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

March 26, 2026

Abstract

We show that large language models (LLMs) can transform Reddit discussions into timely and informative predictors of U.S. inflation. Using inflation-related submissions and time-local comments from major economics-focused subreddits, we construct monthly narrative indicators that capture users' perceived price dynamics and inflation concerns. The indicators are obtained by fine-tuning pre-trained language models—including BERT-, Qwen-, LLaMA-, and Gemma-type architectures—on labels generated by human annotators and ChatGPT, and we benchmark them against an unfine-tuned LLaMA-70B model. All text processing and indicator construction steps rely on strictly backward-looking transformations. Forecasting and nowcasting are implemented in pseudo–real time with recursive expanding windows, a fixed train–test split (2009M1–2012M12 vs. 2013M1–2025M8), and explicit data-availability

*We gratefully acknowledge comments from M. Affinito, R. Sabbatini, P. Soto and the participants of the 7^{th} Conference on "Nontraditional Data, Machine Learning, and Natural Language Processing in Macroeconomics - ECONDAT 2025 Fall Meeting", the Workshop on Machine Learning Economic Forecasting and Nowcasting held at Copenhagen Business School on April 25, 2025, the 2025 ECB FIAT and AI in Economics workshop on "Harnessing Artificial Intelligence for inflation assessment", and the 6^{th} Conference on "Nontraditional Data, Machine Learning, and Natural Language Processing in Macroeconomics - ECONDAT 2024 Fall Meeting". 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: Juri Marcucci, email: juri.marcucci@bancaditalia.it

constraints, ensuring that results are not affected by look-ahead bias. In a recursive pseudo out-of-sample evaluation with horizons up to 18 months, we produce point and density forecasts for headline CPI and core PCE inflation. We compare a large set of competing models and MSE-weighted forecast combinations using standard accuracy metrics and the Model Confidence Set (MCS) procedure. Reddit–LLM models deliver systematic improvements over conventional benchmarks, including autoregressive specifications augmented with survey-based (Michigan) and market-based (inflation swap) expectations. Up to six months ahead, MSE-weighted combinations of Reddit–LLM models frequently achieve RMSE ratios below one and remain in the final MCS. In operational nowcasting exercises using real-time vintages, Reddit–LLM signals constructed from information available only up to the 5th, 10th, and 14th day of the month (and up to the 22nd day for core PCE) improve nowcasts relative to an AR(1) benchmark and perform competitively with the Cleveland Fed Inflation Nowcast. Predictive regressions indicate that the text indicators add information beyond lagged inflation and standard expectation measures, while density forecasts improve calibration as measured by CRPS. Finally, fine-tuned small language models perform comparably to much larger LLMs, pointing to a scalable approach for integrating narrative indicators into inflation monitoring and forecasting toolkits.

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.

1 Introduction

Accurate forecasting of macroeconomic variables, such as inflation, is essential for policymakers, investors, and economic analysts. Traditional forecasting methods rely on historical data and established indicators derived from surveys run by government agencies. However, social media platforms now provide a rich, real-time source of data that can complement conventional models (Angelico et al., 2022). This paper introduces a novel approach to inflation forecasting by leveraging the dynamic textual data available on Reddit.

Reddit represents a real-time repository of user-generated content, offering a unique opportunity to monitor economic perceptions as they evolve. Unlike traditional survey-based methods–which rely on periodic surveys of limited household samples and are potentially skewed by formal questioning structures–social media allows for the instantaneous expression of opinions and perceptions. Leveraging natural language processing (NLP) and large language models (LLMs), we effectively construct a high-frequency, crowd-sourced barometer of inflation sentiment that provides a "bottom-up" view of price dynamics, significantly reducing the lag between data collection and usable information.

Specifically, we explore the potential of Reddit discussions to nowcast and forecast U.S. headline CPI and core PCE inflation. Reddit hosts numerous subreddits dedicated to economic discourse, where users frequently share their views on inflation and macroeconomic trends. By analyzing these discussions, we extract text-based indicators that reflect public sentiment, potentially serving as early signals of inflation dynamics. Our methodology integrates BERT-based models alongside members of the LLaMA 3, Gemma 2, and Qwen 2.5 model families, fine-tuned using labels generated by ChatGPT and LLaMA 3 through zero-shot learning, subsequently refined by human annotators. We also benchmark against an unfine-tuned LLaMA 70B model. A key concern in applications of LLMs to economic forecasting is look-ahead bias (Sarkar and Vafa, 2024; Bybee, 2023). We take several steps to address this: the LLM is never provided with time stamps or temporal markers; its task is limited to text classification (not forecasting); all indicator construction uses strictly backward-looking moving averages; and all results are produced within a pseudo–real-time design with a fixed train–test split and recursive expanding windows.

Our empirical results demonstrate that LLM-derived indicators enhance forecast accuracy

relative to standard benchmarks, including AR(1) models augmented with Michigan survey expectations and one-year inflation swaps. Improvements are particularly pronounced at short horizons, where inflation is typically dominated by transitory shocks and forecast gains are notoriously difficult to achieve. To address concerns about data snooping given the large number of candidate indicators (288 variants across subreddits, models, and smoothing windows), we aggregate forecasts using MSE-weighted combinations following Clark and McCracken (2010), computed strictly within the estimation sample. We further apply the Model Confidence Set (P. R. Hansen et al., 2011) and the fluctuation test (Giacomini and Rossi, 2010) to assess robustness.

Beyond point forecasting, we also contribute a nowcasting application using real-time data vintages, where Reddit signals available at different intra-month cutoffs improve nowcasts ahead of official CPI and PCE releases. We compare our nowcasts to the Cleveland Fed Inflation Nowcast (Knotek and Zaman, 2017), finding a natural complementarity: the Fed nowcast benefits from fast-moving energy price signals, while Reddit indicators capture narrative information–such as perceived price pressures and expectations–that is especially informative during disinflationary phases. Density forecasts via quantile regressions, predictive regressions assessing marginal information content, and Granger-causality tests linking Reddit signals to Michigan survey expectations provide further economic grounding.

Our findings also reveal substantial heterogeneity across subreddits. Indicators from `r/Economics` tend to dominate at short horizons for CPI, while `r/wallstreetbets` outperforms for PCE and at longer horizons, plausibly reflecting the heightened attentiveness of financially engaged users to inflation dynamics.

The remainder of this paper is structured as follows. Section 2 reviews recent research on text-as-data applications in economic forecasting. Section 3 presents the Reddit dataset and provides descriptive analysis. Section 4 discusses the LLMs used, details the fine-tuning methodology, and outlines the construction of time-series indicators. Section 5 presents the empirical results for forecasting, nowcasting, and supplementary analyses. Section 6 concludes.

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 S. 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 Data

3.1 Reddit corpus and platform structure

We use textual data from Reddit, a social media and an online discussion platform organized into topic-specific communities ("subreddits"), denoted by '*r/Subreddit_Name*'.[1] Within each subreddit, users ("'redditors") create *submissions* (posts) while others contribute *comments* and replies, generating threaded discussions that form a tree structure (Choi et al., 2015). Both submissions and comments can be upvoted or downvoted; the net 'score' (upvotes minus downvotes) proxies for visibility and engagement.[2]

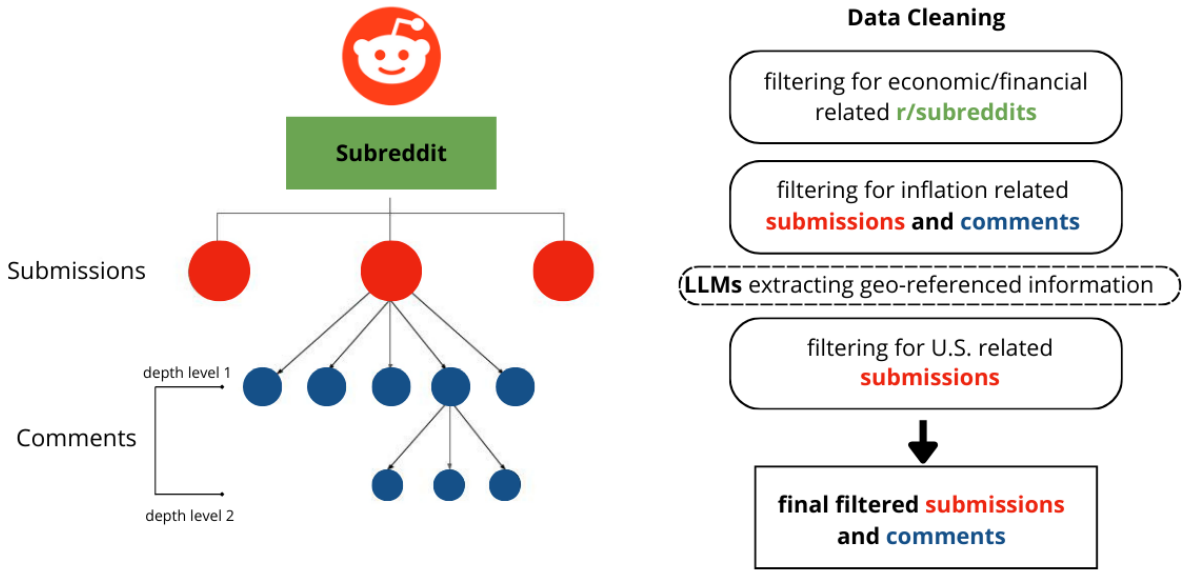
Our baseline corpus comprises three large subreddits on economics and finance that frequently discuss inflation and the cost of living: `r/Economics`, `r/economy`, and `r/wallstreetbets`, with 5.6M, 1.0M, and 19.0M members, respectively, as of July 2025.[3] We distinguish between

[1] According to `https://foundationinc.co/lab/reddit-statistics/` (accessed on February 28, 2026), 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] Submissions vary in format, including text, links to external websites or news articles, images, and videos. Other users can comment, and comments can, in turn, receive replies, creating threaded discussions or comment trees. Submissions and comments can also be upvoted or downvoted, influencing their visibility on the platform.

[3] As of September 2025, Reddit does not publish member counts any more but only the number of weekly visitors and weekly contributions. As of February 2026, `r/Economics`, `r/economy`, and `r/wallstreetbets` have 885K (16K), 230K (7.8K), and 4.2M (171K) weekly visitors (contributions), respectively.

Figure 1: Hierarchical structure of Reddit discussions and overview of the filtering pipeline



Notes: The left panel illustrates the tree structure of Reddit discussions: subreddits (in green) contain submissions (in red), which generate comment threads through replies (in blue). The right panel summarizes our LLM-based multi-step filtering procedure used to isolate timely U.S.-relevant inflation discussions and to construct monthly indicators.

submissions and comments because they play different roles in information production. Submissions typically introduce a topic or link to external news and provide a high-level framing; comments often add granularity, personal experiences, and rebuttals that can sharpen the informational content relevant for inflation monitoring.[4] Figure 1 summarizes the hierarchical structure of Reddit discussions and provides an overview of our filtering pipeline.

Data collection and coverage We assemble the corpus from two data-access regimes reflecting changes in Reddit data availability. Historical content was retrieved from Pushshift snapshots (downloaded in 2023 and updated in early 2024, up to the period when Pushshift access became limited). From mid-2024 onward, new content was collected in real time via the official Reddit API using PRAW. The two sources are merged into a single time-stamped

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

database.

Coverage begins in 2008m1 for `r/Economics`, 2008m3 for `r/economy`, and 2012m4 for `r/wallstreetbets`. The dataset is constructed through 2025m8. The unit of analysis in the forecasting exercise is monthly: we aggregate high-frequency Reddit content into monthly text indicators aligned with monthly inflation data.

Filtering and construction of the inflation-related text sample The raw Reddit corpus is large and heterogeneous, and much of the content is unrelated to aggregate inflation (e.g., asset prices for `r/wallstreetbets`). We therefore implement a transparent multi-step filtering procedure to isolate timely, economically relevant text. The pipeline is summarized in Figure 1 and consists of the following steps.

Step 1: Inflation lexicon filter (submissions and comments). We first identify inflation-related entries using a keyword screen. We retain only submissions and comments containing at least one of the following terms:[5]

$$\{\textit{inflation, deflation, disinflation, hyperinflation, price, prices}\}.$$

This step provides a conservative, easy-to-replicate screen that sharply reduces the volume of irrelevant text before applying more computationally intensive classification.

Step 2: Geographic attribution (submissions). To focus on U.S.-relevant narratives, we apply an LLM-based classifier to infer the country reference of each submission from its title (and associated context where needed). The classifier uses both explicit mentions (e.g., "U.S.", "Fed") and implicit cues (e.g., institutions, policy terms). Submissions classified as U.S.-related, or with no clear geographic attribution, are retained; submissions explicitly linked to non-U.S. countries are removed. The classification protocol is described in Section 4.1.

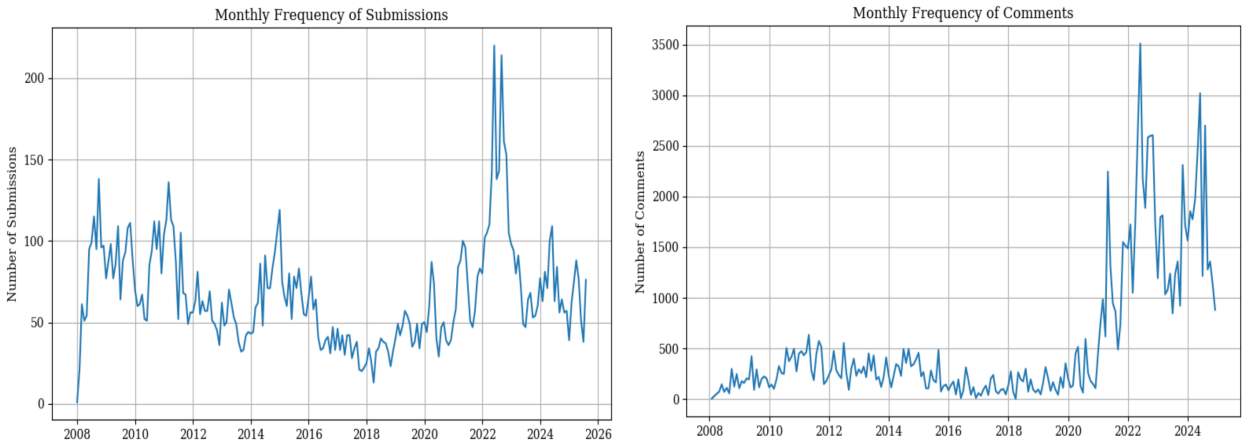
[5] As robustness checks, we expand the lexicon to include terms such as "wages" and "jobs." The resulting indicators and forecasting results are qualitatively unchanged. Furthermore, for `r/wallstreetbets`, we exclude "price" and "prices" because these terms predominantly refer to individual asset prices rather than the aggregate price level.

Step 3: Timeliness restriction for comments. We do not retain the universe of comments. For each retained inflation-related submission, we keep only inflation-related comments posted within two weeks of the submission timestamp, and discard later comments. This restriction is designed to align the comment-based information with a real-time information set and to avoid long-delayed discussions that are less informative for forecasting and nowcasting. Importantly, the restriction is conservative: Appendix A.1 shows that comment activity is heavily front-loaded, with 95% of comments posted within approximately 48 hours of the original submission.

3.2 Descriptive evidence on volume and timing

To gauge the size and time variation of the filtered text sample, Figure 2 plots the monthly number of inflation-related submissions and comments for `r/Economics`. These time series exhibit a marked increase after 2020, with sharp spikes during 2022–2023, consistent with heightened public attention to inflation during the post-pandemic surge. The increase is particularly pronounced for comments, suggesting that user interactions provide a rich and timely source of narrative information.

Figure 2: Monthly frequency of inflation-related submissions and comments in `r/Economics`.



From filtered text to monthly indicators The filtered monthly text sample is subsequently mapped into a set of "educated" narrative indicators using large language models (LLMs). Section 4 details the labeling tasks, fine-tuning strategy, and the aggregation rules that translate model outputs into monthly regressors used for nowcasting and forecasting inflation.

4 LLMs in action: From Text to Inflation Signals

We use LLMs to convert unstructured Reddit discussions into structured, forecast-relevant measures of inflation expectations. The pipeline has three components: (i) geographic filtering to isolate U.S.-relevant content, (ii) human-in-the-loop labeling to construct a high-quality training set, and (iii) large-scale classification of submissions and comments to generate time-series indicators used in our nowcasting and forecasting exercises.

4.1 Country Identification and Data Filtering

We first restrict attention to U.S.-relevant discussions. Using an instruction-tuned LLaMA 70B model, we perform zero-shot classification[6] of each submission into one of three categories: *U.S.-related*, *non-U.S.-related*, or *no clear geographic reference*. We retain submissions classified as U.S.-related and those without a clear geographic attribution, and we discard submissions explicitly linked to non-U.S. countries. The prompt is reported in Appendix B. This step highlights a practical advantage of LLMs for economic measurement: they can recover implicit contextual cues (institutions, policy references, and localized terminology) that are often omitted from short social-media texts.

Table 1 reports the number of submissions and comments by subreddit before and after filtering, illustrating the dimensionality reduction achieved by combining the inflation keyword screen (Section 3.1) with the LLM-based geographic filter.

[6]Zero-shot classification refers to the ability of a model to assign labels without task-specific supervised training.

Table 1: Summary statistics of U.S. inflation-related submissions and comments across subreddits.

Subreddit	Sample	Submissions (overall)	Comments (overall)	Submissions (inflation)	Comments (inflation)
r/Economics	2008m1–2025m8	348,901	6,434,981	14,214	115,276
r/economy	2008m3–2025m8	277,728	2,789,560	14,445	65,170
r/wallstreetbets	2012m4–2025m8	2,550,096	88,480,826	4,801	29,549

Notes: "Overall" counts refer to the full subreddit archives. "Inflation" counts refer to the filtered sample used in the empirical analysis after (i) the inflation keyword screen (Section 3.1) and (ii) the LLM-based country classification (Section 4.1). Comment counts further reflect the timeliness restriction described in Section 3.1.

4.2 Human Labeling Assisted by LLMs

We next construct a labeled training set through a human-in-the-loop procedure described in Appendix C. We define three task-specific labels capturing directional inflation expectations in Reddit submissions: UP, DOWN, and NEUTRAL. The process begins with a manually labeled seed sample (about 500 submissions) and is expanded using a zero-shot LLaMA 70B classifier (about 700 additional submissions). We then fine-tune a smaller LLaMA model (8B) on the resulting labeled dataset and validate it on a held-out test set, where it achieves a weighted F1 score close to 70%.

To improve labeling consistency, we conduct a structured review of disagreement cases and use ChatGPT as an interactive assistant to elicit brief rationales and diagnose ambiguous instances (e.g., implicit references to inflation, sarcasm, or time-horizon confusion). Finally, we extend the labeled dataset with additional ChatGPT-assisted annotation under close human supervision (about 200 submissions). Overall, LLMs serve as annotation accelerators and consistency checkers, while final label assignments remain under human control.

4.3 Inflation signal extraction with fine-tuned LLMs

We fine-tune multiple model families to classify directional inflation expectations at scale across the full Reddit corpus. We consider (i) encoder-based models–BERT base, FinBERT, and InflaBERT[7]–trained via standard supervised fine-tuning, and (ii) decoder-only LLMs–

[7]Developed by Devlin et al. (2018), Araci (2019), and Allard et al. (2024), respectively.

LLaMA, Gemma, and Qwen[8]–adapted using parameter-efficient methods. For the larger models, we employ `QDoRA+`, a quantized, weight-decomposed extension of low-rank adaptation (`QLoRA`) (Liu et al., 2024; Dettmers et al., 2023), and optimize the adapter parameters using LoRA+ (Hayou et al., 2024). We also implement an "extreme" variant, `xQDoRA+`, which further reduces the number of trainable parameters. These configurations enable efficient fine-tuning of billion-parameter models on a single GPU (NVIDIA A100). Appendix D provides additional details on the model families and fine-tuning protocols.

We evaluate classification performance using the weighted F1 score, which accounts for class imbalance and summarizes the precision–recall trade-off. To assess robustness, each model is fine-tuned 20 times using a fixed set of pre-sampled random seeds. For downstream signal extraction, we select the run with the seed achieving the median (high) weighted F1 score across all runs. Figures A.8a–A.8b in the online appendix report the distribution of weighted F1 scores for unfine-tuned ("raw") and fine-tuned models, showing that fine-tuning both improves accuracy and reduces dispersion across runs.

The output of the fine-tuned classifiers–daily shares of UP, DOWN, and NEUTRAL expectations–provides the primary input for our empirical analysis. We aggregate these daily signals into time-series indicators that enter the forecasting and nowcasting models. In this sense, LLMs are not forecasting models per se (cf. Carriero et al. (2024)); rather, they act as *economic sensors* that translate narrative content into measurable predictors.

4.4 Time-series indicator construction

Using the LLM classifications, we construct daily leading indicators that summarize the balance of inflation expectations in Reddit submissions. For each day t, let N_t denote the number of inflation-related submissions. We aggregate submissions labeled as UP or DOWN and compute the daily sum difference ($UP - DOWN$), with NEUTRAL treated as 0, and define the daily submission-based indicator as

$$\widetilde{X}_{\mathrm{LLM},t} = \sum_{n=1}^{N_t} (UP_n - DOWN_n)\,, \tag{1}$$

[8]Developed, respectively, by Touvron et al. (2023), Team (2024), and Qwen et al. (2024).

where the indicator is set to zero on days with no qualifying submissions.[9]

Reddit's threaded structure enables us to incorporate the information contained in comments. For each inflation-related submission i, we link all associated comments via the parent ID, restrict attention to inflation-related comments, and impose a timeliness restriction by retaining only comments posted within two weeks of the submission date. Let $\widetilde{X}^{S_i}_{\text{LLM},t}$ denote the submission signal for thread i on day t and let $\widetilde{X}^{C_{i,j}}_{\text{LLM},t}$ denote the signal for the j-th comment associated with that submission. We construct a thread-level signal as the equally weighted average of the submission and its comments:

$$\widetilde{X}^{S_i,C_i}_{\text{LLM},t} = \frac{\widetilde{X}^{S_i}_{\text{LLM},t} + \sum_{j=1}^{J} \widetilde{X}^{C_{i,j}}_{\text{LLM},t}}{J+1}. \tag{2}$$

We then map this thread-level average back into a discrete label to obtain a comment-adjusted classification for each submission.[10] Figure 3 provides a schematic example where a Gemma 27B signal is re-weighted from NEUTRAL to UP based on the comments discussion.

Combining submissions and comments captures both content generation and user engagement, and it incorporates the social dimension of Reddit discussions. This adjustment is potentially important because submissions often repost news headlines, while comments contain interpretation, personal experiences, and debate that can sharpen the informational content relevant for forecasting.

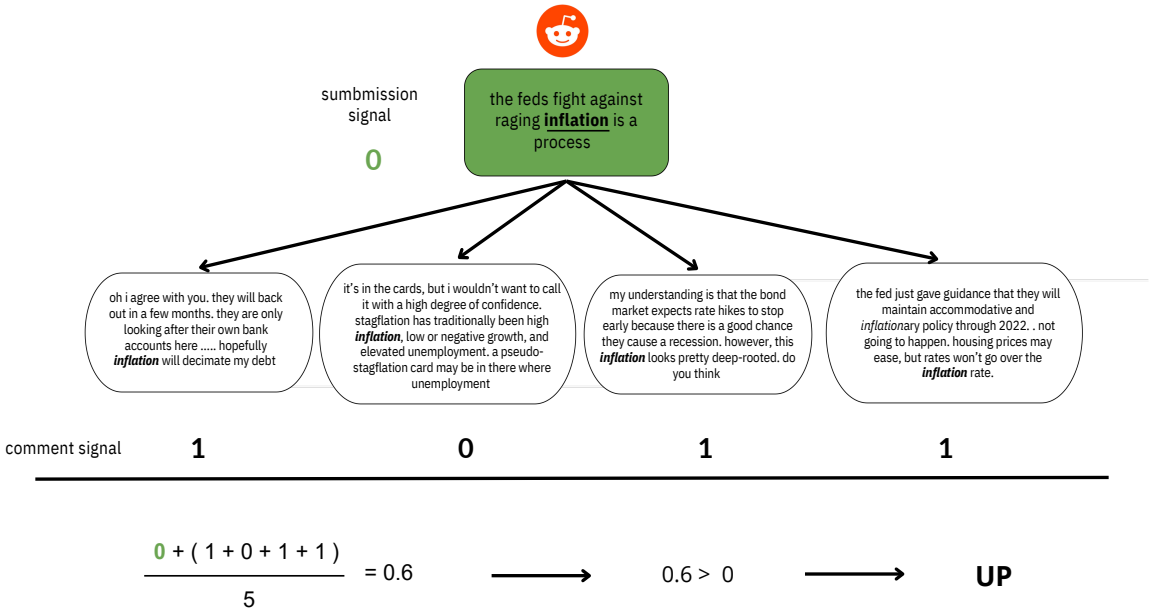
Given limited prior guidance on how Reddit-based indicators should be smoothed before entering forecasting regressions, we consider multiple variants along three dimensions: (i) subreddit (`r/Economics`, `r/economy`, `r/wallstreetbets`); (ii) model used for labeling (BERT, FinBERT, InflaBERT, Qwen 0.5B, Qwen 1.5B, Qwen 7B, LLaMA 1B, LLaMA 3B, LLaMA 8B, Gemma 2B, Gemma 9B, Gemma 27B); and (iii) backward moving-average (MA) windows of length $1, 5, 10, 30, 60, 90, 180$, and 360 days.[11] This yields $3 \times 12 \times 8 = 288$ can-

[9] Equivalently, the daily indicator can be interpreted as the difference between the number of UP and DOWN submissions.

[10] We assign NEUTRAL when $\widetilde{X}^{S_i,C_i}_{\text{LLM}} \in (-thr, thr)$ and UP/DOWN otherwise. In the baseline, we set $thr = 0$, and results are not sensitive to small variations in this threshold. In earlier versions, we used $thr = 0.10$ with similar results. This step can be interpreted as a regularization that adjusts each submission based on the extent to which comments agree or disagree with the original post.

[11] Backward MAs use only current and past observations, ensuring that the transformation does not use any information from the evaluation period.

Figure 3: Example of a submission–comment thread network



Notes: The initial submission is labeled NEUTRAL (0). The thread-level signal averages the submission and comment signals.

didate daily indicators, which are subsequently aggregated to monthly frequency for the forecasting and nowcasting exercises.

Figure 4 plots representative time series for the distribution of the 288 Reddit indicators alongside monthly inflation measured by CPI and core PCE.[12] All series are z-scored for visualization only; in the empirical analysis, Reddit indicators enter in their original scale. Importantly, no transformation is applied using information from the evaluation period: the only permitted smoothing is the backward moving average, computed using in-sample information at each forecast origin.

5 The predictive ability of Reddit-based educated signals

This section evaluates whether the Reddit-based indicators constructed from LLM-interpreted discussions contain incremental information for forecasting U.S. inflation. We focus on year-over-year inflation for headline CPI (all items) and core PCE, the Federal Reserve's preferred inflation gauge. Inflation is measured using the log approximation

$$\pi_t \;=\; 100 \times \ln\!\left(\frac{P_t}{P_{t-12}}\right),$$

and the analysis uses monthly data. The key question is whether text-based signals can complement standard macroeconomic predictors in a pseudo out-of-sample setting.

Reddit-based indicators summarize the directional tone of inflation-related submissions and comments (Section 4.4). We aggregate daily signals to monthly frequency and compare their predictive content to established benchmarks: lagged inflation, the University of Michigan inflation expectations measure, and the one-year inflation swap rate.[13]

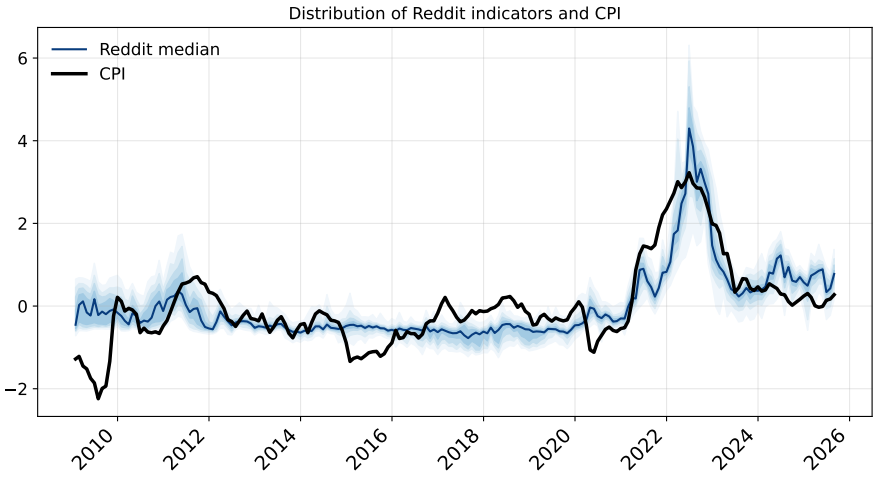
The evaluation proceeds in four complementary steps. First, we conduct a pseudo out-of-sample point-forecasting exercise (Section 5.1), estimating each model using information available up to time t and producing forecasts $\hat{\pi}_{t+h}$ at horizons $h \geq 1$. Second, we imple-

[12] These inflation series are used as dependent variables in the empirical analysis.

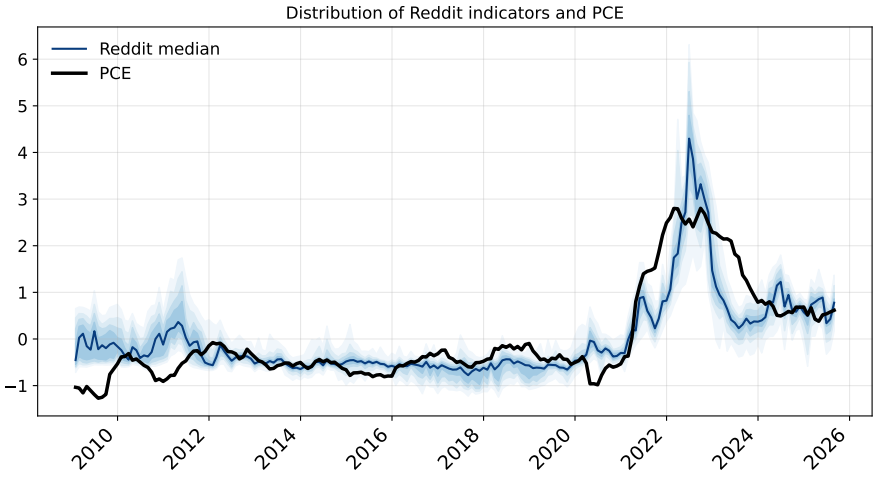
[13] Benchmark and inflation series are obtained from FRED.

Figure 4: Monthly CPI and core PCE inflation alongside Reddit-based signals

(a) CPI (headline)



(b) PCE (core)



Notes: Reddit-based signals are aggregated from daily to monthly frequency. All series are z-scored for comparability.

ment a real-time nowcasting exercise (Section 5.2) that mimics the information flow faced by practitioners: Reddit indicators are updated within the month and used to nowcast the most recently completed month's inflation before the official release. Third, Section 5.3.1 reports full-sample predictive regressions to assess whether Reddit indicators add explanatory power beyond lagged inflation and conventional expectations measures.[14] Fourth, Section 5.3.2 evaluates density forecasts using quantile regression and proper scoring rules. Two additional robustness checks that provide supporting evidence for interpretation and external validity: Section 5.3.3 reports Granger causality results linking Reddit indicators to inflation expectations, and Section 5.3.4 compares our measures with newspaper-based sentiment indices.

The construction of Reddit signals is intentionally exploratory. Because Reddit is rarely used in macroeconomic forecasting, we consider multiple subreddits, model families, and smoothing choices. In particular, each signal is computed for several LLM specifications and multiple backwards-looking moving-average (MA) windows, yielding 288 candidate indicators (Section 4.4). Our objective is to document which modeling and smoothing choices improve signal quality and which do not.

At the same time, the analysis is designed to mitigate data-mining concerns. For forecasting and nowcasting, all model selection and forecast-combination weights are computed strictly within the estimation sample; performance is then evaluated on observations that are not used for selection. The reported results therefore reflect genuine out-of-sample accuracy, including a real-time dimension in the nowcasting exercise.

5.1 A point-forecast application

We begin with a pseudo out-of-sample forecasting exercise comparing a standard autoregressive benchmark, $AR(1)$, to augmented specifications that add one predictor at a time, including Reddit-based indicators. Reddit indicators are computed at daily frequency from LLM outputs (Section 4.4) and then averaged to monthly frequency to match the inflation data. Forecasts are generated recursively: at each forecast origin t (end of month), the model is estimated using data available through t, a forecast is produced for $t+h$, and the estimation

[14] Unlike the pseudo out-of-sample exercises, these regressions use the full available sample and do not split the data into in-sample and evaluation periods.

sample is expanded by one month.

We forecast CPI and PCE inflation (as shown in Figure 4) at horizons $h \in \{1, 2, 3, 4, 5, 6, 9, 12, 18\}$ months. Since Reddit-based measures can be interpreted as a proxy for expectations held by a specific subgroup[15] (Reddit users), we benchmark them against two widely used expectation measures: the University of Michigan inflation expectations series[16] and the one-year inflation swap rate.

Table 2 summarizes the information timing used in the regressions. Michigan expectations are released with a lag and are therefore included with a one-month delay, while the inflation swap rate and Reddit indicators are available in real time.

Table 2: Variables used in the forecasting application

Indicator Name	**Transformation**	**Release Date**	**Month used in regression**
Michigan Expectations	Level	Preliminary (2nd Friday of month t) Revised (4th Friday of month t)	Previous month
1-Year Inflation Swap	Level	Real-time	Current month
Reddit Signals	Smoothed (MA)	Real-time	Current month

Notes: The last column reports how each predictor enters the regression to reflect its information availability. CPI and PCE are the dependent variables and are included for month $t + h$ in the forecasting regression without adjusting for publication lags.

