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

Andrea Del Monaco[1], Luigi Longo[2], Juri Marcucci[3], and Irene Tafani[4]

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

October 8, 2026

**Abstract**

We show that large language models (LLMs) can transform Reddit discussions into timely predictors of U.S. inflation. Using inflation-related submissions and time-local comments from major economics- and finance-focused subreddits, we develop the Reddit Inflation Signal (RIS), a high-frequency text-based measure that summarizes directional inflation-related discourse and perceived price dynamics. The RIS is generated by fine-tuning pre-trained models (BERT-, Qwen-, LLaMA-, and Gemma-type architectures) on labels produced by human annotators and ChatGPT and is benchmarked against an unfine-tuned LLaMA 70B model. Forecasting and nowcasting are implemented in a pseudo-real-time setting with strictly backward-looking transformations, recursive expanding windows, and explicit data-availability constraints. In a recursive pseudo out-of-sample evaluation with horizons up to 18 months, RIS-based models and MSE-weighted forecast combinations improve point and density forecasts of headline CPI and core PCE relative to standard benchmarks, including autoregressive models augmented with Michigan survey expectations and inflation swaps. In real-time nowcasting, the RIS constructed using information available early in the month improves nowcasts and performs competitively with the Cleveland Fed Inflation Nowcast. Importantly, much of the predictive content can be captured with fine-tuned small language models (SLMs), which often perform close to much larger LLMs at a fraction of the computational cost, supporting scalable and resource-efficient deployment.

**JEL**: E31, C32, C53, C55

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

*We would like to thank the Editor, Todd Clark, and three anonymous referees for their valuable comments on earlier versions of the paper. We also gratefully acknowledge comments from M. Affinito, R. Sabbatini, P. Soto and the participants of the 2026 NBER-NSF Time Series Conference (September 11-12, 2026), the CIML Training & Conference (April 23-24, 2026), the 2nd CAM-Risk Conference (April 8-10, 2026), the $7^{th}$ Conference on "Nontraditional Data, Machine Learning, and Natural Language Processing in Macroeconomics - ECONDAT 2025 Fall Meeting" (October 13-15, 2025), the IAAE2025 conference (June 25-27, 2025), the Workshop on Machine Learning Economic Forecasting and Nowcasting (April 25, 2025), the 2025 ECB FIAT and AI in Economics workshop, the $6^{th}$ Conference on "Nontraditional Data, Machine Learning, and Natural Language Processing in Macroeconomics - ECONDAT 2024 Fall Meeting" (November 13-14, 2024), and seminar participants at the Bank of Italy, IIF MacroFor, Amazon's FMF, and QuantCube Technology. The views and ideas expressed herein are those of the authors and do not necessarily reflect those of the Bank of Italy, the European Central Bank, the Eurosystem, or the European Commission. All remaining errors are our own. Corresponding author: Juri Marcucci, email: juri.marcucci@bancaditalia.it

# 1 Introduction

Accurate and timely measures of inflationary pressures are central to monetary policy and macroeconomic analysis. Yet short-horizon inflation forecasting and nowcasting remain notoriously difficult: monthly inflation is dominated by transitory shocks, official releases arrive with lags, and standard predictors—including surveys and market-based expectations—often provide limited incremental content at short forecast horizons. At the same time, public discussion about prices has migrated online. Social-media platforms record, in real time, how households and market participants talk about price changes, interpret news, and form expectations. The challenge is that these data are unstructured, noisy, and difficult to map into economically interpretable objects.

Reddit provides a particularly attractive setting in this respect. It is a large repository of user-generated content and discussion, where views about prices and macroeconomic conditions evolve continuously. Unlike traditional survey-based measures—which rely on periodic questionnaires administered to limited samples and may be shaped by formal question wording—social media allows users to express perceptions and concerns spontaneously and at high frequency. Leveraging modern natural language processing (NLP) tools and large language models (LLMs), Reddit can therefore support a high-frequency, crowdsourced barometer of inflation-related discourse, reducing the lag between information formation and measurement (Angelico et al., 2022).

In this paper, we use Reddit discussions to nowcast and forecast U.S. headline CPI and core PCE inflation. We construct the RIS from inflation-related submissions and time-local comments posted in major economics- and finance-oriented subreddits, capturing perceived price dynamics and inflation concerns. Our methodology integrates fine-tuned models spanning BERT-, Qwen-, LLaMA-, and Gemma-type architectures, trained using labels produced through a human-in-the-loop procedure supported by ChatGPT, and we benchmark performance against an unfine-tuned LLaMA 70B model used in zero-shot mode. A key concern in applications of LLMs to economic measurement is look-ahead bias (Sarkar and Vafa, 2024; Bybee, 2023). We address this directly: the LLMs are used for text classification rather than forecasting; they are not provided with timestamps or temporal markers; all RIS construction relies on strictly backward-looking transformations; and all empirical exercises are implemented in a pseudo-real-time setting with recursive expanding windows, a fixed train–test-split, and explicit data-availability constraints.

This paper asks whether LLMs can convert Reddit discussions into useful predictors of U.S. inflation. We treat LLMs as *economic sensors*: rather than forecasting inflation directly (as in Alam et al., 2026 or Faria-e-Castro and Leibovici (2024)), they extract structured information from

text that can be embedded in standard econometric models. We focus on three large subreddits—`r/Economics`, `r/economy`, and `r/wallstreetbets`—and construct RIS variants from daily *directional* classifications. Each item is classified according to the direction of its inflation- or price-related content: rising or increasing (UP), falling or decreasing (DOWN), or no clear directional signal (NEUTRAL).

