997	0.963***	0.891**
Qwen 2.5 0.5B	1.002	1.007	1.019	1.017	1.011	1.004	1.003	1.012*	0.995
Qwen 2.5 1.5B	1.004	1.002	1.009	1.006	0.998	0.991	0.988*	1.005	0.999
Qwen 2.5 7B	1.003	1.004	1.010	1.004	0.997	0.992	0.997	1.012*	1.001
Gemma 2 2B	1.004	1.006	1.014	1.010	1.005	1.000	0.996	1.010*	1.001
Gemma 2 9B	1.009	1.015	1.029	1.031*	1.031**	1.026**	1.016**	1.015*	0.999
Gemma 2 27B	0.971**	0.951***	0.942***	0.940***	0.948***	0.962***	0.997	1.010**	1.000
LLaMA 3.2 1B	0.986*	0.961*	0.937**	0.918**	0.902***	0.880***	0.804***	0.764***	0.782***
LLaMA 3.2 3B	1.017	1.016	1.023*	1.023**	1.021**	1.018**	1.011**	1.001	0.960*
LLaMA 3.1 8B	1.011	1.007	1.007	0.997	0.987*	0.978**	0.971**	0.976**	1.002
LLaMA 3.1 70B-instruct	0.940	0.876*	0.831**	0.812***	0.819***	0.836***	0.853***	0.821***	0.694***

Table 7: Subreddit: **r/Economics**. Target: **PCE**. RMSE is computed as a ratio: $RMSE(AR_x(1)/RMSE(AR(1))$. r/Economics. First in-sample window:2009m1-2018m12; out-of-sample: 2019m1-2023m12. Forecast with expanding window re-estimating the model each step. Significance levels are based on DM test p-values: $p < 0.001$ (***) $p < 0.01$ (**) $p < 0.05$ (*).

Model	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=9$	$h=12$	$h=18$
Expectations									
Michigan Survey	0.982	0.875***	0.762***	0.676***	0.617***	0.586***	0.558***	0.522***	0.484***
1-Y Inflation Swap	0.981	0.875**	0.791***	0.729***	0.701***	0.676***	0.634***	0.615***	0.641***
Sentiment									
VADER	1.024	1.052	1.033	1.019	1.017	1.010	1.003	1.006	1.016
TextBlob	1.017	1.038**	1.038***	1.029***	1.016**	1.011*	0.998	0.970*	0.978
LLM signals (30-days MA smooth)									
BERT-base	1.030	1.002	0.991	0.973	0.958*	0.947**	0.912***	0.845***	0.840***
InflaBERT	1.033	0.929	0.875*	0.855**	0.838***	0.832***	0.821***	0.770***	0.730***
FinBERT	1.043	1.001	0.988	0.974	0.960*	0.944**	0.907***	0.840***	0.845**
Qwen 2.5 0.5B	1.010	1.075*	1.070**	1.054***	1.045***	1.033**	1.017*	0.987	1.009
Qwen 2.5 1.5B	1.005	1.031	1.025	1.019	1.015	1.009	1.004	0.995	0.987
Qwen 2.5 7B	1.011	1.088**	1.075**	1.056**	1.042**	1.032**	1.021*	0.999	1.020*
Gemma 2 2B	1.013	1.070**	1.058**	1.043**	1.032**	1.021*	1.011	0.994	1.005
Gemma 2 9B	1.018	1.044**	1.052***	1.047***	1.044***	1.045***	1.033**	0.997	1.026*
Gemma 2 27B	0.980	0.963	0.955***	0.967***	0.984***	0.999	1.009**	1.000	1.009**
LLaMA 3.2 1B	0.991	0.959	0.897**	0.856***	0.832***	0.808***	0.739***	0.734***	0.777***
LLaMA 3.2 3B	1.017	1.023**	1.030***	1.022***	1.013*	1.009	0.992	0.948**	0.950*
LLaMA 3.1 70B-instruct	0.934	0.766**	0.696***	0.703***	0.721***	0.739***	0.787***	0.771***	0.678***
LLM signals (60-days MA smooth)									
BERT-base	1.021	1.016	1.005	0.984	0.966	0.954**	0.917***	0.850***	0.834***
InflaBERT	1.030	0.952	0.890*	0.854**	0.831***	0.818***	0.799***	0.746***	0.695***
FinBERT	1.025	1.013	1.002	0.984	0.967	0.952**	0.912***	0.842***	0.834**
Qwen 2.5 0.5B	1.001	1.071*	1.071**	1.055**	1.044**	1.035**	1.022**	0.996	1.013
Qwen 2.5 1.5B	1.000	1.024	1.021	1.015	1.011	1.007	1.004	1.003	0.992
Qwen 2.5 7B	1.003	1.090**	1.079**	1.054**	1.040**	1.034**	1.027**	1.010	1.026**
Gemma 2 2B	1.005	1.066*	1.059**	1.041**	1.029**	1.023*	1.014	1.002	1.009
Gemma 2 9B	1.010	1.045*	1.055***	1.051***	1.050***	1.050***	1.038***	1.006	1.031**
Gemma 2 27B	0.974	0.957	0.942***	0.950***	0.968***	0.986***	1.006***	1.006**	1.011***
LLaMA 3.2 1B	0.991	0.945	0.888**	0.849***	0.823***	0.798***	0.727***	0.718***	0.760***
LLaMA 3.2 3B	1.017	1.023**	1.030***	1.024***	1.017**	1.014*	0.996	0.954**	0.948*
LLaMA 3.1 70B-instruct	0.939	0.775**	0.673***	0.645***	0.658***	0.676***	0.732***	0.727***	0.612***
LLM signals (90-days MA smooth)									
BERT-base	1.014	1.022	1.016	0.996	0.977	0.964*	0.924***	0.860***	0.846**
InflaBERT	1.030	0.971	0.916*	0.867**	0.835***	0.816***	0.783***	0.727***	0.684***
FinBERT	1.015	1.016	1.012	0.995	0.978	0.963*	0.919***	0.850***	0.843**
Qwen 2.5 0.5B	0.997	1.067*	1.069**	1.054**	1.043**	1.037**	1.028**	1.008	1.024*
Qwen 2.5 1.5B	0.998	1.018	1.017	1.011	1.007	1.006	1.005	1.010***	1.002
Qwen 2.5 7B	1.002	1.085**	1.077**	1.052**	1.039*	1.034*	1.032**	1.020*	1.035**
Gemma 2 2B	1.001	1.058*	1.055**	1.037**	1.027*	1.023*	1.017*	1.011*	1.017*
Gemma 2 9B	1.004	1.043*	1.056***	1.054***	1.054***	1.054***	1.042***	1.016	1.042***
Gemma 2 27B	0.976	0.956	0.935***	0.936***	0.951***	0.970***	1.002***	1.012**	1.012***
LLaMA 3.2 1B	0.995	0.942	0.882***	0.846***	0.818***	0.794***	0.720***	0.706***	0.752***
LLaMA 3.2 3B	1.012	1.022**	1.030***	1.026***	1.022***	1.018**	1.000	0.965**	0.960
LLaMA 3.1 70B-instruct	0.955	0.810***	0.696***	0.633***	0.620***	0.629***	0.684***	0.686***	0.571***

Table 8: Subreddit: **r/Economics**. Target: **CPI**. RMSE is computed as a ratio: RMSE($AR_x(1)$)/RMSE(AR(1)). r/Economics. First in-sample window:2009m1-2018m12; oos: 2019m1-2023m12. Forecast with expanding window re-estimating the model each step. Significance levels based on DM tests: $p < 0.001$ (***), $p < 0.01$ (**), $p < 0.05$ (*). The **bold** entry in each column marks the lowest relative RMSE.

The results in Tables 7, 8, 9, and 10 demonstrate the robustness and effectiveness of incorporating Reddit-based indicators into inflation forecasting models. Applying different MA smoothing windows to these indicators yields a notable improvement in forecast accuracy over the baseline AR(1) model. Importantly, the inclusion of Reddit indicators consistently outperforms traditional benchmarks such as Michigan survey expectations and 1-year inflation swaps, particularly for shorter forecast horizons.

Among the models leveraging LLM-generated signals, the LLaMA 3.1 70B model emerges as the most accurate, suggesting its superior ability to capture nuanced patterns in Reddit discussions. However, fine-tuned models also exhibit competitive performance, with some instances where they outperform even the largest models in specific cases. For example, Qwen 2.5 7B delivers the best results for the r/wallstreetbets dataset, underscoring the benefits of task-specific fine-tuning. This finding suggests that fine-tuning smaller models can enable them to rival or even surpass larger third-party models in specialized applications.

For longer forecast horizons, Michigan survey expectations demonstrate superior predictive accuracy compared to Reddit indicators. This suggests that traditional expectations measures, which are anchored in broader economic sentiment and structured survey methodologies, remain valuable for capturing long-term inflation trends. However, the consistent short-term predictive advantage of Reddit-based indicators highlights their real-time utility. The dynamic and frequently updated nature of Reddit data likely makes it well-suited for capturing short-term sentiment shifts, which are crucial for near-term forecasting.

Interestingly, traditional sentiment analysis tools such as VADER and TextBlob do not meaningfully improve forecast accuracy when applied to Reddit data. This is likely due to the highly specific nature of Reddit language, which includes unique jargon, colloquialisms, and cultural references that traditional sentiment models fail to interpret effectively. Consequently, the use of advanced LLMs becomes essential for extracting meaningful signals from Reddit discussions. Appendix A.2 presents cumulative loss function charts for all forecasts. The cumulative loss charts provides a dynamic evaluation of the forecast performance for each of the model. We notice that the Reddit indicators provide remarkable forecast gains during the hyperinflation period.

Finally, given the persistence and high volatility of inflation during the test period, we cannot fully rely on the Diebold-Mariano (DM) test resutls reported in the tables. Therefore, we implement a fluctuation test following Giacomini and Rossi (2010), with results presented in Appendix A.3.

6.2 Predictive regression

We now estimate a predictive regression using the entire sample with an OLS approach:

$$\hat{\pi}_{t+h} = \alpha + \beta_1 \pi_t + \beta_2 \pi_t^e + \beta_3 1YSwap_t + \beta_4 X_t + \varepsilon_t \tag{4}$$

The objective of this application is to evaluate the predictive power of the Reddit indicators. By incorporating Reddit-based indicators, we explicitly test whether these new data sources provide a statistically significant improvement in forecasting performance. In this case, X_t represents the signals generated by all the models.

We employ OLS estimation as a simple framework to assess the predictive power of our indicators, using a 1-month-ahead predictive model. The reason we select a $h = 1$ for the predictive regression is to assess the real-time utility of the Reddit indicators. As it was shown in the forecasting application, the Reddit signals provide a superior performance for short-term forecast horizon and this is in line with the theoretical foundation of indicators available in real-time from social media (Angelico et al., 2022). We provide results for $h = 6$ and $h = 12$ in the appendix A.5. Additionally, we do not account for the lag in release dates, which poses a more stringent test for our constructed indicators, as the signal from traditional expectations measures will generally be stronger.

To assess how the inclusion of additional inflation predictors affects the significance of Reddit indicators, we estimate three versions of the regression in equation 4. The first one (baseline model) includes only the Reddit indicator; the second one (extended model) adds the 1-year inflation swap, while the third one (full model) includes both the Swap and inflation expectations. By progressively expanding the set of explanatory variables, we reduce residuals and address the marginal contribution of Reddit indicators in predicting inflation. All three specifications include lagged inflation as a right-hand-side (RHS) variable.

The plot presented considers a MA(90) and corresponds to the three models mentioned above.

The results presented in Figures A.20 and A.21 illustrate the estimated coefficients for the signals from all LLM models for r/Economics. The red color indicates statistical significance in the predictive regression. The model is estimated over the entire sample period (2009m1-2023m12), where the dependent variable is either PCE or CPI.

For PCE, we observe that the models which previously improved forecast accuracy also yield a significant positive coefficient for future inflation. This is particularly evident for the LLaMA 3.1 70B indicator, which was the best-performing model in the 1-month-ahead forecasting exercise. However, we note that the confidence bands for LLaMA 3.1 70B are much wider compared to other significant indicators. In contrast, Gemma 2 models exhibit 8 out of 9 significant and positive coefficients across the three regression specifications, with significantly narrower confidence bands than LLaMA 3.1 70B.

For CPI, the results follow a similar pattern, though Reddit indicators face greater difficulty achieving statistical significance. The LLaMA 3.1 70B indicator has the largest coefficient magnitude, but as additional explanatory variables are included, its confidence bands

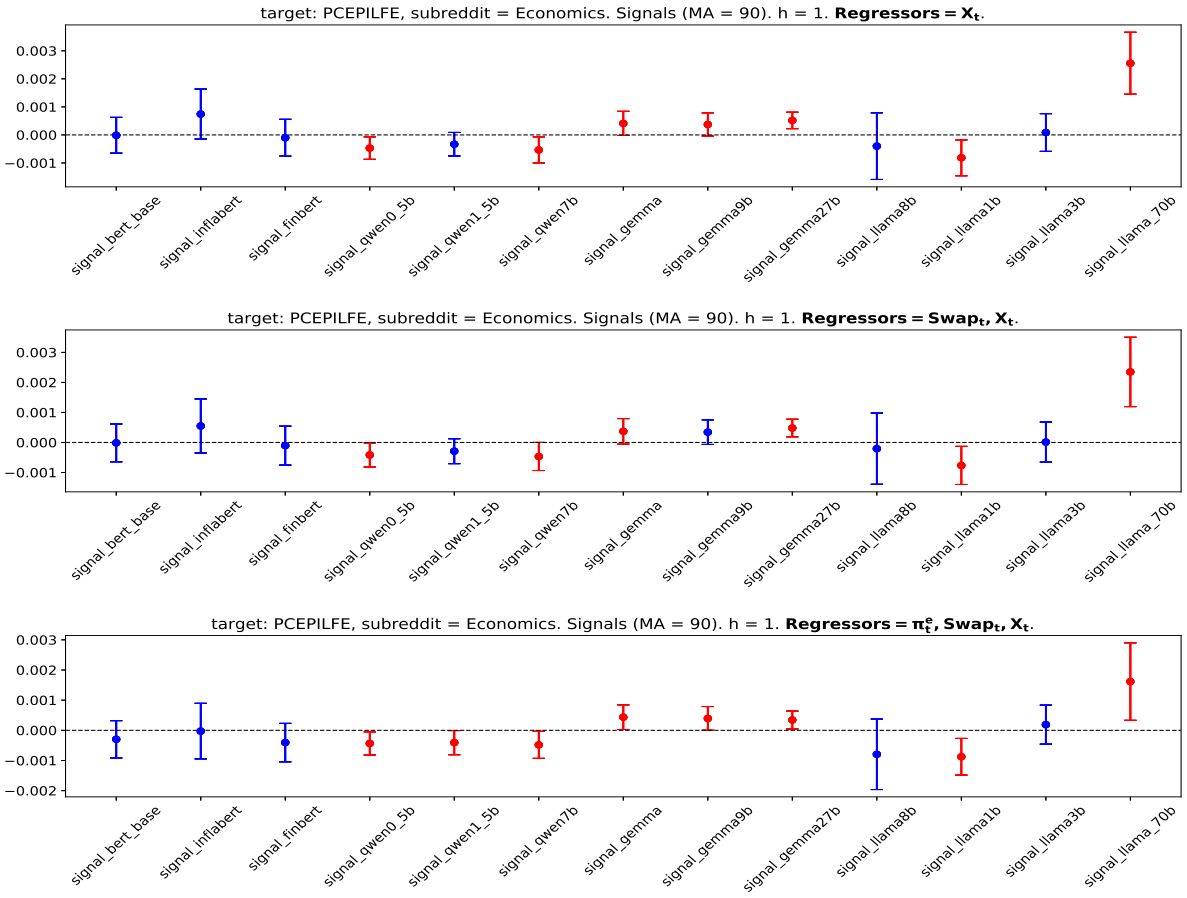


Figure 8: Target = $\mathbf{PCE}_{t+1}$. Coefficients of Reddit signals for different LLMs. Each plot shows coefficients from Reddit for different predictive regression. Top: $\hat{\pi}_{t+1} = \alpha + \beta_1 \pi_t + \beta_4 X_t + \varepsilon_t$; middle: $\hat{\pi}_{t+1} = \alpha + \beta_1 \pi_t + \beta_3 1YSwap_t + \beta_4 X_t + \varepsilon_t$; bottom: $\hat{\pi}_{t+1} = \alpha + \beta_1 \pi_t + \beta_2 \pi^e_t + \beta_3 1YSwap_t + \beta_4 X_t + \varepsilon_t$. Reddit indicators ($X_t$) are used separately in turn in each regression. The red coefficient are significant according to OLS critical values. Subreddit r/Economics.

widen until it loses significance in the full specification. Meanwhile, all Gemma models remain positive and significant, with Gemma 2 27B consistently significant across specifications. This is unsurprising, as it is the largest fine-tuned model among those considered. These results reinforce the finding that fine-tuned proprietary models - despite having smaller parameter sizes - can achieve similar or even superior performance compared to larger third-party models.

We present the corresponding results for r/wallstreetbets in Figures A.22 and A.23. In this case, we observe that the best performing indicator, Qwen 2.5 7B, also yields a positive

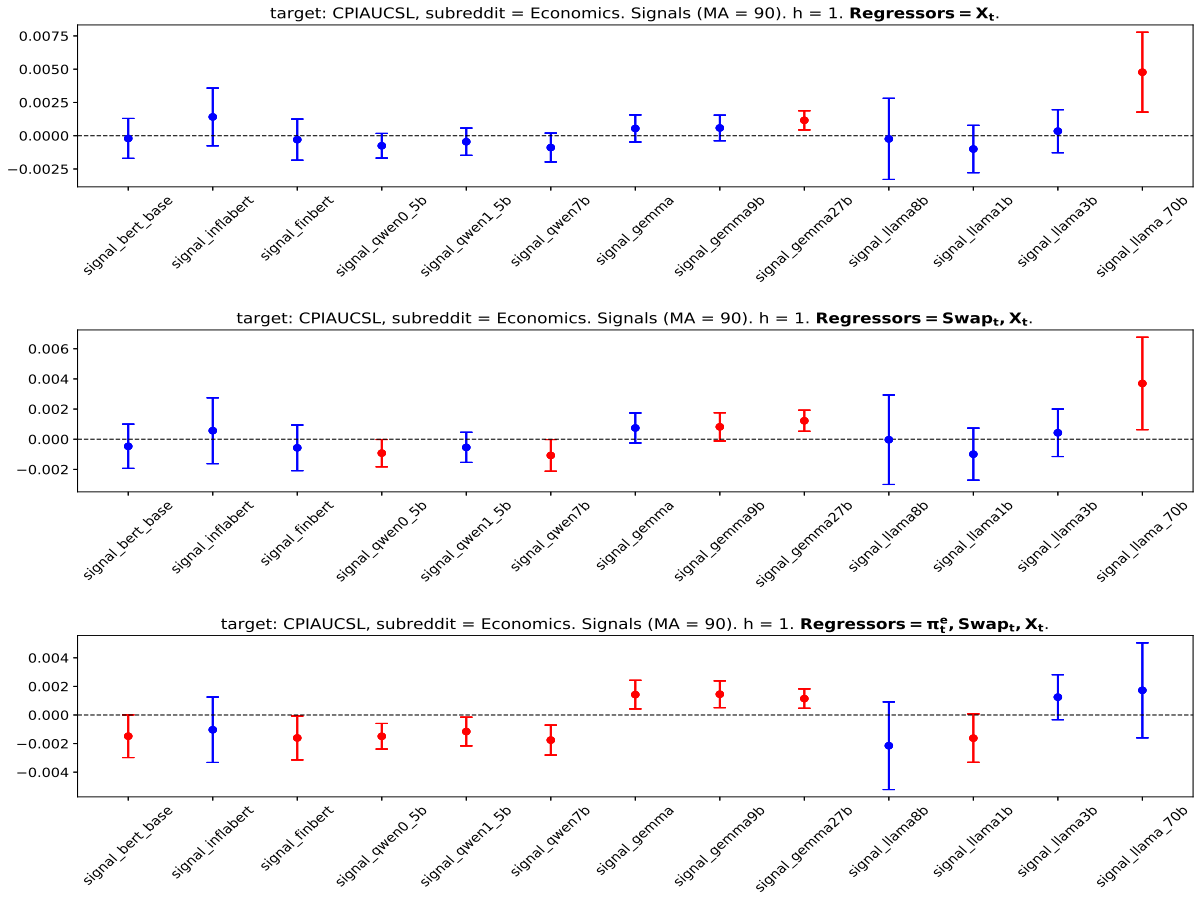


Figure 9: Target = $\mathbf{CPI}_{t+1}$. Coefficients of Reddit signals for different LLMs. Each plot shows coefficients from Reddit for different predictive regression. Top: $\hat{\pi}_{t+1} = \alpha + \beta_1 \pi_t + \beta_4 X_t + \varepsilon_t$; middle: $\hat{\pi}_{t+1} = \alpha + \beta_1 \pi_t + \beta_3 1YSwap_t + \beta_4 X_t + \varepsilon_t$; bottom: $\hat{\pi}_{t+1} = \alpha + \beta_1 \pi_t + \beta_2 \pi^e_t + \beta_3 1YSwap_t + \beta_4 X_t + \varepsilon_t$. Reddit indicators ($X_t$) are used separately in turn in each regression. The red coefficient are significant according to OLS critical values. Subreddit r/Economics.

and significant coefficient. This analysis confirms that Reddit indicators capture residuals future inflation dynamics that remain unexplained by traditional predictors such as Michigan expectations and the 1-year swap

Moreover, this application provides additional validation for Reddit-based indicators. The signal extracted from Reddit can be interpreted as a form of inflation expectations derived from social media discussion. The positive correlation with future inflation aligns well with established expectations-based behavior, further supporting the robustness of our approach.