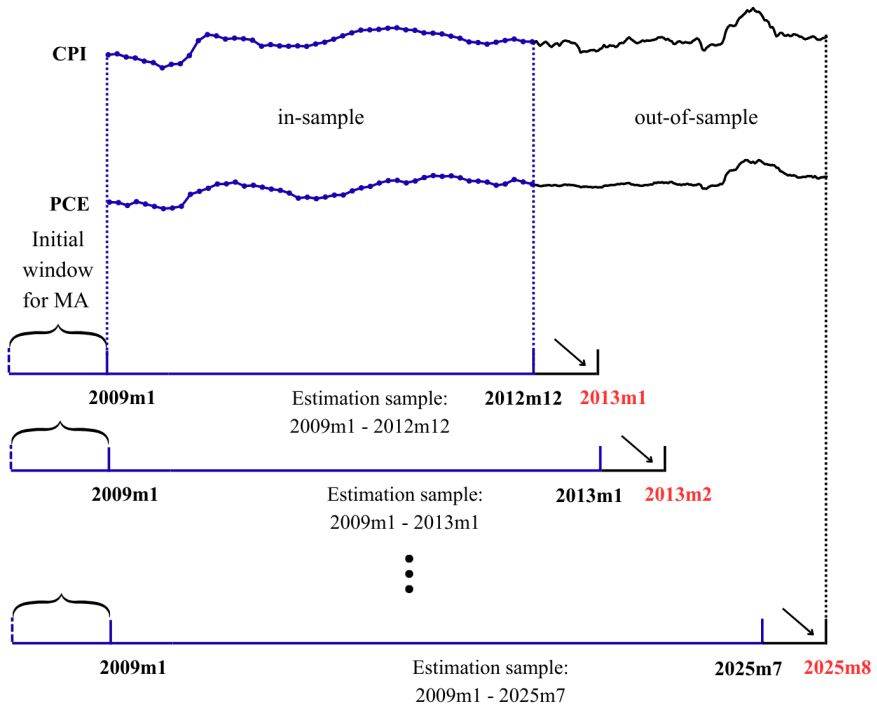
Because we consider many Reddit indicators (288 variants), we summarize results using both (i) the best-performing single indicator within a family and (ii) forecast combinations. In particular, we aggregate the Reddit-based forecasts using MSE-weighted combinations following Clark and McCracken (2010).[17] We construct combinations separately for (a) fine-tuned LLM-based indicators and (b) the unfine-tuned LLaMA 70B indicators to highlight differences between smaller task-adapted models and a large conversational model. As additional comparisons, we also construct lexicon-based sentiment indices using standard tools (TextBlob, VADER) and the Loughran–McDonald dictionary (Loughran and McDonald, 2011).

[15]Following Bybee (2023) who suggests that LLMs can be interpreted as belief-generating mechanisms.

[16]Survey of Professional Forecasters (SPF) indicators are released quarterly and are therefore not included.

[17]We set the discount factor to $\delta = 0.95$ as in Clark and McCracken (2010). Tuning δ can improve performance, but we keep the baseline value for comparability.

Figure 5: In-sample and out-of-sample setup with recursive estimation for $h = 1$.



Notes: The initial estimation sample spans 2009M1–2012M12 (blue); forecasts are evaluated out-of-sample from 2013M1 to 2025M8 (black). At each step, the sample expands by one month and the model is re-estimated before producing the next forecast.

The forecasting regression takes the form

$$\pi_{t+h} = \alpha + \beta_1 \pi_{t-1} + \gamma X_{\tilde{t}} + \varepsilon_{t+h}, \tag{3}$$

where $X_{\tilde{t}}$ denotes an additional predictor (Michigan expectations, the inflation swap rate, or a Reddit-based indicator) observed at the most recent available date $\tilde{t}$.[18]

The initial estimation window spans 2009M1–2012M12 and the evaluation period spans 2013M1–2025M8.[19] Figure 5 illustrates the recursive design for $h = 1$.

Tables 3 and 4 report out-of-sample accuracy using RMSE ratios relative to the $AR(1)$

[18]For Reddit signals and swaps, $\tilde{t} = t$; for Michigan expectations, $\tilde{t} = t - 1$ to reflect release timing.

[19]Reddit data begin in 2008, but we start in 2009M1 to allow for at least one year of observations to compute backward MA smoothers.

Table 3: Forecast results (RMSE ratios) for *CPI*.

	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=9$	$h=12$	$h=18$
Expectations									
Michigan Survey	1.000	0.948	0.839*	0.762*	*0.718**	*0.697**	*0.642**	*0.600**	**0.552***
1-Y Inflation Swap	0.966	0.880**	0.810**	0.781**	0.768*	0.755*	0.717*	0.693*	0.680*
Reddit-sentiment									
Best sentiment	0.945	0.898	0.854	0.807	0.769	0.738	**0.636***	**0.555***	*0.564**
Reddit-LLM									
LLM forecast aggregation	0.900**	*0.829***	*0.783***	*0.750***	**0.726***	*0.709**	0.702	0.712	0.717
Best fine-tuned LLM	0.872**	0.835**	0.813*	0.785**	0.776**	0.771*	0.807	0.810	*0.665*
Llama70B forecast aggregation	*0.813****	**0.733****	**0.689****	**0.653****	**0.632****	**0.623***	0.641	0.655	0.699
Best Llama70B	**0.790****	*0.765****	*0.760****	*0.728***	*0.720***	*0.719***	0.776*	0.758	0.747

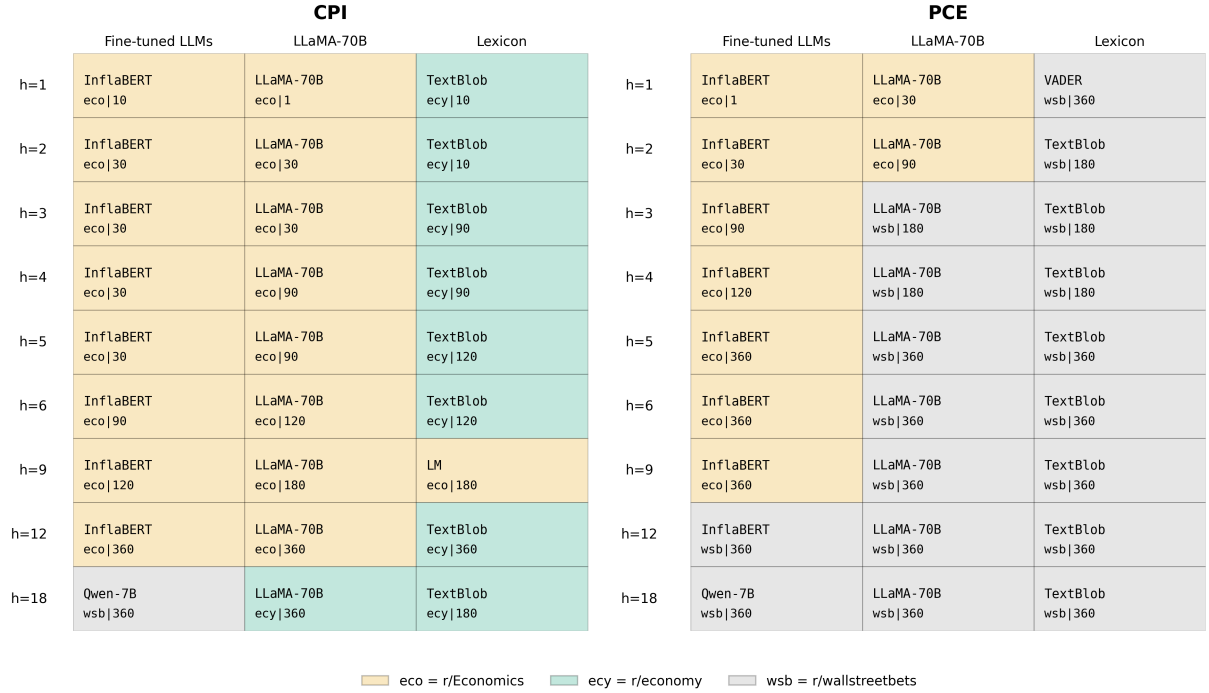
Notes: RMSE is computed as a ratio: $RMSE(AR\text{-}X(1))/RMSE(AR(1))$. The initial in-sample period spans from 2009M1 to 2012M12; the out-of-sample period covers 2013M1 to 2025M8. Forecasts are generated using a recursive window, with the model re-estimated each step (month). Significance levels reflect Diebold–Mariano p-values: $p < 0.01$ (***) $p < 0.05$ (**) $p < 0.1$ (*). Bold (italic) indicates the best (second-best) model for each horizon.

benchmark. For Reddit, we report both the MSE-weighted forecast combination and the best single specification within each family. Overall, Reddit-based LLM indicators deliver the largest gains at short horizons: up to $h = 6$, both the fine-tuned LLM family and the LLaMA 70B family outperform models augmented with Michigan expectations or inflation swaps, and the improvements are often statistically significant. At longer horizons, differences across predictors narrow and expectation-based variables tend to become relatively more competitive, a pattern consistent with the broader inflation-forecasting literature. In terms of comparison among fine-tuned and Llama70B, the latter has the best performance, but the RMSE are really close (in values and significance levels) to the ones of the fine-tuned LLMs.

Figure 6 summarizes, by horizon, which subreddit and MA window are selected among the best-performing specifications within each family. A common pattern is that shorter MA windows tend to be selected at short horizons, while longer windows are more frequently selected at longer horizons, consistent with the idea that long-horizon forecasts benefit from smoother signals. The best fine-tuned model is InflaBERT at all horizons except for $h = 18$, where the Qwen 7B has the lowest forecast error. The signal coming from `r/Economics` seems to dominate for CPI, while for PCE most signal comes from `r/wallstreetbets`.

The results in Table 4 for PCE are very similar to those obtained for CPI. Reddit-based LLM signals outperform the benchmark for all horizons. The fine-tuned models remain

Figure 6: Best models by forecast horizon and model family for CPI and PCE.



Notes: Tile color indicates subreddit; tags report "Subreddit|MA window". Subreddit codes: eco = `r/Economics`, ecy = `r/economy`, wsb = `r/wallstreetbets`. "MA" is the backward-looking moving-average window (days). "Best" is defined by the lowest RMSE ratio relative to $AR(1)$ within each family.

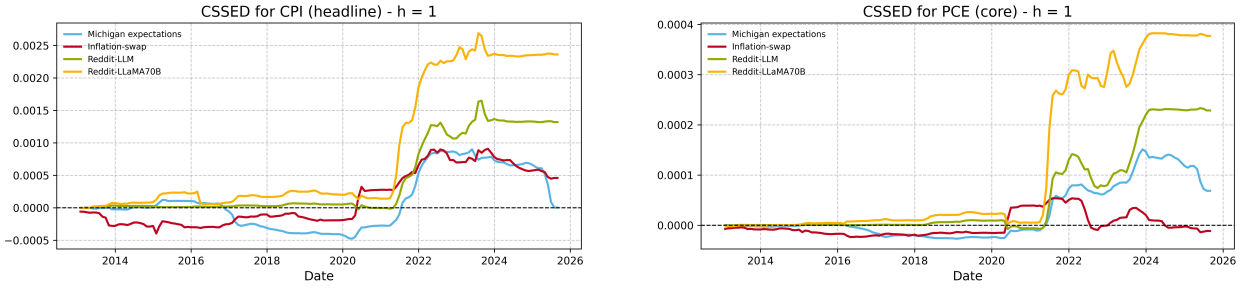
Table 4: Forecast results (RMSE ratios) for *PCE* (PCEPILFE).

	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=9$	$h=12$	$h=18$
Expectations									
Michigan Survey	0.983	0.958	0.908**	0.860**	0.830**	0.819**	0.777*	0.726*	0.682*
1-Y Inflation Swap	1.003	0.986	0.956	0.930*	0.909*	0.897*	0.863*	0.829*	0.780*
Reddit-sentiment									
Best sentiment	0.970	0.952	0.879	0.824	*0.762*	**0.703**	**0.549***	**0.451***	**0.393***
Reddit-LLM									
LLM forecast aggregation	0.941***	0.891**	0.851**	0.816**	0.792**	0.775*	0.731	0.718	0.690
Best fine-tuned LLM	0.924**	0.893**	0.870**	0.851*	0.826**	0.813*	0.788	0.738	*0.587**
Llama70B forecast aggregation	**0.901****	**0.839****	**0.791****	**0.750***	**0.727***	*0.723**	0.720	0.721	0.719
Best Llama70B	*0.917****	*0.874***	*0.832*	*0.793*	0.763	0.739	*0.688*	*0.662*	0.669

Notes: RMSE is computed as a ratio: $RMSE(AR\text{-}X(1))/RMSE(AR(1))$. Significance levels are based on the Diebold–Mariano test: $p < 0.01$ (***) $p < 0.05$ (**) $p < 0.1$ (*). For *Best sentiment*, the aggregated sentiment model is excluded. Bold (italic) indicates the best (second-best) model for each horizon.

Figure 7: CSSED for CPI (headline) and PCE (core) - $h = 1$

(a) CPI **(b)** PCE



competitive, especially the InflaBERT specifications, which outperform expectations at all horizons. Overall, the evidence confirms that Reddit-based LLM indicators contain useful information to forecast core inflation, with both fine-tuned and large LLMs achieving consistent improvements across horizons. In terms of models, InflaBERT remains the best choice for PCE, according to Figure 6.

Figure 7 plots the cumulative sum of squared error differences (CSSED) for CPI and PCE at the one-month-ahead horizon. CSSED provide a time-varying measure of relative forecast performance by cumulating differences in squared forecast errors over the evaluation period. For both CPI and PCE, the MSE-weighted Reddit–LLM forecast combination performs best, followed by the aggregated Reddit–LLaMA-70B forecasts. By contrast, models augmented with Michigan expectations or the one-year inflation swap perform worse overall; the swap-based specification only briefly narrows the gap—and modestly outperforms the Michigan-based benchmark—around 2020.[20]

In Appendix E.1, we also report Tables for MAE and MAD. The results are in line with the RMSE, seeing the LLMs dominating at short-term horizons.

[20]Figures A.22–A.23 in the Online Appendix report CSSED paths for CPI and PCE across all forecast horizons. The patterns largely mirror those in Figure 7 for $h = 1$, except at the longest horizons, where Michigan-expectations-based forecasts tend to perform relatively better.

5.1.1 Additional robustness tests for point forecasts

Because the evaluation period includes episodes of unusually large inflation movements, forecast errors may be heteroskedastic and forecasting relationships may be unstable. In such settings, the standard Diebold-Mariano test may be sensitive to these features. We therefore complement the baseline comparisons with two additional tests.

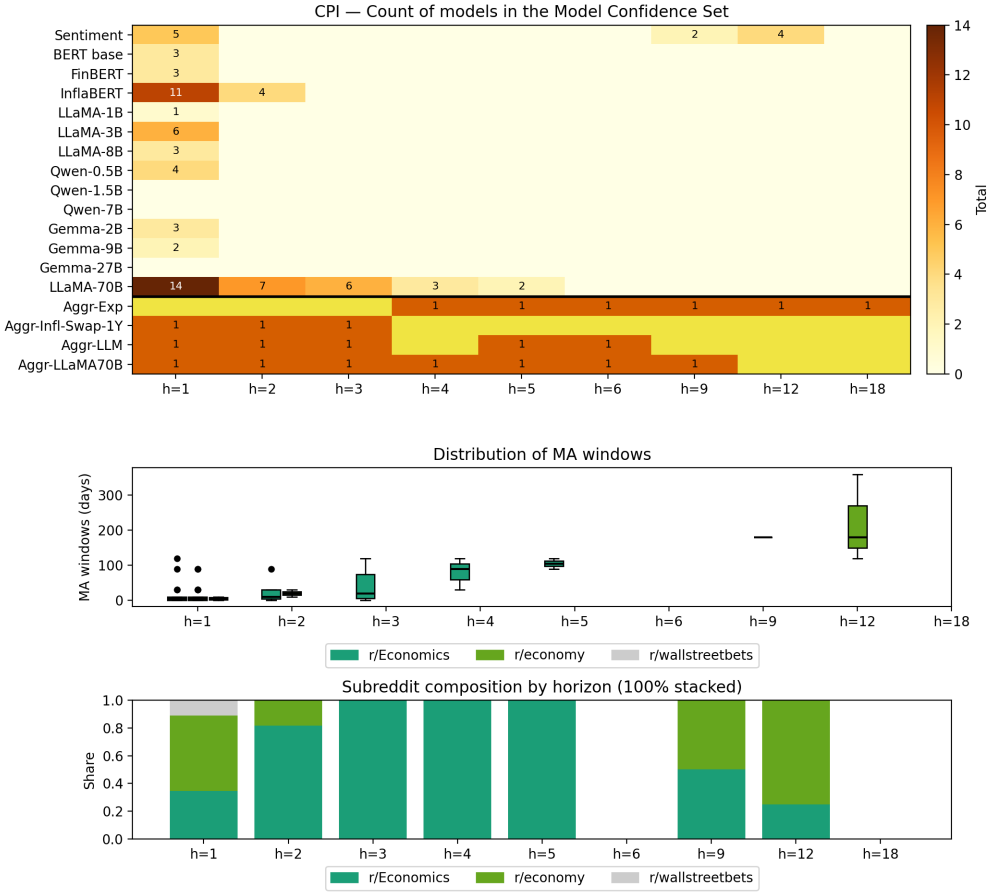
First, we apply the Model Confidence Set (MCS) procedure of P. R. Hansen et al. (2011), which constructs, for each horizon, a set of models that are statistically indistinguishable from the best-performing model at a given confidence level. Unlike pairwise comparisons, the MCS explicitly accounts for multiple testing and allows for the possibility that several models perform similarly.

Figures 8 and 9 summarize MCS inclusion patterns for CPI and PCE, respectively. At short horizons, LLM-based models (including LLaMA 70B, InflaBERT, and selected smaller LLaMA variants) enter the MCS frequently, and the aggregated LLM forecasts remain in the MCS over several horizons. By contrast, Michigan expectations enter the MCS more consistently at longer horizons, and sentiment-based models tend to appear primarily at medium to longer horizons. The composition plots further indicate that longer MA windows are more prevalent in the MCS at longer horizons, and that subreddit composition differs across targets: CPI-relevant models rely more on `r/Economics` and `r/economy`, whereas PCE-relevant models more frequently draw on `r/wallstreetbets` and `r/Economics`.

Second, following Giacomini and Rossi (2010), we apply the fluctuation test, which evaluates forecast differences locally over time and is robust to heteroskedasticity and structural instability. Figures 10 and 11 report fluctuation statistics for $h = 1$ for CPI and PCE, respectively. We focus on the augmented models using inflation swaps, Michigan expectations, the fine-tuned LLM ensemble, and the LLaMA 70B ensemble. Forecasts are compared to $AR(1)$ using a rolling window of 10 observations, and critical values follow Giacomini and Rossi (2010). When the statistic exceeds the threshold, forecast performance differs significantly from the benchmark.

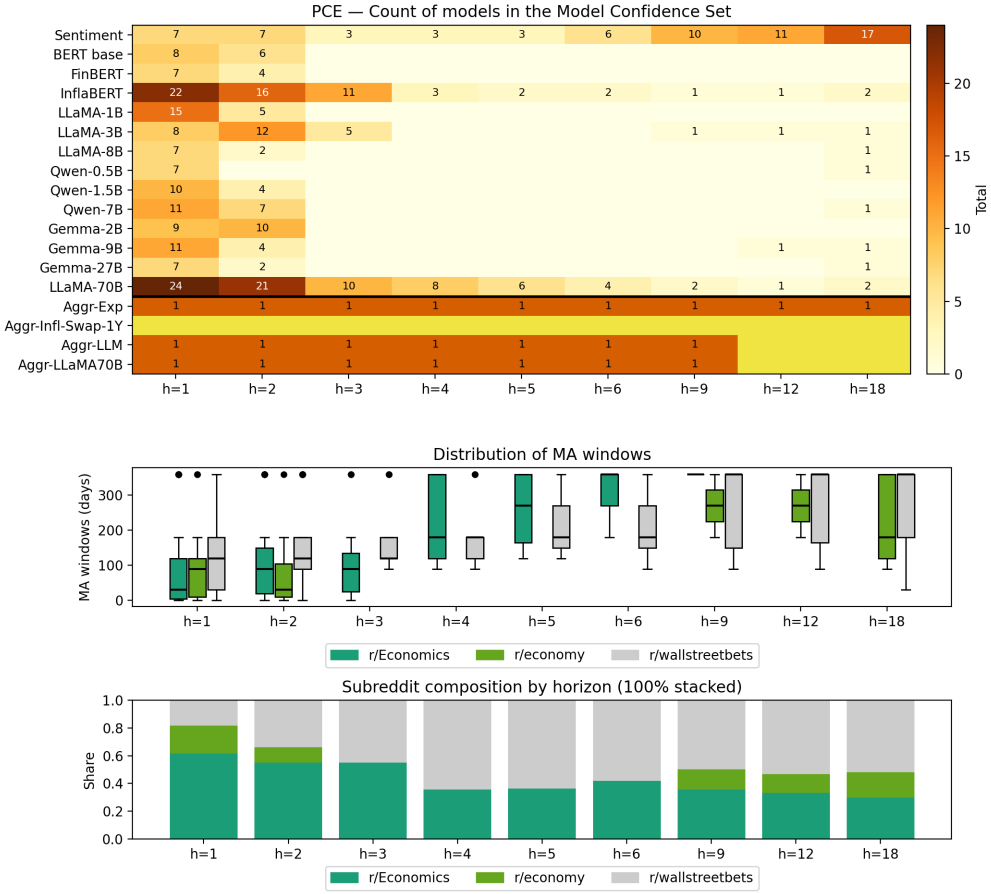
For PCE, the fine-tuned LLM and LLaMA 70B ensembles display similar time variation, while for CPI the fine-tuned ensemble appears slightly more stable. Significant gains are concentrated around the COVID-19 shock, when high-frequency narrative information may

Figure 8: Model confidence set for CPI.



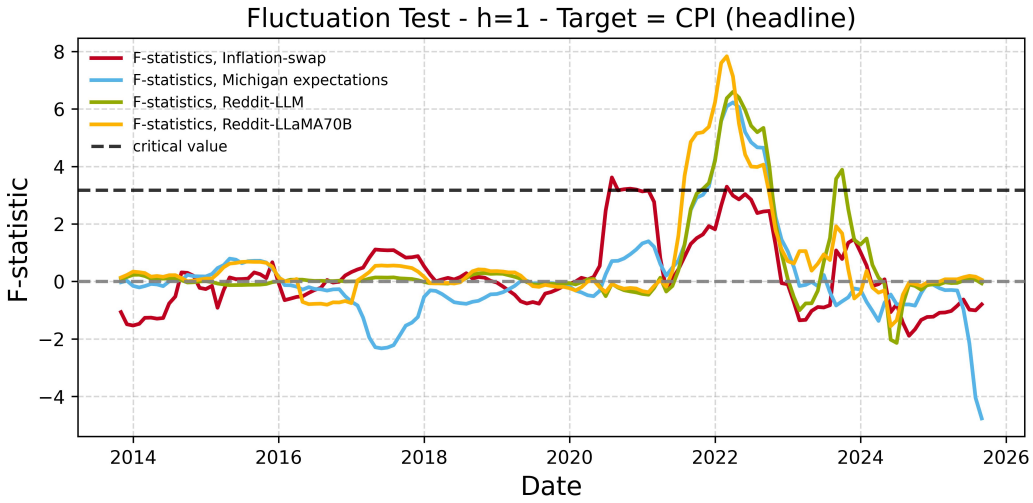
Notes: For each horizon and model family, the top panel reports which models enter the MCS (aggregates are coded as 1 if included). The middle panel reports the distribution of MA windows among MCS models. The bottom panel reports the composition by subreddit.

Figure 9: Model confidence set for PCE.



Notes: For each horizon and model family, the top panel reports which models enter the MCS (aggregates are coded as 1 if included). The middle panel reports the distribution of MA windows among MCS models. The bottom panel reports the composition by subreddit.

Figure 10: Fluctuation test for forecast differences – CPI (*headline*).

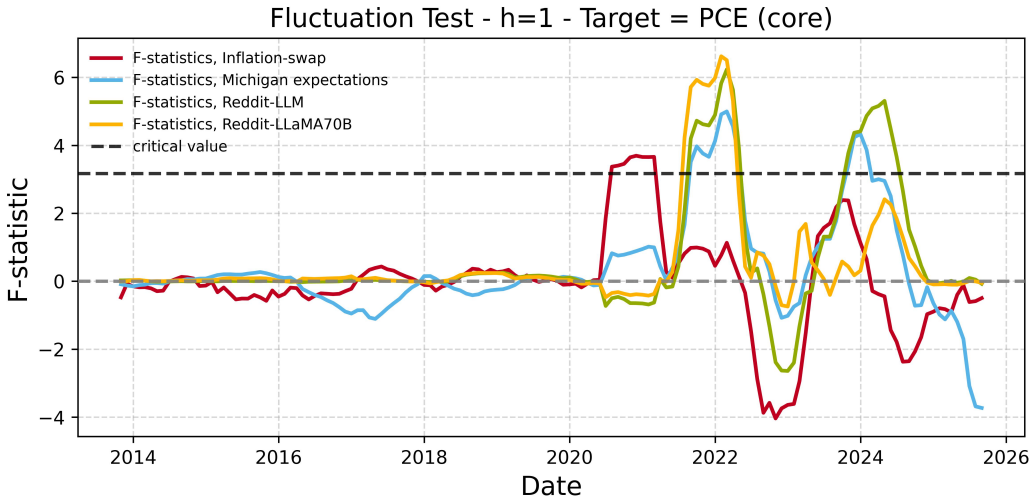


be especially informative. Michigan expectations perform relatively better for PCE, plausibly reflecting release timing: PCE is published later than CPI, and household expectations may already incorporate CPI-related information.

Fluctuation statistics are comparatively flat prior to 2019. This reflects both (i) a period of low and stable inflation, which limits the scope for improvements over an autoregressive benchmark, and (ii) substantially lower Reddit activity in earlier years, which reduces the density of narrative information.

Overall, the evidence indicates that Reddit-based LLM indicators improve point-forecast accuracy for both CPI and PCE, with gains that are statistically supported by global (Diebold–Mariano), multi-model (MCS), and local (fluctuation) comparisons. Importantly, the improvements are not confined to long horizons: Reddit–LLM signals deliver significant gains even at one-month-ahead horizons, where inflation is typically dominated by transitory shocks and forecasting gains are notoriously difficult to achieve. This motivates the nowcasting analysis that follows.

Figure 11: Fluctuation test for forecast differences – PCE (*core*).



5.2 Nowcasting application

We next implement a nowcasting exercise that uses real-time vintages of CPI and PCE from ALFRED.[21] The sample split differs from the forecasting exercise to reflect both vintage availability and the construction of competitive benchmarks. The estimation window is 2009M1–2013M7 and the evaluation window is 2013M8–2025M5.[22]

Nowcasting introduces a ragged-edge setting: Reddit indicators are observed daily, while CPI and PCE are released monthly with a delay. This feature allows us to quantify how much incremental information Reddit provides as the month progresses and before the official release. We therefore construct within-month cutoffs based on the number of days observed in month $t+1$ before the release of π_t. Specifically, we consider cutoffs at +5, +10, and +14 days for both series; for PCE, which is released later, we also consider +22 days. These cutoffs are chosen so that Reddit information is always used strictly prior to the official release.

[21] Real-time vintages are retrieved from `https://alfred.stlouisfed.org/`.

[22] When nowcasting 2013M8, we use vintages available up to 2013M7. For 2013M9, we use vintages up to 2013M8, and so on.

Table 5: Nowcast results (RMSE ratios) for *CPI* (CPIAUCSL).

Metric: RMSE				
Model	+5 days	+10 days	+14 days	+22 days
Reddit-LLM				
LLM forecast aggregation	0.963**	0.956**	0.941**	
Llama70B forecast aggregation	0.859***	0.843***	0.818***	

Metric: MAE				
Model	+5 days	+10 days	+14 days	+22 days
Reddit-LLM				
LLM forecast aggregation	0.971	0.959*	0.941**	
Llama70B forecast aggregation	0.876***	0.852***	0.821***	

Notes: RMSE is computed as a ratio: $loss(AR_x(1))/loss(AR(1))$. For the regression with Reddit, the initial in-sample period spans from 2009M1 to 2013M7; the out-of-sample period covers 2013M8 to 2025M5. Nowcasts are generated using a recursive window, with the model re-estimated at each step (month). Significance levels are based on Diebold-Mariano test p-values: $p < 0.01$ (***) $p < 0.05$ (**) $p < 0.1$ (*).

Operationally, we estimate the following nowcasting regression

$$\pi_t = \alpha + \beta_1 \pi_{t-1} + \gamma X_{t+c} + \varepsilon_t, \tag{4}$$

where $X_{\tilde{t}}$ denotes an additional predictor (Michigan expectations, the inflation swap rate, or a Reddit-based indicator) observed at the most recent available cutoff. For example, the monthly Reddit regressor is constructed using information available up to day c of the following month $(t+1)$, with $t+c$ occurring strictly before the release date.

Tables 5 and 6 report nowcast accuracy in RMSE and MAE ratios relative to the $AR(1)$ benchmark. Two patterns stand out. First, for both CPI and PCE, Reddit-based nowcasts outperform $AR(1)$ across cutoffs. Second, using more within-month Reddit information generally improves accuracy, particularly for PCE, where later cutoffs (+14 and +22 days) deliver the clearest gains.

Table 6: Nowcast results (RMSE ratios) for *PCE* (PCEPILFE).

	metric: RMSE			
Model	+5 days	+10 days	+14 days	+22 days
Reddit-LLM				
LLM forecast aggregation	0.989	0.986	0.981	0.974*
Llama70B forecast aggregation	0.972	0.969	0.955*	0.948*

	metric: MAE			
Model	+5 days	+10 days	+14 days	+22 days
Reddit-LLM				
LLM forecast aggregation	0.980	0.978	0.974*	0.970*
Llama70B forecast aggregation	0.985	0.980	0.971	0.963

Notes: RMSE is computed as a ratio: $loss(AR_x(1))/loss(AR(1))$. For the regression with Reddit, the initial in-sample period spans from 2009M1 to 2013M7; the out-of-sample period covers 2013M8 to 2025M5. Nowcasts are generated using a recursive window, with the model re-estimated at each step (month). Significance levels based on Diebold-Mariano test p-values: $p < 0.01$ (***) $p < 0.05$ (**) $p < 0.1$ (*).

5.2.1 Comparison with the Federal Reserve inflation nowcast

To benchmark the information content of Reddit narratives, we compare our nowcasts to the Federal Reserve's publicly available inflation nowcast (Cleveland Fed), based on the real-time framework of Knotek and Zaman (2017). The Fed nowcast produces daily updates for CPI and PCE using high-frequency energy prices, recent inflation releases, and time-varying weights that adapt to the intra-month information set.

For comparability, we evaluate both approaches at the most informative cutoff for each target: +14 days for CPI and +22 days for PCE. We download the Fed nowcast series from the Cleveland Fed website and align it to our evaluation scheme by selecting, for each target month, the nowcast issued at the corresponding intra-month horizon of the following month.[23]

This comparison should be interpreted cautiously because the information sets differ. The Fed nowcast is estimated on a longer historical sample (beginning around 1999/2000),

[23]For example, when evaluating the nowcast for 2013M8 inflation, we use the nowcast issued on day 14 of 2013M9 for CPI and on day 22 of 2013M9 for PCE. This alignment ensures that both approaches exploit an equivalent amount of within-month information prior to the official release.

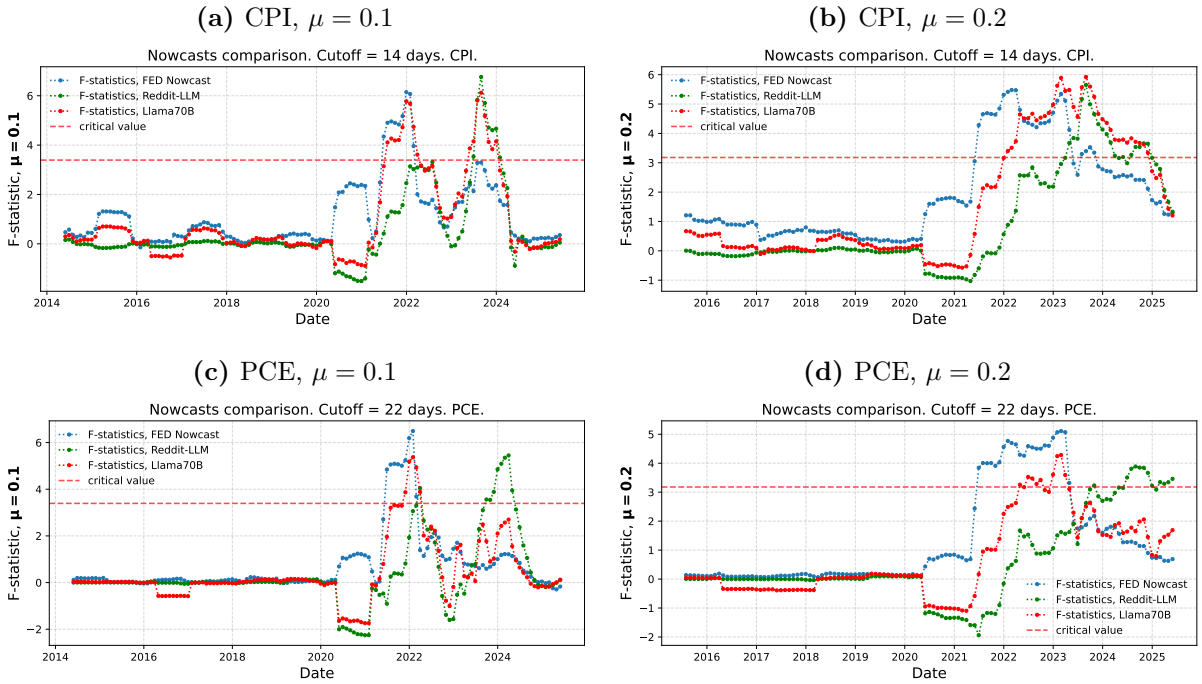


Figure 12: Fluctuation test statistics for nowcasts.

while our Reddit indicators are only available from 2008 onward.[24] Moreover, the Fed nowcast explicitly exploits high-frequency commodity and energy price information, which we deliberately exclude in order to isolate the predictive content of narrative signals.