A first contribution is methodological. We propose a transparent pipeline that converts raw Reddit content into RIS variants suitable for real-time analysis. The pipeline combines (i) keyword filtering to identify inflation-related content, (ii) LLM-based geographic screening to retain U.S.-relevant submissions, (iii) supervised (and parameter-efficient) fine-tuning to classify directional inflation content at scale, and (iv) strictly backward-looking aggregation and smoothing to map daily classifications into monthly RIS variants. All transformations are designed to prevent look-ahead bias. Forecasting and nowcasting are implemented in a pseudo-real-time setting with recursive expanding windows, a fixed train–test split for the forecasting exercise (2009M1–2012M12 vs. 2013M1–2025M8), and explicit constraints reflecting data availability.

A second contribution is empirical. In a recursive pseudo out-of-sample evaluation with horizons up to 18 months, RIS-based models deliver systematic gains for both headline CPI and core PCE inflation. Improvements are largest at short horizons (up to six months ahead), where forecasting gains are usually hardest to achieve. Forecast combinations based on in-sample MSE weights (Clark and McCracken, 2010) further strengthen performance and frequently remain in the final set of superior models under the Model Confidence Set (MCS) procedure. In real-time nowcasting exercises using ALFRED vintages, the RIS constructed using only information available early in the month (e.g., up to day 5, 10, and 14, and up to day 22 for PCE) improves nowcasts relative to an $AR(1)$ benchmark and performs competitively with the Cleveland Fed Inflation Nowcast (Knotek and Zaman, 2017). The two approaches appear complementary: the Fed nowcast benefits from fast-moving energy-price information during the inflation surge, while the RIS adds value during the subsequent disinflation phase.

A third contribution is a comprehensive set of robustness and interpretation exercises. We show that the main results are not sensitive to the choice of benchmark by repeating the analysis using the unobserved-components stochastic-volatility (UCSV) model. Results are also robust to alternative point-forecast loss functions (MAE and MAD) and to restricting the evaluation to a pre-COVID sample, indicating that the gains are not mechanically driven by the COVID-19 and energy-price episodes. We further document incremental predictive content through full-sample predictive regressions (incremental adjusted $R^2$ with HAC inference). For density forecasting, quantile-

regression-based predictive distributions improve calibration as measured by the Continuous Ranked Probability Score (CRPS). Finally, two external-validation exercises support the economic interpretation: (i) the RIS frequently Granger-causes Michigan expectations, consistent with high-frequency online discussions anticipating survey-based expectations; and (ii) the RIS co-moves strongly and primarily contemporaneously with a newspaper-based sentiment index, suggesting that Reddit acts as a real-time amplifier of news while adding a layer of social interpretation through comments.

The paper also speaks to feasibility. Fine-tuned small language models (SLMs) often perform comparably to much larger LLMs, and parameter-efficient fine-tuning delivers large computational savings with little or no loss in classification accuracy. This makes the approach deployable in resource-constrained environments and suitable for routine monitoring.

To enhance the replicability of our results and fine-tuning methodology, we provide public GitHub repositories for the fine-tuning methodology and the daily RIS. We also created the family of fine-tuned models used in the paper, `Reddit Infla-Pulse`, that can be accessed via Hugging Face at this link.

The remainder of the paper is organized as follows. Section 2 reviews recent research on text-as-data applications in economic forecasting. Section 3 describes the Reddit and inflation data and the filtering pipeline. Section 4 explains how LLMs are fine-tuned and used for geolocation, labeling, and RIS construction. Section 5 presents the forecasting and nowcasting results, together with supplementary analyses. Section 6 concludes. The Online Appendix reports additional robustness checks and supporting analyses, including density forecasts, predictive regressions, CSSED and fluctuation tests, the pre-COVID evaluation, the UCSV benchmark comparison, and evidence linking the RIS to expectations and news sentiment.

# 2 Previous studies of forecasting with textual and social-network data

The literature on extracting insights from textual data can broadly be divided into NLP and generative AI/LLM. NLP techniques provide foundations for more advanced generative models such as LLMs, with methods evolving from basic sentiment analysis to sophisticated language interpretation across diverse contexts.

In empirical macroeconomics and 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 used 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 bigrams and trigrams), they developed a measure closely aligned with survey- and market-based expectations.

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 Foundation

Models (TSFMs) 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 online discussion platform organized into topic-specific communities ("subreddits"), denoted by "*r/Subreddit_Name*."[1] Within each subreddit, users

[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.

("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 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 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 RIS variants aligned with monthly inflation data.