We expand the results for $h = 6$ and $h = 12$ in section A.5 of the appendix. For

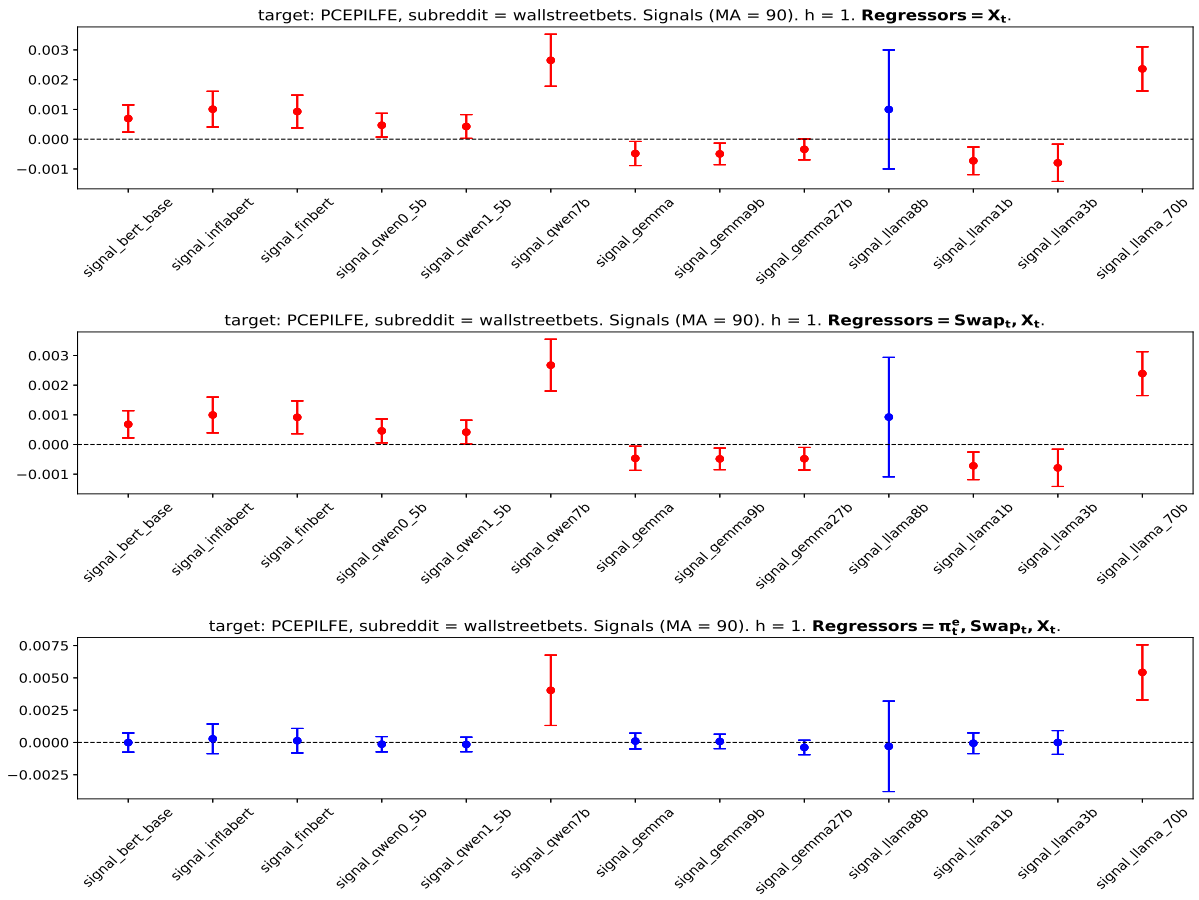


Figure 10: Target = $\mathbf{PCE}_{t+1}$. Coefficients of Reddit signals for different LLMs. Each plot shows coefficients from Reddit for different predictive regression. Top: $\hat{\pi}_{t+1} = \alpha + \beta_1 \pi_t + \beta_4 X_t + \varepsilon_t$; middle: $\hat{\pi}_{t+1} = \alpha + \beta_1 \pi_t + \beta_3 1YSwap_t + \beta_4 X_t + \varepsilon_t$; bottom: $\hat{\pi}_{t+1} = \alpha + \beta_1 \pi_t + \beta_2 \pi^e_t + \beta_3 1YSwap_t + \beta_4 X_t + \varepsilon_t$. Reddit indicators ($X_t$) are used separately in turn in each regression. The red coefficient are significant according to OLS critical values. Subreddit r/wallstreetbets.

r/Economics, we notice that the full regression model (with Michigan, swap) still yields significant positive coefficients of the best Reddit signal (Gemma 2 27B, LLaMA 3.1 70B) for $h = 6$, but the coefficients are not significant for $h = 12$. However, in r/wallstreetbets the best Reddit signal (Qwen 2.5 7B) is always positive and significant for all the horizons considered.

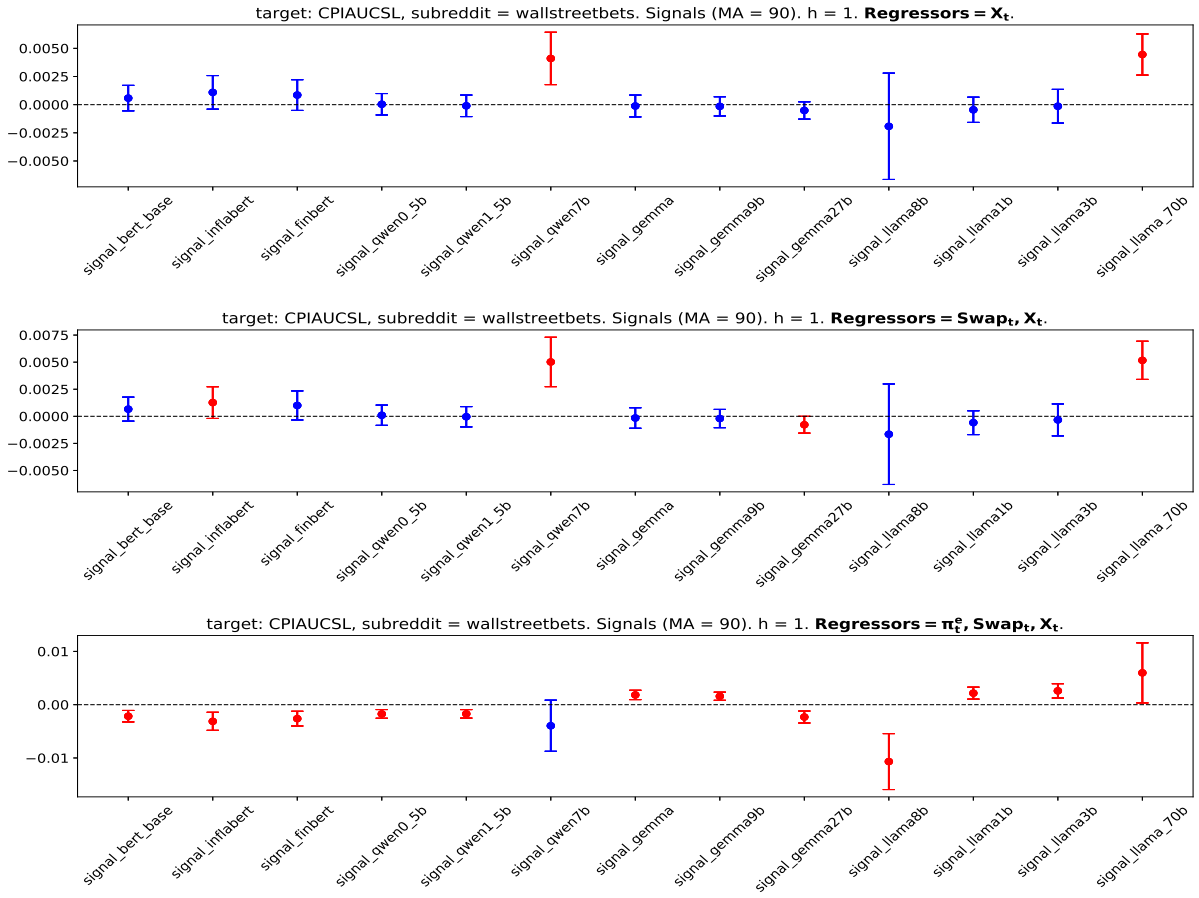


Figure 11: Target = $\mathbf{CPI}_{t+1}$. Coefficients of Reddit signals for different LLMs. Each plot shows coefficients from Reddit for different predictive regression. Top: $\hat{\pi}_{t+1} = \alpha + \beta_1 \pi_t + \beta_4 X_t + \varepsilon_t$; middle: $\hat{\pi}_{t+1} = \alpha + \beta_1 \pi_t + \beta_3 1YSwap_t + \beta_4 X_t + \varepsilon_t$; bottom: $\hat{\pi}_{t+1} = \alpha + \beta_1 \pi_t + \beta_2 \pi^e_t + \beta_3 1YSwap_t + \beta_4 X_t + \varepsilon_t$. Reddit indicators ($X_t$) are used separately in turn in each regression. The red coefficient are significant according to OLS critical values. Subreddit r/wallstreetbets.

6.3 Quantile estimation

We provide a density forecast application following the same structure as the point forecast. Specifically, we estimate the same model as in Equation 6.1, replacing the conditional mean with quantiles, effectively using a quantile regression approach. This allow us to assess whether each indicator contributes individually to projecting the inflation density.

Although we focus on a time series process (subscript t), most of the methodologies explained in this paper can also be applied to cross-sectional or panel data setup.

Quantile regression is a statistical method used to estimate and analyze the dynamics between a set of K regressors $x_t = [x_1, ..., x_K]_t$ and a continuous univariate dependent variable $\mathcal{Y}_t$ at different points of the conditional distribution. Unlike traditional linear regression, which estimates the conditional mean of $\mathcal{Y}_t$, quantile regression allows us to examine diferent quantiles, providing a more comprehensive understanding of the data distribution. A key feature of quantile regression is its ability to capture how predictors influence different parts of the response distribution. This flexibility is particularly valuable when dealing with heteroscedasticity, outliers, or asymmetrical distributions - common characteristics in many empirical economic applications.

$$\mathcal{Y}_{t|\tau} = \boldsymbol{\beta}(\tau)x_t' + \epsilon_{t|\tau}, \quad \tau \in (0,1) \tag{5}$$

In quantile regression, the coefficient vector $\boldsymbol{\beta}(\tau)$ depends on the specific quantile τ and $\mathcal{Y}_t$ is constrained by the quantile condition:

$$\Pr\left[\mathcal{Y}_t \leq x_t'\boldsymbol{\beta}(\tau) \mid x_t\right] = \tau. \tag{6}$$

The estimator of β is obtained by minimizing the quantile loss function, defined as:

$$\rho_\tau\left(u_t\right) = \begin{cases} \tau u_t(\tau), & \text{if } u_t(\tau) \geq 0 \\ (1-\tau)u_t(\tau), & \text{otherwise} \end{cases} \tag{7}$$

where $u_t(\tau) = \mathcal{Y}_t - x_t'\boldsymbol{\beta}(\tau)$. The estimated coefficients are given by:

$$\widehat{\boldsymbol{\beta}}(\tau) = \arg\min_{\boldsymbol{\beta}} \frac{1}{T}\sum_{t=1}^{T} \rho_\tau\left(u_t\right) \tag{8}$$

Unlike linear regression, quantile loss minimization does not have a closed-form solution due to the discontinuity at $u_t(\tau) = 0$. Instead, numerical optimization techniques are typically used to solve this problem.[23].

To reconstruct the forecast distribution, we approximate the target variable's distribution using a non-parametric method proposed by Mitchell et al. (2024):

$$\hat{F}_k\left(\mathcal{Y}_{t+h} \mid x_t\right) = \tau_j + \frac{\tau_{j+1} - \tau_j}{\hat{f}(x_t, \theta_{\tau_{j+1}}) - \hat{f}(x_t, \theta_{\tau_j})}\left(\mathcal{Y}_{t+h} - \hat{f}(x_t, \theta_{\tau_{j+1}})\right) \tag{9}$$

where $\hat{f}(x_t, \theta_{\tau_{j+1}})$ is the predictive function. This approach estimates density by computing quantile distances.

[23]We use a maximum likelihood estimation appraoch.

Once the density is computed, the model's accuracy is assessed by measuring the distance between the predictive distribution and the actual value of the target variable $\mathcal{Y}_{t+h}$. This is achieved using the Continuous Ranked Probability Score (CRPS):

$$CRPS(F, y) = \int \left(F(x) - \mathbf{1}_{\{x \geq y\}}\right)^2 dx$$

The CRPS compares the predictive distribution with a degenerate distribution (the actual observed value of the dependent variable). Lower CRPS values indicate a closer alignment between the predictive distribution and the realized value.

In Tables 11, 12, 13, and 14, we present the density forecasts results in terms of CRPS for moving averages with windows of 30, 60, 90. The results are consistent with the point forecasts, with the notable exception that Reddit indicators exhibit improved performance for longer-term forecast horizons, consistently outperforming Michigan expectations and inflation swap.

Additionally, we estimate the forecast regression in Equation 4 using a quantile regression approach for $h = 1$. We report the estimated coefficients for the Reddit signals in a model where Swap and expectations are included, i.e. $x_t' = \{\pi_{t-1}, \pi_t^e, Swap_t, X_t\}$. We select Gemma 2 27B and LLaMA 3.1 70B as representative Reddit indicators, serving as a comparison between a fine-tuned proprietary model and a third-party model.

Due to the absence of a closed-form solution in quantile regression optimization, we compute confidence bands around the coefficients using a block-bootstrap approach, following Lopez-Salido and Loria (2024). Given the complexity of the estimation, we focus on r/Economics, which provides the longest meaningful sample (2009m1-2023m12).

We adopt the empirical framework of Lopez-Salido and Loria (2024), where the authors examine inflation drivers by analyzing coefficients across quantiles. While our application is less ambitious,[24] this approach allows us to better understand the partial correlation between Reddit signals and future inflation dynamics. Unlike standard regression, quantile regression provides a set of coefficients for each quantile, revealing how the relationship varies across different points in the distribution.

Figure 12 presents the estimated coefficients across quantiles for the Educated signal from the r/Economics subreddit.

[24] Our goal is to test whether Reddit indicators provide additional predictive signal (partial correlation) for future inflation density, rather than analyze the structural determinants of inflation.

Model	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=9$	$h=12$	$h=18$
Expectations									
Michigan Survey	0.985	0.910	0.784**	0.712***	0.652***	0.606***	0.515***	0.449***	0.392***
1-Y Inflation Swap	1.016	1.002	0.989	0.973	0.958*	0.936*	0.858***	0.783***	0.732***
Sentiment									
VADER	1.001	1.012	0.929	0.875	0.811	0.769*	0.826*	0.960	1.041
TextBlob	1.118	1.035	0.939	0.917	0.923	0.911	0.835	0.758**	0.652***
LLM signals (30-days MA smooth)									
BERT-base	1.096	1.053	1.020	1.001	0.977	0.952	0.826	0.745*	0.597***
InflaBERT	1.098	1.041	0.989	0.967	0.943	0.922	0.805	0.741*	0.593***
FinBERT	1.088	1.036	0.997	0.975	0.951	0.930	0.808	0.729*	0.585***
Qwen 0.5B	1.123*	1.085	1.053	1.027	1.002	0.977	0.845	0.751*	0.614***
Qwen 1.5B	1.100**	1.079	1.067	1.052	1.026	1.001	0.872	0.776	0.639***
Qwen 7B	0.966	0.888	0.834	0.815	0.796*	0.780*	0.677**	0.640***	0.579***
Gemma 2B	1.108*	1.076	1.054	1.036	1.012	0.987	0.858	0.770*	0.627***
Gemma 9B	1.120*	1.080	1.044	1.021	0.995	0.970	0.841	0.753*	0.615***
Gemma 2 27B 1.103	1.075	1.045	1.031	1.035	1.039	1.058	1.043	1.035	
LLaMA 1B	1.109*	1.055	1.010	0.985	0.962	0.938	0.817	0.734**	0.596***
LLaMA 3B	1.088**	1.050	1.024	1.009	0.986	0.966	0.848	0.752**	0.618***
LLaMA 8B	1.062**	1.029	1.025	1.040	1.024	1.005	0.945	0.865	0.697***
LLaMA 3.1 70B-instruct	0.996	0.925	0.850	0.818	0.800**	0.782**	0.705***	0.705***	0.640***
LLM signals (60-days MA smooth)									
BERT-base	1.129	1.056	1.015	0.989	0.969	0.942	0.813	0.711*	0.568***
InflaBERT	1.114	1.040	0.985	0.950	0.929	0.905	0.789	0.705**	0.551***
FinBERT	1.122	1.040	0.991	0.961	0.939	0.916	0.794	0.695**	0.552***
Qwen 0.5B	1.142*	1.079	1.044	1.016	0.993	0.966	0.833	0.717*	0.583***
Qwen 1.5B	1.147**	1.085	1.060	1.043	1.023	0.997	0.863	0.743*	0.608***
Qwen 7B	0.966	0.881	0.808	0.764*	0.742*	0.723**	0.630***	0.584***	0.510***
Gemma 2B	1.141*	1.082	1.054	1.033	1.013	0.987	0.854	0.744*	0.603***
Gemma 9B	1.138	1.071	1.033	1.005	0.984	0.957	0.826	0.718*	0.583***
Gemma 2 27B	1.128	1.075	1.045	1.031	1.035	1.039	1.045	1.040	1.046
LLaMA 1B	1.139*	1.062	1.008	0.975	0.953	0.930	0.806	0.705**	0.567***
LLaMA 3B	1.160*	1.077	1.032	1.012	0.993	0.972	0.848	0.732**	0.599***
LLaMA 8B	1.088*	1.043	1.028	1.029	1.024	1.002	0.926	0.812*	0.656***
LLaMA 3.1 70B-instruct	0.993	0.899	0.825	0.774*	0.752**	0.736**	0.663***	0.652***	0.565***
LLM signals (90-days MA smooth)									
BERT-base	1.119	1.054	1.007	0.973	0.948	0.923	0.791	0.667**	0.534***
InflaBERT	1.093	1.029	0.973	0.930	0.901	0.879	0.764*	0.657**	0.511***
FinBERT	1.112	1.038	0.982	0.945	0.919	0.896	0.774	0.653**	0.519***
Qwen 0.5B	1.133	1.075	1.036	1.002	0.975	0.948	0.812	0.673**	0.550***
Qwen 1.5B	1.141*	1.084	1.050	1.024	1.003	0.978	0.838	0.695**	0.570***
Qwen 7B	0.957	0.867	0.789*	0.727**	0.685**	0.669**	0.585***	0.525***	0.451***
Gemma 2B	1.139	1.082	1.047	1.020	0.997	0.972	0.836	0.703**	0.572***
Gemma 9B	1.126	1.064	1.021	0.987	0.961	0.935	0.802	0.671**	0.547***
Gemma 2 27B	1.126	1.064	1.021	0.987	0.961	0.935	0.802	0.671	0.547
LLaMA 1B	1.130	1.055	0.999	0.958	0.932	0.909	0.784*	0.666**	0.534***
LLaMA 3B	1.164*	1.089	1.038	1.007	0.987	0.966	0.837	0.699**	0.574***
LLaMA 8B	1.103	1.067	1.039	1.019	1.007	0.994	0.904	0.761**	0.610***
LLaMA 3.1 70B-instruct	0.962	0.874	0.800*	0.743**	0.709**	0.690***	0.627***	0.603***	0.503***

Table 9: Subreddit: **r/wallstreetbets**. Target: **PCE**. RMSE is computed as a ratio: RMSE($AR_x(1)$)/RMSE(AR(1)). r/Economics, QLoRA. First in-sample window:2015m1-2018m12; oos: 2019m1-2023m12. Forecast with expanding window re-estimating the model each step. Significance levels based on DM tests: $p < 0.001$ (***), $p < 0.01$ (**), $p < 0.05$ (*). The **bold** entry in each column marks the lowest relative RMSE.