Figure 12 reports fluctuation-test statistics for nowcasts using two window lengths ($\mu = 0.1$ and $\mu = 0.2$). While pointwise significance depends on the window choice, the overall patterns are stable. For both CPI and PCE, the Fed nowcast outperforms $AR(1)$ primarily during the initial phase of the post-pandemic inflation surge, when energy prices moved sharply. The LLaMA-based Reddit aggregation shows a similar profile, with gains concentrated in this early high-inflation period. By contrast, the fine-tuned Reddit–LLM aggregation tends to improve upon $AR(1)$ later in the sample, particularly during the disinflation phase in 2023–2024.

Figure 13 reports moving-average differential loss relative to $AR(1)$, confirming that the

[24] This difference may affect comparability in ways that are not fully under our control.

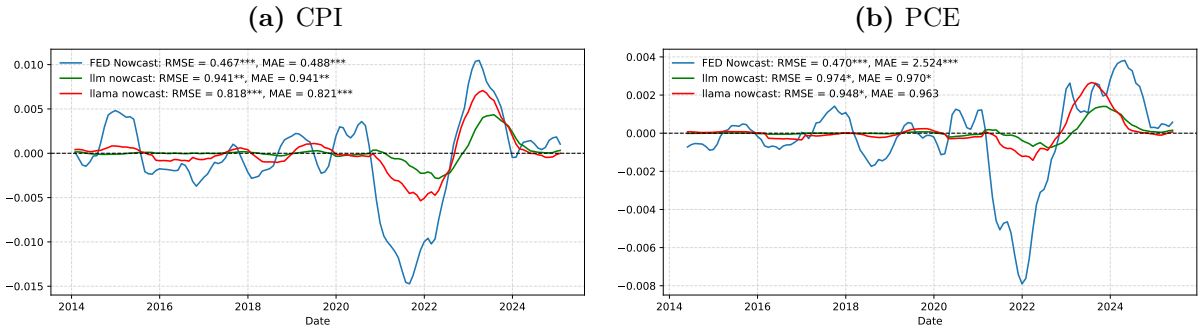


Figure 13: Moving average (10 periods) of differential loss between Reddit/FED nowcasting and an AR(1)). If lower than 0, the nowcast is better than the AR(1).

Fed nowcast achieves the lowest RMSE on average over the full sample. At the same time, for PCE the Reddit-based approaches achieve lower MAE in some configurations, indicating fewer large nowcast errors.

Taken together, the results suggest a natural complementarity. The Fed nowcast benefits from fast-moving energy price signals that are especially valuable during periods of sharp commodity-driven inflation movements. Reddit-based indicators, by construction, capture narrative information–including perceived price pressures and expectations–that may be less tightly linked to commodity prices and appear particularly informative during the subsequent disinflationary phase.

5.3 Supplementary Analyses: Robustness and Economic Interpretation

5.3.1 Predictive regressions (in-sample evidence)

As an additional check–complementary to the pseudo out-of-sample forecasting and nowcasting exercises–we estimate predictive regressions on the full sample to assess whether Reddit-based indicators contain incremental information for future inflation beyond standard predictors. Specifically, we regress h-step-ahead inflation on lagged inflation, the University of Michigan expectations measure, the one-year inflation swap rate, and (one at a time) a Reddit–LLM indicator. We evaluate the marginal contribution of Reddit by testing whether

adding the Reddit regressor yields a statistically significant increase in adjusted R^2 (F-test with HAC standard errors). Across subreddits and horizons, the Reddit indicators deliver systematic and statistically significant improvements in fit for both CPI and PCE. Full results and graphical evidence are reported in the Online Appendix (Appendix E.4).

5.3.2 Density forecasts via quantile regression

To complement the point-forecast results, we also evaluate whether Reddit-based indicators improve *density* forecasts of inflation. We estimate quantile-regression versions of our baseline forecasting model and reconstruct predictive distributions from multiple conditional quantiles. Density accuracy is assessed using the Continuous Ranked Probability Score (CRPS). Overall, Reddit–LLM indicators improve distributional forecasts relative to the $AR(1)$ benchmark, with the largest gains coming from the LLaMA 70B signals at short horizons and from expectation-based measures at longer horizons. We also document that the association between Reddit indicators and inflation varies across the conditional distribution, consistent with state-dependent attention and expectation formation. Full methodological details and additional results are provided in the Online Appendix (Appendix E.5).

5.3.3 Do Reddit narratives anticipate inflation expectations?

To provide additional economic grounding for our Reddit indicators, we examine whether Reddit-based signals help predict survey-based inflation expectations. We estimate VARs including inflation, Michigan expectations, the one-year inflation swap rate, and a Reddit indicator, and we test for Granger-causality between Reddit narratives and expectations across model families, subreddits, and smoothing windows. The results show that Reddit indicators frequently Granger-cause Michigan expectations, while the reverse direction is observed less often; bidirectional causality arises only in a limited number of specifications. These patterns are consistent with the interpretation that high-frequency online discussions both reflect and, in some cases, anticipate shifts in inflation expectations. Full details are provided in the Online Appendix (Appendix F).

5.3.4 Reddit and news sentiment: overlap and differences

To clarify what our Reddit indicators measure, we examine their relationship with established newspaper-based sentiment indices. Two facts stand out. First, subreddit posting patterns indicate that `r/Economics` and `r/economy` largely function as news-sharing hubs (often without explicit flair), whereas `r/wallstreetbets` is dominated by user-generated narratives. Second, despite these different content-generation mechanisms, Reddit-based sentiment closely co-moves with a standard news-sentiment series and exhibits a strong contemporaneous correlation, with limited evidence of systematic leads or lags. Taken together, the results suggest that Reddit acts as a real-time amplifier of news while adding a distinct layer of social interpretation through comments. Full evidence and figures are reported in the Online Appendix (Appendix G).

6 Conclusion

This paper investigates whether textual information from Reddit discussions can produce economically meaningful indicators for forecasting and nowcasting U.S. inflation. Using large language models as measurement tools, we extract structured signals from unstructured social media text and integrate them into standard econometric frameworks.

A first contribution is methodological. We propose a transparent pipeline that converts raw Reddit submissions and comments into time-series indicators suitable for forecasting. Each step–keyword filtering, LLM-based geographic classification, fine-tuned sentiment labeling, and backward-looking aggregation–is designed to avoid look-ahead bias. All exercises are implemented in a pseudo–real-time setting with recursive expanding windows, a fixed train–test split (2009M1–2012M12 for estimation, 2013M1–2025M8 for evaluation), and explicit data-availability constraints.

Given the absence of prior guidance on how Reddit data should be processed for macroeconomic analysis, we deliberately explore multiple dimensions–subreddit, LLM architecture, and smoothing window–yielding 288 candidate indicators. Rather than selecting individual specifications ex post, we rely on MSE-weighted forecast combinations computed strictly within the estimation sample, mitigating concerns about overfitting and data snooping. The

Model Confidence Set procedure and the Giacomini–Rossi fluctuation test provide additional safeguards against multiple testing and instability.

Empirically, the Reddit-based indicators improve forecast accuracy relative to standard benchmarks, including autoregressive models, survey-based expectations, and market-based measures. Improvements are observed across horizons and are particularly evident at short horizons, where inflation dynamics are typically dominated by shocks. In a real-time nowcasting application using actual data vintages, Reddit indicators also provide timely information ahead of official CPI and PCE releases and perform competitively with the Cleveland Fed Inflation Nowcast, with the two approaches displaying a natural complementarity: the Fed nowcast excels during sharp commodity-driven movements, while Reddit signals are more informative during the subsequent disinflationary phase.

The informational content of Reddit discussions varies across communities and over time. Financially oriented subreddits tend to be more informative during periods of elevated inflation, consistent with the idea that attention and expectations formation intensify in high-volatility environments. Supplementary analyses–including Granger-causality tests showing that Reddit signals frequently anticipate Michigan survey expectations, comparisons with newspaper-based sentiment indices, and density forecasts via quantile regressions–further support the economic interpretation of the indicators.

Overall, the findings suggest that Reddit-based text indicators, when processed in a disciplined and transparent way, can complement existing sources of information for inflation monitoring. The framework developed here provides a basis for integrating social media data into applied macroeconomic analysis, and fine-tuned small language models can achieve this at modest computational cost.

References

Allard, Marc-Antoine, Paul Teiletche, and Adam Zinebi (2024). *Enhancing Inflation Nowcasting with LLM: Sentiment Analysis on News*. arXiv: 2410.20198 (cit. on p. 12).

Angelico, Cristina, Juri Marcucci, Marcello Miccoli, and Filippo Quarta (2022). "Can we measure inflation expectations using Twitter?" In: *Journal of Econometrics* 228.2, pp. 259–277. ISSN: 0304-4076. DOI: https://doi.org/10.1016/j.jeconom.2021.12.008 (cit. on pp. 3, 5).

Antweiler, Werner and Murray Z Frank (2004). "Is all that talk just noise? The information content of internet stock message boards". In: *The Journal of finance* 59.3, pp. 1259–1294 (cit. on p. 6).

Aprigliano, Valentina, Simone Emiliozzi, Gabriele Guaitoli, Andrea Luciani, Juri Marcucci, and Libero Monteforte (2023). "The power of text-based indicators in forecasting Italian economic activity". In: *International Journal of Forecasting* 39.2, pp. 791–808 (cit. on p. 5).

Araci, Dogu (2019). *FinBERT: Financial Sentiment Analysis with Pre-trained Language Models*. arXiv: 1908.10063 (cit. on p. 12).

Aruoba, S Borağan and Thomas Drechsel (2024). *Identifying monetary policy shocks: A natural language approach*. Tech. rep. National Bureau of Economic Research (cit. on p. 5).

Ash, Elliott and Stephen Hansen (2023). "Text algorithms in economics". In: *Annual Review of Economics* 15.1, pp. 659–688 (cit. on p. 7).

Audrino, Francesco, Jessica Gentner, and Simon Stalder (2024). "Quantifying uncertainty: a new era of measurement through large language models". In: (cit. on p. 7).

Azqueta-Gavaldón, Andrés, Dominik Hirschbühl, Luca Onorante, and Lorena Saiz (2023). "Sources of Economic Policy Uncertainty in the euro area". In: *European Economic Review* 152, p. 104373 (cit. on p. 5).

Baker, Scott R, Nicholas Bloom, and Steven J Davis (2016). "Measuring economic policy uncertainty". In: *The quarterly journal of economics* 131.4, pp. 1593–1636 (cit. on p. 5).

Barbaglia, Luca, Stefano Consoli, and Stefano Manzan (2022). "Forecasting with Economic News". In: *Journal of Business & Economic Statistics* 41.3, pp. 708–719. DOI: 10.1080/07350015.2022.2060988 (cit. on p. 6).

Bybee, Leland (2023). "Surveying Generative AI's Economic Expectations". In: *arXiv preprint arXiv:2305.02823* (cit. on pp. 3, 6, 19).

Carriero, Andrea, Davide Pettenuzzo, and Shubhranshu Shekhar (2024). "Macroeconomic Forecasting with Large Language Models". In: *arXiv preprint arXiv:2407.00890* (cit. on pp. 6, 13).

Choi, Daejin, Jinyoung Han, Taejoong Chung, Yong-Yeol Ahn, Byung-Gon Chun, and Ted Taekyoung Kwon (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*, pp. 233–243 (cit. on p. 7).

Clark, Todd E and Michael W McCracken (2010). "Averaging forecasts from VARs with uncertain instabilities". In: *Journal of Applied Econometrics* 25.1, pp. 5–29 (cit. on pp. 4, 19).

Dettmers, Tim, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer (2023). *QLoRA: Efficient Finetuning of Quantized LLMs*. arXiv: 2305.14314 (cit. on p. 13).

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv: 1810.04805 (cit. on p. 12).

Faria-e-Castro, Miguel and Fernando Leibovici (2024). *Artificial Intelligence and Inflation Forecasts*. Tech. rep. (cit. on p. 6).

Giacomini, Raffaella and Barbara Rossi (2010). "Forecast comparisons in unstable environments". In: *Journal of Applied Econometrics* 25.4, pp. 595–620 (cit. on pp. 4, 24).

Gueta, Almog, Amir Feder, Zorik Gekhman, Ariel Goldstein, and Roi Reichart (2024). "Can LLMs Learn Macroeconomic Narratives from Social Media?" In: *Working Paper* (cit. on p. 6).

Hansen, Peter R, Asger Lunde, and James M Nason (2011). "The Model Confidence Set". In: *Econometrica* 79.2, pp. 453–497 (cit. on pp. 4, 24).

Hayou, Soufiane, Nikhil Ghosh, and Bin Yu (2024). *LoRA+: Efficient Low Rank Adaptation of Large Models*. arXiv: 2402.12354 (cit. on p. 13).

Horton, John J (2023). *Large language models as simulated economic agents: What can we learn from homo silicus?* Tech. rep. National Bureau of Economic Research (cit. on p. 6).

Kalamara, Eleni, Arthur Turrell, Chris Redl, George Kapetanios, and Sujit Kapadia (2022). "Making text count: Economic forecasting using newspaper text". In: *Journal of Applied Econometrics*. First published: 11 May 2022. DOI: 10.1002/jae.2907 (cit. on p. 6).

Knotek, Edward S and Saeed Zaman (2017). "Nowcasting US headline and core inflation". In: *Journal of Money, Credit and Banking* 49.5, pp. 931–968 (cit. on pp. 4, 30).

Liu, Shih-Yang, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen (2024). *DoRA: Weight-Decomposed Low-Rank Adaptation*. arXiv: 2402.09353 (cit. on p. 13).

Loughran, Tim and Bill McDonald (2011). "When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks". In: *The Journal of finance* 66.1, pp. 35–65 (cit. on pp. 5, 19).

Marcucci, Juri (2024). "Macroeconomic Forecasting with Text-Based Data". In: *Handbook of Research Methods and Applications in Macroeconomic Forecasting*. Ed. by Michael P. Clements and Ana Beatriz Galvão. Edward Elgar Publishing. Chap. 16, pp. 425–468 (cit. on p. 5).

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Jun Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu (2024). "Qwen2.5 Technical Report". In: *arXiv preprint arXiv:2412.15115* (cit. on p. 13).

Renault, Thomas (2017). "Intraday online investor sentiment and return patterns in the US stock market". In: *Journal of Banking & Finance* 84, pp. 25–40 (cit. on p. 6).

Sarkar, Suproteem K and Keyon Vafa (2024). "Lookahead Bias in Pretrained Language Models". In: *Available at SSRN 4754678* (cit. on p. 3).

Shapiro, Adam Hale, Moritz Sudhof, and Daniel J Wilson (2022). "Measuring news sentiment". In: *Journal of econometrics* 228.2, pp. 221–243 (cit. on p. 7).

Team, Google AI (2024). *Gemma 2 Technical Report*. Technical Report. Available at https://ai.google.dev/g (cit. on p. 13).

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Baptiste Rozière, Naman Goyal, Eric Hambro, Aidan Azhar, Timothée Rodriguez, et al. (2023). "LLaMA: Open and Efficient Foundation Language Models". In: *arXiv preprint arXiv:2302.13971* (cit. on p. 13).

Online Appendix (Not for Publication)

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

Andrea Del Monaco (Bank of Italy), Luigi Longo (JRC - European Commission),

Juri Marcucci (Bank of Italy), Irene Tafani (IMT School Lucca)

2026-03-27

A Reddit data: Further details

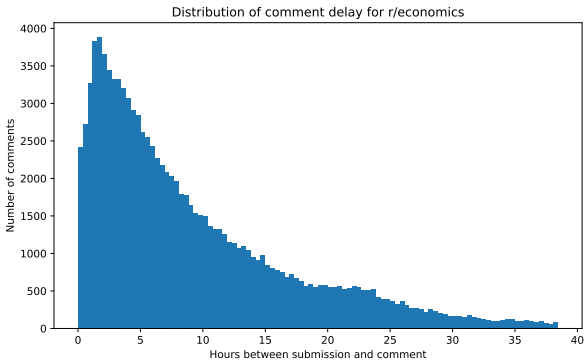
A.1 Comment delay

This section documents the distribution of comment delays and motivates our timeliness restriction, which retains only comments posted within two weeks of the original submission. For each comment, we define the *delay* as the elapsed time (in hours) between the comment timestamp and the timestamp of the associated submission.

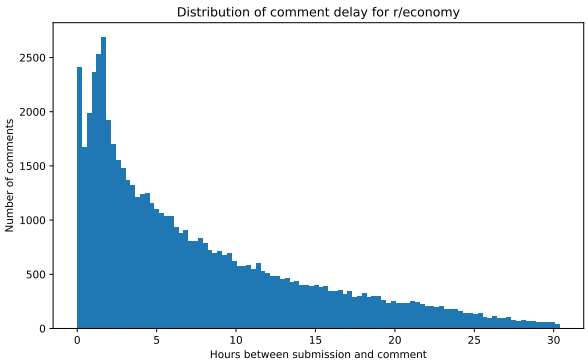
Figure A.1 plots the delay distribution for the central 95% of comments, trimming the upper tail to remove outliers that are not informative for constructing our daily and monthly indicators. Engagement on Reddit is highly front-loaded: most comments are posted shortly after the submission appears, with the bulk arriving within roughly 40 hours. Reaction times are particularly short in `r/wallstreetbets`, where comments are typically posted within one day. Overall, these patterns indicate that our two-week window is conservative and does not materially limit the contemporaneous information content of the Reddit-based signals.

Figure A.1: Distribution of comment "delay" for the three subreddits used

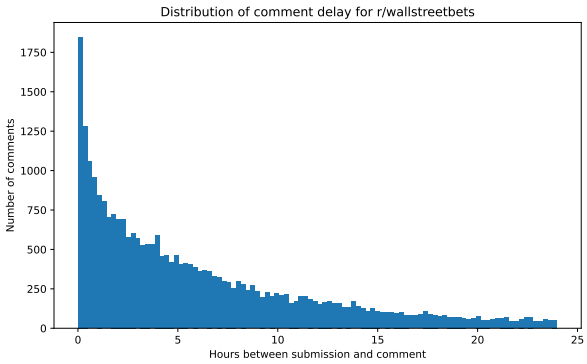
(a) `r/Economics`



(b) `r/economy`



(c) `r/wallstreetbets`



Notes: The figure plots the distribution of comment delays for the central 95% of observations. Delay is defined as the time difference, in hours, between a comment and its associated submission.

B Llama70B (zero-shot) prompt for country identification and price direction

Here we provide the prompts used with the LLaMA70B LLM. The first prompt refers to the geo-reference classification, while the second one refers to the signal extraction from filtered Reddit submissions and comments.

```
You are a forecaster (linguistic interpreter) and you have to analyze
the title of a social media post.  Given the text of the title you have
to classify the country that the user is referring to.  You have to
choose among one of these labels:  US; non-US; other.  US: the title
has to be related specifically to US, not just with keywords, but also
with mentions of people/institutions/events related to the US. non-US:
the title has to be related specifically to a country other than the
US. other:  the title does not have a specific reference to a country.
Return ONLY the classification.  No punctuation at the end; no quotes.
```

Figure A.2: LLM prompt for geo-reference classification used with LLaMA70B-instruct. The title of the submission is given in the user message.

```
You are a linguistic interpreter and want to predict the future
inflation rate in US from textual documents. The following document
will report sentences potentially referring to US: df['title'][i]. You
have to print a signal that can be UP (if the document is signalling
inflation going up in the short/long-term run), DOWN (if the sentence
is signalling inflation going down in the short/long-term run) or
NEUTRAL (if the sentence is neutral or does not signal a particular
direction on inflation). Print only the results of the signal, do
not summarize the sentence nor give any reason on your choice. Even
if more sentences or paragraphs are provided to you, you only have to
print one signal that can be UP, DOWN or NEUTRAL.
```

Figure A.3: LLM prompt for signal extraction used with LLaMA70B-instruct. `df['title'][i]` contains the submission or comment.

C Human Labeling Assisted by LLMs and ChatGPT

LLMs helped scale the construction of our labeled training set, but they did not replace human judgment. Standard off-the-shelf sentiment measures are often too coarse to capture the nuances of inflation-related discourse on Reddit. We therefore designed a task-specific annotation scheme centered on *directional inflation expectations*. Each submission is assigned to one of three mutually exclusive categories: UP (the author conveys an expectation that inflation/prices will rise), DOWN (the author conveys an expectation that inflation/prices will fall), or NEUTRAL (no clear directional signal, mixed statements, or purely descriptive content).

Seed labeling and zero-shot expansion. We began by manually labeling 500 Reddit submissions. To expand coverage efficiently, we then used a 70B-parameter LLaMA model in a zero-shot classification setup to label an additional 700 submissions into the same three categories (UP, DOWN, NEUTRAL) using the prompt shown in Figure A.4. The resulting

1,200 labeled titles formed the initial seed dataset.

```
You are an economist working on the consumer price index and you can
only answer by saying "up", "down" or "neutral".
You have been asked to tell whether a given statement suggests an
increase or decrease in consumer prices or whether it is neutral about
the consumer prices trend.
Here are some examples:
Statement:  a concise explanation of the zero lower bound and how it
stops inflation
Answer:  down
Statement:  hypothetical scenario?  How would hyperinflation play out
in the United States?
Answer:  up
Statement:  the Dow isn't at an all-time high when you adjust for
inflation
Answer:  neutral
```

Figure A.4: Prompt given to LLaMA 70B for labeling as UP, DOWN or NEUTRAL Reddit submissions. The title of the submission is given in a separate user message.

Initial fine-tuning and out-of-sample validation. Next, we fine-tuned a smaller LLaMA model (8B parameters) on the labeled seed dataset and evaluated performance on a held-out test set. The model achieved a weighted F1 score close to 70% on unseen submissions, indicating that the labeling task is learnable and that a compact model can reproduce the annotation scheme with reasonable accuracy.

Interactive review of disagreements. We then conducted a structured error analysis on *mismatched cases*, i.e., submissions in the test set for which the fine-tuned model's prediction differed from the assigned label. Rather than treating these discrepancies as mechanical

prediction errors, we used ChatGPT as an interactive assistant to diagnose the source of disagreement (e.g., implicit versus explicit references to inflation, sarcasm, ambiguity about time horizons, or confusion between current inflation and expected inflation). For each disputed item, we elicited a short rationale for the predicted label and discussed the case jointly, refining the decision rules when needed. This step improved within-team consistency and reduced idiosyncratic labeling.

Dataset expansion with human-in-the-loop ChatGPT labeling. After clarifying the annotation rubric, we further expanded the labeled dataset by adding roughly 200 additional submissions labeled with ChatGPT under close human supervision. We initially queried ChatGPT one post at a time to align terminology and domain usage, then moved to small batches while requiring brief justifications for each label. Batch sizes were gradually increased only after repeated spot checks confirmed stable labeling. Whenever ChatGPT's label conflicted with the authors' judgment, the case was revisited and resolved before being added to the training set.

Outcome and use in the paper. Overall, this iterative procedure produced a labeled dataset of approximately 1,400 submissions with improved internal consistency. In practice, LLMs served as *annotation accelerators* and *consistency checkers*: they reduced the marginal cost of labeling while keeping the final ground-truth assignments under human control. The resulting labeled set is used to fine-tune the LLaMA-based classifiers that generate the text indicators employed to label the remaining submissions and to construct the monthly forecasting regressors used in the main analysis.

C.1 Prompt used to label Reddit submission with ChatGPT

Here we provide the prompt used with ChatGPT to classify the Reddit submissions as UP, DOWN or NEUTRAL.

```
You are an economist and you have to label some posts from social media
as UP for showing or intending increasing prices or inflation, DOWN if
showing or intending decreasing prices or deflation, or NEUTRAL if the
post is meaning neither an increase nor a decrease in inflation/prices.
How would you label the following post?
‘‘Disinflation denial’’
What about this?
‘‘Why is bad to induce deflation to lower inflation?’’   ...
```

Figure A.5: ChatGPT prompt for labeling as UP, DOWN or NEUTRAL Reddit submissions. The title of the submission is given in a separate user message.

D A Primer on Large Language Models (LLMs)

Large Language Models (LLMs) are neural-network-based systems designed to process and generate human language. They operate by converting input text into smaller units called *tokens* (words or subwords) and mapping these tokens into numerical representations through an embedding layer that encodes both semantic content and positional information. The core component is the *attention* mechanism, which enables the model to weigh relationships among all tokens in a sequence so that each token is interpreted in context. Stacked attention and feed-forward blocks, combined with normalization and residual connections, support stable optimization and rich representation learning; dropout is often used as a regularization device during training. In generation, an LLM produces a probability distribution over possible next tokens and selects an output using a decoding rule. Many widely used LLMs are *autoregressive*: they generate text sequentially, conditioning on previously generated tokens, which supports coherent and contextually consistent outputs.

LLMs are trained on very large text corpora and can be adapted to downstream tasks such as classification, summarization, and sentiment analysis. Examples include BERT, GPT-style models, LLaMA, Gemma, and Qwen.

D.1 LLMs adopted

To extract "educated" signals about inflation and price dynamics from Reddit submissions and comments, we consider four model families.

BERT family: 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 large general-language corpora (Toronto BookCorpus and English Wikipedia) using masked language modeling and next-sentence prediction. BERT is available in two versions: the *base* version with approximately 110 million parameters and the *large* version with 340 million parameters, but here we used the *base* version. We also consider two domain-adapted variants of BERT. FinBERT (Araci, 2019) is tailored to financial language and is widely used for tasks such as sentiment classification and risk-related text analysis. InflaBERT (Allard et al., 2024) is a recent variant fine-tuned to infer inflation-related sentiment from news text.

LLaMA family: LLaMA 3 (1B, 3B, 8B, and 70B) LLaMA (Large Language Model Meta AI) is a family of open-source LLMs released by Meta in early 2023 (Grattafiori et al., 2024). The LLaMA 3.x models have been available since April 2024 and feature a substantially expanded vocabulary (more than 128,000 tokens) and training corpus relative to earlier versions.[25] We consider models with 1B, 3B, 8B, and 70B parameters; due to computational constraints, we do not fine-tune the 70B model and instead use it as an unfine-tuned benchmark.

Gemma Family: Gemma 2 (2B, 9B, and 27B) Gemma is a family of open LLMs developed by Google and released in early 2024 (Team et al., 2024). Gemma 2 (released in late June 2024) introduces architectural enhancements such as sliding-window attention and logit soft-capping, and it is pre-trained on a larger corpus than the original release.[26] We consider the 2B, 9B, and 27B versions; Gemma 27B is the largest model we fine-tune.[27]

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

[26] `Italsofeaturesavocabularyof256,128tokens.Link:https://huggingface.co/blog/gemma`

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

Qwen Family: Qwen 2.5 (0.5B, 1.5B, and 7B) Qwen 2.5 is a family of open LLMs developed by the Qwen team at Alibaba Cloud (Qwen et al., 2025). The models are pre-trained on a very large corpus (around 18 trillion tokens) and are designed to perform well across a range of general and specialized tasks.[28] We consider sizes from 0.5B to 7B parameters.

D.2 Why fine-tuning LLMs matters

Although LLMs are trained on broad internet text, off-the-shelf performance may be suboptimal for specialized tasks such as interpreting inflation discussions on Reddit, which involve domain-specific terminology, informal language, and platform-specific conventions. We therefore adapt several models to our downstream classification task using **fine-tuning**, i.e., additional supervised training on a labeled dataset relevant to the target domain.

Fine-tuning *Fine-tuning* is the process of adapting a pre-trained 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 very large models can be computationally costly and time-consuming. For this reason, we primarily rely on *parameter-efficient fine-tuning* (PEFT) methods, which update only a small subset of parameters while leaving the base model largely unchanged. We use two PEFT approaches: `LoRA` (Low-Rank Adaptation) and `DoRA` (Weight-Decomposed Low-Rank Adaptation). We also employ **quantization** to reduce memory requirements and improve throughput during fine-tuning.

Low-Rank Adaptation (`LoRA`) : `LoRA`(Hu et al., 2022) fine-tunes a model by learning low-rank updates to selected weight matrices rather than updating all parameters. It introduces small, trainable components—low-rank matrices, called *adapters*—that adjust the model's behavior without modifying its core architecture. This approach is like adding a custom filter to a camera lens—enhancing specific features without changing the camera itself.

The architecture of large language models typically consists of deep neural networks with numerous dense layers, each involving matrix multiplications during training. These weight

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

matrices generally have full rank. In LoRA, for any layer selected for fine-tuning, the pre-trained weight matrix $W_0 \in \mathbb{R}^{d\times k}$ is augmented with a low-rank decomposition $\Delta W = B \times A$, where $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times k}$, and the rank $r \ll \min(d, k)$.[29]

During fine-tuning, the original weights W_0 are kept frozen, while only the trainable parameters contained in A and B receive gradient updates. The matrices A and B are initialized differently: A with Kaiming-uniform initialization,[30] and B to zero so that $\Delta W = B \times A = 0$ at the start of training. At each forward pass, the weight update takes the form:

$$W' = W_0 + \Delta W = W_0 + (B \times A) \tag{A.1}$$

This approach substantially reduces the number of trainable parameters and significantly lowers the computational burden of adaptation.

Weight-Decomposed Low-Rank Adaptation (DoRA) (Liu et al., 2024) extends LoRAby decomposing the weight matrix into *magnitude* and *direction* components and fine-tuning both. Instead of applying all fine-tuning adjustments in one place, DoRA spreads them out, improving the model's adaptability while maintaining efficiency. DoRAdecomposes high-rank LoRA layers into structured single-rank components and introduces the concept of directional updates, which enhances both learning capacity and stability. Starting from $W = m\,\frac{V}{\|V\|_c} = \|W\|_c\,\frac{W}{\|W\|_c}$, the two components of DoRA are: $m \in \mathbb{R}_{1\times k}$ is the magnitude vector, while $V \in \mathbb{R}_{d\times k}$ is the directional matrix, with $\|\cdot\|_c$ being the vector-wise norm of a matrix across each column. DoRA starts with pre-trained weights W_0, where $m = \|W_0\|_c$, and $V = W_0$, keeping V frozen and m as a trainable vector. The directional component is updated through normal LoRA, while the magnitude is handled by a separate learnable parameter. The resulting effective weights are

$$W' = m\frac{V + \Delta V}{\|V + \Delta V\|_c} = m\frac{W_0 + B \times A}{\|W_0 + B \times A\|_c}. \tag{A.2}$$

[29] This is conceptually related to reduced-rank methods used in econometrics.

[30] Kaiming initialization is a method for setting the initial weights of neural networks, particularly those using ReLU activation functions. It aims to preserve the variance of activations across layers by drawing weights from a distribution scaled by the number of input units, thereby improving training stability and convergence speed.

By separating directional and magnitude updates, DoRAcan improve stability and adaptation capacity while remaining parameter-efficient.

Quantization: QLoRA and QDoRA *Quantization* (Q) reduces memory usage by representing model weights with fewer bits (e.g., 8-bit or 4-bit precision instead of 32-bit). This reduces the model's memory footprint and accelerates 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 or DoRA. When combined with LoRA-style updates, this yields QLoRA (Dettmers et al., 2023), when combined with DoRAupdates, this yields QDoRA. Both enable efficient fine-tuning of relatively large models on commodity hardware. In our setting, quantization and PEFT techniques allow us to adapt multiple LLMs to Reddit posts and comments at a manageable computational cost.

Our fine-tuning techniques. To fine-tune the LLMs, we implement a modified QLoRA pipeline (as in Dettmers et al., 2023). First, we quantize the base models to 4-bit precision to reduce memory requirements and accelerate training. We then apply DoRA (weight-decomposed low-rank adaptation) (Liu et al., 2024), a parameter-efficient fine-tuning method that extends LoRA by separately updating the magnitude and direction of the weight matrices. To further improve training efficiency and stability, we use the LoRA+ optimizer (Hayou et al., 2024). Unlike standard implementations that update the LoRA adapter matrices with a single learning rate, LoRA+ assigns distinct learning rates to the two low-rank matrices (A, B) in equation (A.1).[31] Optimization is performed with AdamW (Loshchilov and Hutter, 2017), which decouples weight decay from the gradient update to improve regularization during training.

We deploy the QDoRA setup with the LoRA+ optimizer in two configurations. The baseline configuration uses a standard adapter rank. The "extreme" configuration reduces the adapter rank by a factor of 2^3, yielding a more aggressive parameter-efficient specification. We refer to these approaches as QDoRA+ and xQDoRA+, respectively.

In the context of this paper, these methods make it feasible to fine-tune language models to recognize the linguistic patterns and narrative cues that are informative for inflation

[31] LoRA+ sets $\eta_B = \lambda \eta_A$ with a fixed $\lambda > 1$ and tunes η_A.

forecasting using Reddit discussions.

Fine-Tuning LLMs to Extract an "Educated" Signal from Reddit posts To construct the text-based leading indicators used in our nowcasting and forecasting exercises, we first build a labeled dataset of approximately 1,400 submissions whose title contains the term *inflation*. Each title is classified into one of three categories: UP, DOWN, or NEUTRAL, indicating whether it conveys increasing inflation/prices, decreasing inflation/prices, or no clear directional signal. Labels are obtained using a combination of human judgment, LLaMA-70B and ChatGPT.

We proceed in two steps. First, we manually label a subset of submission titles from `r/Economics` and we add some labels using LLaMA-70B. We then use most of these labeled examples to train a LLaMA-8B classifier, which is applied to the remaining unlabeled titles. Observations for which the model predictions disagree with the initial labels are reviewed and adjudicated with the support of ChatGPT, yielding a consistent training set for supervised learning.

The title datasets for `r/Economics` and `r/economy` each contain more than 14,000 observations. To capture heterogeneity in users' views on inflation and price dynamics, we fine-tune a range of language models on the labeled titles. In addition to the uncased BERT base model, we consider decoder-only LLMs with substantially more parameters, including LLaMA, Gemma, and Qwen variants. The same training pipeline is applied across models.[32] The preferred specification is selected based on the (weighted) F1 score.