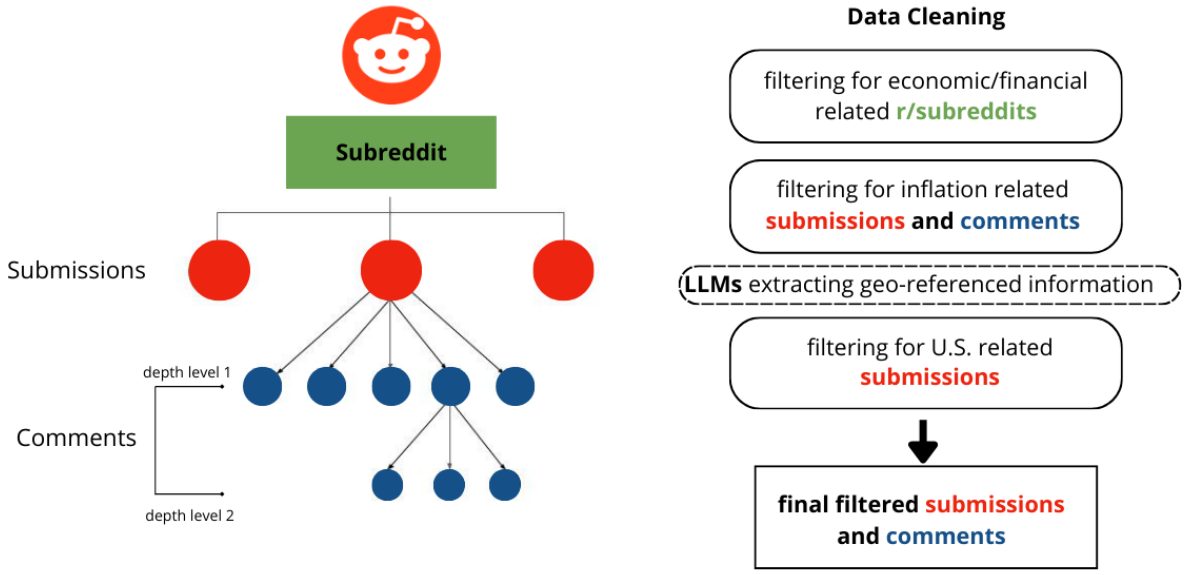
**Filtering and construction of the inflation-related text sample** The raw Reddit corpus is large and heterogeneous, and much of the content is unrelated to aggregate inflation (e.g., asset prices for `r/wallstreetbets`). We therefore implement a transparent multi-step filtering pro-

[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 no longer publishes member counts; instead it reports 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.

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

**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 the RIS.

cedure 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 content, 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.

**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.

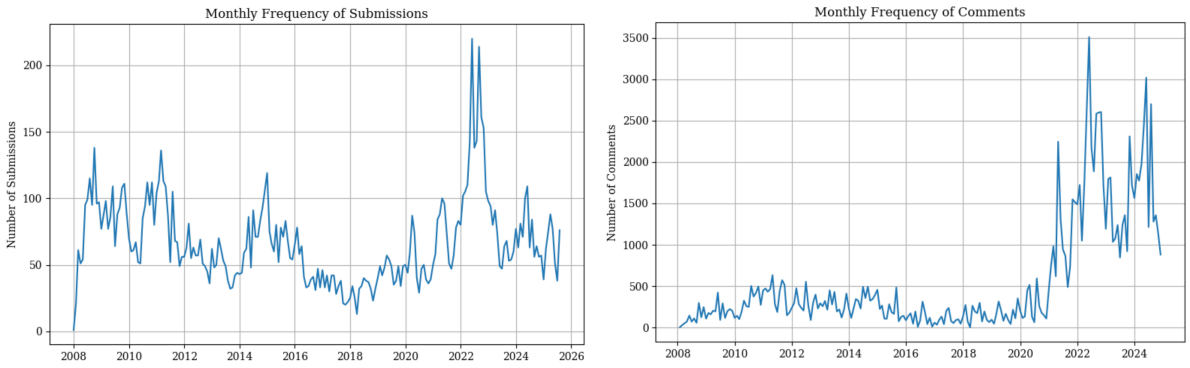
## 3.2 Descriptive evidence on volume and timing

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

[5] As robustness checks, we expand the lexicon to include terms such as "wages" and "jobs." The resulting RIS variants 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.

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 textual information.

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




**From filtered text to the RIS** The filtered text sample is subsequently mapped into RIS variants using large language models. Section 4 details the labeling tasks, fine-tuning strategy, and aggregation rules that translate model outputs into the monthly RIS variants used for nowcasting and forecasting inflation.

# 4 LLMs in action: From text to the RIS

We use LLMs to convert unstructured Reddit discussions into structured input for the RIS. 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 the time-series RIS variants 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 Figure A.2 in the Online Appendix. 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.

**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 that capture the direction of inflation-related content in Reddit submissions: UP, DOWN, and NEUTRAL.[7] The process begins with a manually labeled seed sample (about 500 submissions) and is expanded using a zero-shot LLaMA 70B classifier

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

[7] A submission is labeled UP if it conveys increasing prices/inflation, DOWN if it conveys decreasing prices/inflation, and NEUTRAL if it conveys no clear directional signal or suggests broadly stable prices/inflation.

(about 700 additional submissions).[8] 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).[9] 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 RIS extraction with fine-tuned LLMs

We fine-tune multiple model families to classify the direction of inflation-related content at scale across the full Reddit corpus. We consider (i) encoder-based models–BERT base, FinBERT, and InflaBERT[10]–trained via standard supervised fine-tuning, and (ii) decoder-only LLMs–LLaMA, Gemma, and Qwen[11]–adapted using parameter-efficient fine-tuning (PEFT) 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[12] on a single GPU (NVIDIA A100). Appendix D provides additional details on the model families and fine-tuning protocols.[13]

We evaluate classification performance using the weighted F1 score, which accounts for class

[8]Figure A.3 in the Online Appendix shows the prompt used to label all posts with the unfine-tuned LLaMA 70B, while Figure A.4 shows the prompt used to label the initial 700 posts.

[9]Figure A.5 in the Online Appendix reports the prompt used with ChatGPT.

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

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

[12]Online Appendix D.3 compares the baseline fine-tuning method (`QDoRA+`) with the more parameter-efficient variant (`xQDoRA+`) in a matched design across seeds and model architectures. We find that `xQDoRA+` delivers statistically significant reductions in wall time—especially for larger models—while weighted F1 differences are generally small and typically not statistically distinguishable from zero.