Model	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=9$	$h=12$	$h=18$
Expectations									
Michigan Survey	1.035	0.955	0.857**	0.776***	0.697***	0.627***	0.495***	0.447***	0.438***
1-Y Inflation Swap	1.001	0.949	0.907*	0.860***	0.836***	0.807***	0.733***	0.689***	0.647***
Sentiment									
VADER	1.017	1.034	1.014	0.997	1.026	1.048	1.055	1.069	1.062
TextBlob	1.122	1.058	0.976	0.919	0.873	0.838*	0.776***	0.728***	0.666***
LLM signals (30-days MA smooth)									
BERT-base	1.096	1.068	1.006	0.939	0.879	0.828	0.737**	0.694***	0.623***
InflaBERT	1.102	1.046	0.975	0.901	0.844	0.800*	0.723***	0.686***	0.603***
FinBERT	1.097	1.068	1.004	0.933	0.873	0.825	0.740**	0.693**	0.624***
Qwen 2.5 0.5B	1.083	1.071	1.036	0.979	0.915	0.862	0.765**	0.712**	0.655***
Qwen 2.5 1.5B	1.083	1.071	1.046	0.996	0.935	0.882	0.786*	0.730**	0.667***
Qwen 2.5 7B	0.963	0.857*	0.778**	0.718***	0.685***	0.672***	0.610***	0.614***	0.585***
Gemma 2 2B	1.089	1.074	1.031	0.972	0.911	0.861	0.768**	0.716**	0.663***
Gemma 2 9B	1.087	1.063	1.020	0.958	0.898	0.849	0.758**	0.706**	0.641***
Gemma 2 27B	1.089	1.074	1.032	0.972	0.911	0.862	0.769	0.717	0.664
LLaMA 3.2 1B	1.090	1.082	1.029	0.964	0.904	0.855	0.772**	0.717**	0.655***
LLaMA 3.2 3B	1.071	1.076	1.051	1.008	0.948	0.894	0.805*	0.745**	0.687**
LLaMA 3.1 8B	1.062*	1.032	1.006	0.979	0.930	0.894	0.860**	0.815**	0.710***
LLaMA 3.1 70B-instruct	0.987	0.926	0.854	0.795*	0.764**	0.745***	0.716***	0.724***	0.668***
LLM signals (60-days MA smooth)									
BERT-base	1.140*	1.083	1.011	0.934	0.867	0.811	0.714**	0.659***	0.598***
InflaBERT	1.123	1.050	0.968	0.890	0.825	0.775**	0.694***	0.644***	0.567***
FinBERT	1.140*	1.085	1.010	0.930	0.861	0.806*	0.715**	0.656***	0.595***
Qwen 2.5 0.5B	1.140*	1.092	1.038	0.973	0.905	0.847	0.742**	0.673***	0.620***
Qwen 2.5 1.5B	1.150*	1.113	1.066	1.008	0.942	0.883	0.775*	0.703**	0.647***
Qwen 2.5 7B	0.995	0.875	0.765**	0.692***	0.644***	0.618***	0.581***	0.561***	0.530***
Gemma 2 2B	1.143*	1.104	1.052	0.987	0.921	0.865	0.763**	0.696**	0.648***
Gemma 2 9B	1.141*	1.086	1.026	0.959	0.894	0.839	0.739**	0.672***	0.616***
Gemma 2 27B	1.131*	1.120	1.116	1.113	1.106	1.097	1.096	1.071	1.066
LLaMA 3.2 1B	1.141*	1.109	1.049	0.975	0.908	0.853	0.759**	0.692**	0.637***
LLaMA 3.2 3B	1.135	1.116	1.081	1.026	0.964	0.904	0.801*	0.729**	0.677**
LLaMA 3.1 8B	1.078*	1.055	1.019	0.978	0.928	0.878	0.820**	0.761**	0.679***
LLaMA 3.1 70B-instruct	0.998	0.903	0.814*	0.743**	0.700***	0.680***	0.659***	0.659***	0.598***
LLM signals (90-days MA smooth)									
BERT-base	1.148	1.101	1.021	0.937	0.862	0.801	0.691**	0.622***	0.576***
InflaBERT	1.121	1.061	0.971	0.885	0.812	0.757**	0.662***	0.602***	0.542***
FinBERT	1.146	1.100	1.018	0.932	0.856	0.795*	0.690***	0.621***	0.571***
Qwen 2.5 0.5B	1.150*	1.119	1.056	0.982	0.908	0.846	0.725**	0.643***	0.596***
Qwen 2.5 1.5B	1.164*	1.143	1.085	1.015	0.943	0.879	0.753*	0.667**	0.621***
Qwen 2.5 7B	1.018	0.903	0.783**	0.687***	0.623***	0.587***	0.555***	0.518***	0.492***
Gemma 2 2B	1.165*	1.137	1.074	1.001	0.928	0.867	0.749**	0.669**	0.625***
Gemma 2 9B	1.151	1.113	1.044	0.969	0.897	0.836	0.720**	0.640***	0.594***
Gemma 2 27B	1.091	1.120	1.117	1.110	1.104 1.100	1.099	1.071	1.073	
LLaMA 3.2 1B	1.159*	1.129	1.057	0.977	0.902	0.842	0.730**	0.656***	0.607***
LLaMA 3.2 3B	1.166	1.151	1.101	1.039	0.970	0.909	0.787*	0.702**	0.657**
LLaMA 3.1 8B	1.082	1.076	1.040	0.986	0.928	0.875	0.786***	0.714***	0.639***
LLaMA 3.1 70B-instruct	0.992	0.897	0.793**	0.711***	0.657***	0.630***	0.605***	0.604***	0.546***

Table 10: Subreddit: **r/wallstreetbets**. Target: **CPI**. RMSE is computed as a ratio: RMSE($AR_x(1)$)/RMSE(AR(1)). r/Economics, QLoRA. First in-sample window:2015m1-2018m12; oos: 2019m1-2023m12. Forecast with expanding window re-estimating the model each step. Significance levels based on DM tests: $p < 0.001$ (***), $p < 0.01$ (**), $p < 0.05$ (*). The **bold** entry in each column marks the lowest relative RMSE.

Model	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=9$	$h=12$	$h=18$
Expectations									
Michigan Survey	1.024	1.006	1.000	0.975	0.962	0.940	0.887*	0.839***	0.888*
1-Y Inflation Swap	1.007	0.988	0.999	1.007	0.986	0.970	0.927*	0.931*	0.985
Sentiment									
VADER	1.004	1.016*	1.022*	1.034***	1.025**	1.022**	1.020**	1.005	0.998
TextBlob	1.006	1.008	1.014	1.004	0.993	0.991	0.990	0.984	0.987
LLM signals (30-days MA smooth)									
BERT-base	1.011	1.008	1.014	1.008	0.998	0.989	0.973*	0.951**	0.956***
InflaBERT	0.994	0.978	0.978	0.972	0.963*	0.966*	0.961**	0.955**	0.933***
FinBERT	1.015	1.015	1.022**	1.019	1.005	0.999	0.975*	0.957**	0.954***
Qwen 2.5 0.5B	1.005	1.017*	1.015	1.002	1.004	1.006	1.002	0.983*	0.982**
Qwen 2.5 1.5B	1.010*	1.023***	1.023**	1.023**	1.012	1.014**	1.009	0.990	0.990**
Qwen 2.5 7B	1.001	1.013*	1.010	1.007	1.000	1.002	1.004	0.985**	0.987*
Gemma 2 2B	1.007	1.016*	1.018**	1.022**	1.012	1.010	1.005	0.988*	0.986**
Gemma 2 9B	1.008	1.013	1.019*	1.008	1.006	1.004	0.992	0.978**	0.976***
Gemma 2 27B	0.980	0.963	0.955	0.967***	0.984***	0.999	1.009**	1.000	1.009**
LLaMA 3.2 1B	1.030**	1.038**	1.056***	1.057***	1.058***	1.059**	1.060**	1.053**	0.985
LLaMA 3.2 3B	1.015**	1.019**	1.032***	1.029***	1.009	1.004	0.989	0.979*	0.978**
LLaMA 3.1 8B	1.006	1.003	1.012**	1.012	1.004	1.007	1.020**	1.007	0.999
LLaMA 3.1 70B-instruct	0.945**	0.939***	0.927***	0.914***	0.920***	0.926***	0.932***	0.928***	0.923***
LLM signals (60-days MA smooth)									
BERT-base	1.013*	1.004	1.020*	1.009	0.996	0.988	0.967**	0.956**	0.955***
InflaBERT	0.995	0.985	0.987	0.971	0.962	0.961*	0.956**	0.950**	0.931***
FinBERT	1.013	1.014	1.031***	1.014	1.006	1.000	0.976*	0.959**	0.957***
Qwen 2.5 0.5B	1.002	1.016	1.018	1.004	1.003	1.004	1.001	0.986	0.984**
Qwen 2.5 1.5B	1.009	1.021***	1.028***	1.019**	1.015*	1.011*	1.011	0.991	0.992
Qwen 2.5 7B	1.001	1.012	1.016	1.003	0.998	1.002	1.001	0.987*	0.989*
Gemma 2 2B	1.005	1.015*	1.026**	1.019**	1.010	1.011*	1.005	0.989	0.988**
Gemma 2 9B	1.005	1.011	1.020*	1.007	1.003	1.005	0.993	0.979*	0.980**
Gemma 2 27B	0.974	0.957	0.942***	0.950***	0.968***	0.986***	1.006***	1.006**	1.011***
LLaMA 3.2 1B	1.030*	1.043**	1.065***	1.064***	1.060**	1.062**	1.073**	1.060**	0.968*
LLaMA 3.2 3B	1.016*	1.023***	1.032***	1.022**	1.010	1.006	0.992	0.982	0.982**
LLaMA 3.1 8B	1.007	1.008	1.012*	1.010	1.005	1.011	1.023**	1.010	0.997
LLaMA 3.1 70B-instruct	0.950**	0.940***	0.925***	0.914***	0.911***	0.918***	0.920***	0.928***	0.915***
LLM signals (90-days MA smooth)									
BERT-base	1.011	1.008	1.022**	1.011	1.002	0.991	0.970**	0.955**	0.955***
InflaBERT	1.003	0.991	0.987	0.980	0.963	0.962	0.952**	0.948**	0.932***
FinBERT	1.014*	1.015	1.028**	1.022*	1.004	1.002	0.982	0.961**	0.955***
Qwen 2.5 0.5B	1.000	1.010	1.020	1.006	0.999	1.003	1.002	0.984*	0.988
Qwen 2.5 1.5B	1.008	1.022***	1.025***	1.020**	1.013	1.014*	1.012	0.994	0.996
Qwen 2.5 7B	1.002	1.008	1.019	1.007	0.996	1.004	1.001	0.989	0.990*
Gemma 2 2B	1.006	1.017*	1.029**	1.021***	1.009	1.010	1.005	0.992	0.990*
Gemma 2 9B	1.005	1.010	1.020	1.012	1.001	1.004	0.993	0.980*	0.984*
Gemma 2 27B	0.976	0.956	0.935***	0.936***	0.951***	0.970***	1.002**	1.012***	1.012***
LLaMA 3.2 1B	1.023*	1.040**	1.067***	1.072***	1.065***	1.071***	1.082***	1.066**	0.968*
LLaMA 3.2 3B	1.016*	1.023**	1.029**	1.018*	1.005	1.007	0.996	0.984	0.984*
LLaMA 3.1 8B	1.010	1.013*	1.016**	1.011	1.008	1.017**	1.026**	1.007	1.000
LLaMA 3.1 70B-instruct	0.958**	0.947***	0.935**	0.916***	0.909***	0.912***	0.915***	0.926***	0.911***

Table 11: Subreddit: **r/Economics**. Target: **PCE**. CRPS is computed as a ratio: $CRPS(AR_x(1)/CRPS(AR(1))$. r/Economics. First in-sample window:2009m1-2018m12; oos: 2019m1-2023m12. Forecast with expanding window re-estimating the model each step. Significance levels are based on DM test p-values: $p < 0.001$ (***), $p < 0.01$ (**), $p < 0.05$ (*).

Model	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=9$	$h=12$	$h=18$
Expectations									
Michigan Survey	1.023	1.029	1.017	0.953	0.888*	0.832***	0.760***	0.706***	0.821***
1-Y Inflation Swap	1.030	1.021	0.994	0.953	0.909	0.867**	0.839***	0.845***	0.917
Sentiment									
VADER	1.015	1.016	1.019	1.024	1.009	0.998	0.981	0.979	0.985
TextBlob	0.989	1.011	0.985	0.974	0.963	0.961	0.938*	0.929*	0.943
LLM signals (30-days MA smooth)									
BERT-base	0.982	0.986	0.965	0.949	0.923**	0.908**	0.894**	0.868**	0.872***
InflaBERT	1.006	0.978	0.923	0.903*	0.894**	0.891**	0.879**	0.882	0.824***
FinBERT	0.999	1.004	0.976	0.964	0.946	0.928**	0.905*	0.888*	0.871***
Qwen 2.5 0.5B	0.986	1.009	0.996	0.972	0.960	0.947*	0.928**	0.907**	0.902***
Qwen 2.5 1.5B	0.986	1.006	0.997	0.983	0.968*	0.959**	0.950**	0.946	0.947**
Qwen 2.5 7B	0.981	0.996	0.995	0.982	0.958*	0.947**	0.934**	0.912**	0.896***
Gemma 2 2B	0.990	1.012	0.998	0.981	0.961*	0.951**	0.936**	0.917**	0.920***
Gemma 2 9B	0.996	1.003	0.984	0.962	0.937**	0.940**	0.911***	0.897**	0.905***
Gemma 2 27B	0.997	0.990*	1.000	1.004	1.012*	1.012	1.028***	1.027***	1.013***
LLaMA 3.2 1B	1.071***	1.090***	1.123***	1.118***	1.118***	1.104***	1.149***	1.151***	1.134***
LLaMA 3.2 3B	0.990	0.983	0.982	0.964	0.946**	0.935**	0.918**	0.904**	0.898***
LLaMA 3.1 8B	0.989	0.998	1.006	0.998	0.988	0.980	0.979	0.974	1.017
LLaMA 3.1 70B-instruct	0.943	0.851**	0.798***	0.794***	0.807***	0.806***	0.813***	0.839**	0.816***
LLM signals (60-days MA smooth)									
BERT-base	0.977	0.991	0.968	0.955	0.932*	0.915**	0.899**	0.888*	0.897**
InflaBERT	0.997	0.971	0.928	0.898*	0.889*	0.878**	0.868**	0.887	0.823***
FinBERT	0.989	0.996	0.979	0.972	0.955	0.934*	0.909*	0.894*	0.895***
Qwen 2.5 0.5B	0.987	1.006	0.996	0.972	0.950*	0.943**	0.924***	0.904**	0.918***
Qwen 2.5 1.5B	0.986	1.007	1.007	0.992	0.973	0.960**	0.952**	0.949	0.962
Qwen 2.5 7B	0.978	0.998	0.999	0.977	0.964	0.947**	0.930**	0.905**	0.908***
Gemma 2 2B	0.987	1.008	1.000	0.982	0.959	0.950**	0.933***	0.918**	0.942**
Gemma 2 9B	0.989	1.001	0.988	0.955*	0.940**	0.928**	0.912***	0.898**	0.902***
Gemma 2 27B	0.991	0.988	0.992	0.999	1.006	1.007	1.021***	1.026***	1.016***
LLaMA 3.2 1B	1.074***	1.090***	1.119***	1.122***	1.118***	1.106***	1.152***	1.157***	1.171***
LLaMA 3.2 3B	0.989	0.989	0.982	0.964*	0.950**	0.938**	0.915**	0.911*	0.927**
LLaMA 3.1 8B	0.983	0.996	0.999	1.008	0.986	0.979	0.972**	0.958**	1.019
LLaMA 3.1 70B-instruct	0.971	0.867**	0.794***	0.764***	0.774***	0.779***	0.794***	0.839**	0.802***
LLM signals (90-days MA smooth)									
BERT-base	0.978	0.988	0.966	0.959	0.935*	0.919**	0.908**	0.881*	0.906**
InflaBERT	0.999	0.986	0.935	0.892*	0.895*	0.876**	0.867**	0.882	0.826***
FinBERT	0.984	0.995	0.981	0.968	0.957	0.944	0.924*	0.892*	0.900**
Qwen 2.5 0.5B	0.987	1.001	0.996	0.974	0.953	0.938**	0.920***	0.907**	0.926***
Qwen 2.5 1.5B	0.986	1.009	1.005	0.995	0.980	0.964*	0.952**	0.959	0.962
Qwen 2.5 7B	0.983	1.000	0.996	0.986	0.965	0.949*	0.929***	0.907**	0.908***
Gemma 2 2B	0.987	1.004	1.002	0.981	0.964	0.947**	0.926***	0.919**	0.941**
Gemma 2 9B	0.994	1.001	0.984	0.963	0.944**	0.927**	0.903***	0.898**	0.905***
Gemma 2 27B	0.997	0.988	0.985*	0.992**	1.001	1.005	1.016***	1.019***	1.011***
LLaMA 3.2 1B	1.070***	1.088***	1.125***	1.125***	1.124***	1.109***	1.153***	1.167***	1.172***
LLaMA 3.2 3B	0.989	0.996	0.987	0.967*	0.957**	0.940**	0.917**	0.918*	0.917***
LLaMA 3.1 8B	0.994	1.002	1.010	1.009	0.985	0.981	0.965**	0.957**	1.026
LLaMA 3.1 70B-instruct	0.982	0.912**	0.803***	0.759***	0.765***	0.764***	0.786***	0.840*	0.808***

Table 12: Subreddit: **r/Economics**. Target: **CPI**. CRPS is computed as a ratio: $CRPS(AR_x(1)/CRPS(AR(1))$. r/Economics. First in-sample window:2009m1-2018m12; oos: 2019m1-2023m12. Forecast with expanding window re-estimating the model each step. Significance levels are based on DM test p-values: $p < 0.001$ (***), $p < 0.01$ (**), $p < 0.05$ (*).