To make fine-tuning feasible on a single GPU (NVIDIA A100), we combine 4-bit quantization with PEFT. In particular, we use `DoRA` (Liu et al., 2024), a weight-decomposed variant of LoRA, which introduces low-rank adapters and updates only a small fraction of model parameters while keeping the base weights fixed. Figure A.6 shows the number of trainable parameters for each LLM with `xQDoRA+` is greatly reduced when compared to `QDoRA+`. Quantization further reduces memory requirements by storing weights at lower precision with minimal loss in downstream performance.

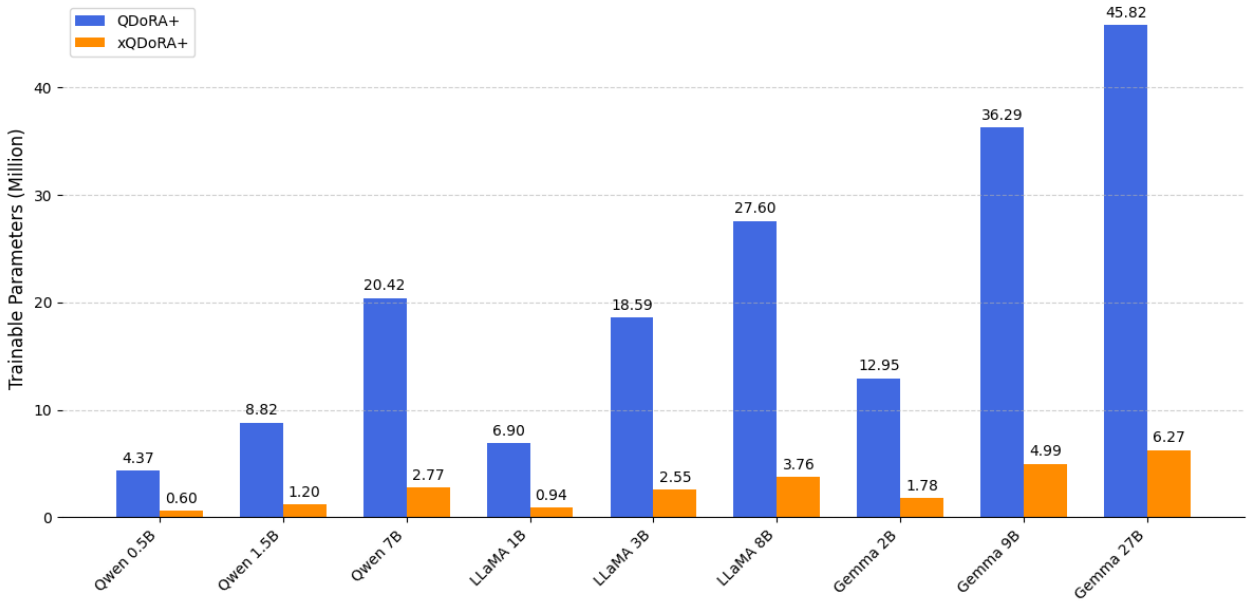
[32]For the LLM training stage, we use 14,214 titles from `r/Economics` spanning 2008–2025. We split the data into 10% test and 90% training, and further split the training sample into 80% training and 20% validation. Labels are encoded into the three classes UP, NEUTRAL, and DOWN. After training, the fitted models are used to generate predicted labels for the 14,445 titles from `r/economy`.

The fine-tuned models are subsequently used to classify not only submission titles but also comments associated with inflation-related submissions.

Reducing the number of trainable parameters does not materially degrade classification performance. Figure A.7 shows that the "extreme" configuration (`xQDoRA+`), which trains substantially fewer adapter parameters than the baseline configuration (`QDoRA+`), achieves comparable weighted F1 scores, and sometimes the weighted F1 is even higher than that obtained with the baseline `QDoRA+`. Additional details are reported in Tables A.1 and A.2. In summary, larger models tend to achieve higher test performance, but strong results are also attainable with smaller, domain-adapted architectures.

To assess the stability of fine-tuning outcomes, we run a Monte Carlo experiment consisting of 20 independent replications of the full training pipeline for each model, varying only the random seed. Figure A.8 reports box plots of weighted F1 scores for unfine-tuned ("raw") models (Figure A.8a) and their fine-tuned counterparts (Figure A.8b). Fine-tuning produces large and systematic improvements in median performance (from roughly 30% for raw models to above 75% after fine-tuning) and substantially reduces dispersion across runs. In addition, the results highlight that smaller, task-adapted models–such as InflaBERT, which is trained on inflation-related news–can match or outperform much larger general-purpose LLMs in this classification setting. Overall, parameter-efficient fine-tuning improves both accuracy and robustness, yielding more reliable text labels for constructing our forecasting indicators.

Figure A.6: Number of trainable parameters (in millions) for each LLM in the fine-tuning process.



Notes: The figure shows the number of trainable parameters (in millions) using the two parameter-efficient fine-tuning techniques QDoRA+ and xQDoRA+.

Table A.1: Percentage of trainable parameters for each model used in the fine-tuning process

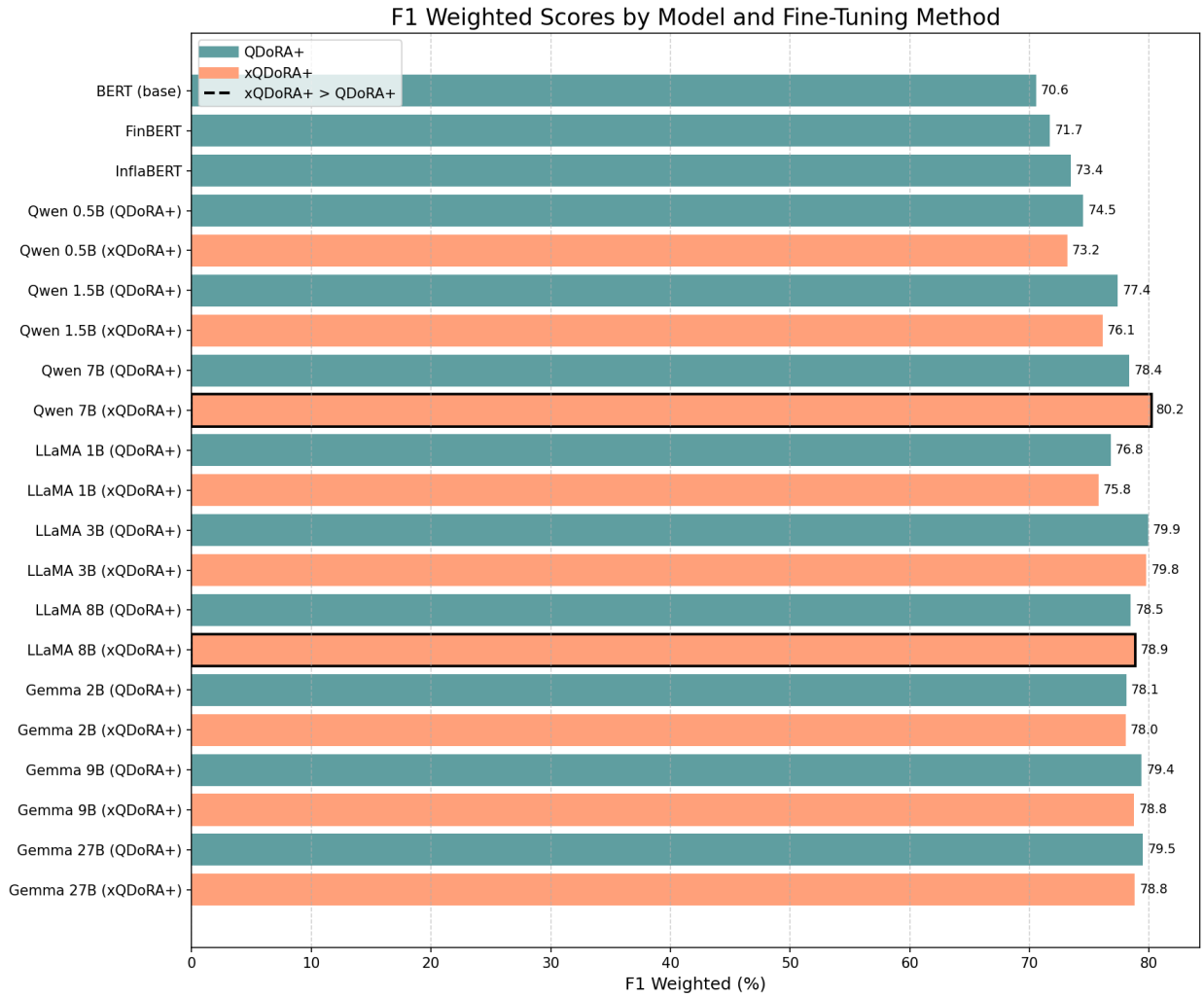
Model	**Size in billion (B) of parameters**	**Percentage of trainable parameters**
BERT (base)	0.109 B	100.0%
FinBERT	0.109 B	100.0%
InflaBERT	0.082 B	100.0%
Qwen 2.5 0.5B	0.498 B	0.8782% (QDoRA+) 0.1198% (xQDoRA+)
Qwen 2.5 1.5B	1.553 B	0.5682% (QDoRA+) 0.0773% (xQDoRA+)
Qwen 2.5 7B	7.091 B	0.2880% (QDoRA+) 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+)

Notes: "Size" reports the number of parameters (in billions). "Trainable parameters" reports the share of parameters updated during fine-tuning using PEFT (`QDoRA+` and `xQDoRA+`.

Table A.2: Performance metrics on the test sample for different LLMs classifying subreddit titles as UP, DOWN, or NEUTRAL

Model	Recall	Precision	F1	weighted F1	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 A.7: Weighted F1 scores for each LLM under different fine-tuning strategies.

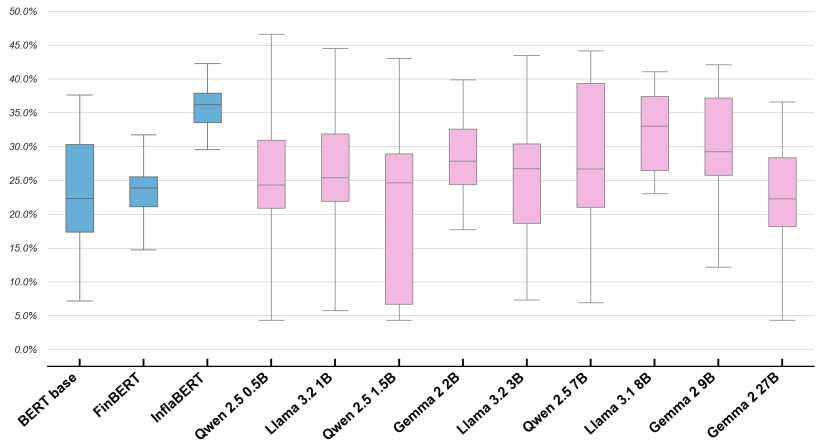


Notes: The figure shows the weighted F1 score for each LLM using the two parameter-efficient fine-tuning techniques, QDoRA+ and xQDoRA+.

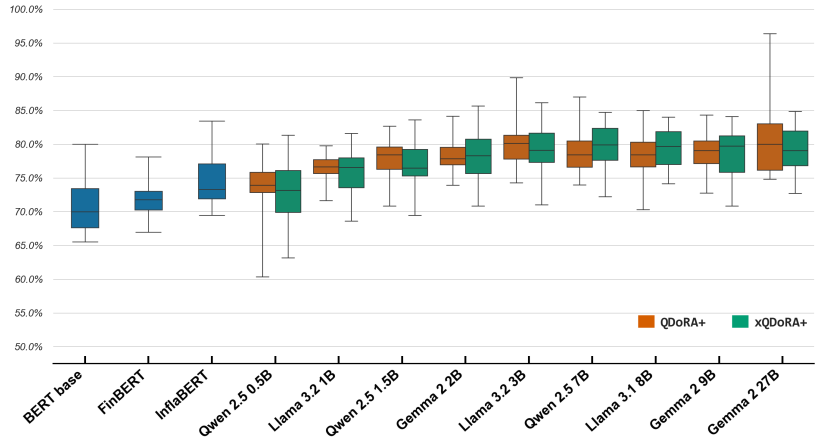
Figure A.8 shows box plots of the F1 score for unfine-tuned models and for the fine-tuned ones, clearly showing a competitive advantage in terms of classification accuracy when the models are fine-tuned.

Figure A.8: Uncertainty of the weighted F1 score of the test set

(a) 'Raw' models



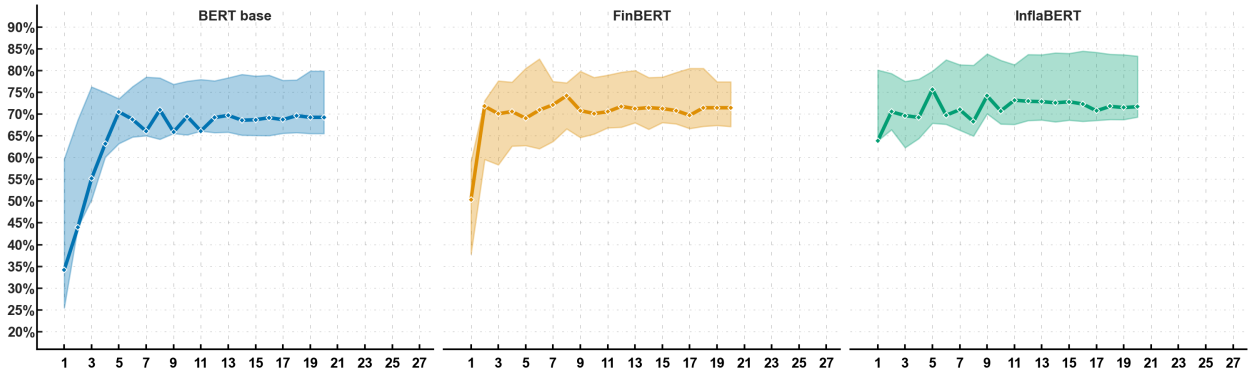
(b) Fine-tuned models



Notes: The figure shows box plots of the F1 scores for unfine-tuned (i.e., "raw") models and for models fine-tuned over 20 independent runs with different random seeds

Figures A.9, A.10, A.11, and A.12 show the evolution of the weighted F1-metric on the validation test over the epochs of training until the process terminates due to no improvements within the latest 5 epochs (the so-called early stopping).

Figure A.9: Weighted F1 scores of the validation set in the training loop: models based on the BERT architecture.



The charts show that the dynamics of the weighted F1-metrics for the BERT-based models are smoother over epochs than those for the `LLaMA`, `Gemma` and `Qwen` families of LLMs. It is also worth noting that the fine-tuning of the Qwen family's LLMs ends earlier than that of the LLMs from the Gemma and LLaMA families.

Figure A.10: Weighted F1 scores of the validation set in the training loop: models from the Qwen 2.5 family.

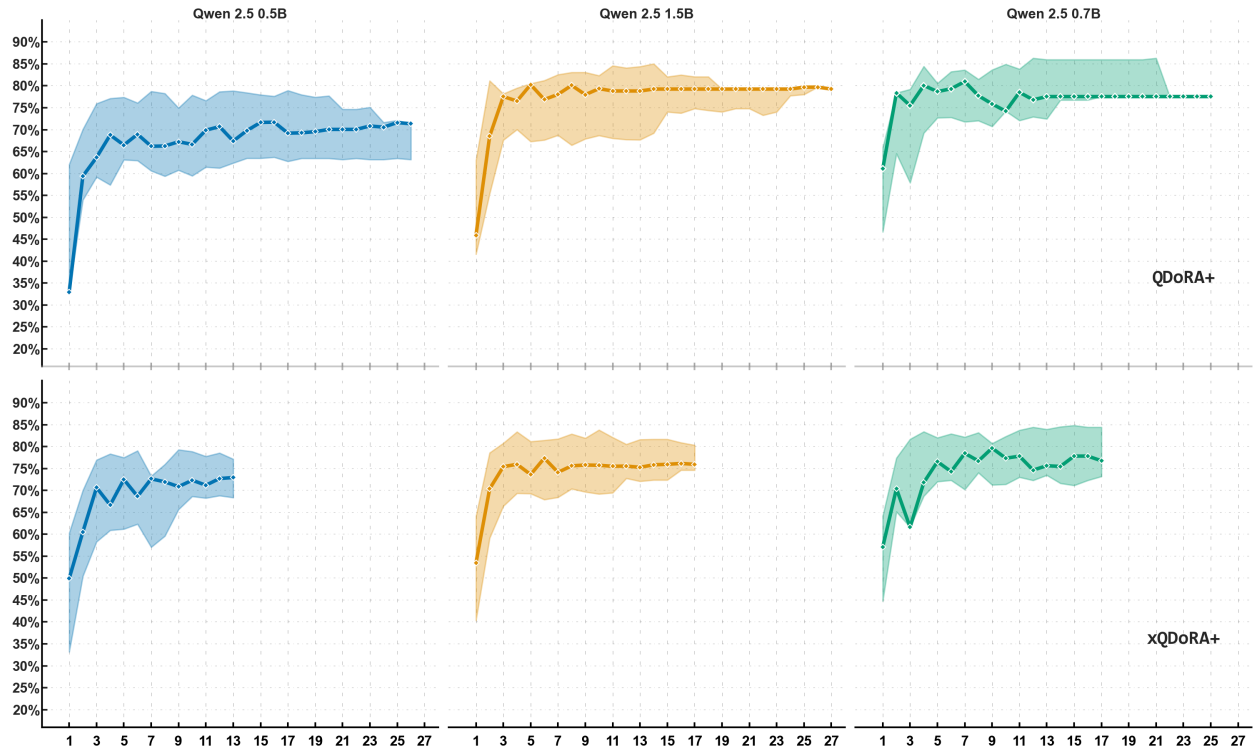


Figure A.11: Weighted F1 scores of the validation set in the training loop: models from the LLaMA 3 family.

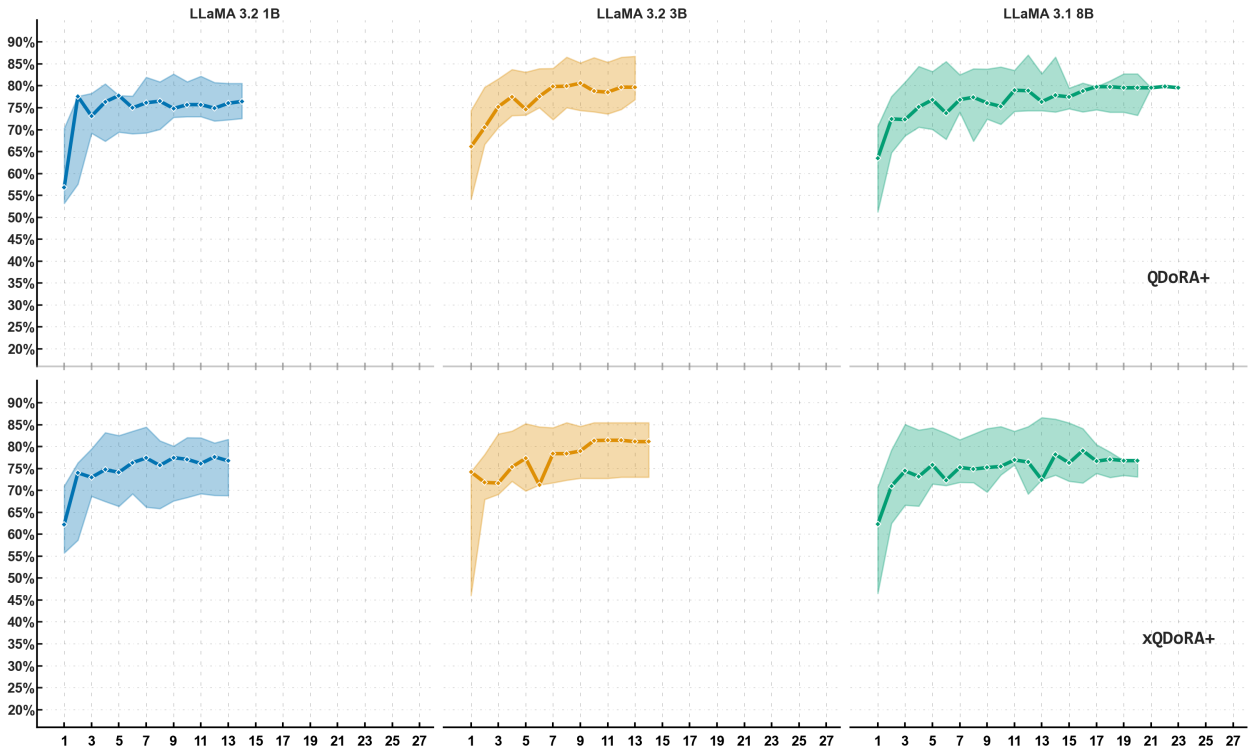
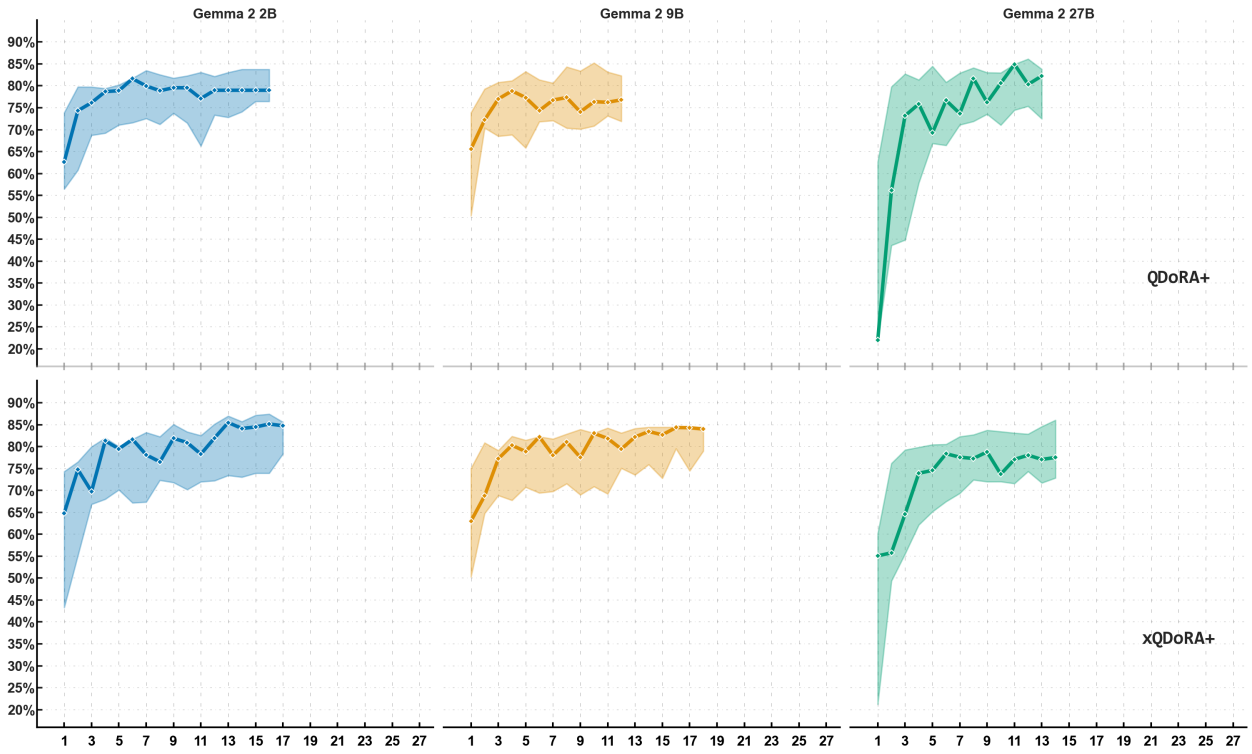


Figure A.12: Weighted F1 scores of the validation set in the training loop: models from the Gemma 2 family.



D.3 Efficiency and effectiveness of the two fine-tuning methods

To quantify the trade-off between computational cost and model quality, we compare our baseline fine-tuning method (`QDoRA+`) to the more parameter-efficient "extreme" variant (`xQDoRA+`). The goal is to determine whether the efficiency gains delivered by `xQDoRA+` come at the expense of classification performance. We focus on two outcomes measured on the held-out test set: (i) inference wall-time (efficiency) and (ii) the weighted F1 score (effectiveness).

Methodology. We exploit a matched experimental design: for each model architecture, both fine-tuning methods are evaluated using the same set of 15 random seeds. This design allows us to difference out seed-specific idiosyncrasies and isolate the incremental effect of `xQDoRA+` relative to `QDoRA+`.

We estimate two fixed-effects specifications. For efficiency, we use a log-linear model so coefficients can be interpreted as percentage changes in wall-time. For effectiveness, we use a linear model for the weighted F1 score to measure absolute performance differences. The generic specification is:

$$Y_{ims} = \alpha_m + \beta_m D_{ims}^{(m)} + \eta_s + \varepsilon_{ims},$$

where i indexes the fine-tuning method, m the model architecture, and s the seed. The term α_m captures the architecture-specific baseline, while η_s denotes seed fixed effects that absorb initialization-specific variation. The indicator $D_{ims}^{(m)}$ equals one for observations produced with `xQDoRA+` under architecture m (and zero otherwise), so β_m measures the architecture-specific effect of moving from `QDoRA+` to `xQDoRA+`.

For efficiency, $Y_{ims} = \log(\text{walltime}_{ims})$, so $\exp(\beta_m) - 1$ is the implied percentage change in wall-time. For effectiveness, $Y_{ims} = \text{F1}_{ims}$, so β_m is the absolute change in weighted F1. Inference relies on cluster-robust standard errors clustered at the seed level ($G = 15$) and t-tests with $G - 1$ degrees of freedom. Because the number of clusters is moderate, we also validate significance using the Wild Cluster Bootstrap (Cameron et al., 2008).

Results. Table A.5 summarizes the joint efficiency–effectiveness comparison. The efficiency estimates show substantial heterogeneity across architectures: small models exhibit

limited gains, while larger models display economically meaningful reductions in latency (e.g., Gemma 2 9B). Table A.3 reports the architecture-specific estimates for the wall-time regression, and Figure A.13 visualizes the implied percentage changes with 95% confidence intervals. Table A.4 confirms that the statistically significant efficiency gains remain significant under the wild bootstrap procedure.

Crucially, the effectiveness results indicate that these wall-time reductions do not translate into economically meaningful losses in predictive performance. For most architectures, weighted F1 differences are statistically indistinguishable from zero. In cases where the point estimate suggests a small decline and conventional inference approaches marginal significance, the wild bootstrap corroborates that we cannot robustly reject the null of equal performance. Overall, the magnitude of any estimated F1 differences is small relative to the typical variation across seeds.

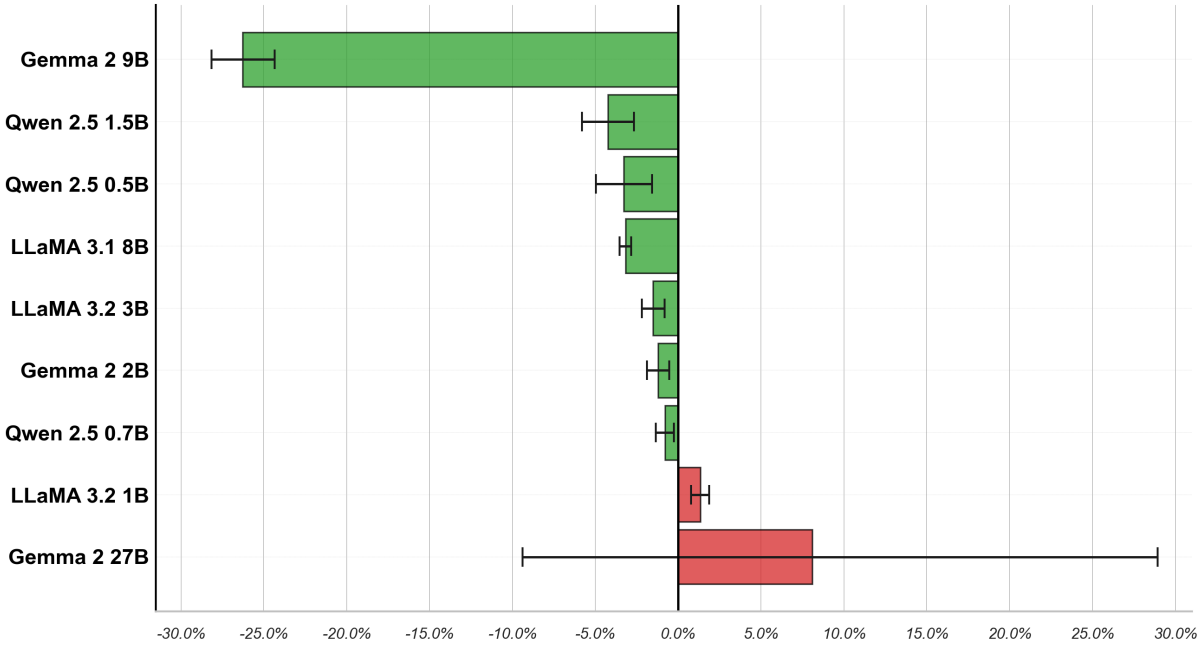
Conclusion. Taken together, the results support `xQDoRA+` as a practical improvement over `QDoRA+`: it delivers measurable reductions in inference wall-time—especially for larger architectures—while maintaining classification performance that is, in statistical terms, comparable to the baseline method. This finding motivates the use of `xQDoRA+` in the large-scale labeling pipeline, where computational constraints are binding.

Table A.3: Fixed-effects estimation of inference efficiency gains

Model architecture	Coeff. (β)	Std. err.	t-stat	p-value	
Gemma 2 2B	-0.012	0.003	-3.758	0.002	***
Gemma 2 27B	0.078	0.082	0.946	0.360	
Gemma 2 9B	-0.305	0.012	-25.301	0.000	***
LLaMA 3.2 1B	0.013	0.002	5.369	0.000	***
LLaMA 3.2 3B	-0.015	0.003	-4.658	0.000	***
LLaMA 3.1 8B	-0.032	0.002	-19.143	0.000	***
Qwen 2.5 1.5B	-0.043	0.008	-5.662	0.000	***
Qwen 2.5 0.5B	-0.033	0.008	-4.055	0.001	***
Qwen 2.5 7B	-0.008	0.003	-3.167	0.007	***

Notes: Fixed-effects regression of log(walltime) ($N = 270$) with seed fixed effects ($G = 15$). Inference uses the t-distribution with $G - 1$ degrees of freedom.

Figure A.13: Estimated efficiency gains by model architecture (95% confidence intervals)



Notes: Implied percentage change in wall-time when switching from `QDoRA+` to `xQDoRA+`. Negative values indicate faster inference under `xQDoRA+`.

Table A.4: Robustness of significant efficiency gains (wild cluster bootstrap)

	p-values		
Model architecture	**Cluster-robust OLS**	**Wild bootstrap**	**Robust at 5%?**
Gemma 2 2B	0.002	0.003	Yes
Gemma 2 9B	0.000	0.000	Yes
LLaMA 3.2 1B	0.000	0.000	Yes
LLaMA 3.2 3B	0.000	0.000	Yes
LLaMA 3.1 8B	0.000	0.000	Yes
Qwen 2.5 1.5B	0.000	0.000	Yes
Qwen 2.5 0.5B	0.001	0.001	Yes
Qwen 2.5 7B	0.007	0.008	Yes

Notes: The wild cluster bootstrap uses $B = 999$ replications and accounts for the moderate number of clusters ($G = 15$). The table reports robustness checks for architectures with statistically significant gains in Table A.3.

Table A.5: Pareto analysis: efficiency versus effectiveness

	Efficiency (time)			**Effectiveness (weighted F1)**			
Model	**Chg. (%)**	***p*-val**		**Diff.**	**Std. err.**	***p*-val**	
Gemma 2 2B	-1.2019	0.0021	**	-0.0317	0.0053	0.0000	***
Gemma 2 27B	8.0896	0.3603		0.0101	0.0088	0.2701	
Gemma 2 9B	-26.2634	0.0000	***	-0.0647	0.0130	0.0002	***
LLaMA 3.2 1B	1.3365	0.0001	***	0.0004	0.0103	0.9729	
LLaMA 3.2 3B	-1.5036	0.0004	***	-0.0096	0.0057	0.1133	
LLaMA 3.1 8B	-3.1756	0.0000	***	-0.0158	0.0065	0.0290	*
Qwen 2.5 1.5B	-4.2364	0.0001	***	0.0023	0.0079	0.7733	
Qwen 2.5 0.5B	-3.2683	0.0012	**	-0.0103	0.0063	0.1213	
Qwen 2.5 7B	-0.7940	0.0069	**	-0.0127	0.0046	0.0142	*

Notes: "Chg. (%)" is the implied percentage change in wall-time from the log-linear specification; negative values indicate faster inference under `xQDoRA+`. "Diff." is the estimated change in weighted F1 (`xQDoRA+` minus `QDoRA+`); negative values indicate a decrease under `xQDoRA+`.

E Additional Forecast Results

In this section of the Appendix, we report additional results of the point forecast exercise, including the results using different loss functions (Section E.1) and results for the RMSE ratios for single Reddit-based indicators over the whole out-of-sample and in the sample before the COVID pandemic (Section E.2).

E.1 Additional forecast losses for point-forecast

In this section we provide additional results on the out-of-sample forecasting application, reporting accuracy measured with mean absolute error (MAE) and mean absolute deviation (MAD). These alternative loss functions complement the RMSE results in the main text by placing less weight on large forecast errors, thereby offering a more balanced view of forecast accuracy across all periods.

The MAE results in Tables A.6 and A.7 are broadly consistent with the RMSE findings. Reddit-based LLM indicators outperform the AR(1) benchmark at most horizons, with the gains concentrated at short-to-medium horizons ($h = 1$ to $h = 6$). The LLaMA 70B forecast aggregation achieves the lowest MAE ratios at short horizons for CPI, while the fine-tuned LLM aggregation is highly competitive throughout. At longer horizons, expectation-based measures become relatively more informative, consistent with the idea that the long-run inflation trend is well captured by survey and market expectations. Table A.8 reports the best individual models within each family based on MAE. The patterns mirror those obtained under RMSE (Figure 6): InflaBERT dominates the fine-tuned family at nearly all horizons, with Qwen 7B entering at $h = 18$; LLaMA 70B specifications from `r/economy` perform best for CPI, while `r/wallstreetbets` dominates for PCE at medium-to-long horizons. Shorter MA windows are preferred at short horizons and longer windows at long horizons, reinforcing the evidence that the optimal degree of smoothing depends on the forecast horizon.