[13]Figures A.6 and A.7 show that moving from `QDoRA+` to the "extreme" configuration `xQDoRA+` reduces the number of trainable parameters sharply across all LLMs, yet yields weighted F1 scores that are essentially unchanged—and in some cases slightly higher—indicating that the computational savings do not come at a meaningful cost in classification accuracy.

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 RIS construction, we select the run whose seed yields 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–the UP, DOWN, and NEUTRAL classifications–provides the primary input for our empirical analysis. We aggregate these classifications into time-series RIS variants that enter the forecasting and nowcasting models. In this sense, LLMs are not forecasting models per se, as in Carriero et al. (2024); rather, they act as *economic sensors* that translate textual content into measurable predictors. We emphasize that the text underlying a directional label need not correspond to a forward-looking expectation in the strict sense: it may reflect a forward-looking belief about future prices, a perception or report of current price changes, or a reaction to a contemporaneous news item, components that cannot be cleanly disentangled at the post level. The RIS should therefore be interpreted broadly rather than as a pure measure of inflation expectations, in line with text-based measures that carry predictive content for inflation and real activity without mapping one-to-one into expectations (Hong et al., 2025; Jiang et al., 2025).

## 4.4 RIS construction

Using the LLM classifications, we construct the Reddit Inflation Signal (RIS), which summarizes the balance of directional inflation content in Reddit discussions. We distinguish between an intermediate submissions-only measure and our baseline RIS, which incorporates both submissions and comments. We denote the daily submissions-only measure by $\widetilde{\text{RIS}}_t^S$. After incorporating comments according to their actual posting dates, the resulting daily RIS is denoted by $\widetilde{\text{RIS}}_t$. Unless otherwise stated, RIS refers to this submissions-and-comments measure.

Let $\tau_i$ denote the calendar date on which submission $i$ is posted, and let $s_i \in \{-1, 0, 1\}$ denote its directional classification, where $-1$, 0, and 1 correspond to DOWN, NEUTRAL, and UP, respectively. The daily submissions-only RIS is then

$$\widetilde{\text{RIS}}_t^S = \sum_{i:\, \tau_i = t} s_i, \tag{1}$$

where $\widetilde{\text{RIS}}_t^S$ is set to zero on days with no qualifying submissions.[14]

Reddit's threaded structure enables us to enrich this submissions-only measure with the information contained in comments. For each inflation-related submission $i$, we link all associated comments via the parent ID and retain only inflation-related comments posted within two weeks of the submission. Importantly, this two-week restriction determines only whether a comment is eligible for inclusion; it does not assign the comment to the submission date. Each comment enters the information set on its own posting date.

Specifically, let $\tau_{ij}$ denote the calendar date on which the $j$-th comment associated with submission $i$ is posted, and let $c_{ij} \in \{-1, 0, 1\}$ denote its directional classification. For any calendar day $t \geq \tau_i$, define the set of eligible comments on submission $i$ that are actually available by day $t$ as

$$\mathcal{C}_i(t) = \{j : \tau_i \leq \tau_{ij} \leq \min(t, \tau_i + 14 \text{ days})\}, \tag{2}$$

and let

$$J_i(t) = |\mathcal{C}_i(t)|$$

denote the number of such comments. The real-time thread-level score is

$$\bar{s}_i(t) = \frac{s_i + \sum_{j \in \mathcal{C}_i(t)} c_{ij}}{1 + J_i(t)}. \tag{3}$$

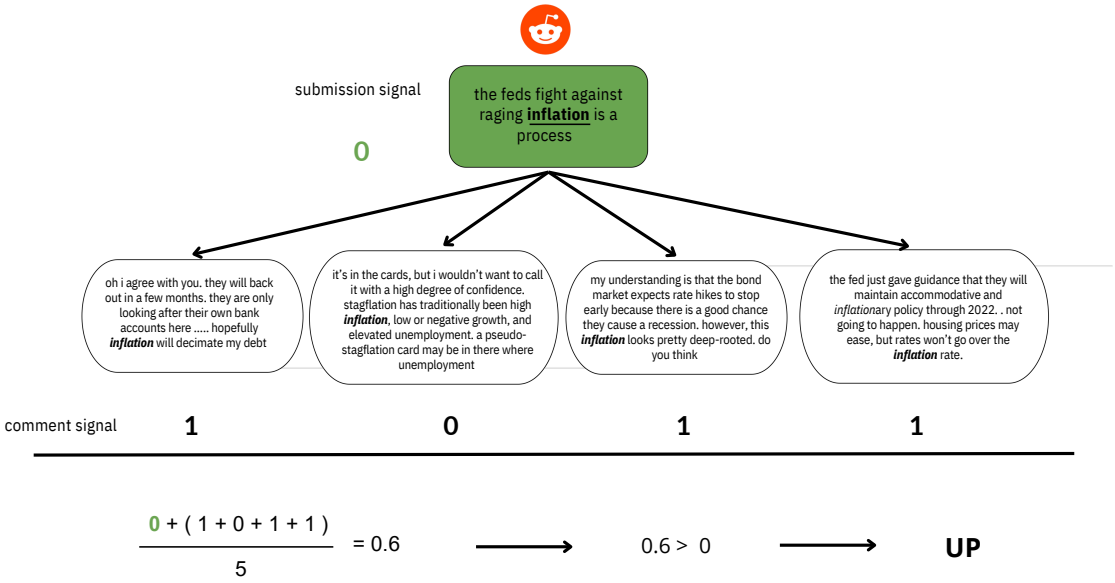
Thus, $\bar{s}_i(t)$ uses only information that is available by calendar day $t$. On the submission date, it reflects the submission together with any comments already posted that day; as additional comments arrive, the thread-level score is updated using those comments only from their actual posting dates onward. A comment posted on day $\tau_{ij}$ can therefore affect the RIS no earlier than $\tau_{ij}$ and is never backdated to the parent submission date $\tau_i$. This real-time dating is particularly important for the nowcasting exercise in Section 5.2, where the information set is truncated at within-month cutoffs preceding the official inflation release.