Model	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=9$	$h=12$	$h=18$
Expectations									
Michigan Survey	0.990	0.986	0.981	0.953	0.929	0.918	0.820***	0.748***	0.849**
1-Y Inflation Swap	0.985	0.992	0.991	0.978	0.975	0.963	0.925**	0.919**	0.977
Sentiment									
VADER	0.933*	0.943	0.956	0.924	0.918	0.932	0.957	0.971	1.002
TextBlob	0.986	0.968	0.966	0.937	0.957	0.956	0.957	0.939	0.848***
LLM signals (30-days MA smooth)									
BERT-base	0.956	0.987	0.968	0.992	0.997	1.010	1.014	1.006	0.945
InflaBERT	0.958	0.977	0.963	0.955	0.966	0.977	0.978	0.992	0.944
FinBERT	0.978	0.970	0.964	0.964	0.979	0.992	1.003	0.999	0.943
Qwen 2.5 0.5B	0.989	1.002	1.004	1.000	1.004	1.012	1.021	1.001	0.930
Qwen 2.5 1.5B	0.981	1.008	1.013	1.013	1.024	1.025	1.030	1.021	0.962
Qwen 2.5 7B	0.915*	0.902*	0.886	0.886	0.896	0.931	0.906	0.906	0.855**
Gemma 2 2B	0.985	0.997	1.004	0.998	1.011	1.023	1.025	1.010	0.969
Gemma 2 9B	0.977	0.999	0.995	0.988	1.001	1.007	1.010	0.994	0.937
Gemma 2 27B	0.982	0.955	0.915	0.903	0.934	0.948	0.963	0.996	1.012
LLaMA 3.2 1B	0.983	0.999	0.982	0.991	0.997	1.011	1.002	0.983	0.947
LLaMA 3.2 3B	0.982	1.007	0.986	0.996	1.013	1.026	1.037	1.019	0.976
LLaMA 3.1 8B	1.013	1.018	1.029	1.016	1.034	1.017	0.997	0.980	0.941*
LLaMA 3.1 70B-instruct	0.907**	0.903*	0.860**	0.866*	0.877*	0.897*	0.886*	0.914	0.920
LLM signals (60-days MA smooth)									
BERT-base	0.967	0.984	0.983	0.978	0.996	1.029	1.025	1.027	0.937
InflaBERT	0.964	0.977	0.969	0.953	0.975	0.987	0.986	0.998	0.936
FinBERT	0.968	0.982	0.970	0.967	0.993	1.013	1.017	1.015	0.932
Qwen 2.5 0.5B	0.965	1.013	1.013	0.992	1.007	1.027	1.019	1.014	0.920
Qwen 2.5 1.5B	0.986	1.014	1.023	1.008	1.025	1.028	1.039	1.027	0.952
Qwen 2.5 7B	0.911*	0.910*	0.879	0.864	0.885	0.925	0.897	0.902	0.842***
Gemma 2 2B	0.972	1.006	1.010	0.988	1.013	1.033	1.023	1.018	0.960
Gemma 2 9B	0.962	1.003	0.997	0.980	0.997	1.023	1.020	1.004	0.926
Gemma 2 27B	0.965	0.962	0.914	0.886	0.903	0.933	0.946	0.987	0.99
LLaMA 3.2 1B	0.967	0.999	1.001	0.974	0.992	1.025	1.013	0.989	0.942
LLaMA 3.2 3B	0.994	1.014	1.015	1.010	1.023	1.037	1.050	1.024	0.971
LLaMA 3.1 8B	1.003	1.008	1.021	1.013	1.021	1.019	1.016	0.989	0.927**
LLaMA 3.1 70B-instruct	0.885*	0.890*	0.855**	0.834*	0.866*	0.888	0.883*	0.901	0.907
LLM signals (90-days MA smooth)									
BERT-base	0.975	0.984	0.990	0.978	1.017	1.035	1.039	1.035	0.934
InflaBERT	0.966	0.967	0.958	0.946	0.970	0.996	0.988	0.988	0.931
FinBERT	0.971	0.976	0.974	0.968	0.998	1.030	1.033	1.022	0.939
Qwen 2.5 0.5B	0.983	0.992	0.989	0.983	1.007	1.035	1.042	1.011	0.908
Qwen 2.5 1.5B	0.997	1.006	1.018	1.005	1.029	1.044	1.053	1.037	0.943
Qwen 2.5 7B	0.915*	0.896*	0.874	0.842	0.873	0.911	0.896	0.891	0.832***
Gemma 2 2B	0.984	0.989	0.998	0.989	1.023	1.044	1.043	1.020	0.944
Gemma 2 9B	0.977	0.986	0.987	0.976	1.005	1.034	1.033	1.005	0.907
Gemma 2 27B	0.987	0.961	0.906	0.857	0.872	0.903	0.919	0.965	0.991
LLaMA 3.2 1B	0.978	0.981	0.991	0.976	1.006	1.027	1.025	0.990	0.940
LLaMA 3.2 3B	1.003	1.006	1.031	1.017	1.054	1.066	1.067	1.036	0.978
LLaMA 3.1 8B	0.992	0.998	1.014	0.997	1.007	1.019	1.013	0.998	0.914**
LLaMA 3.1 70B-instruct	0.890*	0.878*	0.848*	0.836*	0.858*	0.888	0.882*	0.905	0.900*

Table 13: Subreddit: **r/wallstreetbets**. Target: **PCE**. CRPS is computed as a ratio: $CRPS(AR_x(1)/CRPS(AR(1))$. r/Economics. First in-sample window:2009m1-2018m12; oos: 2019m1-2023m12. Forecast with expanding window re-estimating the model each step. Significance levels are based on DM test p-values: $p < 0.001$ (***), $p < 0.01$ (**), $p < 0.05$ (*).

Model	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=9$	$h=12$	$h=18$
Expectations									
Michigan Survey	0.996	0.989	0.994	0.978	0.944	0.910**	0.813***	0.741***	0.886**
1-Y Inflation Swap	1.033	1.016	1.001	0.975	0.944	0.918*	0.872***	0.877**	0.976
Sentiment									
VADER	0.943	0.936	0.925	0.900	0.914	0.926	0.909*	0.934	1.043
TextBlob	1.033	1.046	1.003	0.934	0.922	0.940	0.939	0.938	0.913**
LLM signals (30-days MA smooth)									
BERT-base	1.031	1.030	0.995	0.964	0.976	0.993	1.023	1.086	0.961
InflaBERT	1.011	1.029	0.984	0.944	0.954	0.968	0.986	1.062	0.922
FinBERT	1.032	1.055	1.010	0.962	0.969	0.999	1.016	1.079	0.942
Qwen 2.5 0.5B	1.049	1.052	1.021	0.973	0.989	1.012	1.039	1.141	0.957
Qwen 2.5 1.5B	1.039	1.075	1.037	1.001	1.016	1.020	1.054	1.151	0.980
Qwen 2.5 7B	0.962	0.940	0.884*	0.842**	0.851**	0.873**	0.885*	0.925	0.882**
Gemma 2 2B	1.037	1.051	1.010	0.978	0.991	1.006	1.030	1.162	1.019
Gemma 2 9B	1.027	1.052	1.024	0.977	0.975	0.996	1.019	1.116	0.965
Gemma 2 27B	1.004	1.007	0.962	0.925	0.941	0.968	0.952	0.999	0.972
LLaMA 3.2 1B	1.036	1.047	1.017	0.985	0.983	1.011	1.027	1.122	0.950
LLaMA 3.2 3B	1.038	1.095*	1.067	1.019	1.024	1.033	1.066	1.158	1.009
LLaMA 3.1 8B	1.045	1.042	1.000	0.973	0.992	1.001	1.019	1.068	0.927**
LLaMA 3.1 70B-instruct	0.951	0.923*	0.875**	0.837***	0.854**	0.891**	0.891*	0.952	0.875**
LLM signals (60-days MA smooth)									
BERT-base	1.020	1.042	1.023	0.985	0.978	1.010	1.044	1.113	0.936
InflaBERT	1.015	1.032	1.007	0.961	0.958	0.965	0.994	1.088	0.909
FinBERT	1.029	1.064	1.032	0.999	0.975	0.995	1.039	1.089	0.919
Qwen 2.5 0.5B	1.047	1.056	1.035	0.998	0.980	1.003	1.055	1.153	0.925
Qwen 2.5 1.5B	1.055	1.079	1.067	1.020	1.022	1.026	1.085	1.188	0.963
Qwen 2.5 7B	0.980	0.953	0.880**	0.826**	0.812**	0.847**	0.878*	0.932	0.886
Gemma 2 2B	1.043	1.050	1.039	1.007	0.992	1.011	1.062	1.174	0.976
Gemma 2 9B	1.044	1.050	1.034	1.000	0.975	0.996	1.034	1.138	0.947
Gemma 2 27B	0.964	1.008	0.964	0.912	0.910	0.929	0.933	0.982	0.956
LLaMA 3.2 1B	1.043	1.050	1.037	0.998	0.979	1.000	1.048	1.139	0.911
LLaMA 3.2 3B	1.065	1.093	1.093	1.046	1.028	1.043	1.075	1.186	0.991
LLaMA 3.1 8B	1.027	1.034	1.018	1.001	0.986	1.002	1.032	1.086	0.918**
LLaMA 3.1 70B-instruct	0.949	0.904	0.880**	0.825**	0.834**	0.857**	0.897	0.935	0.856**
LLM signals (90-days MA smooth)									
BERT-base	1.018	1.042	1.041	1.002	0.987	1.009	1.066	1.124	0.919
InflaBERT	1.026	1.031	1.008	0.964	0.955	0.966	1.016	1.081	0.884
FinBERT	1.032	1.050	1.040	1.007	0.985	0.999	1.061	1.090	0.898
Qwen 2.5 0.5B	1.035	1.051	1.039	0.996	0.977	1.010	1.082	1.152	0.892
Qwen 2.5 1.5B	1.048	1.071	1.078	1.049	1.033	1.033	1.132	1.202	0.950
Qwen 2.5 7B	0.998	0.970	0.893*	0.824**	0.802**	0.834**	0.887	0.924	0.850*
Gemma 2 2B	1.033	1.062	1.052	1.012	1.005	1.015	1.100	1.176	0.935
Gemma 2 9B	1.028	1.044	1.035	0.996	0.979	1.004	1.063	1.120	0.922
Gemma 2 27B	0.993	1.004	0.956	0.894	0.879	0.897	0.920	0.939	0.92
LLaMA 3.2 1B	1.041	1.065	1.043	1.007	0.984	1.006	1.067	1.139	0.902
LLaMA 3.2 3B	1.068	1.104	1.119	1.078	1.064	1.059	1.118	1.211*	0.975
LLaMA 3.1 8B	1.028	1.036	1.038	1.006	1.002	1.009	1.068	1.140	0.918
LLaMA 3.1 70B-instruct	0.964	0.916	0.856**	0.811**	0.823**	0.836**	0.902	0.909	0.825***

Table 14: Subreddit: **r/wallstreetbets**. Target: **CPI**. CRPS is computed as a ratio: $CRPS(AR_x(1)/CRPS(AR(1))$. r/Economics. First in-sample window:2009m1-2018m12; oos: 2019m1-2023m12. Forecast with expanding window re-estimating the model each step. Significance levels are based on DM test p-values: $p < 0.001$ (***), $p < 0.01$ (**), $p < 0.05$ (*).

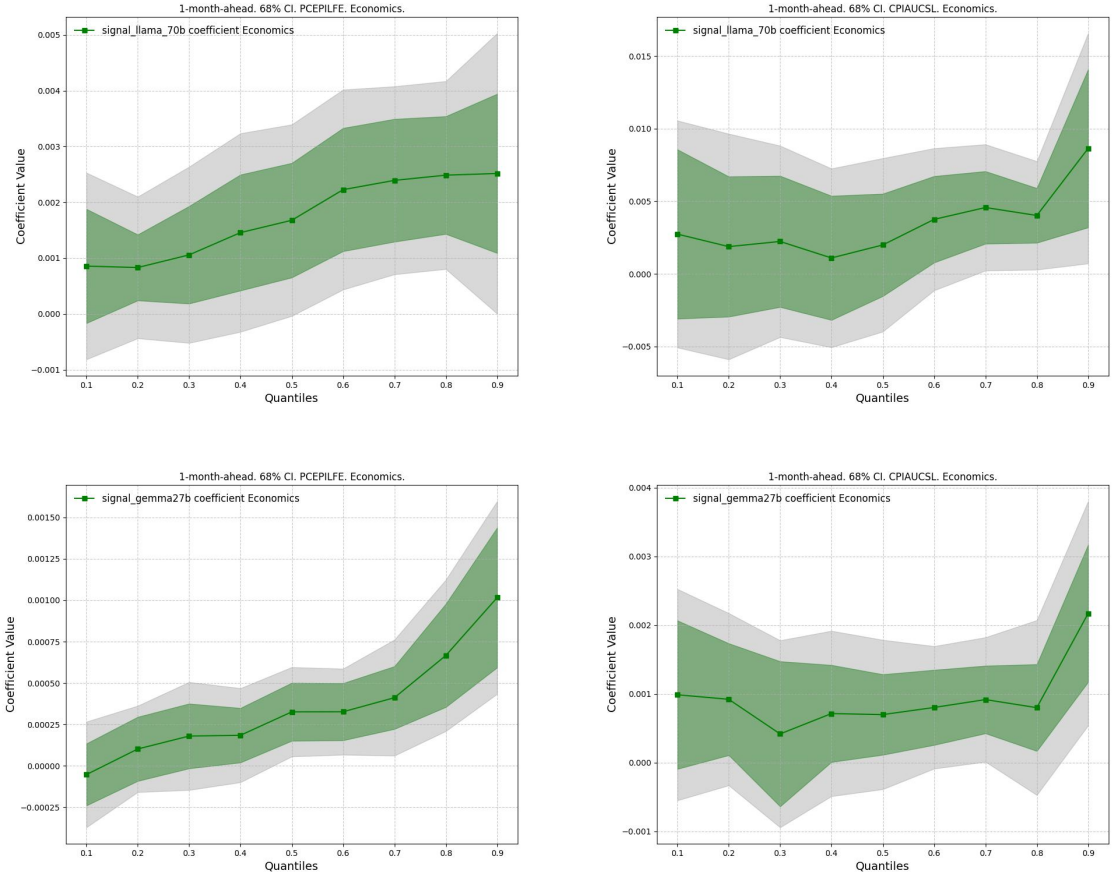


Figure 12: Coefficients for different quantiles of the dependent variable. A one-step-ahead quantile regression is run on the following variables $x_t' = \{\pi_{t-1}, \pi_t^e, Swap_t, X_t\}$, where X_t are in turn the reddit indicators Gemma 2 27B and LLaMA 3.1 70B. We show coefficients for different quantiles of the dependent variables. Confidence intervals are computed with block bootstrap.

The Educated signal behaves similarly to inflation expectations in terms of coefficient estimates. As documented by Goldstein, 2023, belief rigidity tends to decrease during periods of rising inflation, meaning that expectations and sentiment indicators contribute significantly to inflation projection. This finding aligns with existing research on the role of expectations in inflation forecasting. When inflation accelerates, economic agents become more attentive, making expectations a better predictor of future inflation.

In comparing LLaMA 3.1 70B and Gemma 2 27B, we observe that the confidence bands for Gemma 2 27B are narrower and more centered, confirming its superior statistical performance relative to LLaMA 3.1 70B. This finding is consistent with our previoius point-predictive regression results, further reinforcing the robustness of our approach.

7 Conclusion

In this study, we explored the potential of utilizing social media data, specifically from Reddit, to forecast US inflation. By employing a combination of standard text mining techniques and advanced Large Language Models, we derived sentiment indicators that significantly enhance the forecasting performance of traditional time series models.

Our findings indicate a strong correlation between the signal extracted with our LLMs from Reddit discussions and the economic trends reflected in the Personal Consumption Expenditure (PCE) and Consumer Price Index (CPI). This relationship underscores the value of public sentiment as a leading indicator for inflation, providing real-time insights that can complement conventional economic data sources.

The integration of LLMs allowed for a nuanced understanding of the discourse surrounding inflation and price dynamics, capturing the complexities of public opinion that traditional methods like sentiment analysis may overlook. By fine-tuning these LLMs with human-annotated labels discussed with another LLM like CHatGPT, we improved the accuracy of sentiment classification, sometimes increasing the overall forecasting performance compared to larger third-party LLMs.

Overall, this research highlights the importance of incorporating modern computational techniques and big data from social media into economic forecasting. As the landscape of information continues to evolve, leveraging platforms like Reddit can offer valuable insights into public sentiment and its implications for economic dynamics and policymaking. Future work will aim to expand this analysis by incorporating additional data sources and exploring the potential for real-time policy applications based on social media sentiment.

References

Allard, M.-A., Teiletche, P., and Zinebi, A. (2024). Enhancing Inflation Nowcasting with LLM: Sentiment Analysis on News. 13

Angelico, C., Marcucci, J., Miccoli, M., and Quarta, F. (2022). Can we measure inflation expectations using Twitter? *Journal of Econometrics*, 228(2):259–277. 2, 3, 32

Antweiler, W. and Frank, M. Z. (2004). Is all that talk just noise? the information content of internet stock message boards. *The Journal of finance*, 59(3):1259–1294. 4

Aprigliano, V., Emiliozzi, S., Guaitoli, G., Luciani, A., Marcucci, J., and Monteforte, L. (2023). The power of text-based indicators in forecasting italian economic activity. *International Journal of Forecasting*, 39(2):791–808. 4

Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. 13

Aruoba, S. B. and Drechsel, T. (2024). Identifying monetary policy shocks: A natural language approach. Technical report, National Bureau of Economic Research. 3

Ash, E. and Hansen, S. (2023). Text algorithms in economics. *Annual Review of Economics*, 15(1):659–688. 5

Audrino, F., Gentner, J., and Stalder, S. (2024). Quantifying uncertainty: a new era of measurement through large language models. 5

Azqueta-Gavaldón, A., Hirschbühl, D., Onorante, L., and Saiz, L. (2023). Sources of economic policy uncertainty in the euro area. *European Economic Review*, 152:104373. 3

Baker, S. R., Bloom, N., and Davis, S. J. (2016). Measuring economic policy uncertainty. *The quarterly journal of economics*, 131(4):1593–1636. 3

Barbaglia, L., Consoli, S., and Manzan, S. (2022). Forecasting with economic news. *Journal of Business & Economic Statistics*, 41(3):708–719. 4

Bybee, L. (2023). Surveying Generative AI's Economic Expectations. *arXiv preprint arXiv:2305.02823*. 4, 24

Carriero, A., Pettenuzzo, D., and Shekhar, S. (2024). Macroeconomic forecasting with large language models. *arXiv preprint arXiv:2407.00890*. 4

Choi, D., Han, J., Chung, T., Ahn, Y.-Y., Chun, B.-G., and Kwon, T. T. (2015). Characterizing conversation patterns in reddit: From the perspectives of content properties and user participation behaviors. In *Proceedings of the 2015 acm on conference on online social networks*, pages 233–243. 5

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms. 15

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 13

Faria-e Castro, M. and Leibovici, F. (2024). Artificial intelligence and inflation forecasts. Technical report. 4

Giacomini, R. and Rossi, B. (2010). Forecast comparisons in unstable environments. *Journal of Applied Econometrics*, 25(4):595–620. 31, A5

Goldstein, N. (2023). Tracking inattention. *Journal of the European Economic Association*, 21(6):2682–2725. 46

Gueta, A., Feder, A., Gekhman, Z., Goldstein, A., and Reichart, R. (2024). Can LLMs Learn Macroeconomic Narratives from Social Media? *Working Paper*. 4

Hayou, S., Ghosh, N., and Yu, B. (2024). Lora+: Efficient low rank adaptation of large models. 15

Horton, J. J. (2023). Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research. 4

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. (2022). LoRA: Low-rank adaptation of large language models. *ICLR*, 1(2):3. 12

Kalamara, E., Turrell, A., Redl, C., Kapetanios, G., and Kapadia, S. (2022). Making text count: Economic forecasting using newspaper text. *Journal of Applied Econometrics*. First published: 11 May 2022. 4

Liu, S.-Y., Wang, C.-Y., Yin, H., Molchanov, P., Wang, Y.-C. F., Cheng, K.-T., and Chen, M.-H. (2024). DoRA: Weight-Decomposed Low-Rank Adaptation. 12, 15, 17

Long, S., Lucey, B., Xie, Y., and Yarovaya, L. (2023). "I just like the stock": The role of Reddit sentiment in the GameStop share rally. *Financial Review*, 58(1):19–37. 8

Lopez-Salido, D. and Loria, F. (2024). Inflation at risk. *Journal of Monetary Economics*, page 103570. 38

Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. 15

Loughran, T. and McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of finance*, 66(1):35–65. 3

Marcucci, J. (2024). Macroeconomic Forecasting with Text-Based Data. In Michael P. Clements and Ana Beatriz Galvão, editor, *Handbook of Research Methods and Applications in Macroeconomic Forecasting*, chapter 16, pages 425–468. Edward Elgar Publishing. 3, 4

Mitchell, J., Poon, A., and Zhu, D. (2024). Constructing density forecasts from quantile regressions: Multimodality in macrofinancial dynamics. *Journal of Applied Econometrics*, 39(5):790–812. 37

Renault, T. (2017). Intraday online investor sentiment and return patterns in the us stock market. *Journal of Banking & Finance*, 84:25–40. 4

Shapiro, A. H., Sudhof, M., and Wilson, D. J. (2022). Measuring news sentiment. *Journal of econometrics*, 228(2):221–243. 5

A Online Appendix (Not for publication)

A.1 Traditional Sentiment Analysis

Sentiment analysis, also known as opinion mining, is a natural language processing (NLP) technique used to determine the sentiment or emotional tone behind a body of text. This analysis can identify the polarity of the text, whether it is positive, negative, or neutral, and often extends to determining the intensity of the sentiment. Sentiment analysis is widely used in various applications, including social media monitoring, customer feedback analysis, and market research. It leverages machine learning algorithms, lexicon-based approaches, or a combination of both to analyze and interpret textual data.