Table A.6: Forecast results (MAE ratios) for *CPI*.

Model	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=9$	$h=12$	$h=18$
Expectations									
Best expectations	1.034	1.023	0.967	0.923	0.893	0.877	0.823	0.784	0.702
1-Y Inflation Swap (best)	1.020	0.939	0.905	0.904	0.895	0.866	0.807	0.821	0.764
Reddit-sentiment									
Best sentiment	0.992	0.927	0.904	0.862	*0.820*	*0.787*	**0.717**	**0.700**	**0.658**
Reddit-LLM									
LLM forecast aggregation	0.897	*0.884*	0.873	0.847	0.827	0.809	0.820	0.840	0.794
Best fine-tuned LLM	*0.895*	0.863	0.865	0.852	0.854	0.862	0.887	0.897	**0.751**
Llama70B forecast aggregation	**0.809**	**0.771**	**0.771**	**0.743**	**0.729**	**0.722**	0.746	*0.761*	0.757
Best Llama70B	0.884	0.881	*0.897*	*0.883*	*0.875*	0.861	*0.855*	0.847	*0.753*

Notes: MAE is computed as a ratio: $MAE(AR\text{-}X(1))/MAE(AR(1))$, where the benchmark is $AR(1)$. The initial in-sample period spans from 2009M1 to 2012M12; the out-of-sample period covers 2013M1 to 2025M8. Forecasts are generated using a recursive window, with the model re-estimated each step (month). For *Best sentiment*, the aggregated sentiment model is excluded. Bold (italic) indicates the best (second-best) model for each horizon.

Table A.7: Forecast results (MAE ratios) for *PCE*.

Model	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=9$	$h=12$	$h=18$
Expectations									
Best expectations	1.016	0.977	0.939	0.899	0.868	0.857	0.841	0.814	0.766
1-Y Inflation Swap (best)	1.008	0.972	0.935	0.923	0.909	0.886	0.841	0.883	0.822
Reddit-sentiment									
Best sentiment	0.978	0.962	0.915	0.857	*0.793*	**0.745**	**0.621**	**0.536**	**0.465**
Reddit-LLM									
LLM forecast aggregation	0.902	0.854	0.840	0.812	0.801	0.795	*0.779*	0.788	0.723
Best fine-tuned LLM	*0.896*	0.866	0.869	0.845	0.834	0.829	0.797	*0.732*	**0.629**
Llama70B forecast aggregation	**0.878**	**0.838**	**0.823**	**0.809**	**0.784**	*0.785*	0.813	0.840	0.758
Best Llama70B	0.956	*0.895*	*0.856*	*0.831*	*0.821*	0.810	0.783	*0.767*	*0.687*

Notes: MAE is computed as a ratio: $MAE(AR\text{-}X(1))/MAE(AR(1))$, where the benchmark is $AR(1)$. The initial in-sample period spans from 2009M1 to 2012M12; the out-of-sample period covers 2013M1 to 2025M8. Forecasts are generated using a recursive window, with the model re-estimated each step (month). Bold (italic) indicates the best (second-best) model for each horizon.

Table A.8: Best forecasting models by horizon (CPI vs PCE) within each family, based on MAE ratios.

	Fine-tuned LLMs			LLaMA-70B (not fine-tuned)			Lexicon / sentiment		
h	Model	Subreddit	MA	Model	Subreddit	MA	Model	Subreddit	MA
CPI									
1	InflaBERT	r/Economics	1	LLaMA-70B	r/economy	1	VADER	r/wallstreetbets	360
2	InflaBERT	r/Economics	30	LLaMA-70B	r/economy	5	TextBlob	r/economy	10
3	InflaBERT	r/Economics	30	LLaMA-70B	r/economy	30	TextBlob	r/economy	10
4	InflaBERT	r/Economics	30	LLaMA-70B	r/economy	10	LM	r/Economics	90
5	InflaBERT	r/Economics	30	LLaMA-70B	r/economy	90	LM	r/Economics	90
6	InflaBERT	r/Economics	30	LLaMA-70B	r/economy	90	LM	r/Economics	90
9	InflaBERT	r/economy	90	LLaMA-70B	r/economy	120	LM	r/Economics	180
12	InflaBERT	r/economy	120	LLaMA-70B	r/economy	360	TextBlob	r/economy	360
18	Qwen-7B	r/wallstreetbets	360	LLaMA-70B	r/economy	360	TextBlob	r/economy	180
PCE									
1	InflaBERT	r/Economics	90	LLaMA-70B	r/economy	30	VADER	r/wallstreetbets	360
2	InflaBERT	r/Economics	180	LLaMA-70B	r/wallstreetbets	180	TextBlob	r/economy	90
3	InflaBERT	r/Economics	180	LLaMA-70B	r/wallstreetbets	180	TextBlob	r/wallstreetbets	120
4	InflaBERT	r/wallstreetbets	180	LLaMA-70B	r/wallstreetbets	180	TextBlob	r/wallstreetbets	360
5	InflaBERT	r/wallstreetbets	180	LLaMA-70B	r/wallstreetbets	360	TextBlob	r/wallstreetbets	360
6	InflaBERT	r/wallstreetbets	180	LLaMA-70B	r/wallstreetbets	360	TextBlob	r/wallstreetbets	360
9	InflaBERT	r/wallstreetbets	360	LLaMA-70B	r/wallstreetbets	360	TextBlob	r/wallstreetbets	360
12	InflaBERT	r/wallstreetbets	360	LLaMA-70B	r/wallstreetbets	360	TextBlob	r/wallstreetbets	360
18	Qwen-7B	r/wallstreetbets	360	LLaMA-70B	r/economy	360	TextBlob	r/wallstreetbets	360

Notes: "Subreddit" indicates the subreddit used to construct the indicator; "MA" is the backward-looking moving-average window (days). Best is defined by the lowest MAE ratio relative to AR(1) within each family. In the lexicon/sentiment family, "TextBlob" corresponds to entries of the form `sentiment_subreddit_MA`, "VADER" to `sentiment_vader_subreddit_MA`, and "LM" to `sentiment_lm_subreddit_count_MA`.

Table A.9: Forecast results (MAD ratios) for *CPI*.

Model	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=9$	$h=12$	$h=18$
Expectations									
Best expectations	1.032	0.951	0.876	0.814	0.777	0.762	0.700	*0.658*	**0.598**
1-Y Inflation Swap (best)	1.011	0.929	0.896	0.875	0.857	0.823	0.767	0.783	0.732
Reddit-sentiment									
Best sentiment	0.982	0.927	0.899	0.849	*0.796*	*0.753*	**0.665**	**0.647**	*0.600*
Reddit-LLM									
LLM forecast aggregation	0.893	*0.839*	*0.817*	*0.792*	0.773	0.761	0.766	0.771	0.750
Best fine-tuned LLM	*0.891*	0.856	0.845	0.835	0.837	0.838	0.843	0.848	**0.695**
Llama70B forecast aggregation	**0.799**	**0.727**	**0.713**	**0.688**	**0.675**	**0.671**	*0.681*	0.696	*0.699*
Best Llama70B	0.880	0.852	0.860	0.850	0.835	0.811	0.809	0.808	0.703

Notes: MAD is computed as a ratio: $MAD(AR\text{-}X(1))/MAD(AR(1))$, where the benchmark is $AR(1)$. The initial in-sample period spans from 2009M1 to 2012M12; the out-of-sample period covers 2013M1 to 2025M8. Forecasts are generated using a recursive window, with the model re-estimated each step (month). For *Best sentiment*, the aggregated sentiment model is excluded. Bold (italic) indicates the best (second-best) model for each horizon.

Table A.10: Forecast results (MAD ratios) for *PCE*.

Model	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=9$	$h=12$	$h=18$
Expectations									
Best expectations	1.014	0.958	0.903	0.861	0.822	0.807	*0.742*	0.703	0.649
1-Y Inflation Swap (best)	1.002	0.974	0.942	0.924	0.900	0.875	0.834	0.853	0.794
Reddit-sentiment									
Best sentiment	0.985	0.970	0.918	0.862	0.788	**0.722**	**0.573**	**0.471**	**0.395**
Reddit-LLM									
LLM forecast aggregation	*0.899*	*0.846*	*0.819*	*0.794*	*0.782*	0.780	0.755	0.751	0.659
Best fine-tuned LLM	0.904	0.870	0.855	0.850	0.836	0.831	0.789	0.718	*0.571*
Llama70B forecast aggregation	**0.880**	**0.826**	**0.796**	**0.776**	**0.762**	*0.767*	0.770	0.752	0.682
Best Llama70B	0.953	0.916	0.864	0.835	0.810	0.776	0.743	*0.697*	0.627

Notes: MAD is computed as a ratio: $MAD(AR\text{-}X(1))/MAD(AR(1))$, where the benchmark is $AR(1)$. The initial in-sample period spans from 2009M1 to 2012M12; the out-of-sample period covers 2013M1 to 2025M8. Forecasts are generated using a recursive window, with the model re-estimated each step (month). Bold (italic) indicates the best (second-best) model for each horizon.

Table A.11: Best forecasting models by horizon (CPI vs PCE) within each family, based on MAD ratios.

	Fine-tuned LLMs			LLaMA-70B (not fine-tuned)			Lexicon / sentiment		
h	Model	Subreddit	MA	Model	Subreddit	MA	Model	Subreddit	MA
CPI									
1	InflaBERT	r/Economics	1	LLaMA-70B	r/economy	1	VADER	r/wallstreetbets	360
2	Aggregated pred	–	–	LLaMA-70B	r/economy	30	TextBlob	r/economy	10
3	Aggregated pred	–	–	LLaMA-70B	r/economy	30	TextBlob	r/economy	10
4	Aggregated pred	–	–	LLaMA-70B	r/economy	5	TextBlob	r/economy	5
5	Aggregated pred	–	–	LLaMA-70B	r/economy	90	LM	r/Economics	90
6	Aggregated pred	–	–	LLaMA-70B	r/economy	90	LM	r/Economics	90
9	Aggregated pred	–	–	LLaMA-70B	r/economy	120	LM	r/Economics	180
12	Aggregated pred	–	–	LLaMA-70B	r/economy	360	LM	r/Economics	180
18	Qwen-7B	r/wallstreetbets	360	Aggregated pred	–	–	TextBlob	r/economy	180
PCE									
1	InflaBERT	r/Economics	90	LLaMA-70B	r/economy	30	VADER	r/Economics	30
2	InflaBERT	r/Economics	180	LLaMA-70B	r/wallstreetbets	180	TextBlob	r/wallstreetbets	1
3	Qwen-7B	r/economy	90	LLaMA-70B	r/wallstreetbets	180	TextBlob	r/wallstreetbets	10
4	Qwen-7B	r/economy	180	LLaMA-70B	r/wallstreetbets	180	TextBlob	r/wallstreetbets	180
5	InflaBERT	r/wallstreetbets	180	LLaMA-70B	r/wallstreetbets	360	TextBlob	r/wallstreetbets	360
6	InflaBERT	r/wallstreetbets	180	LLaMA-70B	r/wallstreetbets	360	TextBlob	r/wallstreetbets	360
9	InflaBERT	r/wallstreetbets	360	LLaMA-70B	r/wallstreetbets	360	TextBlob	r/wallstreetbets	360
12	InflaBERT	r/wallstreetbets	360	LLaMA-70B	r/wallstreetbets	360	TextBlob	r/wallstreetbets	360
18	Qwen-7B	r/wallstreetbets	360	LLaMA-70B	r/economy	360	TextBlob	r/wallstreetbets	360

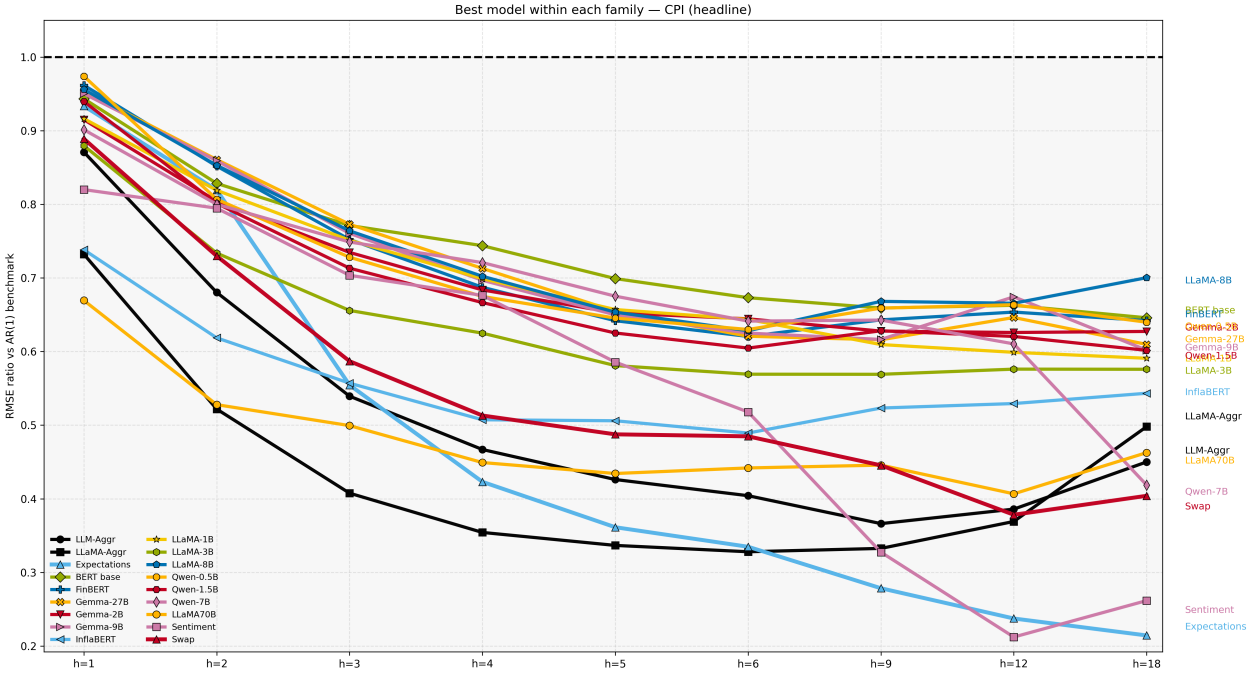
Notes: "Subreddit" indicates the subreddit used to construct the indicator; "MA" is the backward-looking moving-average window (days). Best is defined by the lowest MAD ratio relative to AR(1) within each family. "Aggregated pred" denotes the aggregated forecast model; for these rows, subreddit and MA are not applicable. In the lexicon/sentiment family, "TextBlob" corresponds to entries of the form `sentiment_subreddit_MA`, "VADER" to `sentiment_vader_subreddit_MA`, and "LM" to `sentiment_lm_subreddit_count_MA`. The aggregated sentiment model is excluded.

Table A.12: Best forecasting models by horizon (CPI vs PCE) within each family, based on RMSE ratios.

	Fine-tuned LLMs			LLaMA-70B (not fine-tuned)			Lexicon / sentiment		
h	Model	Subreddit	MA	Model	Subreddit	MA	Model	Subreddit	MA
CPI									
1	InflaBERT	r/Economics	10	LLaMA-70B	r/Economics	1	TextBlob	r/economy	10
2	InflaBERT	r/Economics	30	LLaMA-70B	r/Economics	30	TextBlob	r/economy	10
3	InflaBERT	r/Economics	30	LLaMA-70B	r/Economics	30	TextBlob	r/economy	90
4	InflaBERT	r/Economics	30	LLaMA-70B	r/Economics	90	TextBlob	r/economy	90
5	InflaBERT	r/Economics	30	LLaMA-70B	r/Economics	90	TextBlob	r/economy	120
6	InflaBERT	r/Economics	90	LLaMA-70B	r/Economics	120	TextBlob	r/economy	120
9	InflaBERT	r/Economics	120	LLaMA-70B	r/Economics	180	LM	r/Economics	180
12	InflaBERT	r/Economics	360	LLaMA-70B	r/Economics	360	TextBlob	r/economy	360
18	Qwen-7B	r/wallstreetbets	360	LLaMA-70B	r/economy	360	TextBlob	r/economy	180
PCE									
1	InflaBERT	r/Economics	1	LLaMA-70B	r/Economics	30	VADER	r/wallstreetbets	360
2	InflaBERT	r/Economics	30	LLaMA-70B	r/Economics	90	TextBlob	r/wallstreetbets	180
3	InflaBERT	r/Economics	90	LLaMA-70B	r/wallstreetbets	180	TextBlob	r/wallstreetbets	180
4	InflaBERT	r/Economics	120	LLaMA-70B	r/wallstreetbets	180	TextBlob	r/wallstreetbets	180
5	InflaBERT	r/Economics	360	LLaMA-70B	r/wallstreetbets	360	TextBlob	r/wallstreetbets	360
6	InflaBERT	r/Economics	360	LLaMA-70B	r/wallstreetbets	360	TextBlob	r/wallstreetbets	360
9	InflaBERT	r/Economics	360	LLaMA-70B	r/wallstreetbets	360	TextBlob	r/wallstreetbets	360
12	InflaBERT	r/wallstreetbets	360	LLaMA-70B	r/wallstreetbets	360	TextBlob	r/wallstreetbets	360
18	Qwen-7B	r/wallstreetbets	360	LLaMA-70B	r/wallstreetbets	360	TextBlob	r/wallstreetbets	360

Notes: "Subreddit" indicates the subreddit used to construct the indicator; "MA" is the backward-looking moving-average window (days). Best is defined by the lowest RMSE ratio relative to AR(1) within each family.

Figure A.14: RMSE ratios for best models by forecast horizon for CPI



Notes:

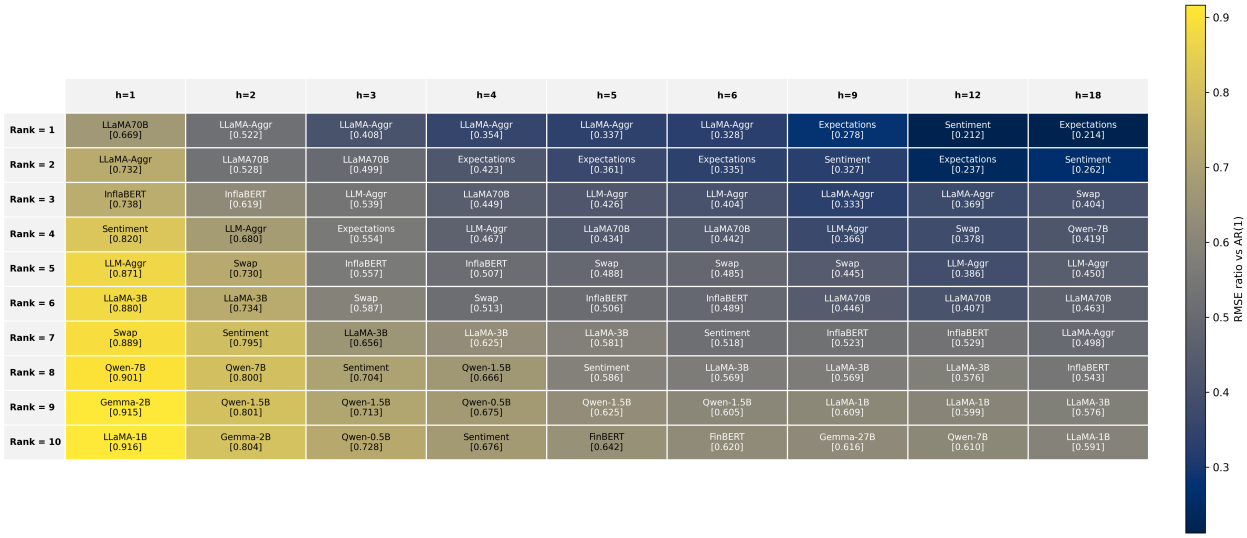
E.2 Point forecasts - Pre-COVID vs Full Out-of-sample

E.3 Model Confidence Set (MCS)

In this section we comment on the results of the Model Confidence Set by Hansen et al. (2011). The Model Confidence Set (MCS) procedure of Hansen et al. (2011) provides a data-driven way to identify a subset of *superior* forecasting models. The key idea is to run a sequence of hypothesis tests for *equal predictive ability* (EPA) across a set of competing models under a user-chosen loss function. Starting from an initial collection of models, the algorithm sequentially removes the worst-performing specification until the null of equal predictive ability can no longer be rejected at a pre-specified confidence level.

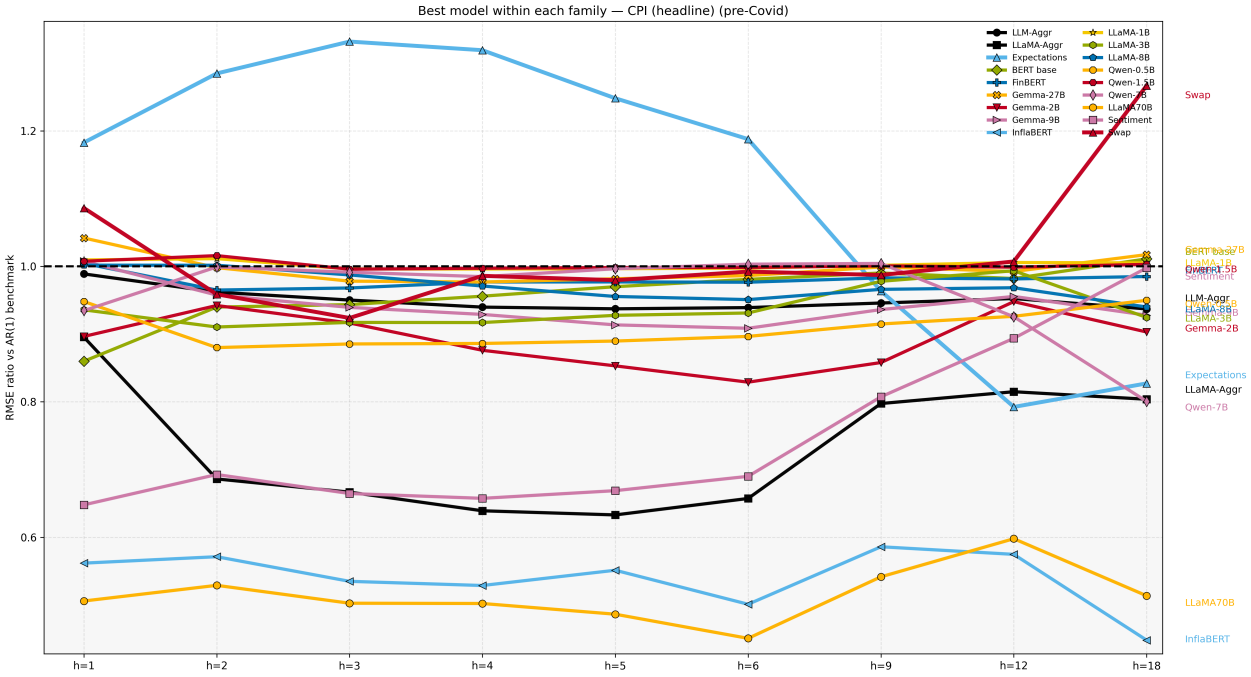
Let $\mathcal{M}_0$ denote the initial model set with cardinality $|\mathcal{M}_0| = M$. For a confidence level $1-\alpha$, the MCS procedure delivers a (possibly smaller) set $\mathcal{M}^*_{1-\alpha} \subseteq \mathcal{M}_0$ with size $M^* \leq M$ such that the EPA hypothesis is *not rejected* for all models in $\mathcal{M}^*_{1-\alpha}$. The most informative

Figure A.15: Best models by forecast horizon for CPI



Notes:

Figure A.16: RMSE ratios for best models by forecast horizon for CPI - (pre-Covid)



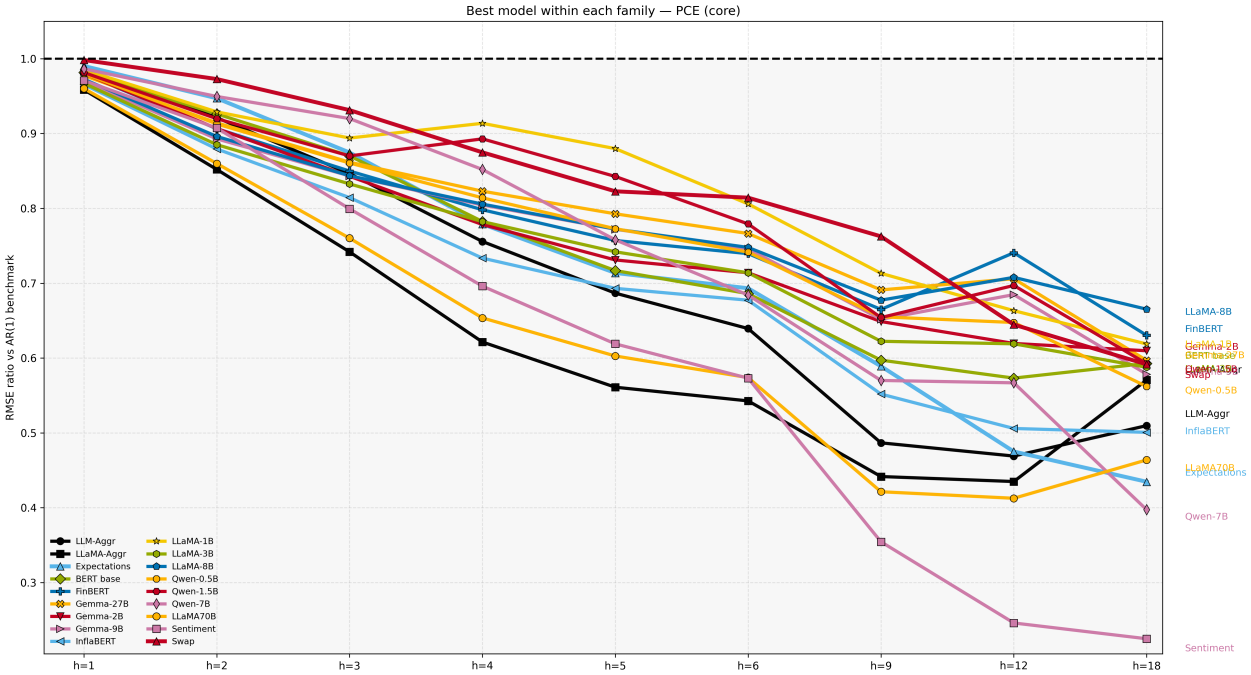
Notes:

Figure A.17: Best models by forecast horizon for CPI - (pre-Covid)



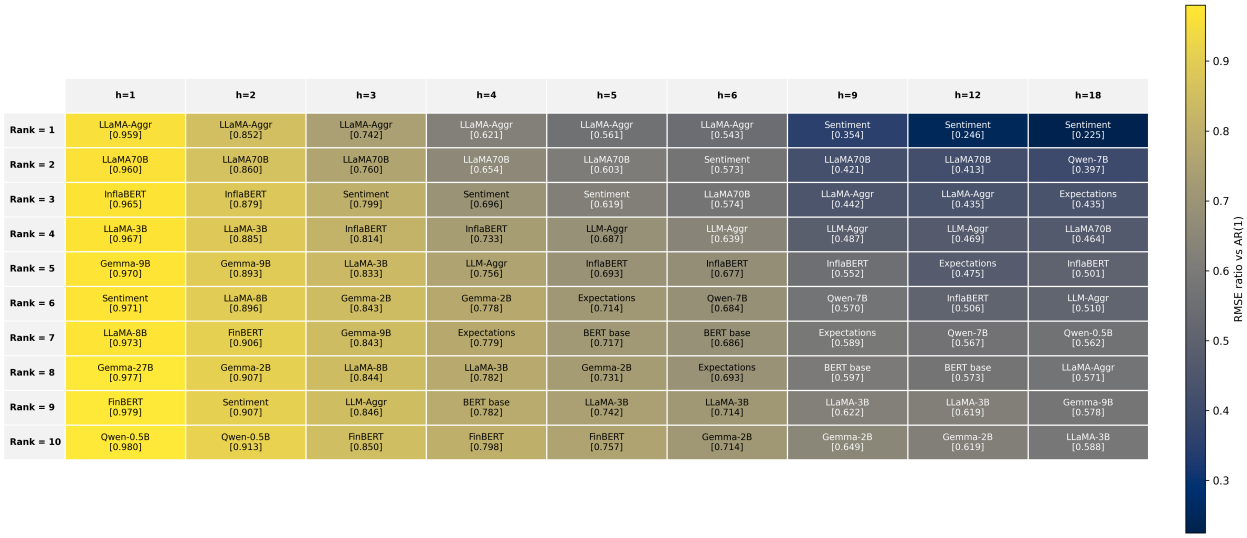
Notes:

Figure A.18: RMSE ratios for best models by forecast horizon for PCE



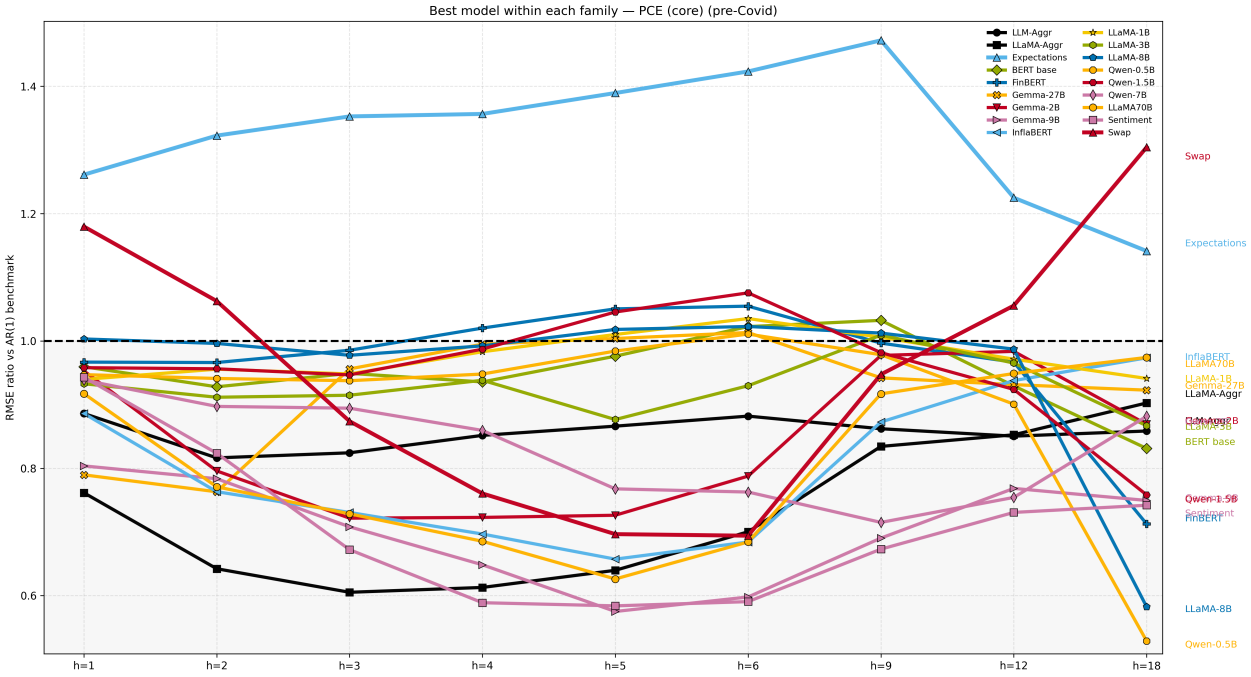
Notes:

Figure A.19: Best models by forecast horizon for CPI



Notes:

Figure A.20: RMSE ratios for best models by forecast horizon for CPI - (pre-Covid)



Notes:

Figure A.21: Best models by forecast horizon for CPI - (pre-Covid)



Notes:

Figure A.22: CSSED plots for CPI (headline) at different forecast horizons.