We map the real-time thread-level score $\bar{s}_i(t)$ back into a discrete directional classification. In the baseline, a positive score is classified as UP, a negative score as DOWN, and a zero score as NEUTRAL.[15]

[14] Equivalently, $\widetilde{\text{RIS}}_t^S$ is the difference between the number of UP and DOWN submissions posted on day $t$, with NEUTRAL submissions contributing zero.

[15] More generally, with threshold $thr \geq 0$, we assign UP when $\bar{s}_i(t) > thr$, DOWN when $\bar{s}_i(t) < -thr$,

**Figure 3:** Example of a submission–comment thread network




*Notes*: The initial submission is labeled NEUTRAL (0). As comments become available, the thread-level score $\bar{s}_i(t)$ is updated using only comments posted by day $t$. In the example, the resulting thread classification changes from NEUTRAL to UP.

The resulting comment-adjusted classifications are aggregated using their actual information dates to obtain our baseline daily RIS, $\widetilde{\text{RIS}}_t$. In particular, any revision to a thread classification induced by a newly posted comment enters the daily series when that comment becomes available, rather than being assigned retrospectively to the submission date. Henceforth, unless explicitly stated otherwise, RIS refers to this submissions-and-comments measure.

Figure 3 provides a schematic example in which a Gemma 27B classification is adjusted from NEUTRAL to UP as information from the associated comments becomes available.

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

and NEUTRAL when $|\bar{s}_i(t)| \leq thr$. Our baseline sets $thr = 0$. Results are not sensitive to small variations in this threshold; earlier versions used $thr = 0.10$ with similar results. This step can be interpreted as a regularization that adjusts the submission classification according to the directional content of the discussion that has become available by day $t$.

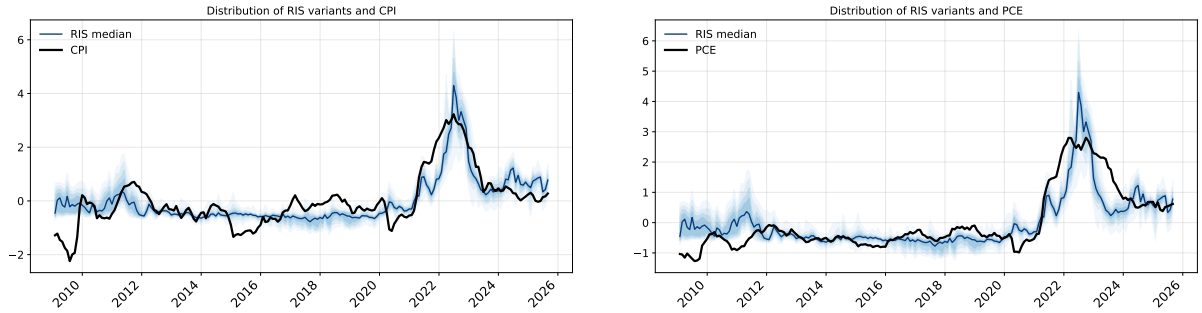
Given limited prior guidance on how the baseline daily RIS should be smoothed before entering forecasting regressions, we consider variants along three dimensions: (i) subreddit (`r/Economics`, `r/economy`, `r/wallstreetbets`); (ii) model used for labeling (BERT, FinBERT, InflaBERT, Qwen 0.5B, Qwen 1.5B, Qwen 7B, LLaMA 1B, LLaMA 3B, LLaMA 8B, Gemma 2B, Gemma 9B, Gemma 27B); and (iii) backward-looking moving-average (MA) windows of length $1, 5, 10, 30, 90, 120, 180$, and 360 days.[16] This yields $3 \times 12 \times 8 = 288$ candidate RIS variants.

More formally, let $\widetilde{\text{RIS}}^{(k)}_{r,m,t}$ denote the daily RIS for subreddit $r$, classifier $m$, and backward MA window $k$. After aggregation to monthly frequency, we denote the corresponding predictor by $\text{RIS}^{(k)}_{r,m,t}$. To keep the notation uncluttered, we suppress the subreddit, classifier, and smoothing indices whenever the particular variant is clear from context and simply write $\text{RIS}_t$. These monthly RIS variants are the measures used in the forecasting and nowcasting exercises below.

Figure 4 plots representative time series for the distribution of the 288 RIS variants alongside monthly inflation measured by headline CPI and core PCE.[17] All series are z-scored for visualization only; in the empirical analysis, RIS variants enter the models on their original scale. Importantly, no transformation is applied using future information: the backward moving average at each date uses only the current and preceding daily observations.

**Figure 4:** Monthly CPI and core PCE inflation alongside RIS variants

**(a)** CPI (*headline*) **(b)** PCE (*core*)




*Notes*: RIS variants are aggregated from daily to monthly frequency. All series are z-scored for comparability.

[16] Backward MAs use only current and past observations, ensuring that the transformation does not use future information.