A.1.1 Textblob

TextBlob's sentiment analysis relies on a straightforward pattern-matching approach, making it more straight forward. The model is pre-trained on a dataset with labelled examples, associating specific words or phrases with sentiment labels such as positive, neutral, or negative. During analysis, TextBlob identifies these patterns in a given text, assigning sentiment polarity scores based on the context and presence of these indicative words. The aggregated score falls within the range of $[-1.0, 1.0]$, where negative values represent negativity, positive values denote positivity, and 0.0 indicates neutrality. This approach prioritizes ease of use and transparency, allowing researchers to grasp the sentiment assignment rationale based on identified patterns. While effective for basic sentiment tasks, it may lack the nuance needed for more complex language nuances like sarcasm or subtle expressions.

A.1.2 VADER

VADER (Valence Aware Dictionary and sEntiment Reasoner) is a sentiment analysis tool which is specifically attuned to sentiments expressed in social media. It is a lexicon and rule-based sentiment analysis tool that is sensitive to both polarity (positive/negative) and intensity (strength) of emotion. VADER relies on a dictionary which maps lexical features to emotion intensities known as valence scores. For example, words like "happy" or "excellent" are assigned positive valence scores, whereas words like "sad" or "terrible" are given negative scores. The tool also considers contextual modifiers, such as degree modifiers (e.g., "very", "extremely") which alter the intensity of the sentiment. VADER is particularly effective in analyzing the sentiment of short texts from social media platforms, making it a suitable choice for sentiment analysis in this study.

Figure A.1: Weighted F1 scores of the validation set in the training loop: models based on the BERT architecture.

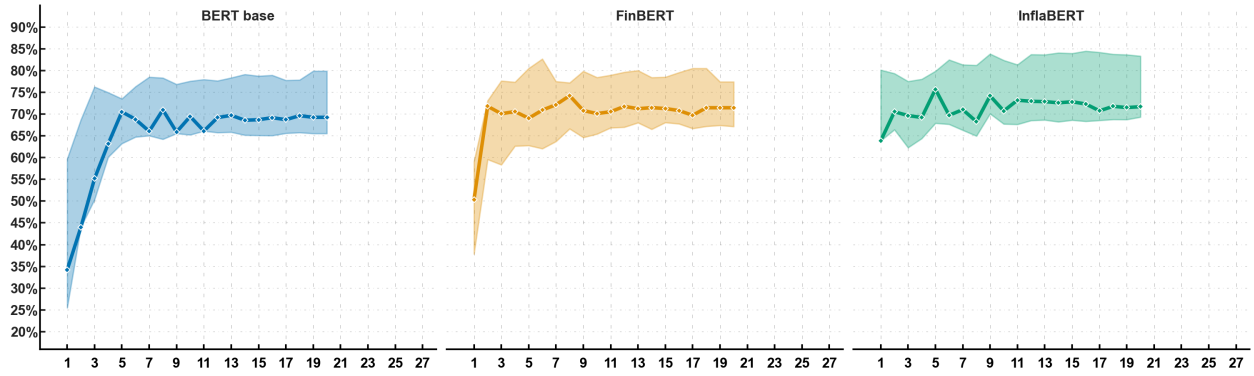


Figure A.2: Weighted F1 scores of the validation set in the training loop: models from the Qwen 2.5 family.

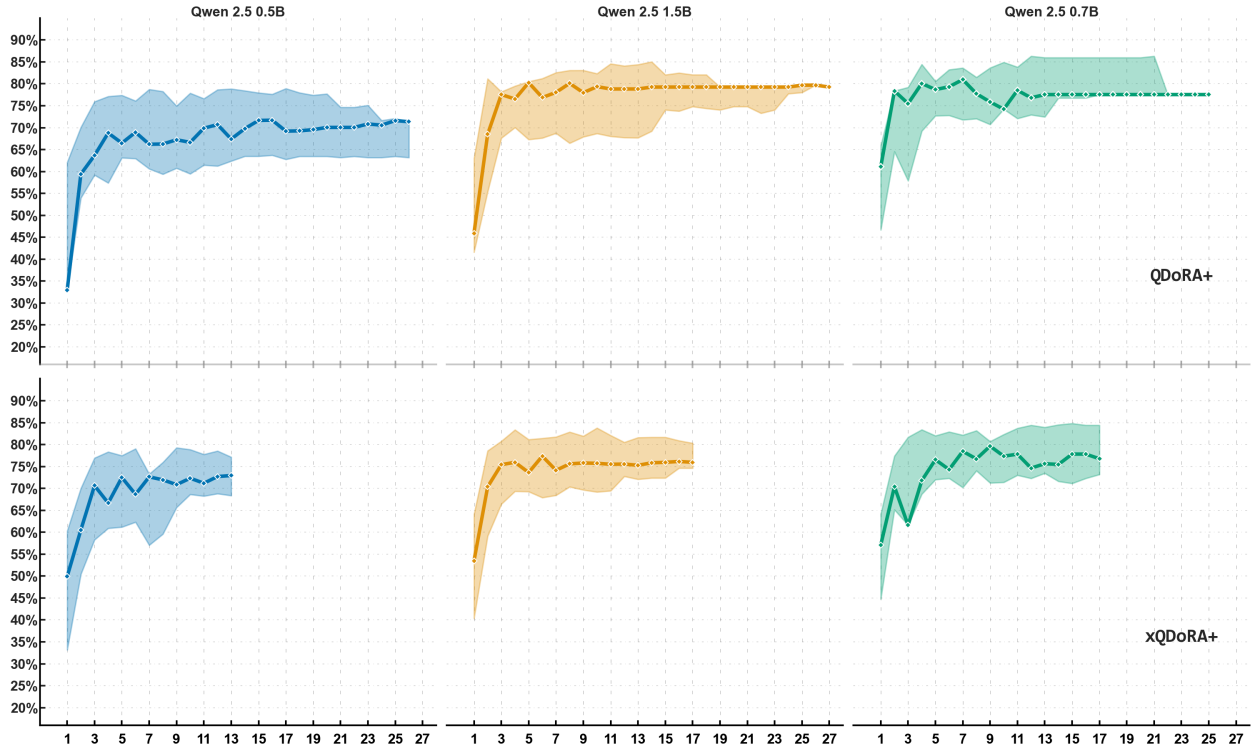


Figure A.3: Weighted F1 scores of the validation set in the training loop: models from the LLaMA 3 family.

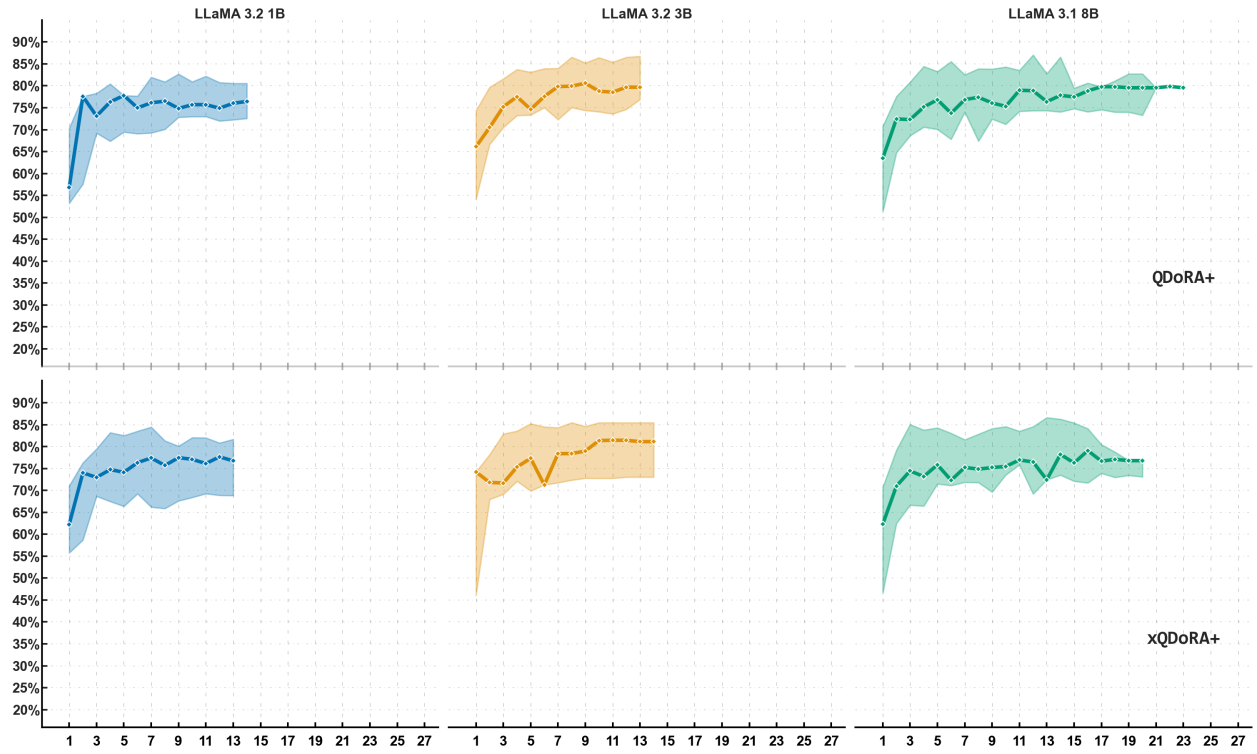
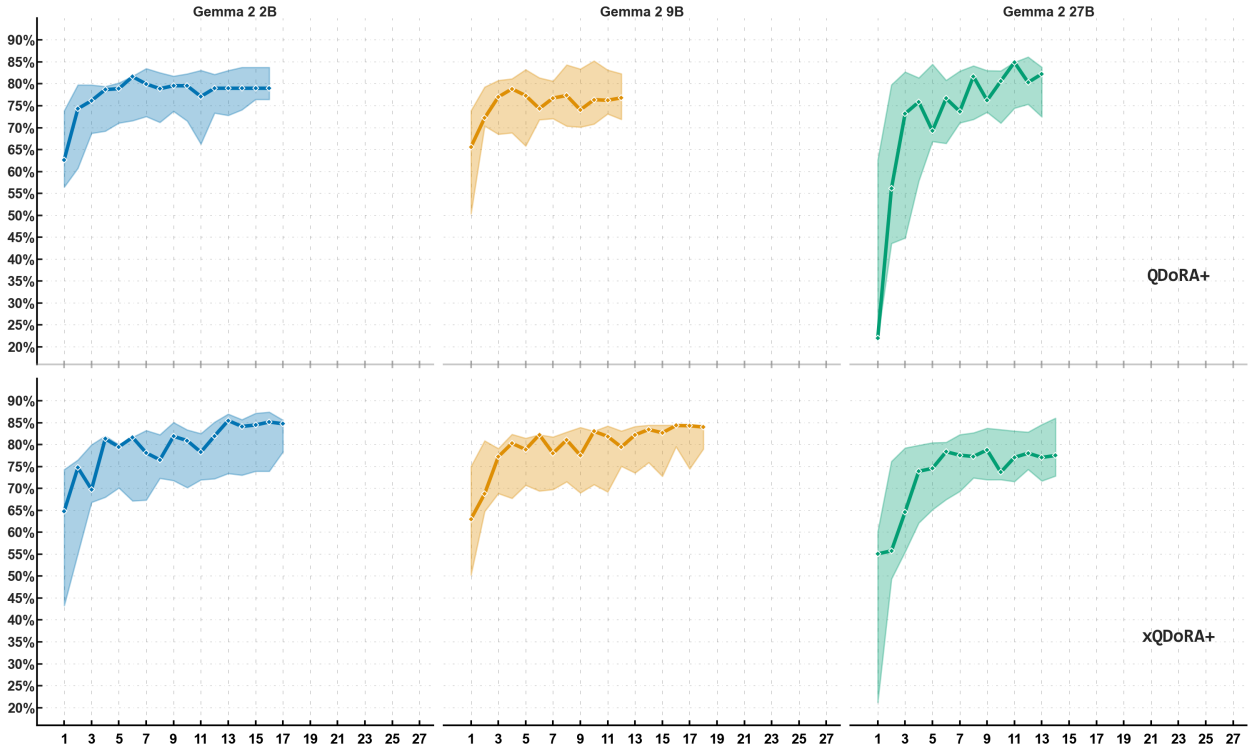


Figure A.4: Weighted F1 scores of the validation set in the training loop: models from the Gemma 2 family.



A.2 Cumulative loss functions

In this section we provide charts for cumulative MSE for the best-performing indicators, at different horizons and MA window. We show cumulative loss for the best 5 models in each MA window, considering also the expectations.

A.3 Fluctuation test

In this section we provide results from the fluctuation test of Giacomini and Rossi (2010) for the forecast application of the subreddit r/Economics. By referring to the methodology of Giacomini and Rossi (2010), we select the window parameter m = 10 and the relative critical value for significance. In general, positive values of the fluctuation test statistic imply that the corresponding model outperforms the benchmark AR(1).

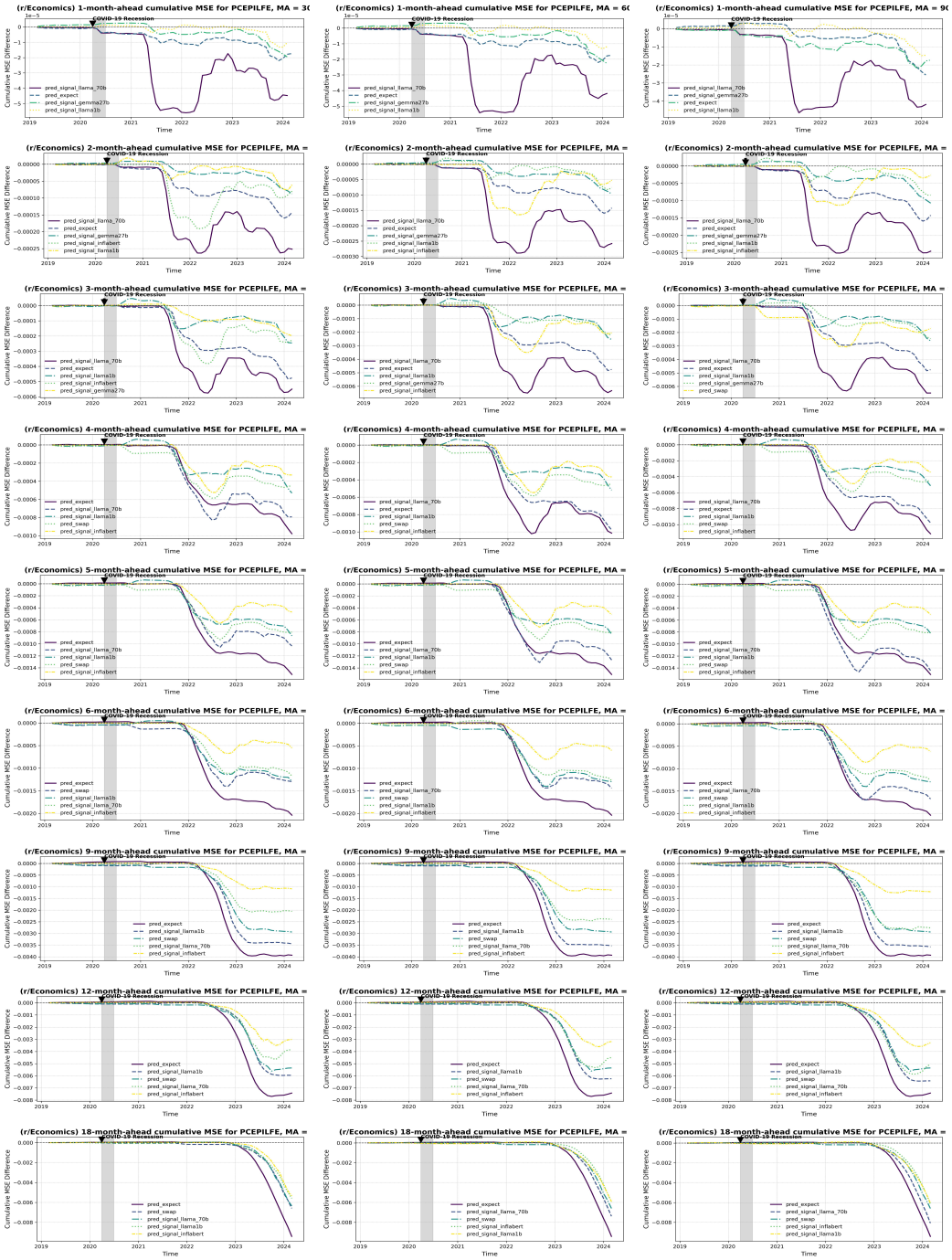
A.4 Robustness check: different MA windows and r/economy results

In this section we provide robustness results for r/economy, as well as additional MA windows for the other subreddits. To facilitate results visualization, we report all the results in histogram format.

A.5 Predictive regression for $h = 6$ and $h = 12$

We show the results for $h = 6$ and $h = 12$. For r/Economics, we notice that the full regression model (with Michigan and swap) yields significant positive coefficients of the best Reddit signal (Gemma 2 27B, LLaMA 3.1 70B). In r/wallstreetbets the best Reddit signal (Qwen 2.5 7B) is always positive and significant for all the horizons considered.