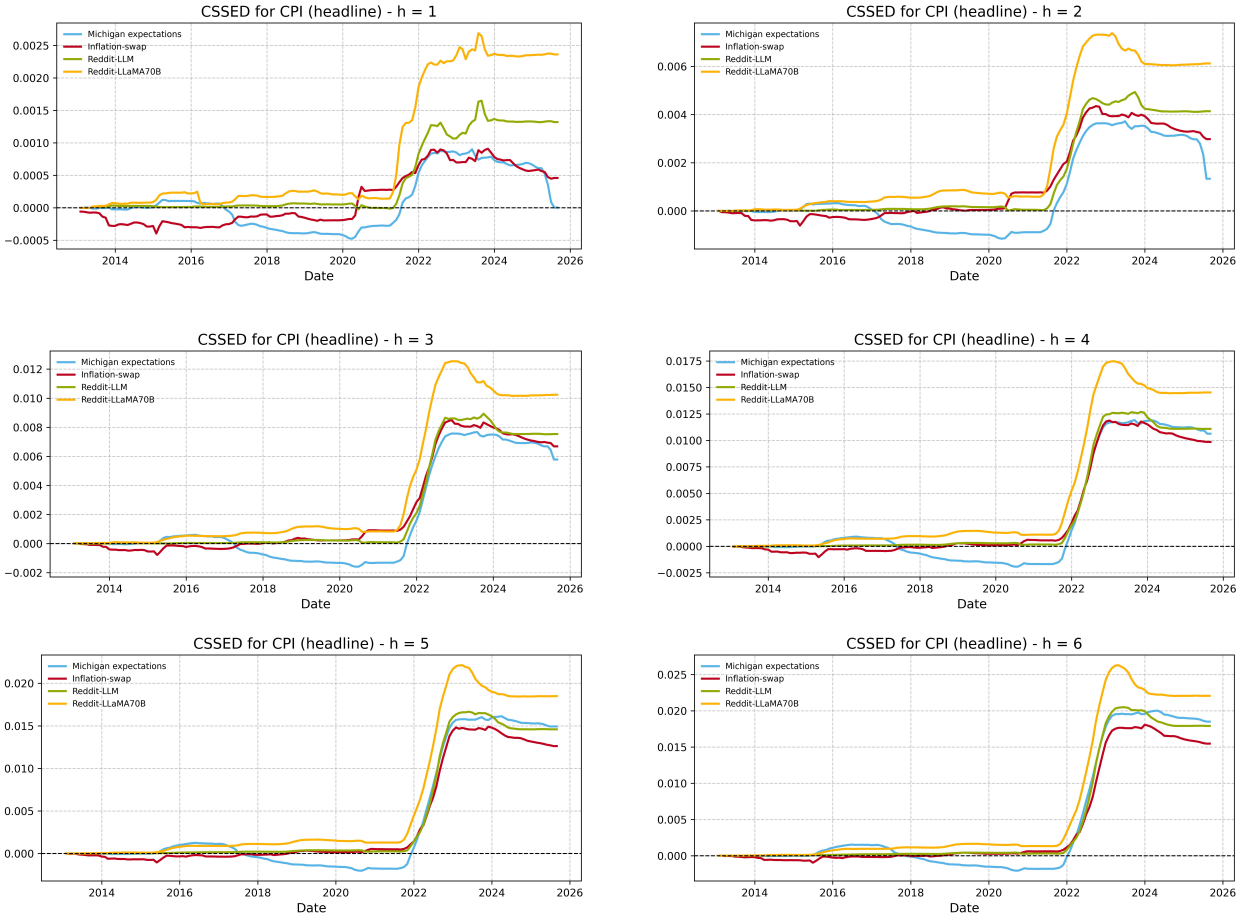
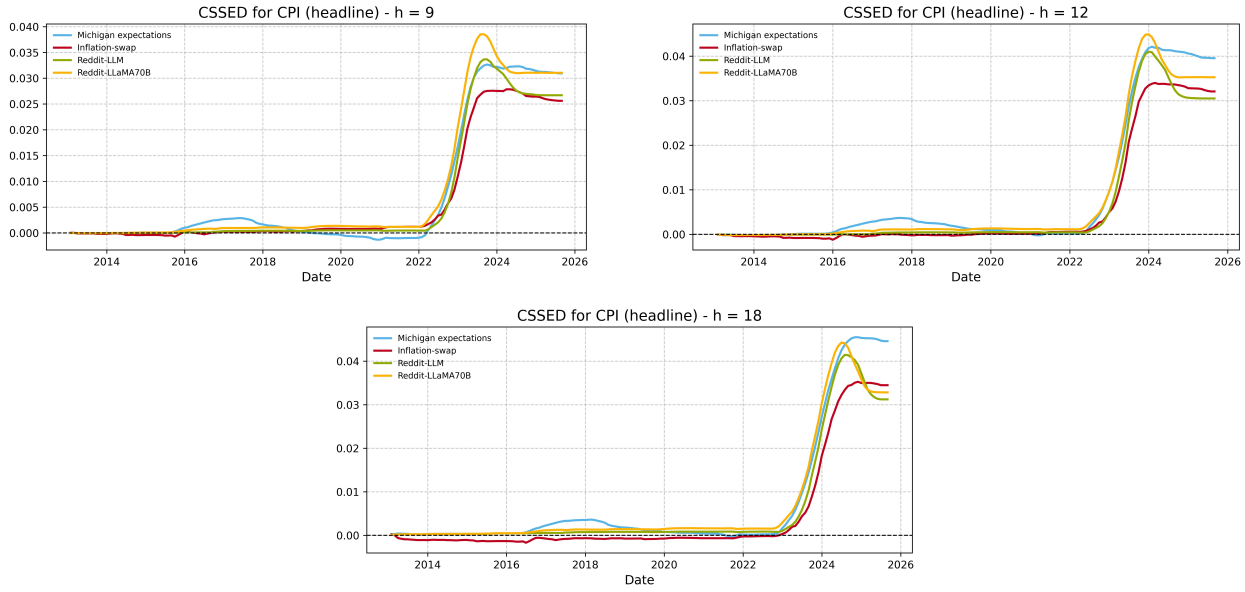


Figure A.22: CSSED plots for CPI (headline) at different forecast horizons (cont.).



Notes: This figure reports the cumulative sum of squared forecast error differences (CSSED) for CPI (headline) forecasts across alternative horizons. Negative values indicate that the alternative model outperforms the autoregressive benchmark.

Figure A.23: CSSED plots for PCE (core) at different forecast horizons.

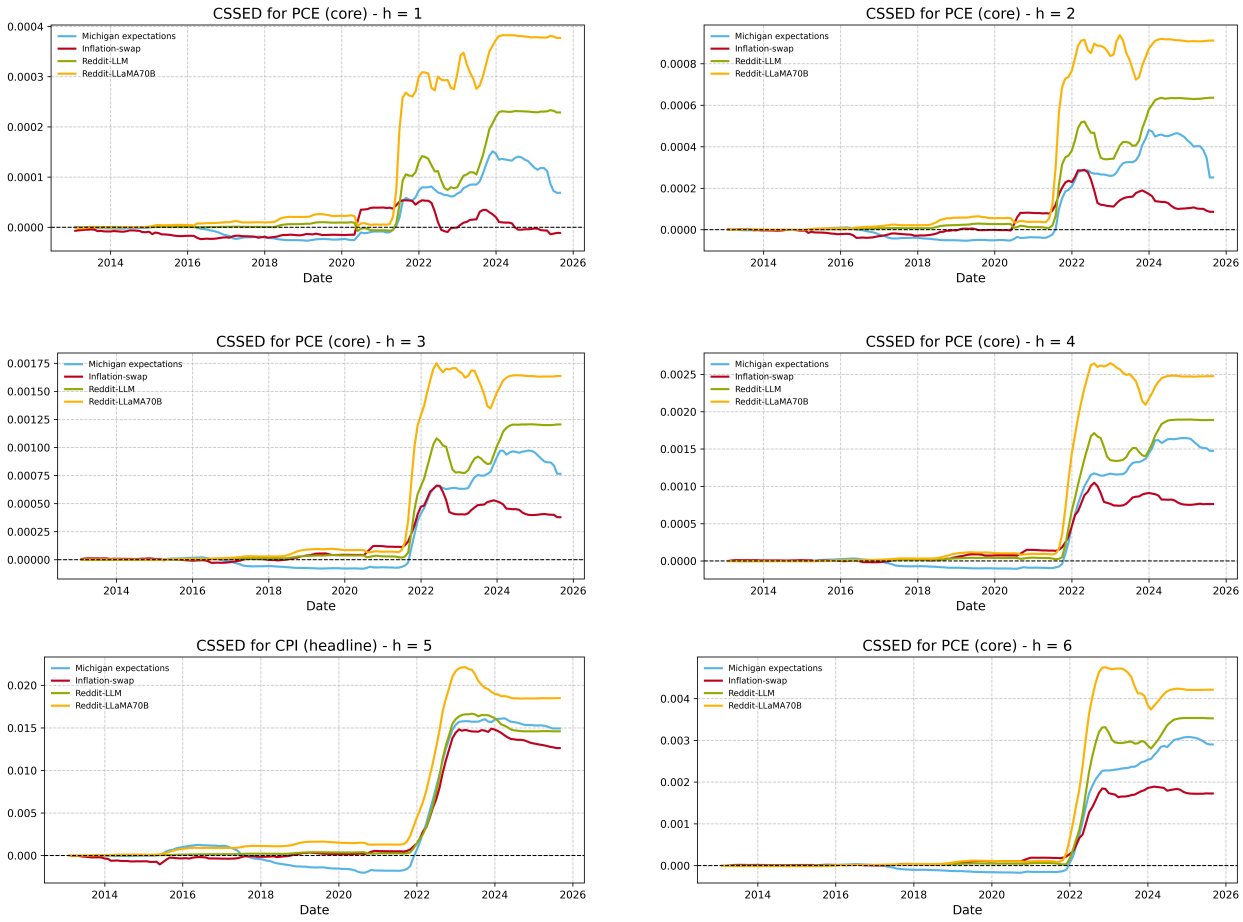
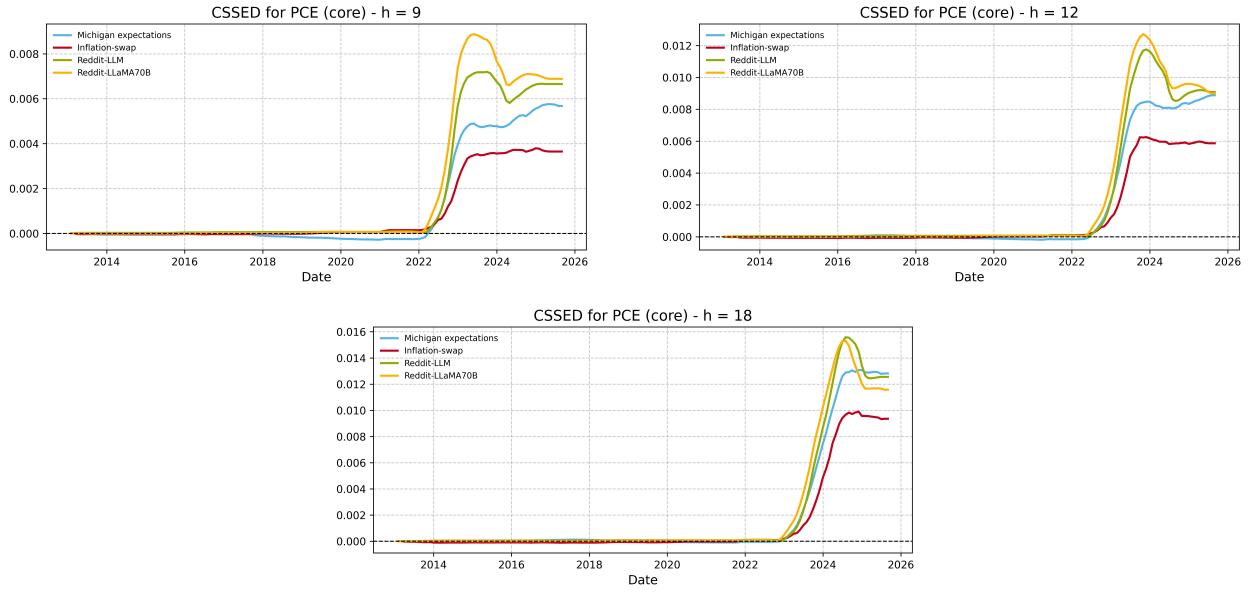


Figure A.23: CSSED plots for PCE (core) at different forecast horizons (cont.).



Notes: This figure reports the cumulative sum of squared forecast error differences (CSSED) for PCE (core) forecasts across alternative horizons. Negative values indicate that the alternative model outperforms the autoregressive benchmark.

Figure A.24: Fluctuation test plots for CPI (headline) at different forecast horizons.

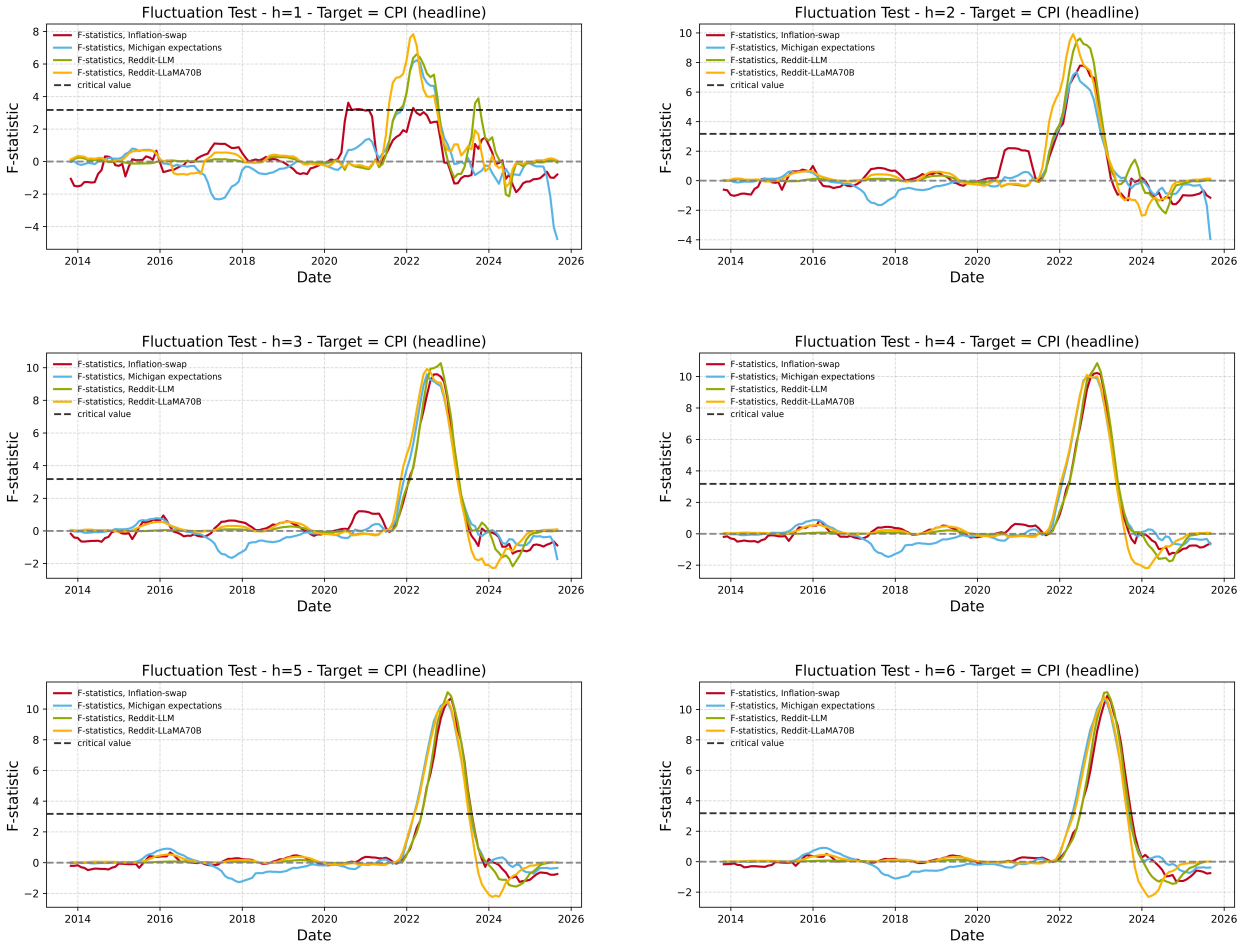
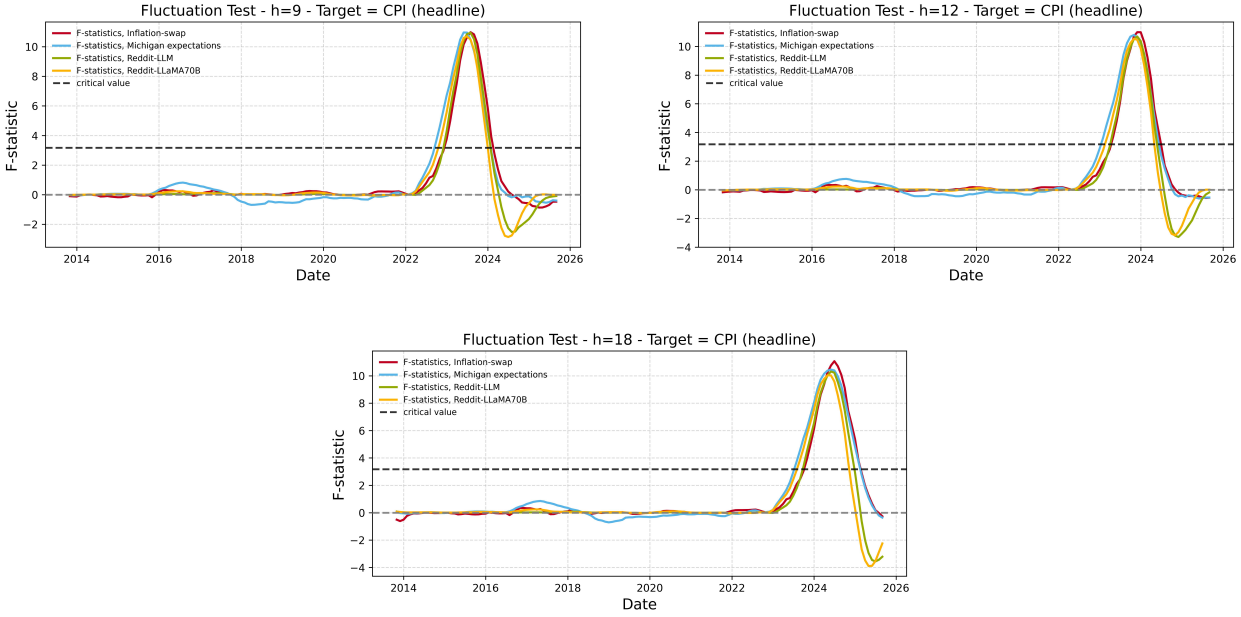


Figure A.24: Fluctuation test plots for CPI (headline) at different forecast horizons (cont.).

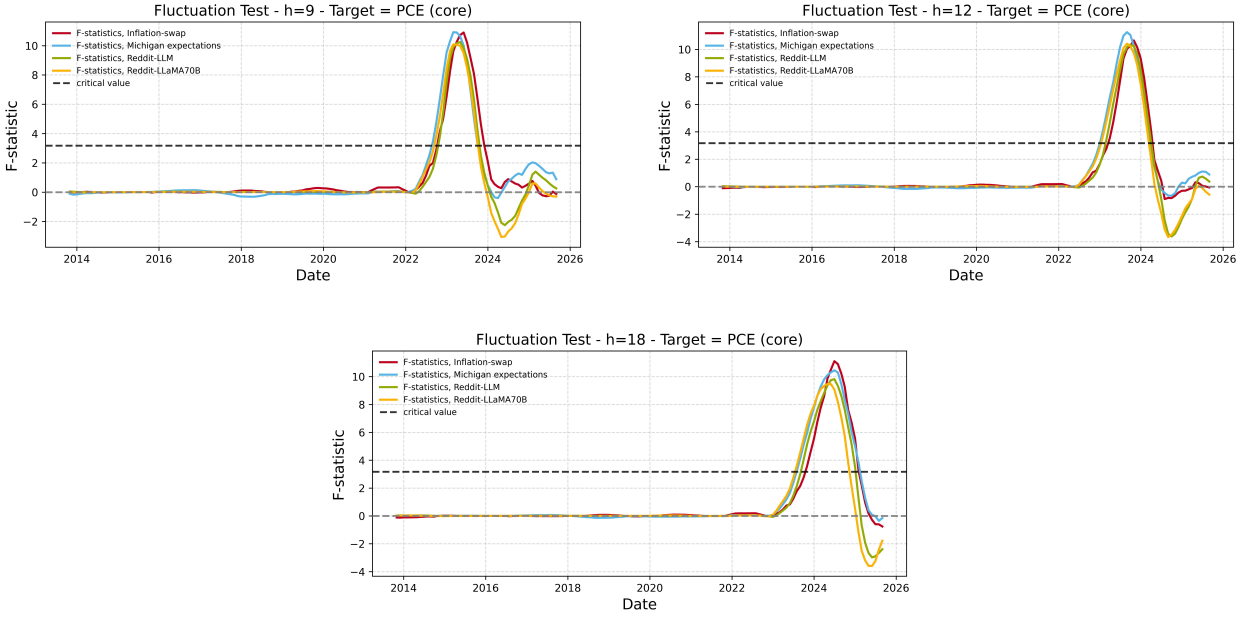


Notes: This figure reports fluctuation test plots for CPI (headline) forecasts across alternative horizons.

Figure A.25: Fluctuation test plots for PCE (core) at different forecast horizons.



Figure A.25: Fluctuation test plots for PCE (core) at different forecast horizons (cont.).



Notes: This figure reports fluctuation test plots for PCE (core) forecasts across alternative horizons.

case is when the final set contains a single element, i.e. $M^* = 1$.

Loss differentials. Let $L(m, t+h)$ be the loss incurred by model m when forecasting at horizon h for time $t+h$, where $m \in \{1, \ldots, M\}$ and $t \in \{R, \ldots, T-h\}$. Define the loss differential between models m and n as

$$d_{mn,t+h} \;=\; L(m, t+h) - L(n, t+h), \qquad m, n = 1, \ldots, M, \;\; t = R, \ldots, T-h. \tag{A.3}$$

For each model m, its average loss relative to the other models in a set $\mathcal{M}$ is

$$d_{m\cdot,t+h} \;=\; \frac{1}{|\mathcal{M}| - 1} \sum_{n \in \mathcal{M}} d_{mn,t+h}, \qquad m \in \mathcal{M}. \tag{A.4}$$

EPA hypotheses. For a given set $\mathcal{M}$, the EPA null can be expressed in two equivalent ways. The pairwise formulation is

$$H_{0,\mathcal{M}} : \; c_{mn} = 0, \quad \forall\, m, n \in \mathcal{M}, \tag{A.5}$$

$$H_{A,\mathcal{M}} : \; c_{mn} \neq 0, \quad \text{for some } m, n \in \mathcal{M}, \tag{A.6}$$

while the "against-the-average" formulation is

$$H_{0,\mathcal{M}} : \; c_{m\cdot} = 0, \quad \forall\, m \in \mathcal{M}, \tag{A.7}$$

$$H_{A,\mathcal{M}} : \; c_{m\cdot} \neq 0, \quad \text{for some } m \in \mathcal{M}, \tag{A.8}$$

where

$$c_{mn} = \mathbb{E}(d_{mn,t+h}), \qquad c_{m\cdot} = \mathbb{E}(d_{m\cdot,t+h}).$$

Studentized statistics. Let P_h denote the number of available forecast evaluation observations at horizon h (e.g. $P_h = T - h - R + 1$). Define the sample mean loss differentials

$$\bar{d}_{mn} \;:=\; \frac{1}{P_h} \sum_{t=R}^{T-h} d_{mn,t+h}, \qquad \bar{d}_{m\cdot} \;:=\; \frac{1}{|\mathcal{M}| - 1} \sum_{n \in \mathcal{M}} \bar{d}_{mn}. \tag{A.9}$$

The corresponding studentized statistics are

$$t_{mn} \;=\; \frac{\bar{d}_{mn}}{\widehat{\sigma}(\bar{d}_{mn})}, \qquad t_{m\cdot} \;=\; \frac{\bar{d}_{m\cdot}}{\widehat{\sigma}(\bar{d}_{m\cdot})}, \tag{A.10}$$

where $\widehat{\sigma}(\bar{d}_{mn})$ and $\widehat{\sigma}(\bar{d}_{m\cdot})$ are standard error estimates, typically obtained via a block bootstrap to accommodate serial dependence in the loss differentials.

Test statistics. The pairwise EPA hypothesis in (A.5) naturally leads to the range statistic

$$T_{R,\mathcal{M}} \;=\; \max_{m,n\in\mathcal{M}} |t_{mn}|, \tag{A.11}$$

while the "against-the-average" EPA hypothesis in (A.7) maps to

$$T_{\max,\mathcal{M}} \;=\; \max_{m\in\mathcal{M}} t_{m\cdot}. \tag{A.12}$$

Sequential elimination rule. The MCS is constructed through sequential testing: at each iteration, test EPA on the current set $\mathcal{M}$; if EPA is rejected, remove the worst model according to an elimination rule and repeat. Under the range statistic in (A.11), the eliminated model is

$$e_{R,\mathcal{M}} \;=\; \arg\max_{m\in\mathcal{M}} \left(\sup_{n\in\mathcal{M}} t_{mn} \right), \tag{A.13}$$

whereas under the max statistic in (A.12) one removes

$$e_{\max,\mathcal{M}} \;=\; \arg\max_{m\in\mathcal{M}} t_{m\cdot}. \tag{A.14}$$

Algorithm (MCS procedure).

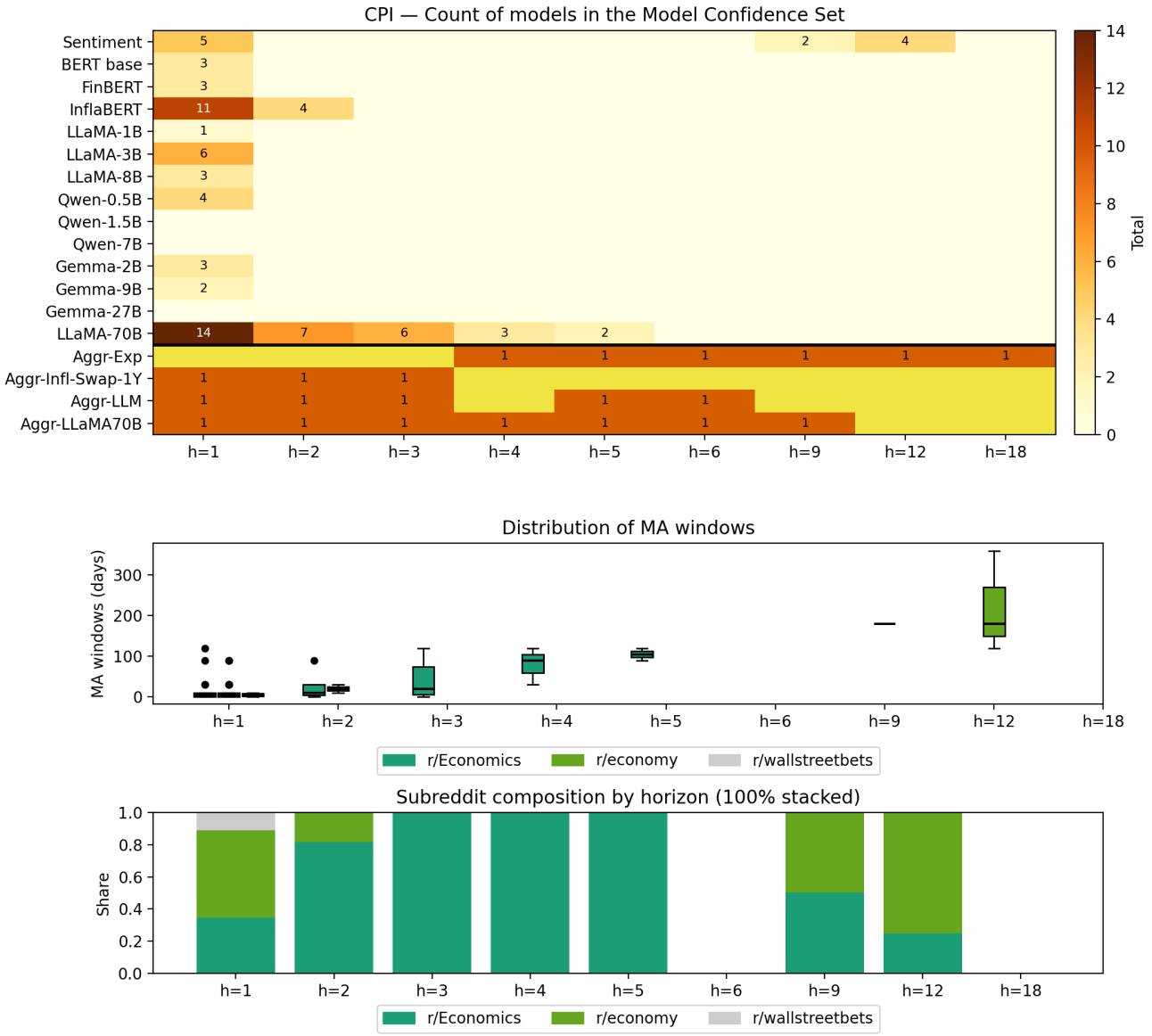
1. Initialize $\mathcal{M} \leftarrow \mathcal{M}_0$.
2. Test the EPA null $H_{0,\mathcal{M}}$ using either $T_{R,\mathcal{M}}$ in (A.11) or $T_{\max,\mathcal{M}}$ in (A.12). If EPA is *not* rejected at level α, stop and set $\mathcal{M}^*_{1-\alpha} = \mathcal{M}$.
3. If EPA is rejected, determine the worst model using (A.13) or (A.14), remove it from the set, $\mathcal{M} \leftarrow \mathcal{M} \setminus \{e_{\cdot,\mathcal{M}}\}$, and return to step 2.

MCS empirical results for CPI and PCE Figure A.26 depicts the MCS results for our forecasting exercise, where the target variable is CPI. The top chart shows the heatmap related to the number of Reddit-and-LLM-based indicators across the forecast horizons from one to 18 months ahead. For each forecast horizon we report the number of forecasting models that enter in the MCS at each forecast horizon. At 1-month ahead, for example, the Reddit-based indicators that enter more often in the MCS are those built using an unfine-tuned LLAMA-70B (14 models), a fine-tuned InflaBERT (11 models), a fine-tuned LLaMA-3B (6 models), a standard lexicon (5 models) or a fine-tuned Qwen-0.5B (4 models). We have to add that the models using the forecast aggregation built using only LLaMA-70B-based indicators (Aggr-LLaMA-70B) or other smaller but fine-tuned LLMs (Aggr-LLM) enter almost always in the final MCS except for horizons greater than nine months. This means that the forecasts built using aggregations of fine-tuned small LMMs or the unfine-tuned LLaMA-70B are superior to both those obtained by aggregating the forecasts of Michigan expectations (which work fine only at forecast horizons from four to 18 months) and those obtained by aggregating the signal from the 1-year inflation swaps (which enter the MCS only for forecast horizons up to three months). At nine and 12 months ahead, a few models using lexicon-based indicators enter the MCS. The middle chart of Figure A.26 shows the distribution of MA windows used across subreddits. At one-month horizon the models falling into the MCS use all the three subreddits with MA windows up to 100 days. For longer forecast horizons up to nine months, they use only `r/Economics` with MA below 200 days. At 12-months-ahead they use mainly `r/economy` and MA windows greater than 100 days. The bottom chart of Figure A.26 shows for each forecast horizon which subreddit contributes more to the inclusion in the MCS set. At one month ahead, all three subreddits contribute to the inclusion in the MCS, while at longer horizons, only `r/Economics` and `r/economy` give indicators that help the forecasting model to enter in the MCS.

Figure A.27 shows the MCS results for our forecasting exercise, where the target variable is PCE. The top chart shows the heatmap related to the number of Reddit-and-LLM-based indicators across the forecast horizons from one to 18 months ahead. For each forecast horizon we report the number of forecasting models that enter in the MCS at each forecast horizon. At 1-month ahead, for example, the Reddit-based indicators that enter more often in the MCS are those built using an unfine-tuned LLAMA-70B (14 models), a fine-tuned InflaBERT (11

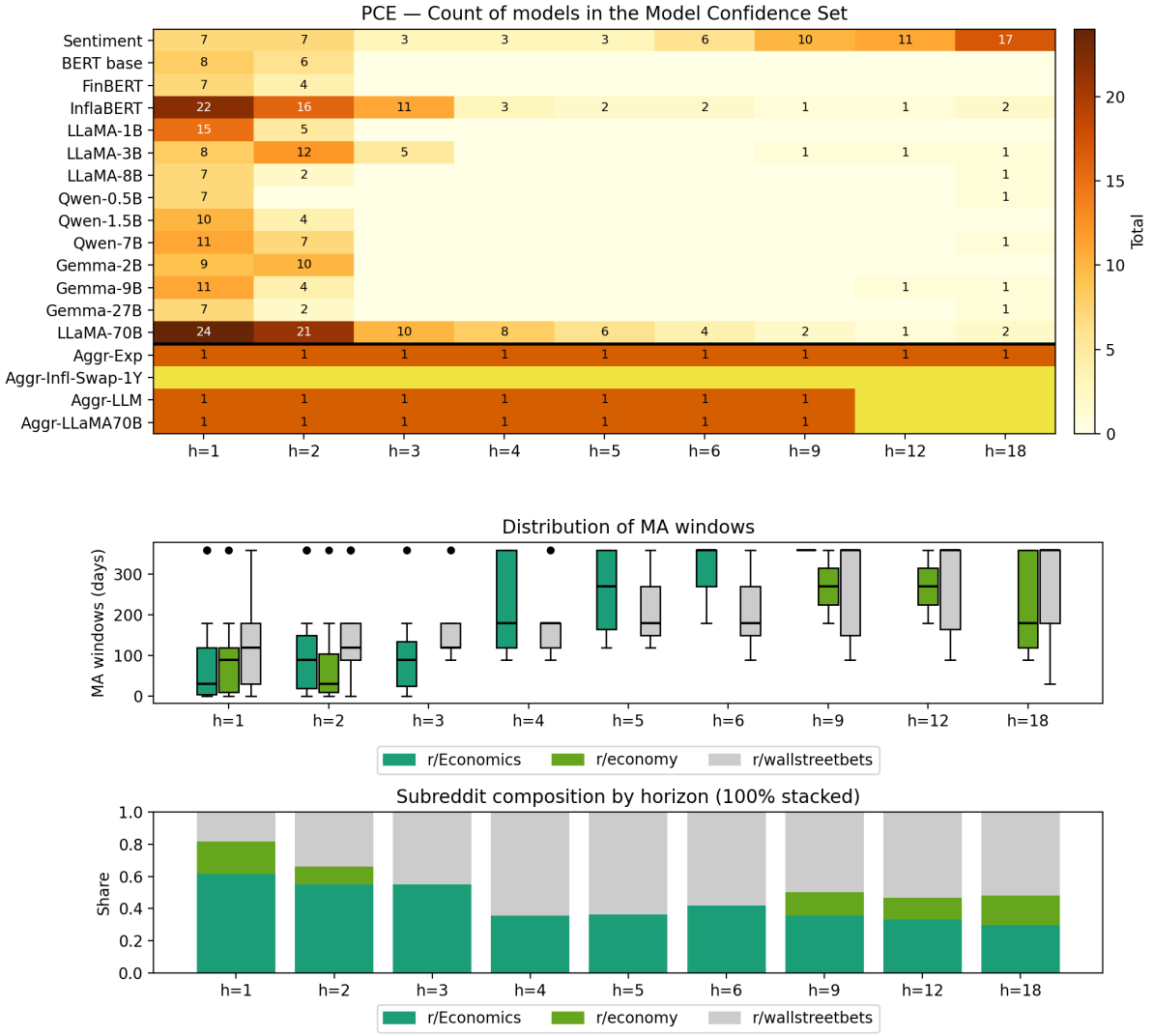
models), a fine-tuned LLaMA-3B (6 models), a standard lexicon (5 models) or a fine-tuned Qwen-0.5B (4 models). We have to add that the models using the forecast aggregation built using only LLaMA-70B-based indicators (Aggr-LLaMA-70B) or other smaller but fine-tuned LLMs (Aggr-LLM) enter almost always in the final MCS except for horizons greater than nine months. This means that the forecasts built using aggregations of fine-tuned small LMMs or the unfine-tuned LLaMA-70B are superior to both those obtained by aggregating the forecasts of Michigan expectations (which work fine only at forecast horizons from four to 18 months) and those obtained by aggregating the signal from the 1-year inflation swaps (which enter the MCS only for forecast horizons up to three months). At nine and 12 months ahead, a few models using lexicon-based indicators enter the MCS. The middle chart of Figure A.26 shows the distribution of MA windows used across subreddits. At one-month horizon the models falling into the MCS use all the three subreddits with MA windows up to 100 days. For longer forecast horizons up to nine months, they use only `r/Economics` with MA below 200 days. At 12-months-ahead they use mainly `r/economy` and MA windows greater than 100 days. The bottom chart of Figure A.26 shows for each forecast horizon which subreddit contributes more to the inclusion in the MCS set. At one month ahead, all three subreddits contribute to the inclusion in the MCS, while at longer horizons, only `r/Economics` and `r/economy` give indicators that help the forecasting model to enter in the MCS.