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

# 5 Predictive performance of the RIS

This section evaluates whether the RIS contains 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 the RIS can complement standard macroeconomic predictors in a pseudo out-of-sample setting.

RIS variants summarize the directional content of inflation-related submissions and comments (Section 4.4). We aggregate the daily variants to monthly frequency and compare their predictive content with established benchmarks: lagged inflation, the University of Michigan inflation expectations measure, and the one-year inflation swap rate.[18]

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 implement a real-time nowcasting exercise (Section 5.2) that mimics the information flow faced by practitioners: the RIS is 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 the RIS adds explanatory power beyond lagged inflation and conventional expectations measures.[19] Fourth, Section 5.3.2 evaluates density forecasts using quantile regression and proper scoring rules. Two additional robustness checks provide supporting evidence for interpretation and external validity: Section 5.3.3 reports Granger-causality results linking the RIS to inflation expectations, and Section 5.3.4 compares the RIS with newspaper-based sentiment indices.

The construction of the RIS is intentionally exploratory. Because Reddit is rarely used in macroeconomic forecasting, we consider multiple subreddits, model families, and smoothing choices. In particular, the RIS is computed for several LLM specifications and multiple backward-looking moving-average (MA) windows, yielding 288 candidate variants (Section 4.4). Our objective is to document which modeling and smoothing choices improve predictive performance 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

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

[19] 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.

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, $AR(1)$, with augmented specifications that add one predictor at a time, including RIS variants. These variants 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 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. Following Bybee (2023), RIS variants can be interpreted as containing expectation-related information from a specific subgroup (Reddit users), while also reflecting perceptions of current prices. We therefore benchmark them against two widely used expectation measures: the University of Michigan inflation expectations series[20] 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 the RIS 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 $t$)<br>Revised (4th Friday of month $t$) | Previous month |
| One-Year Inflation Swap | Level | Real-time | Current month |
| RIS | 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 288 RIS variants, we summarize results using both (i) the best-performing single variant within a family and (ii) forecast combinations. In particular, we aggregate the RIS-

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

based forecasts using MSE-weighted combinations following Clark and McCracken (2010).[21] We construct two combinations: *Aggr-RIS*, the MSE-weighted aggregation across fine-tuned LLM/SLM RIS variants, and *Aggr-LLaMA 70B*, the MSE-weighted aggregation using the unfine-tuned LLaMA 70B benchmark. As additional comparisons, we also construct lexicon-based sentiment indices using standard tools (TextBlob, VADER) and the Loughran–McDonald (LM) dictionary (Loughran and McDonald, 2011).

The forecasting regression takes the form

$$\pi_{t+h} = \alpha + \beta_1 \pi_{t-1} + \gamma X_{\tilde{t}} + \varepsilon_{t+h}, \tag{4}$$

where $X_{\tilde{t}}$ denotes an additional predictor (Michigan expectations, the inflation swap rate, or a RIS variant $\text{RIS}_{\tilde{t}}$) observed at the most recent available date $\tilde{t}$.[22]

The initial estimation window spans 2009M1–2012M12, and the evaluation period spans 2013M1–2025M8.[23] Figure A.14 in the Online Appendix illustrates the recursive design for $h = 1$.

Tables 3 and 4 report out-of-sample accuracy using RMSE ratios relative to the $AR(1)$ benchmark for headline CPI and core PCE, respectively. For the RIS, we report Aggr-RIS and Aggr-LLaMA70B together with the best single variant within each model family (fine-tuned LLMs or unfine-tuned LLaMA 70B). Overall, RIS-based models deliver the largest gains at short horizons: up to six months ahead, 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.[24] In comparing fine-tuned LLMs with LLaMA 70B, the latter performs best overall, but RMSE ratios are often close in magnitude and statistical significance to those of fine-tuned LLMs. In addition, fine-tuned LLMs give significantly better forecasts than Michigan

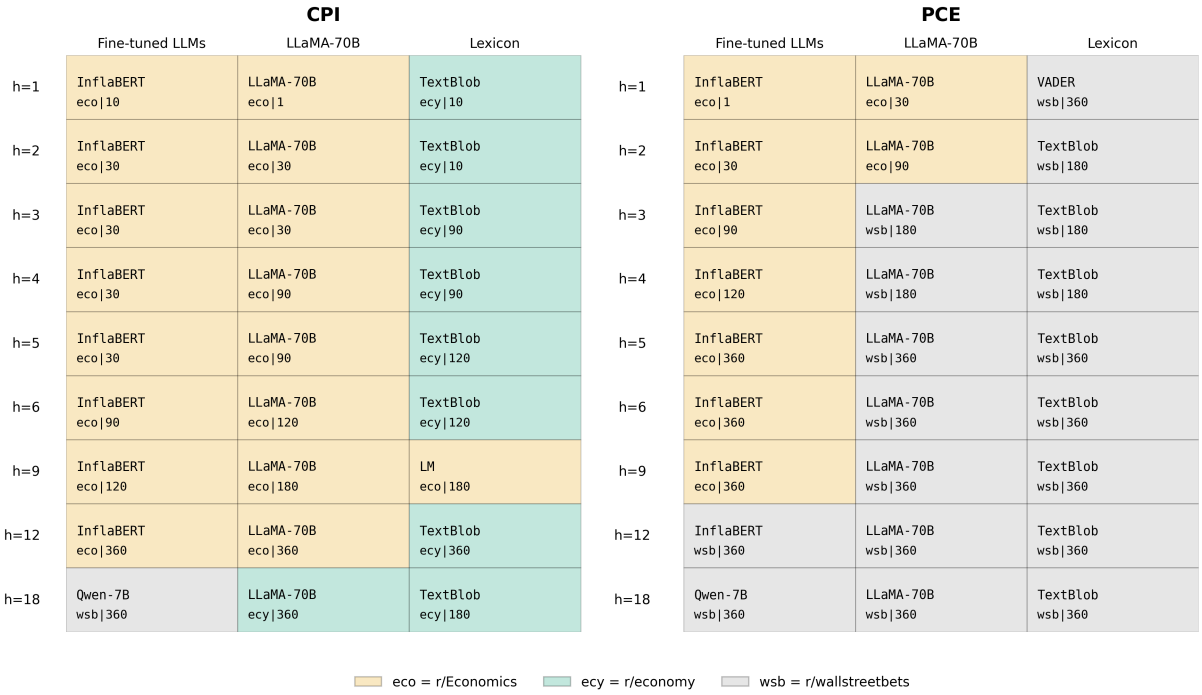
[21] We set the discount factor to $\delta = 0.95$ as in Clark and McCracken (2010). Tuning $\delta$ can improve performance, but we keep the baseline value for comparability.