Figure A.5: r/Economics cumulative MSE for PCE



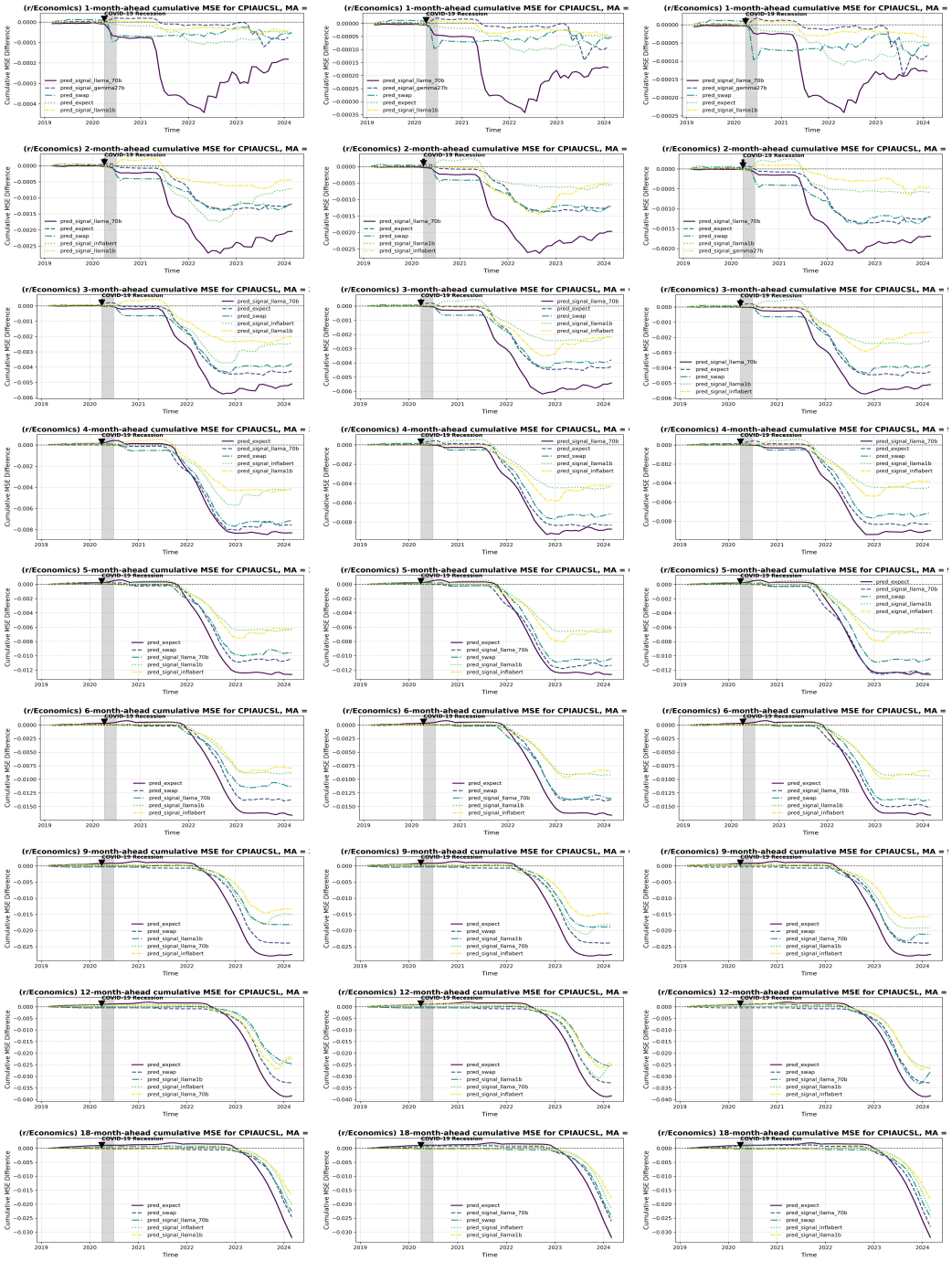


Figure A.6: r/Economics cumulative MSE. CPI.

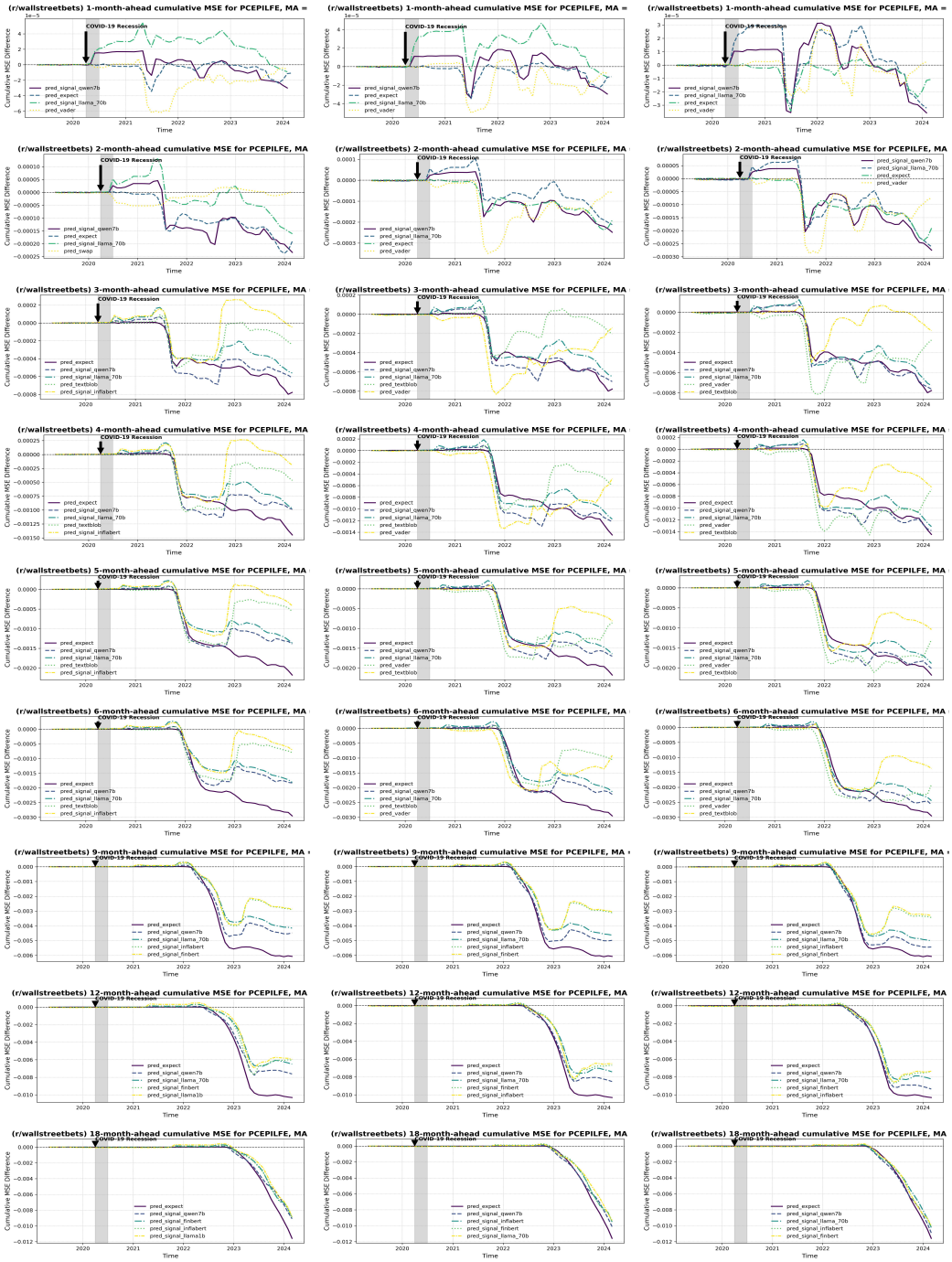


Figure A.7: r/wallstreetbets cumulative MSE. PCE.

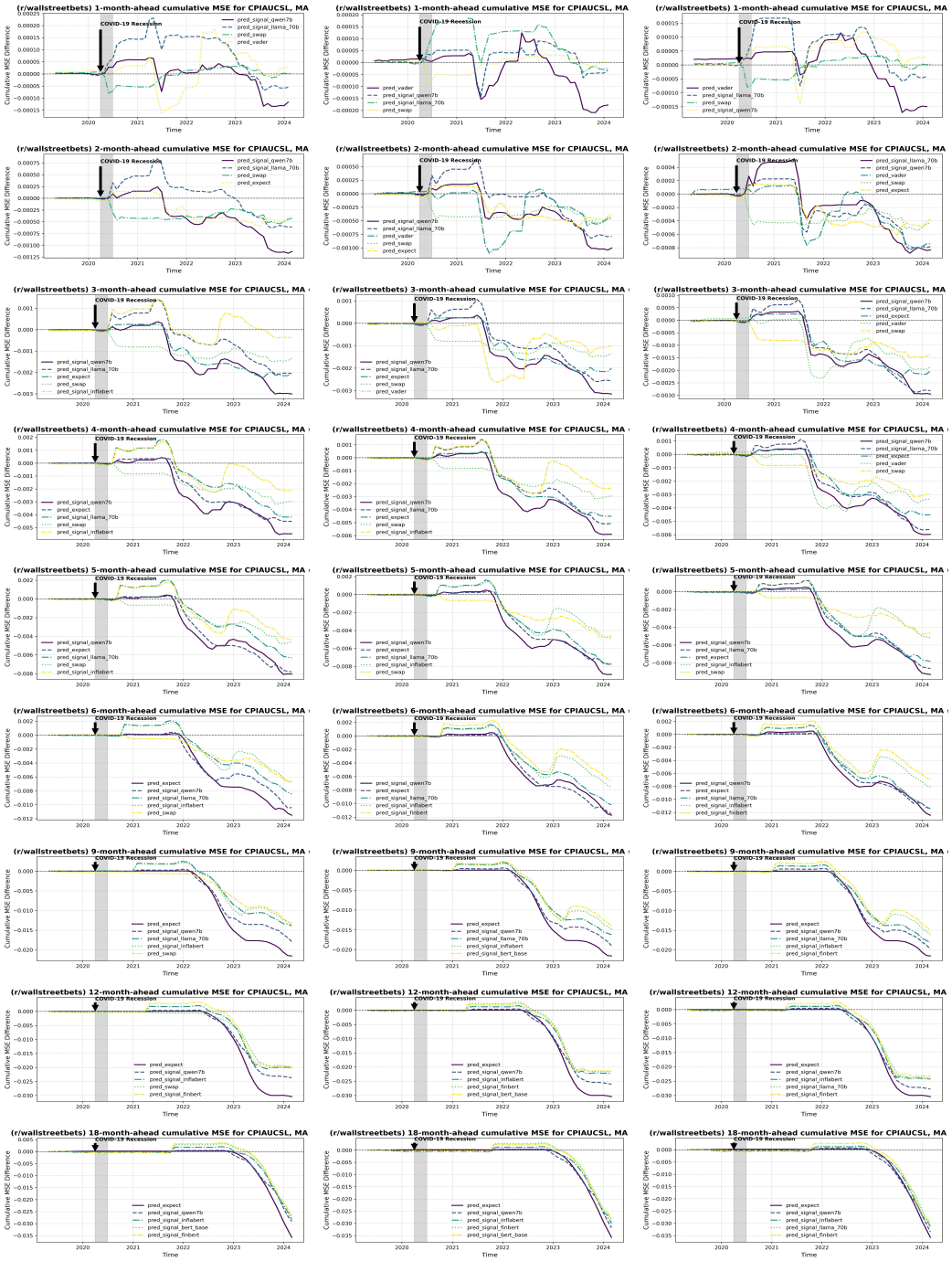
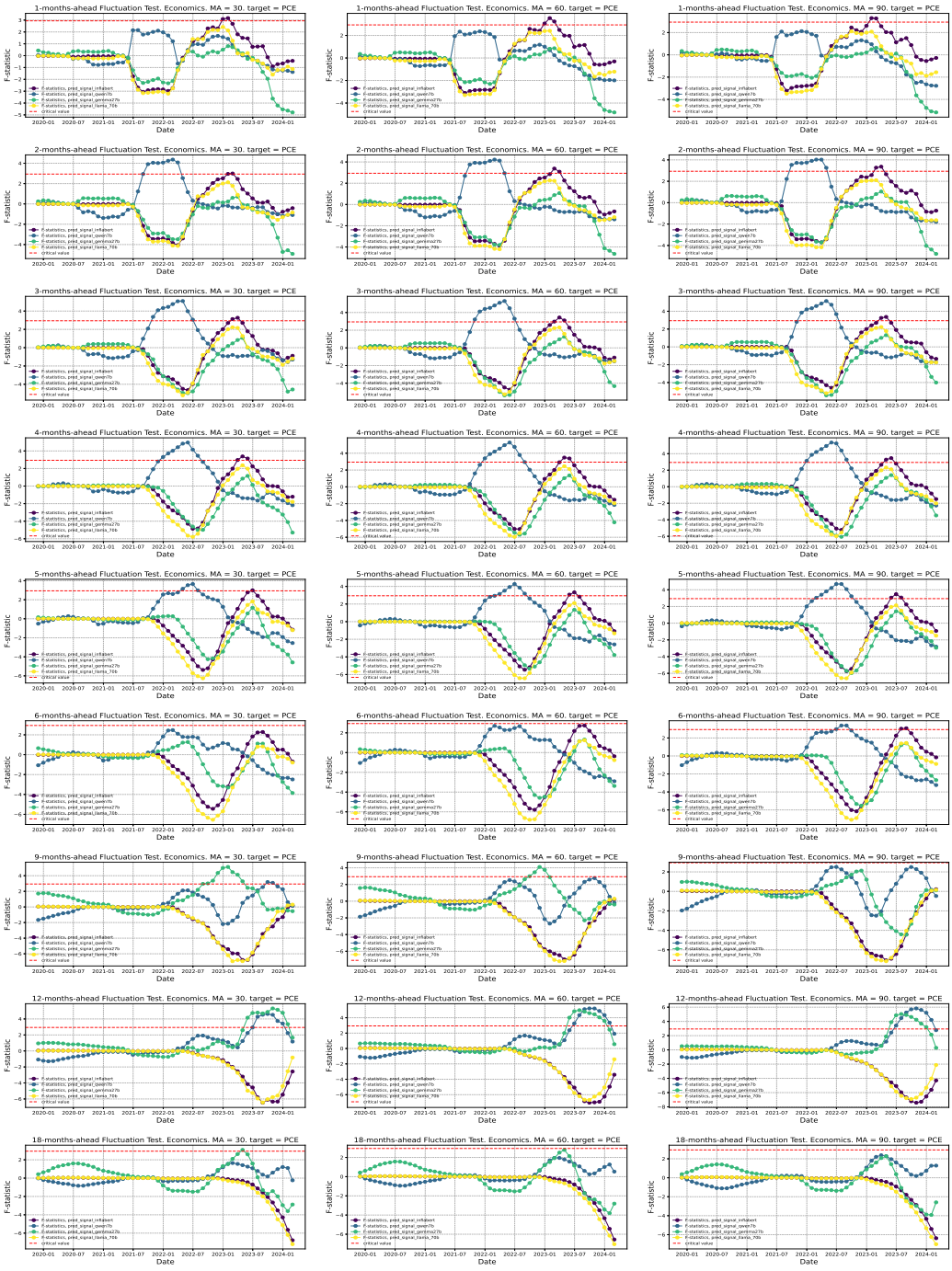


Figure A.8: r/wallstreetbets cumulative MSE. CPI.

Figure A.9: r/Economics fluctuation test statistics for PCE.



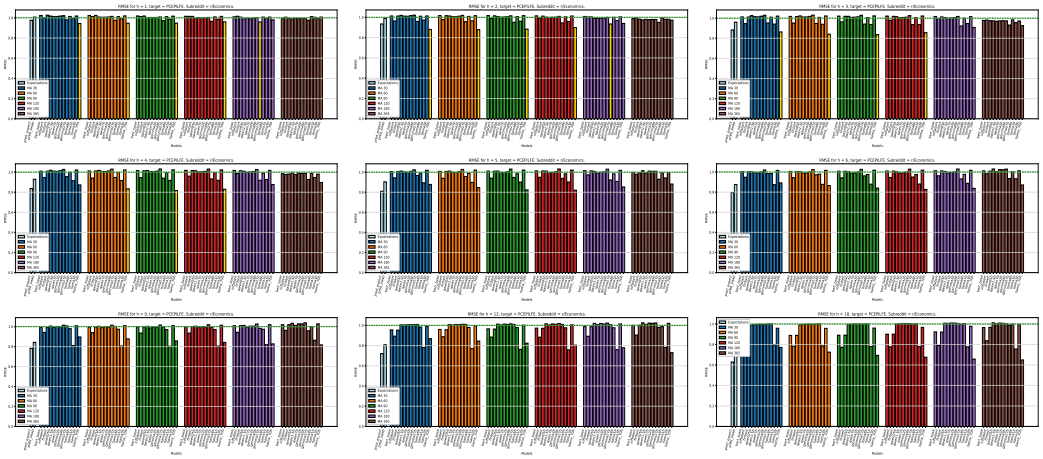


Figure A.10: RMSE for different MA window size from smoothing the Reddit signals. r/Economics. Target: PCE

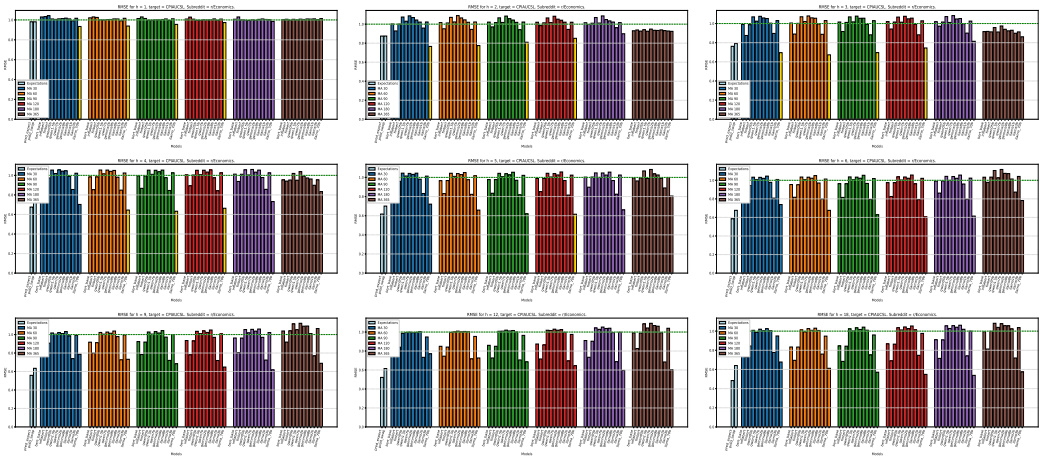


Figure A.11: RMSE for different MA window size from smoothing the Reddit signals. r/Economics. Target: CPI

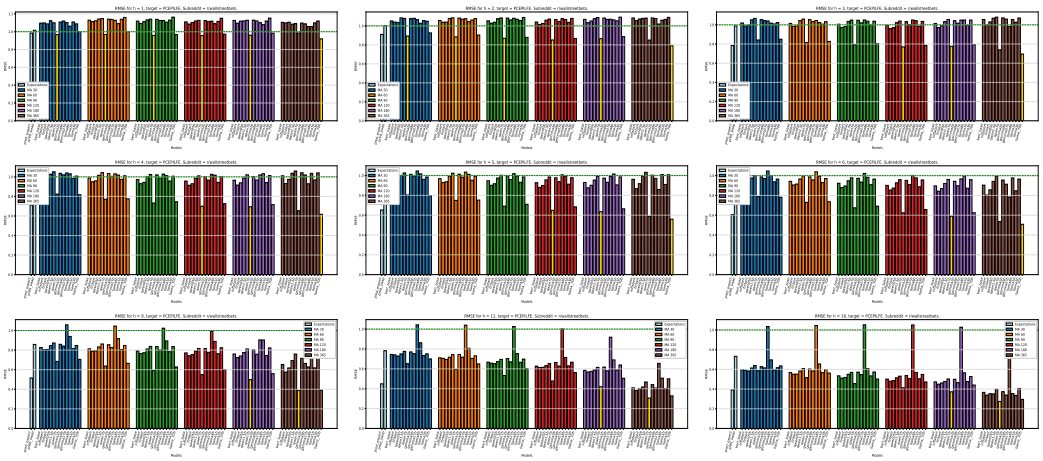


Figure A.12: RMSE for different MA window size from smoothing the Reddit signals. r/wallstreetbets. Target: PCE