Figure A.26: Results for MCS test for CPI ($h = 1, \ldots, 18$)



Notes: The top chart shows the heatmap of the number of models that belong to the final MCS for CPI for each forecast horizon from one to 18-months ahead (here, the total model for each Reddit-based indicator computed with a single LLM or an aggregation). The middle chart shows the boxplots for each subreddit and forecast horizon of the MA windows used in the best models selected in the final MCS. The bottom chart shows the share of the best models using each subreddit by forecast horizon. Number of bootstrap samples $B = 1,000$.

Figure A.27: Results for MCS test for PCE ($h = 1, \ldots, 18$)



Notes: The top chart shows the heatmap of the number of models that belong to the final MCS for PCE for each forecast horizon from one to 18-months ahead (here, the total model for each Reddit-based indicator computed with a single LLM or an aggregation). The middle chart shows the boxplots for each subreddit and forecast horizon of the MA windows used in the best models selected in the final MCS. The bottom chart shows the share of the best models using each subreddit by forecast horizon. Number of bootstrap samples $B = 1,000$.

E.4 Predictive Regressions

This section reports an *in-sample* predictive-regression exercise designed to quantify the incremental explanatory power of Reddit-based indicators when the full sample is used for estimation. This analysis complements the pseudo out-of-sample forecasting and nowcasting results in the main text, which are constructed to mimic real-time information flow. Here, the objective is narrower: to measure whether Reddit–LLM indicators add statistically significant predictive content for inflation beyond conventional expectation measures and lagged inflation.

Specification. For each horizon h, we estimate the following OLS regression:

$$\pi_{t+h} \;=\; \alpha \;+\; \beta_1 \pi_t \;+\; \beta_2 \pi_t^e \;+\; \beta_3 \text{1YSwap}_t \;+\; \beta_4 X_t^R \;+\; \varepsilon_t, \tag{A.15}$$

where π_t denotes inflation, π_t^e is survey-based inflation expectations (Michigan), 1YSwap_t is the one-year inflation swap rate, and X_t^R is a Reddit-based indicator constructed from LLM classifications. The regression is estimated separately for each Reddit indicator (fine-tuned LLM signals and the LLaMA 70B-based signals), and for each subreddit and moving-average (MA) specification.

Hypothesis and inference. While individual coefficient estimates for X_t^R are often significant, our main focus is whether Reddit indicators improve overall explanatory power. We therefore test the null hypothesis $H_0 : \beta_4 = 0$ using an F-test and report the associated change in adjusted R^2 when X_t^R is added to a baseline specification that includes lagged inflation, Michigan expectations, and the one-year swap. Standard errors are heteroskedasticity- and autocorrelation-robust (HAC).

Samples. For `r/Economics` and `r/economy`, the estimation sample spans 2009M1–2025M8 (193 monthly observations), aligned with the nowcasting sample. For `r/wallstreetbets`, the available history is shorter; we therefore estimate the predictive regressions over 2016M1–

2025M8 (116 observations).[33]

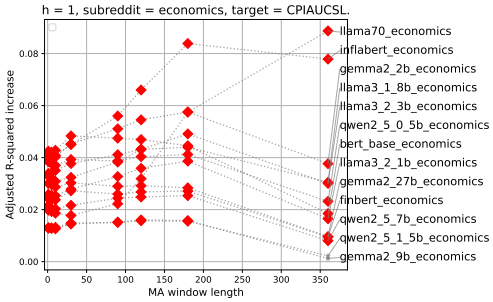
Graphical summary. Figures A.28 and A.29 summarize results for CPI and PCE, respectively, focusing on horizons $h = 1$ and $h = 6$ for readability.[34] In each panel, the vertical axis reports the incremental change in adjusted R^2 from adding X_t^R to the baseline specification, while the horizontal axis indexes the MA window used to smooth the Reddit indicator. Markers indicate whether the F-test rejects $H_0 : \beta_4 = 0$ at the 5% level.

Results. Figures A.28 and A.29 show that Reddit-based indicators provide statistically significant incremental explanatory power across model families, subreddits, and horizons, and for both CPI and PCE. The improvements in fit are broadly robust to the choice of MA window: while the magnitude of the adjusted R^2 gain varies with smoothing, the corresponding F-tests frequently reject $H_0 : \beta_4 = 0$. No single model family or subreddit dominates uniformly, suggesting that the information captured by Reddit narratives is distributed across heterogeneous sources rather than concentrated in a single platform or specification.

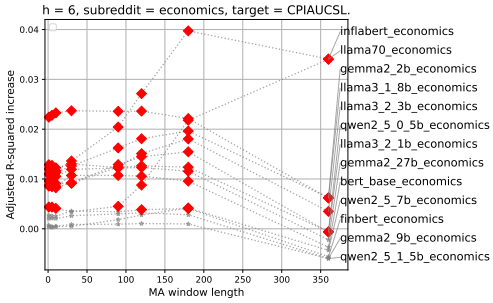
Results for `r/wallstreetbets` should be interpreted with caution given the shorter estimation sample, which reduces statistical power and makes comparisons less precise. Nonetheless, even in this shorter sample the Reddit indicators often remain significant. Overall, the predictive-regression evidence is consistent with the out-of-sample results: Reddit–LLM indicators contain incremental information about inflation dynamics beyond lagged inflation and standard survey- and market-based expectations measures.

[33]The shorter sample reflects the later emergence of the subreddit and implies that results for `r/wallstreetbets` are not directly comparable in precision to those for the other subreddits. This limitation is less consequential in the forecasting/nowcasting applications because the forecast-combination procedure tends to select `r/wallstreetbets` indicators primarily in later subsamples.

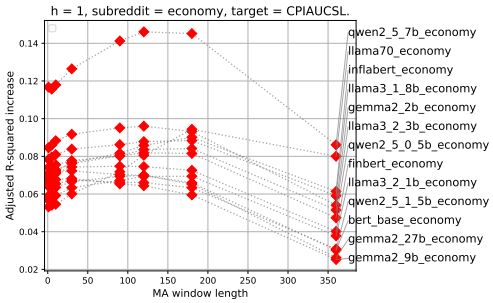
[34]We omit intermediate horizons to preserve clarity; patterns are similar across the remaining horizons.



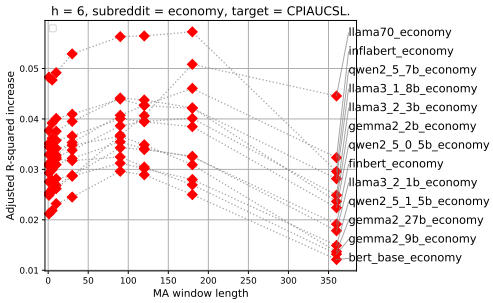
(a) `r/Economics` ($h = 1$), CPI.



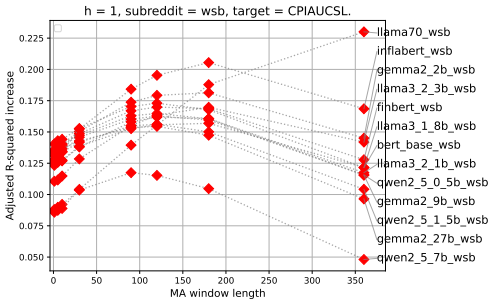
(b) `r/Economics` ($h = 6$), CPI.



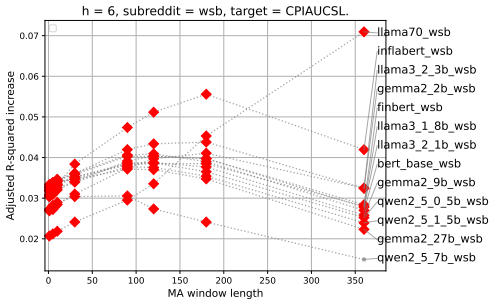
(c) `r/economy` ($h = 1$), CPI.



(d) `r/economy` ($h = 6$), CPI.

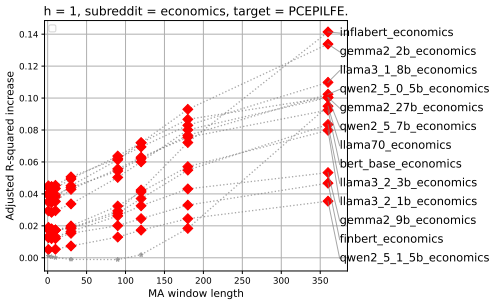


(e) `r/wallstreetbets` ($h = 1$), CPI.

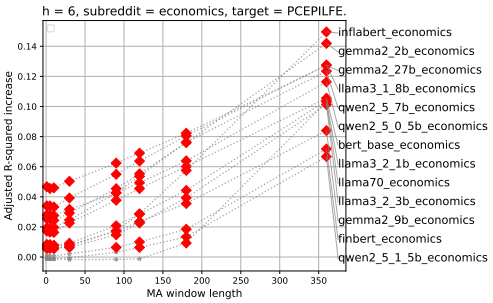


(f) `r/wallstreetbets` ($h = 6$), CPI.

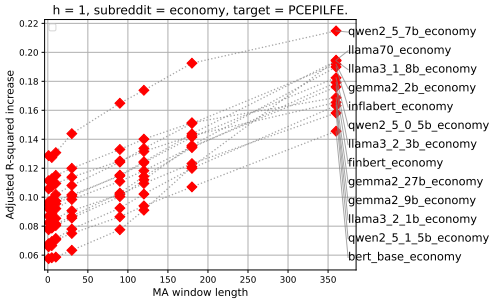
Figure A.28: CPI predictive regressions: incremental adjusted R^2 from adding a Reddit indicator X_t^R to a baseline regression including lagged inflation, Michigan expectations, and the one-year inflation swap rate. Markers denote rejection of $H_0 : \beta_4 = 0$ at the 5% level (HAC inference). Samples: 2009M1–2025M8 for `r/Economics` and `r/economy`; 2016M1–2025M8 for `r/wallstreetbets`.



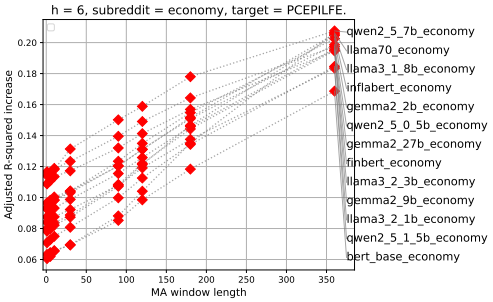
(a) `r/Economics` ($h = 1$), PCE.



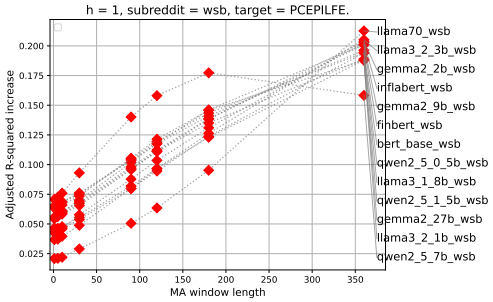
(b) `r/Economics` ($h = 6$), PCE.



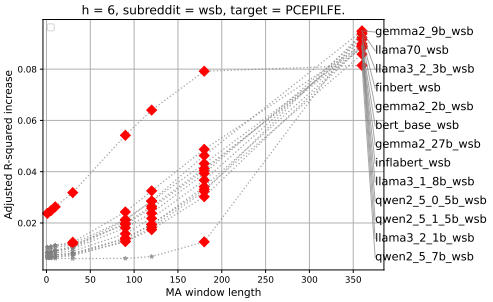
(c) `r/economy` ($h = 1$), PCE.



(d) `r/economy` ($h = 6$), PCE.



(e) `r/wallstreetbets` ($h = 1$), PCE.



(f) `r/wallstreetbets` ($h = 6$), PCE.

Figure A.29: PCE predictive regressions: incremental adjusted R^2 from adding a Reddit indicator X_t^R to a baseline regression including lagged inflation, Michigan expectations, and the one-year inflation swap rate. Markers denote rejection of $H_0 : \beta_4 = 0$ at the 5% level (HAC inference).

E.5 Quantile estimation

This section extends the point-forecast analysis by evaluating *density* forecasts of inflation. Rather than modeling only the conditional mean, we estimate conditional quantiles of inflation using quantile regression and reconstruct the predictive distribution from a grid of quantile forecasts. This approach allows us to assess whether Reddit-based indicators improve forecasts not only on average, but also in the tails of the inflation distribution.

Quantile regression framework. Let $\mathcal{Y}_t$ denote the inflation outcome (CPI or PCE) and let $x_t = [x_{1,t}, \ldots, x_{K,t}]$ be a vector of predictors. For a quantile index $\tau \in (0,1)$, quantile regression models the conditional quantile function as

$$\mathcal{Y}_{t|\tau} = x_t' \boldsymbol{\beta}(\tau) + \epsilon_{t|\tau}, \qquad \tau \in (0,1), \tag{A.16}$$

so that

$$\Pr(\mathcal{Y}_t \leq x_t' \boldsymbol{\beta}(\tau) \mid x_t) = \tau. \tag{A.17}$$

The estimator minimizes the quantile ("check") loss

$$\rho_\tau(u) = \begin{cases} \tau u, & u \geq 0, \\ (\tau - 1)u, & u < 0, \end{cases} \tag{A.18}$$

where $u_t(\tau) = \mathcal{Y}_t - x_t' \boldsymbol{\beta}(\tau)$, yielding

$$\widehat{\boldsymbol{\beta}}(\tau) = \arg\min_{\boldsymbol{\beta}} \frac{1}{T} \sum_{t=1}^{T} \rho_\tau(u_t(\tau)) \,. \tag{A.19}$$

Because the objective is not differentiable at $u = 0$, estimation is performed numerically.[35]

Reconstructing the predictive distribution. Given estimated conditional quantiles $\widehat{q}_{\tau_j}(x_t) = x_t' \widehat{\boldsymbol{\beta}}(\tau_j)$ on a grid $\{\tau_j\}$, we reconstruct an approximate conditional distribution using the nonparametric interpolation approach of Mitchell et al. (2024). Intuitively, the

[35] We implement numerical estimation using a likelihood-based routine for quantile regression.

method interpolates between adjacent quantile forecasts to obtain a smooth approximation to the conditional CDF.

Density forecast evaluation. We evaluate predictive distributions using the Continuous Ranked Probability Score (CRPS),

$$\text{CRPS}(F, y) = \int \left(F(x) - \mathbf{1}_{\{x \geq y\}}\right)^2 dx, \tag{A.20}$$

which measures the distance between the forecast CDF $F(\cdot)$ and the realized outcome y (treated as a degenerate distribution). Lower CRPS values indicate better calibrated and sharper density forecasts.

Implementation choices. Tables A.13 and A.14 report CRPS ratios for CPI and PCE, respectively, relative to the $AR(1)$ benchmark. For comparability with the point-forecast exercise, we use the same set of Reddit-based indicators but focus on signals smoothed with a 90-day backward moving average. For computational reasons, we do not compute forecast combinations for density forecasts.[36] We verified that results are qualitatively similar for alternative MA window lengths.

To facilitate numerical estimation, we adopt a slightly longer in-sample window (and shorter evaluation window) than in the point-forecast exercise: in-sample 2009M1–2018M12 and out-of-sample 2019M1–2025M8. Forecasts are generated with an expanding-window scheme.

CRPS results. The tables show systematic improvements in density forecast accuracy relative to $AR(1)$, particularly at short horizons. For CPI, LLaMA 70B signals deliver the lowest (best) CRPS ratios for most horizons up to $h = 9$, with Michigan expectations performing best at $h = 1$ (although not always significantly). For PCE, LLaMA 70B remains the best performer up to $h = 6$, while Michigan expectations becomes relatively more competitive at longer horizons. Among the fine-tuned models, BERT-based specifications–especially

[36] We also restrict attention to `r/Economics` and `r/economy`. The sample for `r/wallstreetbets` is shorter and triggers convergence issues in the quantile-estimation routine.

InflaBERT–are consistently among the strongest performers.

Quantile-dependent coefficients. To illustrate how Reddit signals relate to different parts of the inflation distribution, we also estimate quantile regressions for $h = 6$ using predictors

$$x_t' = \{\pi_{t-1}, \pi_t^e, \text{Swap}_t, X_t\},$$

where X_t is either the InflaBERT-based indicator or the LLaMA 70B-based indicator. These two cases serve as representative examples of a fine-tuned, task-adapted model and a large unfine-tuned model. Because inference in quantile regression is non-standard, we compute confidence bands using a block-bootstrap procedure following Lopez-Salido and Loria (2024). Due to computational cost, we focus on `r/Economics`, which provides the longest available sample (2009M1–2025M8).

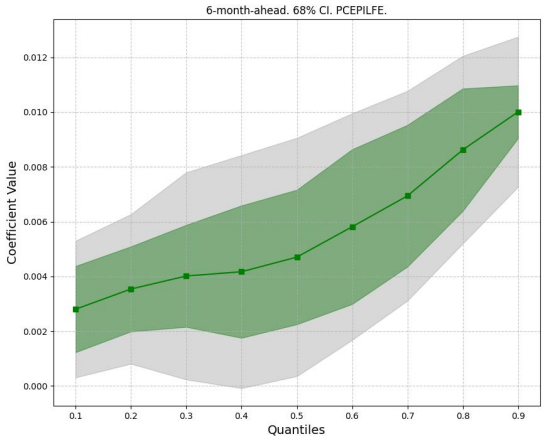
Our approach is inspired by Lopez-Salido and Loria (2024), who study inflation dynamics through quantile-dependent coefficients. Our objective here is not structural identification,[37] but rather to document whether Reddit signals are more informative in particular regions of the inflation distribution.

Figure A.31 plots the estimated coefficients across quantiles. The Reddit "educated" signal displays a profile similar to conventional expectation measures: its predictive contribution tends to be stronger in parts of the distribution associated with elevated inflation outcomes, consistent with state-dependent attention and expectation formation documented in the literature (Goldstein, 2023; Coibion and Gorodnichenko, 2025; Bracha and Tang, 2025; Weber et al., 2025). Moreover, InflaBERT yields tighter and more stable confidence bands across quantiles than LLaMA 70B, consistent with its superior performance in the point-forecast evaluation.

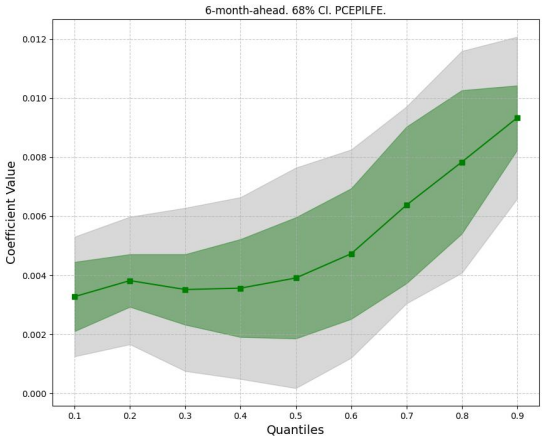
[37] We aim to quantify the incremental predictive content (partial correlation) of Reddit indicators for density forecasting rather than identify structural drivers of inflation.

Figure A.30: Estimated coefficients across quantiles of the dependent variable (PCE or CPI).

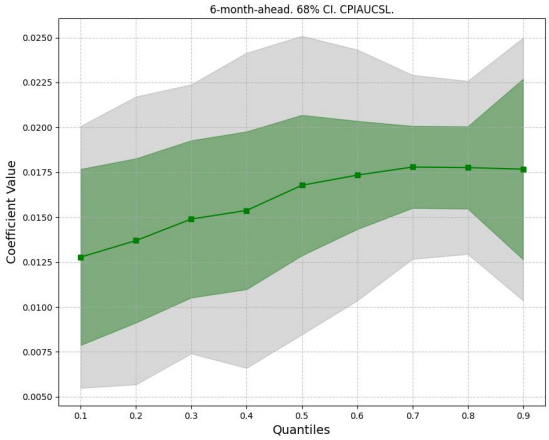
(a) PCE / `LLaMA 3.1 70B`



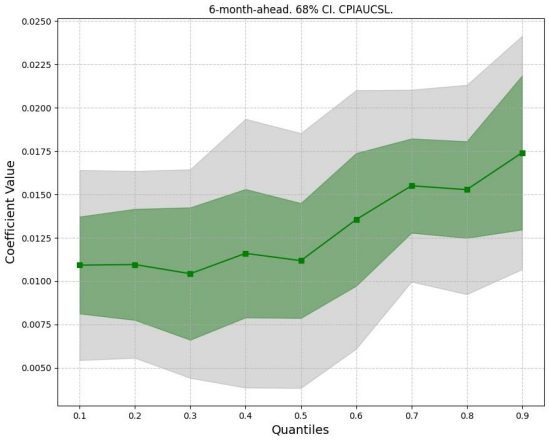
(b) CPI / `LLaMA 3.1 70B`



(c) PCE / `InflaBERT`



(d) CPI / `InflaBERT`



Notes: The plots report quantile-regression coefficients for $h = 6$ from regressions including lagged inflation, Michigan expectations, the one-year inflation swap, and a Reddit indicator X_t (InflaBERT or LLaMA 70B). Coefficients are reported across quantiles of the conditional distribution of the dependent variable (PCE or CPI). Confidence intervals are computed using a block bootstrap.

E.6 Quantile estimation (OLD)

We extend our forecasting framework by providing a density forecast application that follows the same structure as the point forecast analysis. Specifically, we estimate the model in Equation 4, replacing the conditional mean with conditional quantiles, thereby employing a quantile regression approach. This allows us to assess whether each indicator contributes meaningfully to the prediction of the entire conditional distribution of inflation.

While our focus is on a univariate time series process (indexed by subscript t), the methodological framework we describe is general and applicable to cross-sectional or panel data settings as well.

Quantile regression is a statistical method used to analyze the relationship between a set of K regressors $x_t = [x_1, \ldots, x_K]_t$ and a continuous univariate dependent variable $\mathcal{Y}_t$ at different points of its conditional distribution. Unlike standard linear regression, which estimates the conditional mean, quantile regression provides estimates for conditional quantiles, allowing for a more comprehensive understanding of the distributional dynamics. A key feature of quantile regression is its ability to capture how predictors influence different parts of the distribution, making it particularly useful in the presence of heteroscedasticity, outliers, or asymmetric distributions - features common in empirical macroeconomic data.

$$\mathcal{Y}_{t|\tau} = \boldsymbol{\beta}(\tau)x_t' + \epsilon_{t|\tau}, \quad \tau \in (0,1) \tag{A.21}$$

In this framework, the coefficient vector $\boldsymbol{\beta}(\tau)$ varies across quantiles τ, and the conditional quantile function satisfies:

$$\Pr\left[\mathcal{Y}_t \leq x_t'\boldsymbol{\beta}(\tau) \mid x_t\right] = \tau. \tag{A.22}$$

The estimator is obtained by minimizing the quantile loss function:

$$\rho_\tau\left(u_t\right) = \begin{cases} \tau u_t(\tau), & \text{if } u_t(\tau) \geq 0 \\ (1-\tau)u_t(\tau), & \text{otherwise} \end{cases} \tag{A.23}$$

where $u_t(\tau) = \mathcal{Y}_t - x_t'\boldsymbol{\beta}(\tau)$. The quantile regression estimator is then given by:

$$\widehat{\boldsymbol{\beta}}(\tau) = \arg\min_{\boldsymbol{\beta}} \frac{1}{T} \sum_{t=1}^{T} \rho_\tau \left(u_t \right) \tag{A.24}$$

Unlike ordinary least squares, this minimization problem does not have a closed-form solution due to the non-differentiability of the loss function at $u_t(\tau) = 0$. As a result, numerical optimization methods are typically employed.[38].

To reconstruct the predictive distribution, we adopt the nonparametric quantile interpolation method proposed by Mitchell et al. (2024):

$$\hat{F}_k \left(\mathcal{Y}_{t+h} \mid x_t \right) = \tau_j + \frac{\tau_{j+1} - \tau_j}{\hat{f}(x_t, \theta_{\tau_{j+1}}) - \hat{f}(x_t, \theta_{\tau_j})} \left(\mathcal{Y}_{t+h} - \hat{f}(x_t, \theta_{\tau_{j+1}}) \right) \tag{A.25}$$

where $\hat{f}(x_t, \theta_{\tau_j})$ denotes the predicted quantile at level τ_j. This approach estimates the density by interpolating between quantile estimates.

To evaluate the accuracy of the predictive distribution, we use the Continuous Ranked Probability Score (CRPS), which measures the distance between the forecast distribution and the observed outcome:

$$\text{CRPS}(F, y) = \int \left(F(x) - \mathbf{1}_{\{x \geq y\}} \right)^2 dx \tag{A.26}$$

The CRPS measures the distance between the predictive distribution and the realized outcome, which is treated as a degenerate distribution centered at the observed value of the dependent variable. Lower CRPS values indicate a better alignment between the forecast distribution and the actual realization.

In Table A.13 and A.14 we report density forecast performance in terms of CRPS ratios respectively for CPI and PCE. In this setup, we use the same for Reddit-based indicators used for the point-forecast, smoothed with a 90-day moving averages window. For computational reasons, we do not aggregate the predictions as we did for the point-forecast evaluation[39]. However, we checked for different size of the backward MA-smoother, finding that the results do not drastically change to the ones shown for 90-day MA.

[38]We use a maximum likelihood-based approach for numerical estimation.

[39]We only show results for r/Economics and r/economy because the sample for r/wallstreetbets is too short and gives convergence issues in the MLE algorithm for quantile regression.

In order to facilitate the MLE algorithm for quantile estimation, we shortened the sample out-of-sample set to give more observations in the in-sample window, mitigating convergence issues (in-sample: 2009M1-2018M12, out-of-sample: 2019M1-2025M8).

For CPI, the Llama70B overperforms all the model for all the horizons until $h = 9$, except for 1-month ahead forecast, where Michigan expectations provides a lower CRPS, even though is not statistically significant. For PCE, the Llama70B is always the best model up to horizon $h = 6$, and the Michigan expectations improve the performance for long-term forecasts. The BERT-base models, in particular Inflabert, confirms to be the best among the fine-tuned LLMs.

We further estimate the forecast regression in Equation A.15 using a quantile regression framework for $h = 6$[40]. We report estimated coefficients for Reddit indicators from a model that includes expectations and swaps, i.e., $x_t' = \{\pi_{t-1}, \pi_t^e, \text{Swap}_t, X_t\}$, where X_t refers to either `Inflabert` or `LLaMA 3.1 70B`. These serve as representative cases for Reddit indicators from fine-tuned proprietary and third-party models, respectively.

Given the absence of closed-form solutions in quantile regression, we compute confidence bands using a block-bootstrap approach following Lopez-Salido and Loria (2024). Due to computational complexity, we restrict our analysis to r/Economics, which provides the longest available sample (2009m1-2025m8).

Our approach is inspired by the empirical framework of Lopez-Salido and Loria (2024), who study inflation dynamics by evaluating quantile-dependent regression coefficients. While our goal is less structural,[41] this framework allows us to uncover how the relationship between Reddit indicators and inflation evolves across the distribution.

Figure A.31 presents the estimated coefficients across quantiles for the Educated signal extracted from the r/Economics subreddit.

The Educated signal exhibits behavior similar to that of conventional expectation-based indicators. As noted by Goldstein (2023), belief rigidity tends to decline during periods of elevated inflation, which amplifies the contribution of expectations to forecasting models. This finding is consistent with the view that economic agents pay greater attention (i.e. they

[40] We do not report $h = 1$ because the signal at 1-month horizon is not statistically significant for density forecast in a quantile setup.

[41] We aim to assess the incremental predictive power - i.e., partial correlation - of Reddit signals for inflation density forecasting, rather than to identify structural drivers of inflation.

Table A.13: Forecast results (CRPS ratios) for *CPI* (90-day MA signals).

Model	$h = 1$	$h = 2$	$h = 3$	$h = 4$	$h = 5$	$h = 6$	$h = 9$	$h = 12$	$h = 18$
Expectations									
Michigan Survey	**0.976**	1.004	1.016	0.957	0.889*	0.833***	0.731***	**0.650*****	**0.715*****
1-Y Inflation Swap	1.017	1.017	1.010	0.972	0.933	0.885*	0.792***	0.767***	0.808***
Sentiment									
VADER (r/Economics)	1.008	1.006	1.018	1.031**	1.026*	1.019	0.975**	0.960**	0.984
TextBlob (r/Economics)	1.014	1.013	1.001	0.979	0.953	0.927	0.862***	0.866***	0.872***
LM (r/Economics)	1.144	1.112	1.092	1.036	1.043	1.062	1.086	1.017	1.118
VADER (r/economy)	1.074**	1.063***	1.067***	1.064***	1.058***	1.037*	1.058*	1.036	1.013
TextBlob (r/economy)	1.018	1.004	0.951	0.925	0.935	0.944	0.936	0.972	1.084
LM (r/economy)	1.145	1.112	1.091	1.035	1.042	1.062	1.087	1.018	1.117
LLM signals (90-day MA smooth)									
BERT-base (r/Economics)	1.029	1.010	0.971	0.922*	0.872***	0.844***	0.811***	0.777***	0.859***
InflaBERT (r/Economics)	1.023	0.986	0.931	0.873*	0.843*	0.838**	0.830*	0.847	0.892
FinBERT (r/Economics)	1.003	0.995	0.974	0.934*	0.893***	0.860***	0.824***	0.787***	0.815***
Qwen 2.5 0.5B (r/Economics)	1.028	1.007	0.970	0.927	0.878**	0.842***	0.792***	0.772***	0.814***
Qwen 2.5 1.5B (r/Economics)	1.013	1.010	0.992	0.954	0.913**	0.881***	0.844***	0.825***	0.845***
Qwen 2.5 7B (r/Economics)	1.043**	1.042*	1.035	1.012	0.968	0.937	0.903**	0.888**	0.916*
Gemma 2 2B (r/Economics)	1.047	1.000	0.953	0.899*	0.860**	0.836**	0.786***	0.755**	0.847**
Gemma 2 9B (r/Economics)	1.018	1.011	0.998	0.965	0.929*	0.898***	0.860***	0.820***	0.844***
Gemma 2 27B (r/Economics)	1.047**	1.025	1.002	0.972	0.938	0.908*	0.853**	0.826**	0.871**
LLaMA 3.2 1B (r/Economics)	1.043*	1.012	0.977	0.936	0.892**	0.860**	0.829**	0.803**	0.860**
LLaMA 3.2 3B (r/Economics)	1.035	0.997	0.961	0.917*	0.870**	0.849**	0.817***	0.802**	0.860**
LLaMA 3.1 8B (r/Economics)	1.057*	1.003	0.976	0.940	0.906	0.889	0.849**	0.809**	0.845**
LLaMA 3.1 70B-instruct (r/Economics)	0.993	**0.923****	**0.830*****	**0.764*****	**0.721*****	**0.701*****	**0.705*****	0.729***	0.771***
BERT-base (r/economy)	1.117**	1.058	1.033	1.012	0.997	1.002	1.043	1.040	1.192
InflaBERT (r/economy)	1.055	1.062	1.034	0.977	0.943	0.927	0.893	0.919	1.053
FinBERT (r/economy)	1.142**	1.077	1.037	0.982	0.948	0.946	0.934	0.910	1.024
Qwen 2.5 0.5B (r/economy)	1.141**	1.068	1.030	0.979	0.940	0.928	0.929	0.922	1.071
Qwen 2.5 1.5B (r/economy)	1.144**	1.091	1.060	1.027	0.981	0.970	0.986	0.998	1.138
Qwen 2.5 7B (r/economy)	1.110*	1.035	0.987	0.932	0.894	0.878	0.873	0.922	1.094
Gemma 2 2B (r/economy)	1.174**	1.098	1.063	1.013	0.989	0.966	0.934	0.916	1.075
Gemma 2 9B (r/economy)	1.146**	1.096	1.061	1.029	0.998	0.997	0.991	0.972	1.089
Gemma 2 27B (r/economy)	1.151**	1.103	1.082	1.040	1.025	1.027	1.031	1.016	1.156
LLaMA 3.2 1B (r/economy)	1.139**	1.088	1.048	0.991	0.966	0.948	0.980	0.984	1.124
LLaMA 3.2 3B (r/economy)	1.138*	1.068	1.025	0.986	0.957	0.951	0.953	0.962	1.105
LLaMA 3.1 8B (r/economy)	1.133*	1.058	1.019	0.961	0.918	0.913	0.909	0.871	0.989
LLaMA 3.1 70B-instruct (r/economy)	1.077	1.025	0.962	0.910	0.870	0.855	0.833	0.882	1.024

Notes: CRPS is computed as a ratio: $CRPS(AR_x(1))/CRPS(AR(1))$. The initial in-sample period spans from 2009M1 to 2018M12; the out-of-sample period covers 2019M1 to 2025M8. Forecasts use an expanding window and are re-estimated each step. The **bold** entry in each column marks the lowest relative CRPS among the reported models for that horizon. Stars $(*, **, ***)$ are the significance indicators you provided. LLM rows use a **90-day moving-average** smoothing of signal features.

are more attentive) to price dynamics during inflationary episodes, thereby improving the informativeness of sentiment-based signals (Coibion and Gorodnichenko, 2025; Bracha and Tang, 2025; Weber et al., 2025).