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

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

[24] As a robustness check, Online Appendix E.8 repeats the point-forecast exercise using only the volume of inflation-related submissions (post counts)—smoothed with the same backward-looking MA windows—in place of the LLM-based directional component of the RIS. Post counts provide at best modest improvements at medium horizons and rarely at $h = 1$, with limited statistical significance; they are consistently dominated by the RIS-based measures, especially at short horizons. This suggests that predictive gains stem primarily from the semantic content extracted by LLMs rather than from attention intensity alone.

**Figure 5:** 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.

expectations until $h = 4$ for CPI and until $h = 12$ for PCE.[25]

Figure 5 summarizes, by horizon, which subreddit and MA window are selected among the best-performing specifications within each family (fine-tuned LLMs and unfine-tuned LLaMA 70B). 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 RIS variants. The best fine-tuned model is InflaBERT at all horizons except $h = 18$, where Qwen 7B has the lowest forecast error. RIS variants from `r/Economics` tend to dominate for CPI, while for PCE most predictive content comes from `r/wallstreetbets`.

[25]Tables A.15–A.16 in the Online Appendix report the point-forecast comparison using the unobserved-components stochastic-volatility (UCSV) model, a standard univariate benchmark in the U.S. inflation-forecasting literature (Stock and Watson, 2007; Faust and Wright, 2013), as an alternative benchmark. While $AR(1)$ tends to perform better at very short horizons and UCSV is more competitive at longer horizons, the main ranking is unchanged: RIS-based models—especially Aggr-LLaMA70B—remain among the best performers at short horizons for both headline CPI and core PCE.

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

| | $h=1$ | $h=2$ | $h=3$ | $h=4$ | $h=5$ | $h=6$ | $h=9$ | $h=12$ | $h=18$ |
|---|---|---|---|---|---|---|---|---|---|
| *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* |
| *Lexicon-based sentiment* | | | | | | | | | |
| Best sentiment | 0.945 | 0.898 | 0.854 | 0.807 | 0.769 | 0.738 | **0.636*** | **0.555*** | *0.564** |
| *RIS-based models* | | | | | | | | | |
| Aggr-RIS | 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 |
| Aggr-LLaMA70B | *0.813*** | **0.733*** | **0.689*** | **0.653*** | **0.632*** | **0.623*** | *0.641* | 0.655 | 0.699 |
| Best LLaMA 70B | **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.

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

| | $h=1$ | $h=2$ | $h=3$ | $h=4$ | $h=5$ | $h=6$ | $h=9$ | $h=12$ | $h=18$ |
|---|---|---|---|---|---|---|---|---|---|
| *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* |
| *Lexicon-based sentiment* | | | | | | | | | |
| Best sentiment | 0.970 | 0.952 | 0.879 | 0.824 | *0.762* | **0.703** | **0.549*** | **0.451*** | **0.393*** |
| *RIS-based models* | | | | | | | | | |
| Aggr-RIS | 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** |
| Aggr-LLaMA70B | **0.901**** | **0.839**** | **0.791**** | **0.750*** | **0.727*** | *0.723** | 0.720 | 0.721 | 0.719 |
| Best LLaMA 70B | *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.

The results in Table 4 for PCE are very similar to those obtained for CPI. RIS-based models outperform the benchmark at all horizons. The fine-tuned models remain competitive, especially the InflaBERT specifications, which outperform expectations at all horizons. Overall, the evidence confirms that the RIS contains useful information for forecasting 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 5.

Figure 6 plots the cumulative sum of squared error differences (CSSED) for CPI and PCE at the one- and six-month-ahead horizons. CSSED paths 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, Aggr-LLaMA70B performs best, followed by Aggr-RIS. 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.[26]

Section E.1 in the Online Appendix reports forecasting results evaluated with MAE and MAD losses (which downweight large errors). The RMSE ranking is confirmed: RIS-based models deliver the largest gains at short horizons (typically $h \leq 6$), while traditional benchmarks become relatively more competitive at longer horizons. Across loss functions, Aggr-LLaMA70B is generally strongest at short horizons, with Aggr-RIS close behind and typically outperforming Michigan-expectations-based forecasts. The best *individual* RIS variants are often generated by fine-tuned small language models (e.g., Qwen 7B or LLaMA 3B), highlighting a resource-efficient route to deployment. Section E.2 in the Online Appendix shows that these conclusions persist in a pre-COVID sample: LLaMA 70B (single or Aggr-LLaMA70B) remains particularly strong at short-to-medium horizons, InflaBERT is the most reliable fine-tuned model (with Qwen 7B emerging at $h = 18$), and the most informative subreddit differs by target (`r/Economics` for CPI vs. `r/wallstreetbets` for PCE).