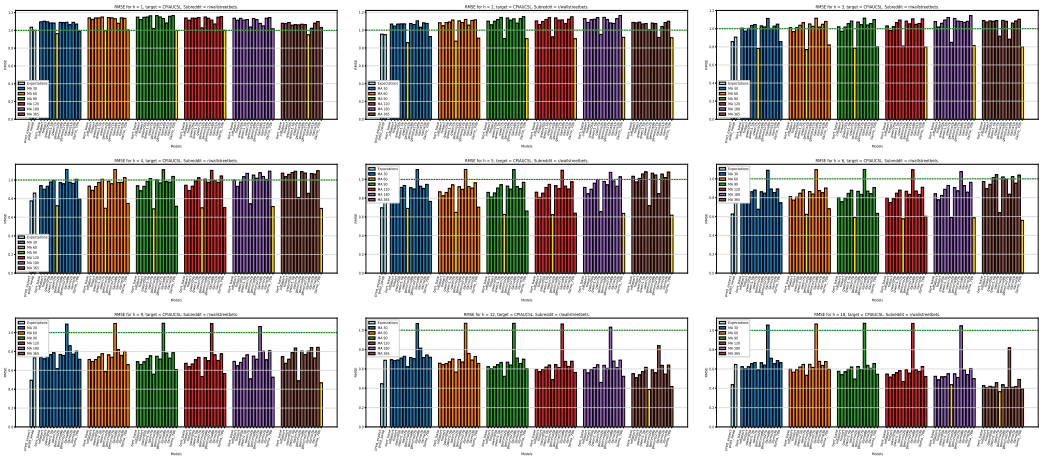


Figure A.13: RMSE for different MA window size from smoothing the Reddit signals. r/wallstreetbets. Target: CPI

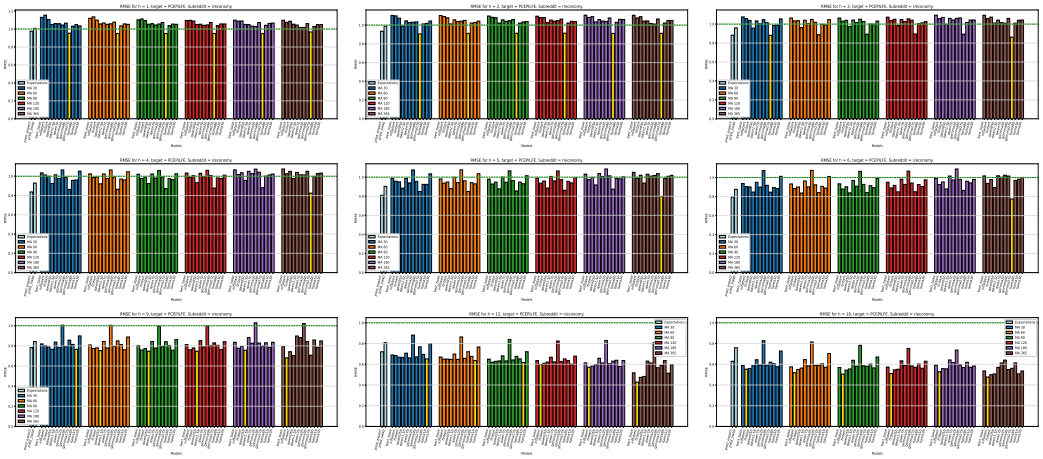


Figure A.14: RMSE for different MA window size from smoothing the Reddit signals. r/economy. Target: PCE

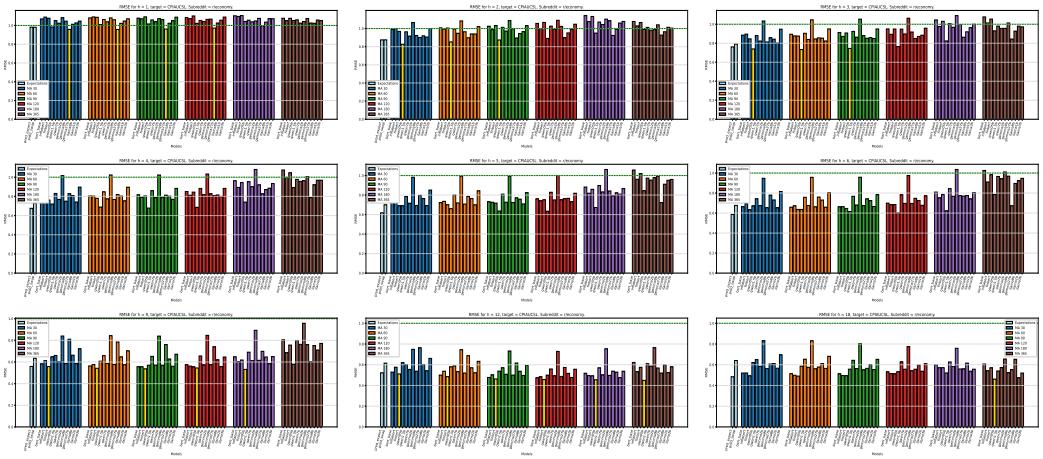


Figure A.15: RMSE for different MA window size from smoothing the Reddit signals. r/economy. Target: CPI

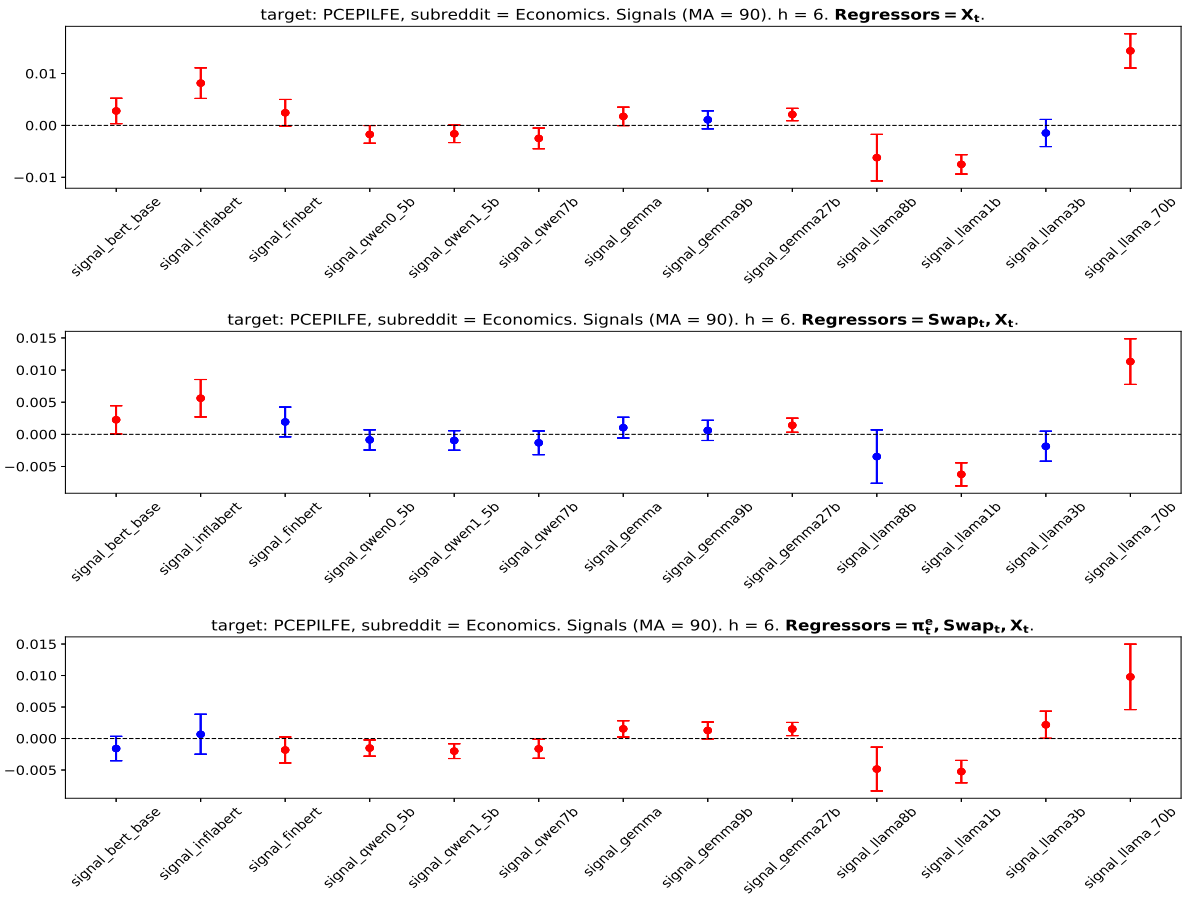


Figure A.16: Target = $\mathbf{PCE}_{t+1}$. Coefficients of Reddit signals for different LLMs. Each plot shows coefficients from Reddit for different predictive regression. Top: $\hat{\pi}_{t+1} = \alpha + \beta_1 \pi_t + \beta_4 X_t + \varepsilon_t$; middle: $\hat{\pi}_{t+1} = \alpha + \beta_1 \pi_t + \beta_3 1YSwap_t + \beta_4 X_t + \varepsilon_t$; bottom: $\hat{\pi}_{t+1} = \alpha + \beta_1 \pi_t + \beta_2 \pi^e_t + \beta_3 1YSwap_t + \beta_4 X_t + \varepsilon_t$. Reddit indicators ($X_t$) are used separately in turn in each regression. The red coefficient are significant according to OLS critical values. Subreddit r/Economics.

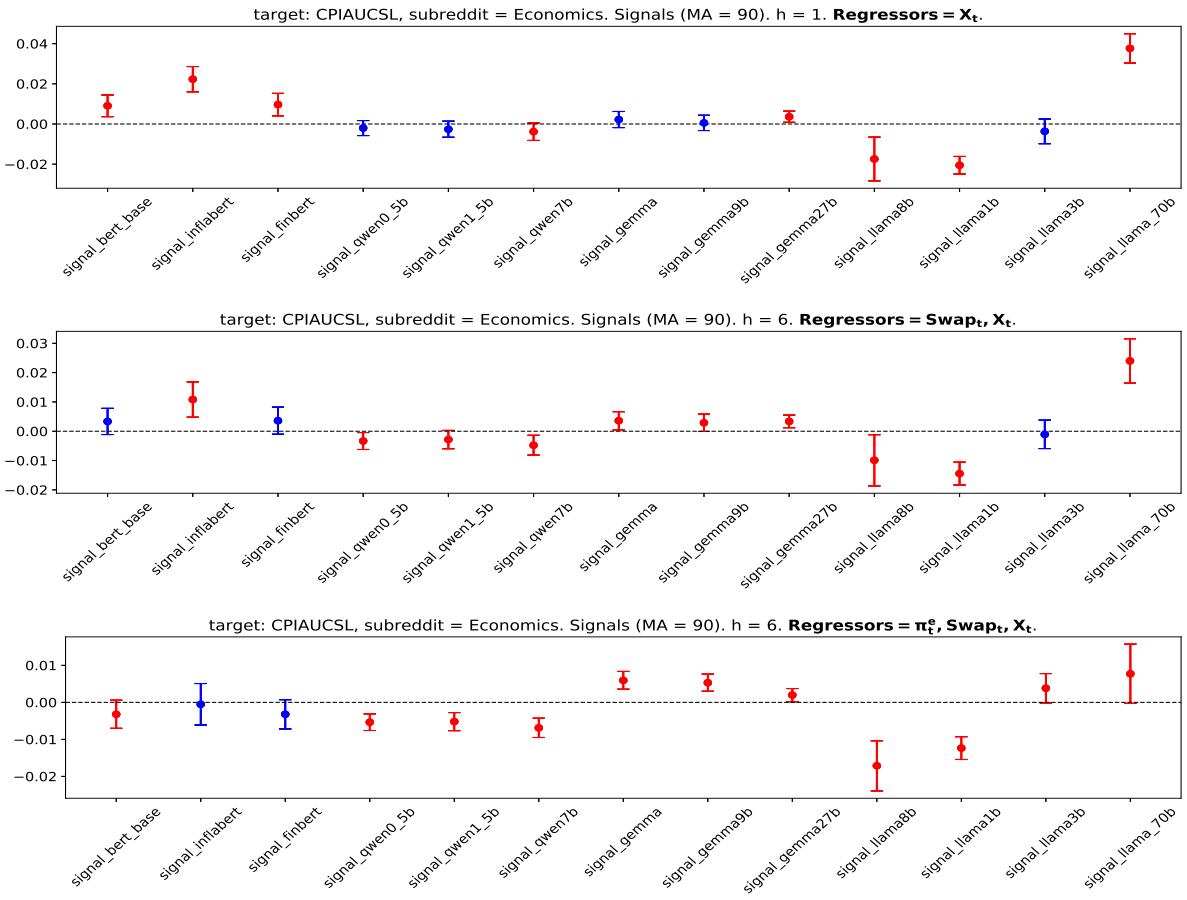


Figure A.17: Target = $\mathbf{CPI}_{t+1}$. Coefficients of Reddit signals for different LLMs. Each plot shows coefficients from Reddit for different predictive regression. Top: $\hat{\pi}_{t+1} = \alpha + \beta_1 \pi_t + \beta_4 X_t + \varepsilon_t$; middle: $\hat{\pi}_{t+1} = \alpha + \beta_1 \pi_t + \beta_3 1Y Swap_t + \beta_4 X_t + \varepsilon_t$; bottom: $\hat{\pi}_{t+1} = \alpha + \beta_1 \pi_t + \beta_2 \pi^e_t + \beta_3 1Y Swap_t + \beta_4 X_t + \varepsilon_t$. Reddit indicators ($X_t$) are used separately in turn in each regression. The red coefficient are significant according to OLS critical values. Subreddit r/Economics.

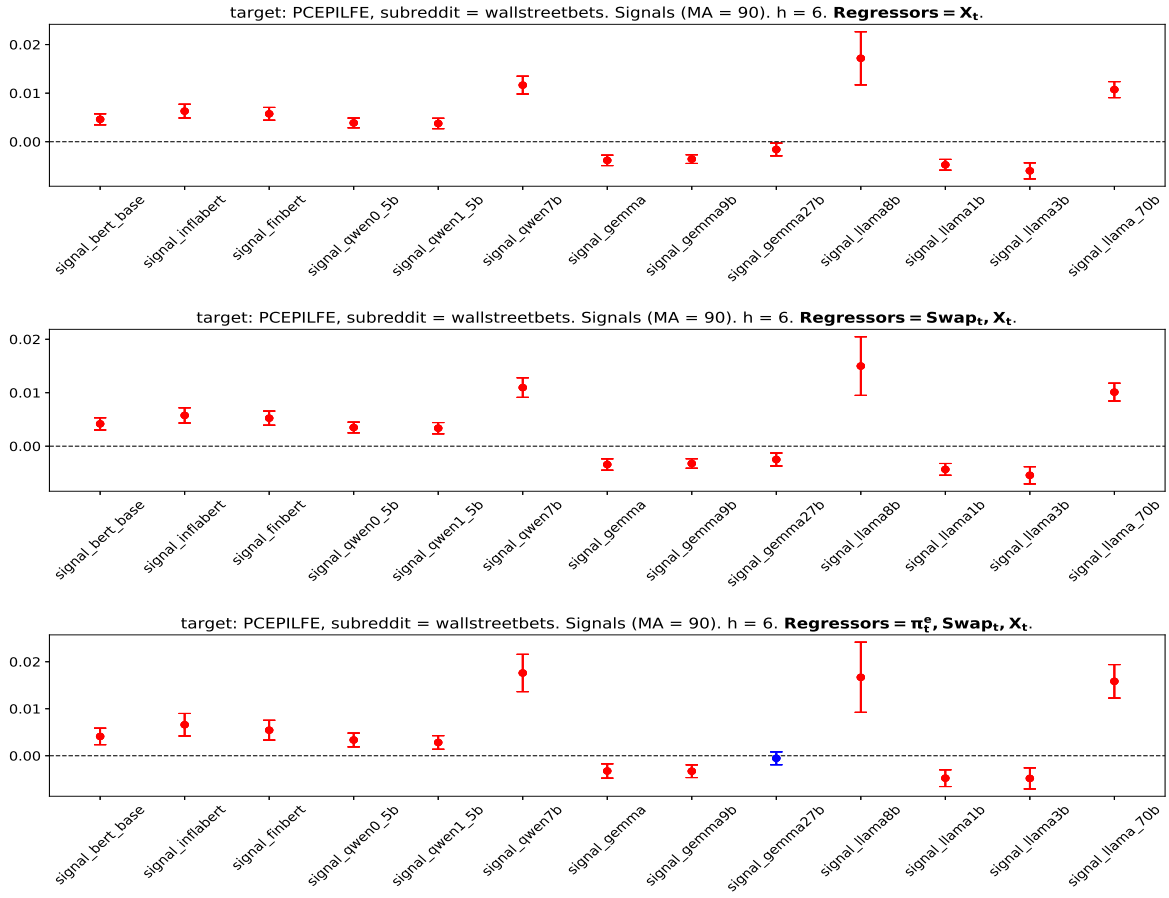


Figure A.18: Target = $\mathbf{PCE}_{t+1}$. Coefficients of Reddit signals for different LLMs. Each plot shows coefficients from Reddit for different predictive regression. Top: $\hat{\pi}_{t+1} = \alpha + \beta_1 \pi_t + \beta_4 X_t + \varepsilon_t$; middle: $\hat{\pi}_{t+1} = \alpha + \beta_1 \pi_t + \beta_3 1YSwap_t + \beta_4 X_t + \varepsilon_t$; bottom: $\hat{\pi}_{t+1} = \alpha + \beta_1 \pi_t + \beta_2 \pi^e_t + \beta_3 1YSwap_t + \beta_4 X_t + \varepsilon_t$. Reddit indicators ($X_t$) are used separately in turn in each regression. The red coefficient are significant according to OLS critical values. Subreddit r/wallstreetbets.