In comparing `LLaMA 3.1 70B` with `Inflabert`, we find that the latter produces narrower and more centered confidence bands across quantiles, confirming its superior statistical per-

Table A.14: Forecast results (CRPS ratios) for *PCE* (90-day MA signals).

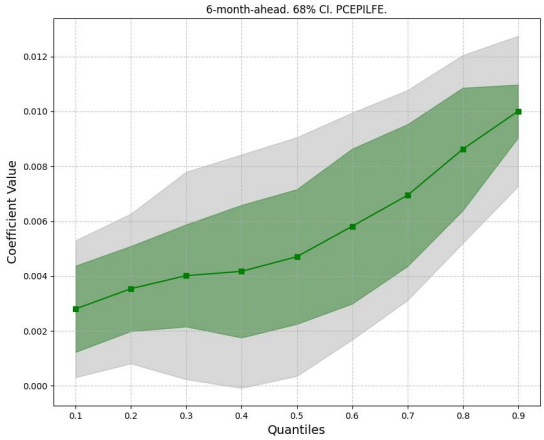
Model	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=9$	$h=12$	$h=18$
Expectations									
Michigan Survey	1.006	0.997	0.993	0.957	0.942	**0.910***	**0.818*****	**0.718*****	**0.760*****
1-Y Inflation Swap	1.020	0.994	0.989	0.988	0.965	0.940	0.852*****	0.806*****	0.860*****
Sentiment									
VADER (r/Economics)	1.001	1.019**	1.034*****	1.043*****	1.052*****	1.055*****	1.030*****	1.012*	1.015*****
TextBlob (r/Economics)	1.002	1.013*	1.008	0.994	0.983	0.974	0.955**	0.940*****	0.939**
LM (r/Economics)	1.100	1.026	1.022	1.019	1.028	1.057	1.047	1.021	1.136
VADER (r/economy)	1.019**	1.014	1.019**	1.016**	1.014**	1.011	0.992	0.958**	0.950*
TextBlob (r/economy)	1.015	1.003	0.989	0.985	1.001	1.005	0.970	0.987	1.041
LM (r/economy)	1.101	1.025	1.022	1.018	1.028	1.056	1.047	1.022	1.136
LLM signals (90-day MA smooth)									
BERT-base (r/Economics)	1.005	1.000	0.998	0.979	0.969	0.972	0.947**	0.913*****	0.925*****
InflaBERT (r/Economics)	0.994	0.982	0.968	0.943	0.922*	0.927*	0.925**	0.912**	0.891*****
FinBERT (r/Economics)	1.005	1.010	1.012	1.005	0.990	0.985	0.950*****	0.918*****	0.919*****
Qwen 2.5 0.5B (r/Economics)	1.011	1.003	1.001	0.993	0.980	0.972	0.936**	0.893*****	0.907*****
Qwen 2.5 1.5B (r/Economics)	1.006	1.014*	1.023**	1.014	1.000	0.997	0.966*****	0.941*****	0.936*****
Qwen 2.5 7B (r/Economics)	1.012	1.010	1.004	0.991	0.980	0.972	0.940**	0.923*****	0.942*
Gemma 2 2B (r/Economics)	1.005	0.985	0.970	0.963	0.947	0.944*	0.926**	0.897*****	0.918**
Gemma 2 9B (r/Economics)	1.011	1.015	1.025*	1.023	1.005	0.999	0.954*****	0.919*****	0.923*****
Gemma 2 27B (r/Economics)	1.017*	1.005	1.004	0.996	0.986	0.979	0.951*	0.912*****	0.922**
LLaMA 3.2 1B (r/Economics)	1.008	1.000	1.003	0.986	0.973	0.970	0.949**	0.919*****	0.920*****
LLaMA 3.2 3B (r/Economics)	1.005	0.993	0.987	0.966	0.953*	0.951*	0.933**	0.912*****	0.917*****
LLaMA 3.1 8B (r/Economics)	1.012	0.986	0.986	0.978	0.971	0.965	0.951*	0.918*****	0.934**
LLaMA 3.1 70B-instruct (r/Economics)	**0.969****	**0.956*****	**0.927*****	**0.890*****	**0.874*****	**0.871*****	0.871*****	0.855*****	0.845*****
BERT-base (r/economy)	1.044	1.013	1.020	1.001	1.004	0.997	1.022	1.003	1.087
InflaBERT (r/economy)	1.061	1.033	1.024	1.008	1.001	0.986	0.966	0.977	1.012
FinBERT (r/economy)	1.054	1.025	1.040	1.011	1.010	0.991	0.952	0.914	0.987
Qwen 2.5 0.5B (r/economy)	1.070	1.038	1.036	1.010	1.005	0.986	0.955	0.936	0.993
Qwen 2.5 1.5B (r/economy)	1.051	1.023	1.021	1.002	0.993	0.979	0.959	0.942	1.030
Qwen 2.5 7B (r/economy)	1.073	1.050	1.029	1.012	0.992	0.968	0.916	0.932	1.018
Gemma 2 2B (r/economy)	1.117	1.035	1.029	1.006	1.016	1.002	0.929	0.913	0.996
Gemma 2 9B (r/economy)	1.054	1.029	1.033	1.024	1.018	1.006	0.997	0.972	1.042
Gemma 2 27B (r/economy)	1.069	1.039	1.046	1.050	1.051	1.042	1.012	0.989	1.058
LLaMA 3.2 1B (r/economy)	1.068	1.022	1.013	1.002	0.999	0.992	0.957	0.941	1.031
LLaMA 3.2 3B (r/economy)	1.057	1.033	1.011	0.982	0.984	0.983	0.945	0.927	0.995
LLaMA 3.1 8B (r/economy)	1.062	1.016	1.014	1.009	0.999	0.982	0.955	0.944	0.954
LLaMA 3.1 70B-instruct (r/economy)	1.064	1.023	0.997	0.969	0.956	0.943	0.929	0.933	0.948

Notes: CRPS is computed as a ratio: $CRPS(AR_x(1))/CRPS(AR(1))$. The initial in-sample period spans from 2009M1 to 2018M12; the out-of-sample period covers 2019M1 to 2025M8. Forecasts are generated using an expanding window, with the model re-estimated at each step. The **bold** entry in each column marks the lowest relative CRPS among the reported models for that horizon. Stars $(*, **, ***)$ denote significance levels as provided in the evaluation outputs. LLM rows here use a **90-day moving-average** smoothing of signal features.

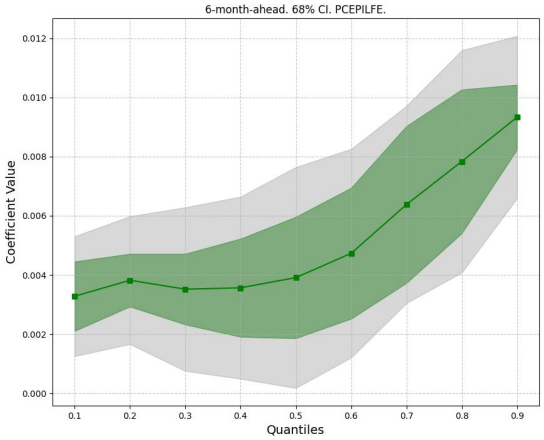
formance. This result is consistent with our earlier findings from the point forecast analysis and further supports the robustness of fine-tuned Reddit-based indicators.

Figure A.31: Estimated coefficients across quantiles of the dependent variable (PCE or CPI)

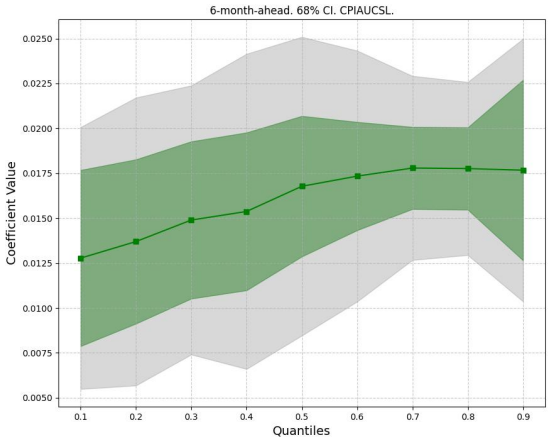
(a) PCE / `LLaMA 3.1 70B`



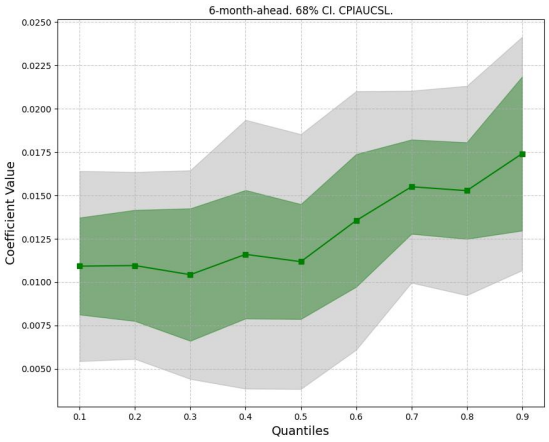
(b) CPI / `LLaMA 3.1 70B`



(c) PCE / `Inflabert`



(d) CPI / `Inflabert`



Notes: The plots display estimated coefficients from one-step-ahead quantile regressions of the form $x_t' = \{\pi_{t-1}, \pi_t^e, \text{Swap}_t, X_t\}$, where X_t represents Reddit-based indicators derived from the `Inflabert` and `LLaMA 3.1 70B` models. Coefficients are shown across quantiles of the conditional distribution of the dependent variable (PCE or CPI). Confidence intervals are computed using block bootstrapping.

F Reddit transmission to expectations

Social-media signals have been shown to capture many of the same features as survey-based expectations (Daas and Puts, 2014; Angelico et al., 2022; Born et al., 2024; Gorodnichenko et al., 2024). Unlike surveys, however, online narratives are available at high frequency and without publication lags, offering policymakers timely and potentially lower-cost information about inflation sentiment.

This appendix provides additional evidence supporting our interpretation of Reddit indicators as expectation-related measures. The central question is whether Reddit discussions merely *co-move* with expectations or whether they also *anticipate* survey-based inflation expectations.

F.1 VAR setup and Granger-causality tests

We estimate a VAR that includes inflation, survey expectations, market-based expectations, and (one at a time) a Reddit indicator extracted using different LLMs:

$$\left[\pi_t,\ \pi_t^e,\ \text{1YSwap}_t,\ X_t^R\right],$$

where π_t is inflation (PCE), π_t^e denotes Michigan inflation expectations, 1YSwap_t is the one-year inflation swap rate, and X_t^R is a Reddit-based indicator.[42] The VAR lag order is selected using the Akaike Information Criterion (AIC) with a maximum of 12 lags.

For each subreddit and each MA smoothing window, we test Granger-causality between Reddit sentiment and survey expectations, focusing on two directions: (i) whether Reddit sentiment Granger-causes Michigan expectations ($X_t^R \rightarrow \pi_t^e$), and (ii) whether Michigan expectations Granger-cause Reddit sentiment ($\pi_t^e \rightarrow X_t^R$). Bidirectional causality is recorded when both tests reject.

Tables A.15–A.17 report the results by subreddit; Table A.18 repeats the exercise for an aggregate Reddit indicator constructed by averaging the three subreddit indicators (within each model family).

[42] We use PCE inflation because it is more closely correlated with expectation measures than CPI in our sample.

F.2 Summary of results

Across subreddits and model families, the dominant pattern is that Reddit indicators more frequently Granger-cause Michigan expectations than the reverse. The strength of this pattern varies by subreddit.

`r/Economics`. For `r/Economics`, we identify 64 cases with statistically significant Granger-causality in at least one direction. Of these, 47 indicate $X_t^R \to \pi_t^e$ (Reddit anticipates expectations), 10 indicate bidirectional causality, and 7 indicate $\pi_t^e \to X_t^R$ (expectations anticipate Reddit).

`r/economy`. For `r/economy`, the evidence is weaker but remains informative. We identify 25 significant cases in total: 17 correspond to $X_t^R \to \pi_t^e$, 1 indicates bidirectional causality, and 7 correspond to $\pi_t^e \to X_t^R$.

`r/wallstreetbets`. For `r/wallstreetbets`, we detect 32 significant causality relations. Among these, 20 correspond to $X_t^R \to \pi_t^e$, 5 indicate bidirectional causality, and 7 correspond to $\pi_t^e \to X_t^R$. Despite the noisier and more speculative nature of this subreddit, a non-trivial share of the extracted signal remains forward-looking with respect to survey expectations.

Aggregate across subreddits. Table A.18 reports results for an aggregate Reddit indicator (the average across the three subreddits, computed within each model family). The conclusions are consistent with the subreddit-level evidence: in most specifications, X_t^R Granger-causes π_t^e; the reverse direction occurs less frequently; and bidirectional relations are comparatively rare.

Overall, the strongest and most consistent forward-looking pattern arises from the large, thematically coherent community `r/Economics`, but both `r/economy` and `r/wallstreetbets` also show meaningful evidence that Reddit narratives can anticipate shifts in survey-based inflation expectations. These results provide additional support for interpreting Reddit-based measures as expectation-related signals and help rationalize their forecasting performance in the main analysis.

Model	1	5	10	30	90	120	180	360
BERT base	–	–	–	–	$\pi_t^e \to X_t^R$	$\pi_t^e \to X_t^R$	$X_t^R \to \pi_t^e$	–
FinBERT	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	$X_t^R \leftrightarrow \pi_t^e$	$\pi_t^e \to X_t^R$	$X_t^R \to \pi_t^e$	–
Gemma 27B	–	–	–	–	–	$\pi_t^e \to X_t^R$	–	$X_t^R \to \pi_t^e$
Gemma 2B	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	–
Gemma 9B	–	–	–	–	–	$\pi_t^e \to X_t^R$	–	–
InflaBERT	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	$X_t^R \leftrightarrow \pi_t^e$	$X_t^R \to \pi_t^e$	$X_t^R \leftrightarrow \pi_t^e$	$X_t^R \leftrightarrow \pi_t^e$	–
LLaMA 8B	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	$\pi_t^e \to X_t^R$	$X_t^R \leftrightarrow \pi_t^e$	$X_t^R \to \pi_t^e$	–
LLaMA 1B	$X_t^R \to \pi_t^e$	–	–	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$
LLaMA 3B	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	$X_t^R \leftrightarrow \pi_t^e$	$X_t^R \leftrightarrow \pi_t^e$	$X_t^R \leftrightarrow \pi_t^e$	–
LLaMA 70B	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	$X_t^R \leftrightarrow \pi_t^e$	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$
Qwen 0.5B	–	–	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$
Qwen 1.5B	–	–	–	$X_t^R \to \pi_t^e$	$X_t^R \leftrightarrow \pi_t^e$	$\pi_t^e \to X_t^R$	–	–
Qwen 7B	–	–	–	–	–	–	–	–

Table A.15: Granger causality between survey expectations π_t^e and Reddit sentiment X_t^R for subreddit `r/Economics`. Yellow cells indicate cases where Reddit sentiment Granger-causes expectations. Green cells indicate bidirectional causality, – indicates no causality in any direction.

Model	1	5	10	30	90	120	180	360
BERT base	–	–	–	–	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	–	–
FinBERT	–	–	–	–	–	–	–	–
Gemma 27B	–	–	–	–	–	–	–	–
Gemma 2B	–	–	–	–	–	$X_t^R \to \pi_t^e$	–	–
Gemma 9B	–	–	–	–	–	–	–	$\pi_t^e \to X_t^R$
InflaBERT	$\pi_t^e \to X_t^R$	–	–	–	–	$X_t^R \to \pi_t^e$	–	–
LLaMA 8B	–	–	–	–	–	–	–	–
LLaMA 1B	–	–	–	–	–	–	–	–
LLaMA 3B	–	–	–	–	–	–	$X_t^R \to \pi_t^e$	–
LLaMA 70B	$\pi_t^e \to X_t^R$	$\pi_t^e \to X_t^R$	$\pi_t^e \to X_t^R$	$\pi_t^e \to X_t^R$	$\pi_t^e \to X_t^R$	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$
Qwen 0.5B	–	–	–	–	–	–	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$
Qwen 1.5B	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	–
Qwen 7B	–	–	–	–	–	$X_t^R \leftrightarrow \pi_t^e$	–	–

Table A.16: Granger causality between survey expectations π_t^e and Reddit sentiment X_t^R for subreddit `r/economy`. Yellow cells indicate cases where Reddit sentiment Granger-causes expectations. Green cells indicate bidirectional causality, – indicates no causality in any direction.

Model	1	5	10	30	90	120	180	360
BERT base	–	$X_t^R \to \pi_t^e$	–	–	–	–	–	–
FinBERT	–	–	–	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	–
Gemma 27B	–	–	–	–	–	–	–	–
Gemma 2B	–	–	–	–	$X_t^R \to \pi_t^e$	–	$X_t^R \to \pi_t^e$	–
Gemma 9B	–	–	–	–	$X_t^R \to \pi_t^e$	–	–	–
InflaBERT	–	–	–	–	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	$\pi_t^e \to X_t^R$
LLaMA 8B	–	–	–	–	$\pi_t^e \to X_t^R$	–	$\pi_t^e \to X_t^R$	–
LLaMA 1B	–	–	–	–	–	–	–	–
LLaMA 3B	–	–	–	–	$X_t^R \leftrightarrow \pi_t^e$	$X_t^R \leftrightarrow \pi_t^e$	$X_t^R \leftrightarrow \pi_t^e$	$\pi_t^e \to X_t^R$
LLaMA 70B	–	–	–	–	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	–	$\pi_t^e \to X_t^R$
Qwen 0.5B	–	–	–	$\pi_t^e \to X_t^R$	$X_t^R \to \pi_t^e$	$X_t^R \leftrightarrow \pi_t^e$	$X_t^R \leftrightarrow \pi_t^e$	–
Qwen 1.5B	–	–	–	–	$X_t^R \to \pi_t^e$	–	$\pi_t^e \to X_t^R$	–
Qwen 7B	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	–	–	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	–

Table A.17: Granger causality between survey expectations π_t^e and Reddit sentiment X_t^R for subreddit `r/wallstreetbets`. Yellow cells indicate cases where Reddit sentiment Granger-causes expectations. Green cells indicate bidirectional causality, – indicates no causality in any direction.

Model	1	5	10	30	90	120	180	360
BERT base	–	–	–	–	–	–	–	$\pi_t^e \to X_t^R$
FinBERT	$X_t^R \to \pi_t^e$	–	–	$X_t^R \to \pi_t^e$	–	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	$\pi_t^e \to X_t^R$
Gemma 27B	–	–	–	–	–	–	–	$\pi_t^e \to X_t^R$
Gemma 2B	–	–	–	–	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	–	$\pi_t^e \to X_t^R$
Gemma 9B	–	–	–	–	$X_t^R \to \pi_t^e$	$\pi_t^e \to X_t^R$	$X_t^R \to \pi_t^e$	–
InflaBERT	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	$X_t^R \leftrightarrow \pi_t^e$	$X_t^R \to \pi_t^e$	$X_t^R \leftrightarrow \pi_t^e$	$X_t^R \to \pi_t^e$	$\pi_t^e \to X_t^R$
LLaMA 8B	–	–	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	–	–	–	–
LLaMA 1B	–	–	–	–	$X_t^R \to \pi_t^e$	$\pi_t^e \to X_t^R$	–	$\pi_t^e \to X_t^R$
LLaMA 3B	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	$X_t^R \leftrightarrow \pi_t^e$	$X_t^R \to \pi_t^e$	$\pi_t^e \to X_t^R$
LLaMA 70B	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	$X_t^R \leftrightarrow \pi_t^e$	$X_t^R \leftrightarrow \pi_t^e$	$X_t^R \leftrightarrow \pi_t^e$	$X_t^R \to \pi_t^e$	$\pi_t^e \to X_t^R$
Qwen 0.5B	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	$X_t^R \leftrightarrow \pi_t^e$	$X_t^R \to \pi_t^e$	$\pi_t^e \to X_t^R$	$X_t^R \to \pi_t^e$	–
Qwen 1.5B	$X_t^R \to \pi_t^e$	–	–	$X_t^R \to \pi_t^e$	$X_t^R \to \pi_t^e$	$X_t^R \leftrightarrow \pi_t^e$	$X_t^R \to \pi_t^e$	$\pi_t^e \to X_t^R$
Qwen 7B	–	–	–	–	–	–	–	–

Table A.18: Granger causality between survey expectations π_t^e and Reddit sentiment X_t^R for the aggregate of the three subreddits. Yellow cells indicate cases where Reddit sentiment Granger-causes expectations. Green cells indicate bidirectional causality, – indicates no causality in any direction.

G The relationship between Reddit and news

This appendix investigates how Reddit discussions relate to traditional news sentiment. Establishing this link is important for interpreting our Reddit-based indicators: if Reddit mainly echoes news, the signal may reflect a socially filtered version of the news cycle; if it diverges, it may capture additional narrative content generated by users.

G.1 What types of content are posted on Reddit?

Table A.19 summarizes selected composition statistics for submissions in the three subreddits considered. Two patterns are informative.

First, explicit news-related flairs account for a non-negligible share of submissions in `r/Economics` (14.15%) and `r/economy` (16.86%), consistent with these subreddits functioning as hubs for sharing external economic information. By contrast, `r/wallstreetbets` does not use a "news" flair in a comparable way, consistent with its focus on user-driven commentary and community-specific content rather than traditional news posting norms.

Second, the prevalence of submissions with *selftext* (i.e., posts containing original written content) differs sharply across subreddits. In `r/Economics` (12.67%) and `r/economy` (17.71%), only a minority of submissions include substantial self-authored text, suggesting that discussion frequently centers on externally sourced material (e.g., links and headlines). In contrast, `r/wallstreetbets` exhibits a much higher share of selftext submissions (61.86%), indicating that a large portion of its content is composed of user-generated narratives, opinions, and trading-related commentary.

Overall, these patterns suggest that `r/Economics` and `r/economy` primarily operate as news-sharing and discussion venues, while `r/wallstreetbets` is more oriented toward original user content.

	r/Economics	r/economy	r/wallstreetbets
News flair	14.15%	16.86%	–
Selftext (not a flair)	12.67%	17.71%	61.86%

Table A.19: Selected submission characteristics by subreddit. Percentages may not sum to 100% because additional characteristics with negligible shares are omitted.

G.2 Why Reddit is not "just news"

Reddit is not a passive feed: users actively choose what to post and what to engage with. As a result, the content that appears on Reddit is a selected subset of the broader news and information environment. Our indicators therefore capture two elements simultaneously: (i) *selection*–which items users deem relevant enough to share–and (ii) *interpretation*–how those items are discussed and reframed through comments.

This distinguishes Reddit-based signals from traditional news-sentiment measures in two key ways. First, Reddit acts as a filtering mechanism: only content that attracts attention is posted and therefore enters the signal. Second, discussion adds a social layer that can amplify, contest, or regularize the sentiment associated with a given news item. Deviations between Reddit and news sentiment can therefore arise precisely because of selection and discussion dynamics.

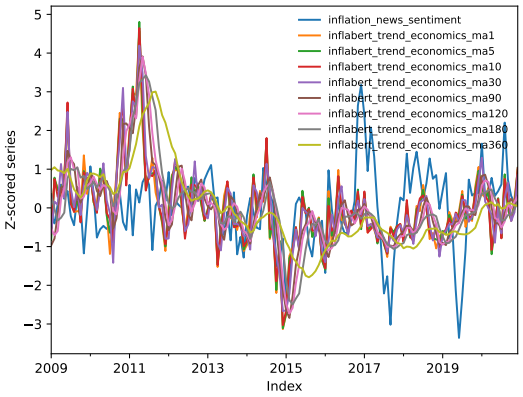
G.3 Empirical relationship between Reddit and news sentiment

We compare our Reddit-based sentiment series with an established newspaper-based sentiment measure from Barbaglia et al. (2022). Figure A.32 plots the two series (z-scored for comparability) for each subreddit using InflaBERT-based Reddit signals. Across all three subreddits, Reddit sentiment closely tracks the news-sentiment index, indicating a strong informational connection between the two sources.

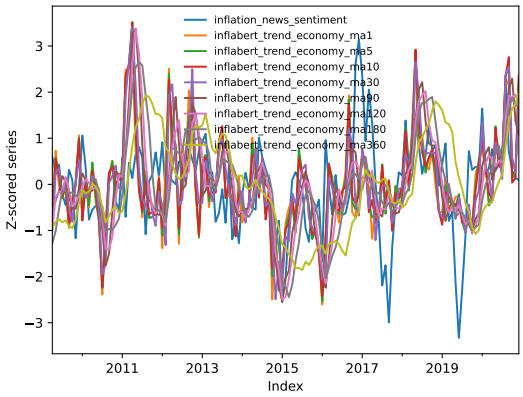
To examine timing, Figure A.33 reports lead–lag correlations between Reddit sentiment and the news-sentiment series. In all three panels, the peak correlation occurs at lag zero, indicating that the strongest association is contemporaneous. Correlations at positive and negative lags are generally smaller, providing limited evidence that Reddit systematically

Figure A.32: Time series (z-scored) of newspaper-based news sentiment and InflaBERT Reddit signals by subreddit.

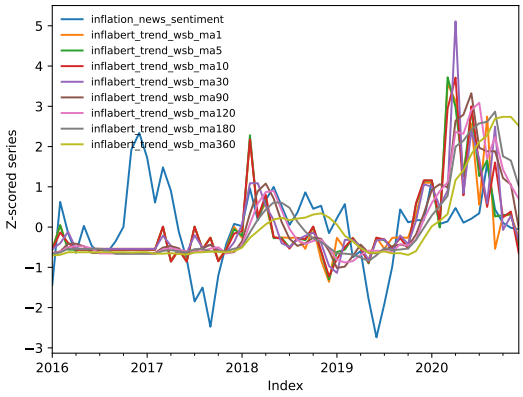
(a) `r/Economics`



(b) `r/economy`



(c) `r/wallstreetbets`



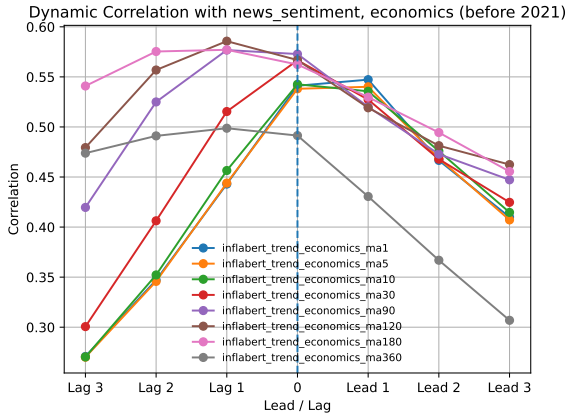
Notes: Each panel reports the standardized time series of newspaper-based news sentiment and InflaBERT Reddit signals for the corresponding subreddit.

leads or lags traditional news sentiment.

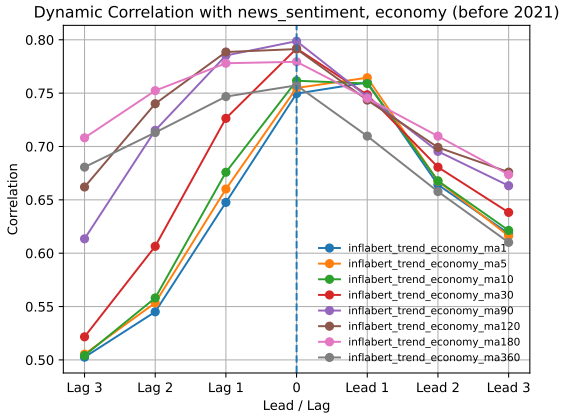
Taken together, the evidence indicates that Reddit sentiment strongly co-moves with traditional news sentiment and is primarily contemporaneous with it. This suggests that Reddit operates largely as a real-time *amplifier* of the news cycle rather than a systematic leading or lagging indicator of newspaper sentiment. At the same time, Reddit does not merely relay news: it embeds information within a layer of social interaction. Through comments, users interpret, reinforce, and sometimes contest posted content, so that the sentiment extracted from Reddit reflects both the underlying news and the community's collective framing. In this sense, our Reddit indicators can be viewed as a socially filtered counterpart to standard news-sentiment measures.

Figure A.33: Lead–lag correlations between Reddit sentiment and newspaper-based sentiment.

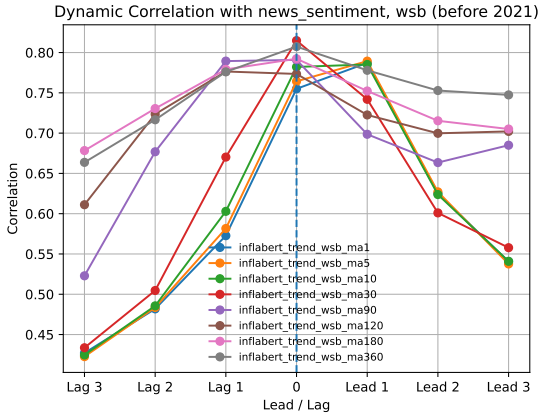
(a) `r/Economics`



(b) `r/economy`



(c) `r/wallstreetbets`



References (Online Appendix)

Allard, Marc-Antoine, Paul Teiletche, and Adam Zinebi (2024). *Enhancing Inflation Nowcasting with LLM: Sentiment Analysis on News*. arXiv: 2410.20198 (cit. on p. A8).

Angelico, Cristina, Juri Marcucci, Marcello Miccoli, and Filippo Quarta (2022). "Can we measure inflation expectations using Twitter?" In: *Journal of Econometrics* 228.2, pp. 259–277 (cit. on p. A69).

Araci, Dogu (2019). *FinBERT: Financial Sentiment Analysis with Pre-trained Language Models*. arXiv: 1908.10063 (cit. on p. A8).

Barbaglia, Luca, Stefano Consoli, and Stefano Manzan (2022). "Forecasting with Economic News". In: *Journal of Business & Economic Statistics* 41.3, pp. 708–719. DOI: 10.1080/07350015.2022.2060988 (cit. on p. A74).

Born, Benjamin, Hrishbh Dalal, Nora Lamersdorf, Jana-Lynn Schuster, and Sascha Steffen (Aug. 2024). "Inflation Expectations in the Social Media Age". Manuscript (cit. on p. A69).

Bracha, Anat and Jenny Tang (2025). "Inflation levels and (in) attention". In: *Review of Economic Studies* 92.3, pp. 1564–1594 (cit. on pp. A61, A66).

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller (Aug. 2008). "Bootstrap-Based Improvements for Inference with Clustered Errors". In: *The Review of Economics and Statistics* 90.3, pp. 414–427 (cit. on p. A23).

Coibion, Olivier and Yuriy Gorodnichenko (May 2025). *"Inflation, Expectations and Monetary Policy: What Have We Learned and to What End?"* Working Paper 33858. National Bureau of Economic Research. DOI: 10.3386/w33858 (cit. on pp. A61, A66).

Daas, Piet J. H. and Marco J. H. Puts (Sept. 2014). *Social Media Sentiment and Consumer Confidence*. Statistics Paper Series 5. The Hague, Netherlands: Statistics Netherlands (CBS) (cit. on p. A69).

Dettmers, Tim, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer (2023). *QLoRA: Efficient Finetuning of Quantized LLMs*. arXiv: 2305.14314 (cit. on p. A11).

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv: 1810.04805 (cit. on p. A8).

Goldstein, Nathan (2023). "Tracking inattention". In: *Journal of the European Economic Association* 21.6, pp. 2682–2725 (cit. on pp. A61, A65).

Gorodnichenko, Yuriy, Tho Pham, and Oleksandr Talavera (2024). "Central bank communication on social media: What, to whom, and how?" In: *Journal of Econometrics*, p. 105869 (cit. on p. A69).

Grattafiori, Aaron et al. (2024). *The Llama 3 Herd of Models*. arXiv: 2407.21783 [cs.AI] (cit. on p. A8).

Hansen, Peter R, Asger Lunde, and James M Nason (2011). "The Model Confidence Set". In: *Econometrica* 79.2, pp. 453–497 (cit. on p. A33).

Hayou, Soufiane, Nikhil Ghosh, and Bin Yu (2024). *LoRA+: Efficient Low Rank Adaptation of Large Models*. arXiv: 2402.12354 (cit. on p. A11).

Hu, Edward J, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. (2022). "LoRA: Low-rank adaptation of large language models." In: *ICLR* 1.2, p. 3 (cit. on p. A9).

Liu, Shih-Yang, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen (2024). *DoRA: Weight-Decomposed Low-Rank Adaptation*. arXiv: 2402.09353 (cit. on pp. A10, A11, A12).

Lopez-Salido, David and Francesca Loria (2024). "Inflation at risk". In: *Journal of Monetary Economics*, p. 103570 (cit. on pp. A61, A65).

Loshchilov, Ilya and Frank Hutter (2017). *Decoupled Weight Decay Regularization*. arXiv: 1711.05101 (cit. on p. A11).

Mitchell, James, Aubrey Poon, and Dan Zhu (2024). "Constructing density forecasts from quantile regressions: Multimodality in macrofinancial dynamics". In: *Journal of Applied Econometrics* 39.5, pp. 790–812 (cit. on pp. A59, A64).

Qwen, : An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang,

and Zihan Qiu (2025). *Qwen2.5 Technical Report*. arXiv: 2412.15115 [cs.CL] (cit. on p. A9).

Team, Gemma et al. (2024). *Gemma 2: Improving Open Language Models at a Practical Size*. arXiv: 2408.00118 [cs.CL] (cit. on p. A8).

Weber, Michael, Bernardo Candia, Hassan Afrouzi, Tiziano Ropele, Rodrigo Lluberas, Serafin Frache, Brent Meyer, Saten Kumar, Yuriy Gorodnichenko, Dimitris Georgarakos, et al. (2025). "Tell Me Something I Don't Already Know: Learning in Low-and High-Inflation Settings". In: *Econometrica* 93.1, pp. 229–264 (cit. on pp. A61, A66).