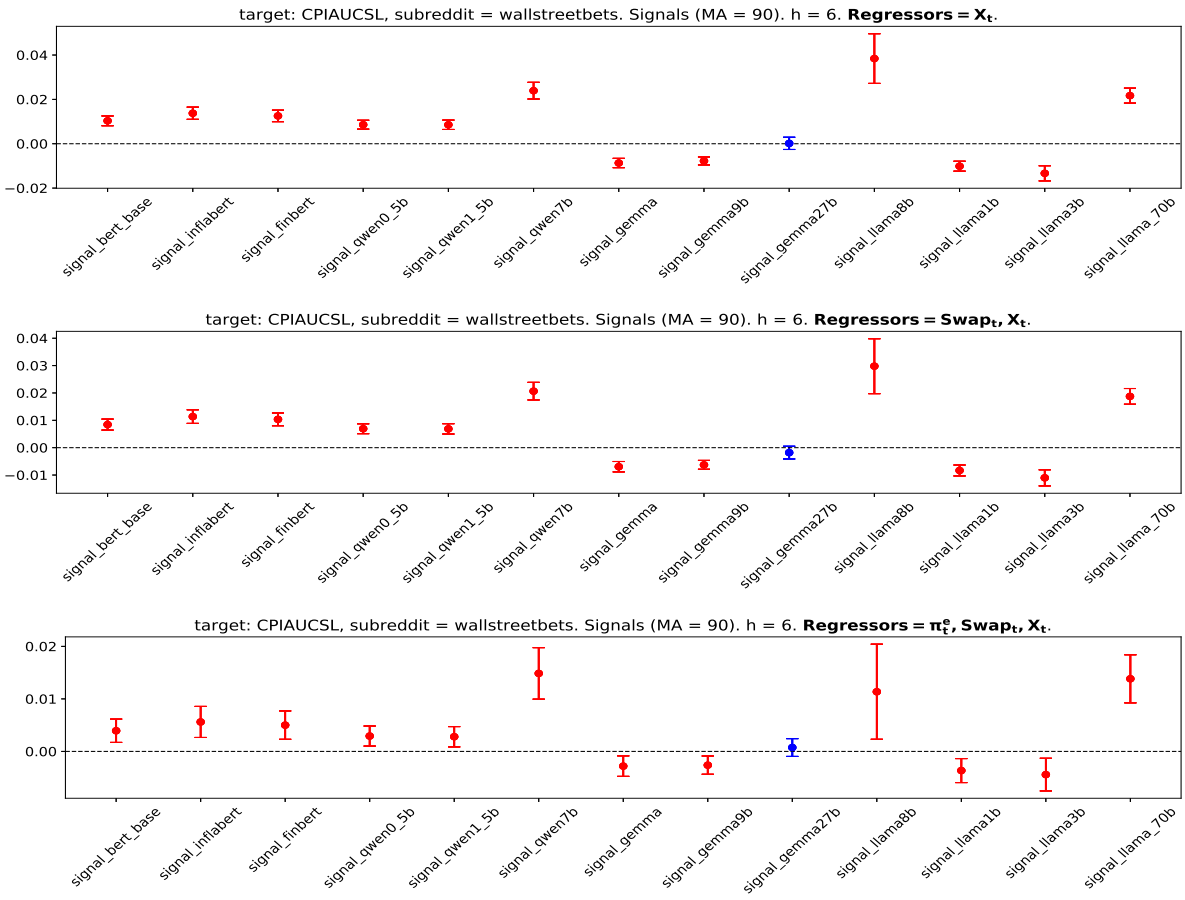


Figure A.19: Target = $\mathbf{CPI}_{t+1}$. Coefficients of Reddit signals for different LLMs. Each plot shows coefficients from Reddit for different predictive regression. Top: $\hat{\pi}_{t+1} = \alpha + \beta_1 \pi_t + \beta_4 X_t + \varepsilon_t$; middle: $\hat{\pi}_{t+1} = \alpha + \beta_1 \pi_t + \beta_3 1YSwap_t + \beta_4 X_t + \varepsilon_t$; bottom: $\hat{\pi}_{t+1} = \alpha + \beta_1 \pi_t + \beta_2 \pi^e_t + \beta_3 1YSwap_t + \beta_4 X_t + \varepsilon_t$. Reddit indicators ($X_t$) are used separately in turn in each regression. The red coefficient are significant according to OLS critical values. Subreddit r/wallstreetbets.

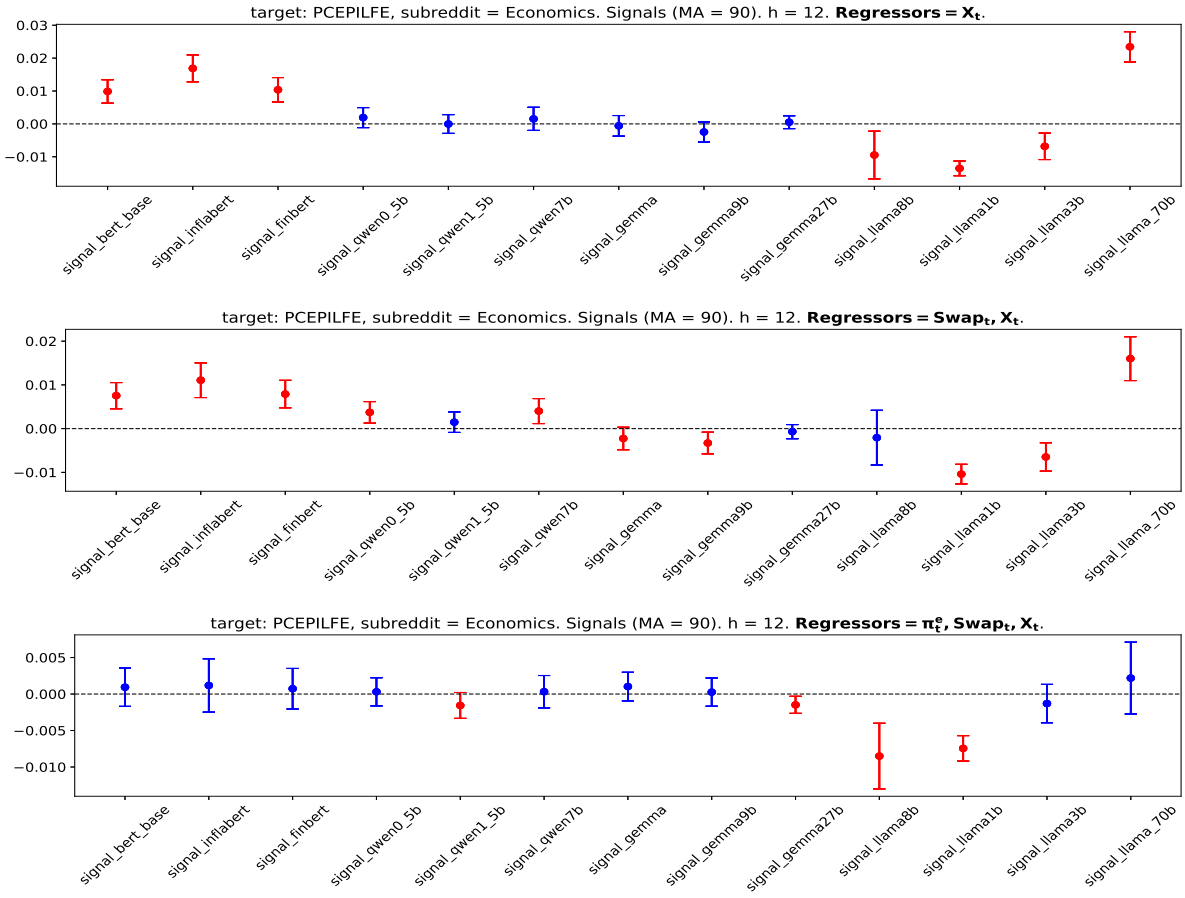


Figure A.20: Target = $\mathbf{PCE}_{t+1}$. Coefficients of Reddit signals for different LLMs. Each plot shows coefficients from Reddit for different predictive regression. Top: $\hat{\pi}_{t+1} = \alpha + \beta_1 \pi_t + \beta_4 X_t + \varepsilon_t$; middle: $\hat{\pi}_{t+1} = \alpha + \beta_1 \pi_t + \beta_3 1YSwap_t + \beta_4 X_t + \varepsilon_t$; bottom: $\hat{\pi}_{t+1} = \alpha + \beta_1 \pi_t + \beta_2 \pi^e_t + \beta_3 1YSwap_t + \beta_4 X_t + \varepsilon_t$. Reddit indicators ($X_t$) are used separately in turn in each regression. The red coefficient are significant according to OLS critical values. Subreddit r/Economics.

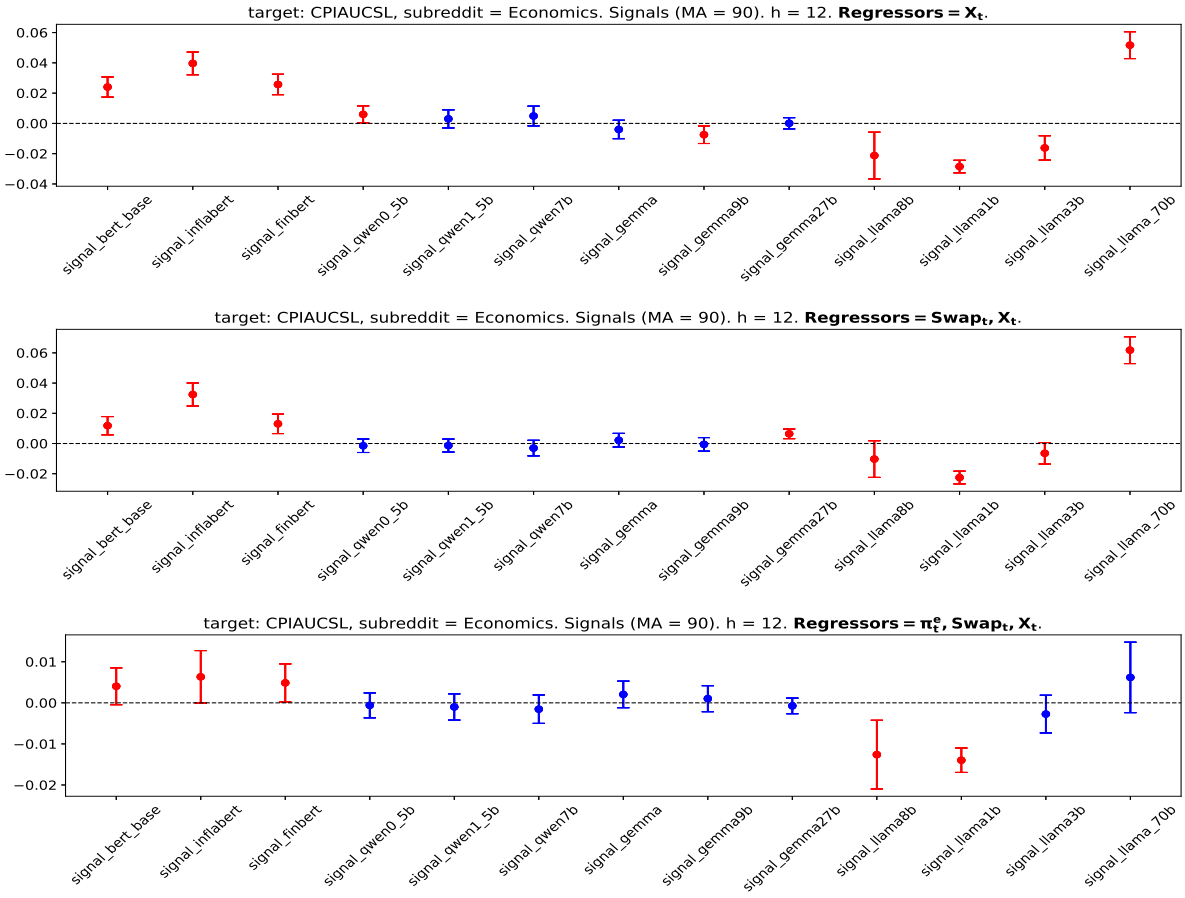


Figure A.21: Target = $\mathbf{CPI}_{t+1}$. Coefficients of Reddit signals for different LLMs. Each plot shows coefficients from Reddit for different predictive regression. Top: $\hat{\pi}_{t+1} = \alpha + \beta_1 \pi_t + \beta_4 X_t + \varepsilon_t$; middle: $\hat{\pi}_{t+1} = \alpha + \beta_1 \pi_t + \beta_3 1YSwap_t + \beta_4 X_t + \varepsilon_t$; bottom: $\hat{\pi}_{t+1} = \alpha + \beta_1 \pi_t + \beta_2 \pi^e_t + \beta_3 1YSwap_t + \beta_4 X_t + \varepsilon_t$. Reddit indicators ($X_t$) are used separately in turn in each regression. The red coefficient are significant according to OLS critical values. Subreddit r/Economics.

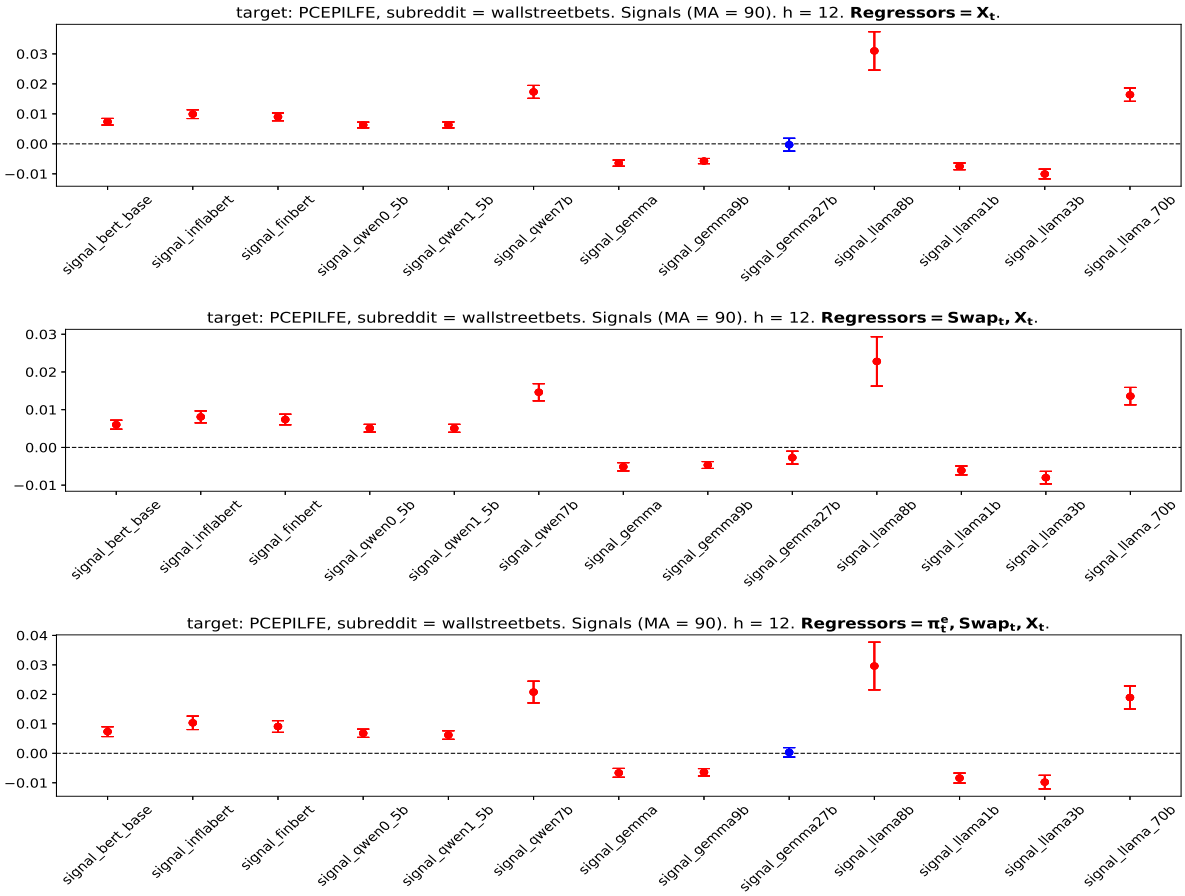


Figure A.22: Target = $\mathbf{PCE}_{t+1}$. Coefficients of Reddit signals for different LLMs. Each plot shows coefficients from Reddit for different predictive regression. Top: $\hat{\pi}_{t+1} = \alpha + \beta_1 \pi_t + \beta_4 X_t + \varepsilon_t$; middle: $\hat{\pi}_{t+1} = \alpha + \beta_1 \pi_t + \beta_3 1YSwap_t + \beta_4 X_t + \varepsilon_t$; bottom: $\hat{\pi}_{t+1} = \alpha + \beta_1 \pi_t + \beta_2 \pi^e_t + \beta_3 1YSwap_t + \beta_4 X_t + \varepsilon_t$. Reddit indicators ($X_t$) are used separately in turn in each regression. The red coefficient are significant according to OLS critical values. Subreddit r/wallstreetbets.

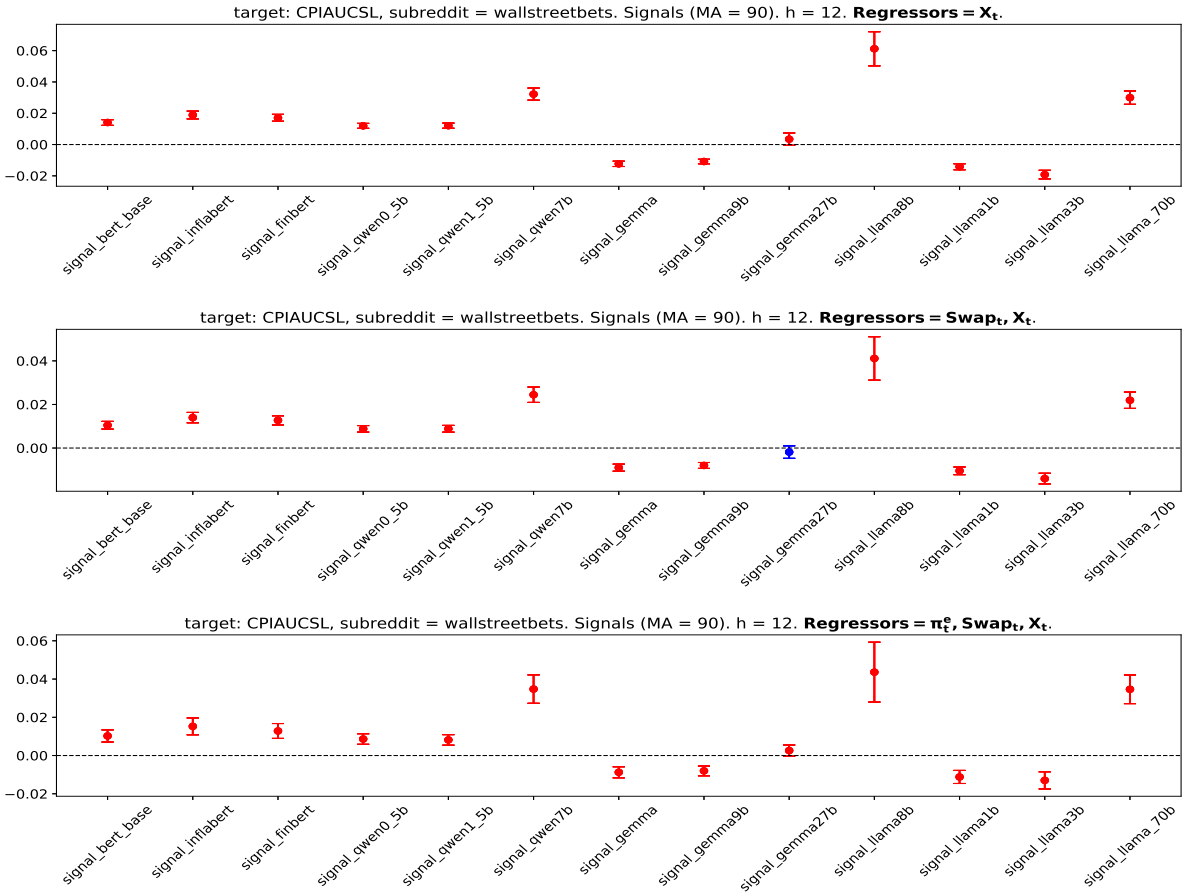


Figure A.23: Target = $\mathbf{CPI}_{t+1}$. Coefficients of Reddit signals for different LLMs. Each plot shows coefficients from Reddit for different predictive regression. Top: $\hat{\pi}_{t+1} = \alpha + \beta_1 \pi_t + \beta_4 X_t + \varepsilon_t$; middle: $\hat{\pi}_{t+1} = \alpha + \beta_1 \pi_t + \beta_3 1YSwap_t + \beta_4 X_t + \varepsilon_t$; bottom: $\hat{\pi}_{t+1} = \alpha + \beta_1 \pi_t + \beta_2 \pi^e_t + \beta_3 1YSwap_t + \beta_4 X_t + \varepsilon_t$. Reddit indicators ($X_t$) are used separately in turn in each regression. The red coefficient are significant according to OLS critical values. Subreddit r/wallstreetbets.