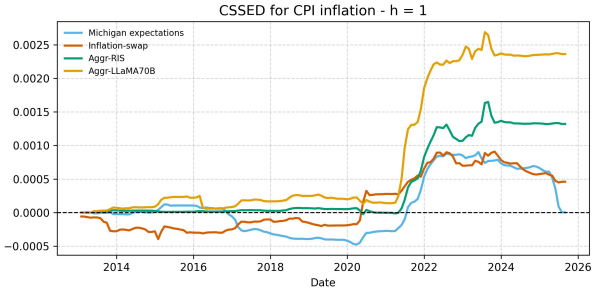
### 5.1.1 Additional robustness tests for point forecasts

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

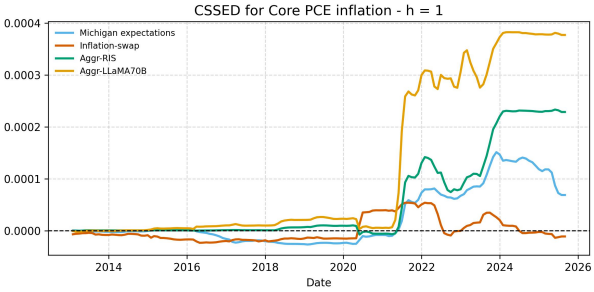
[26] Figures A.23–A.24 in the Online Appendix report CSSED paths for CPI and PCE across all forecast horizons. The qualitative ranking is consistent with Figure 6: Aggr-RIS and Aggr-LLaMA70B accumulate the largest gains at short-to-medium horizons, with improvements most pronounced around the COVID and energy-price episodes. At the longest horizons, the gap narrows and Michigan-expectations-based forecasts become relatively more competitive.

**Figure 6:** CSSED for CPI (headline) and PCE (core) - $h = (1, 6)$

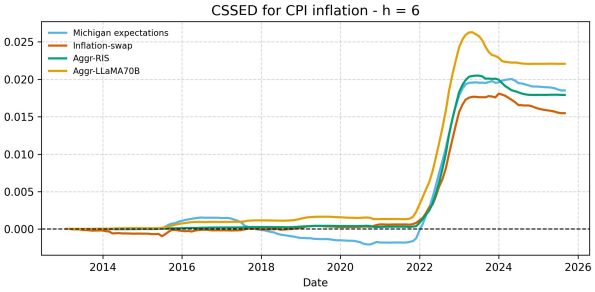
**(a)** $CPI, h = 1$



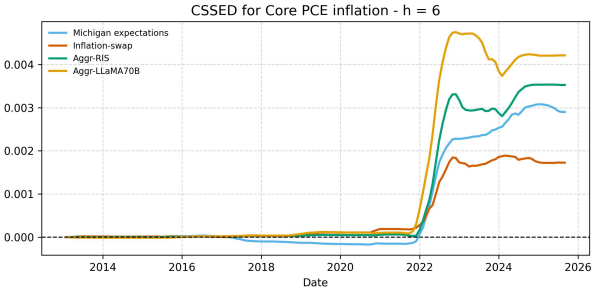

**(b)** $PCE, h = 1$




**(c)** $CPI, h = 6$




**(d)** $PCE, h = 6$




*Notes*: CSSED is computed relative to the AR(1) benchmark. Aggr-RIS and Aggr-LLaMA70B are the corresponding MSE-weighted forecasts. Values above (below) zero indicate better (worse) performance of the alternative model.

standard Diebold-Mariano test may be sensitive to these features. We therefore complement the baseline comparisons with two additional tests.

First, 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 7a and 7b 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 RIS 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 Reddit information may 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 available information.[27] Together with the CSSED, these time-local diagnostics corroborate the headline RMSE results and clarify that the RIS is most informative when inflation dynamics are shifting rapidly.

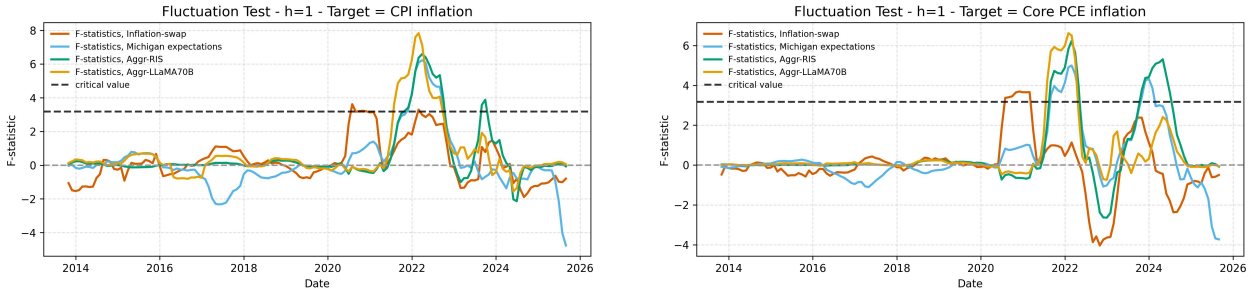
Second, we apply the MCS procedure of P. R. Hansen et al. (2011), which constructs, for each horizon, a set of models that are statistically indistinguishable from the best-performing model at a given confidence level. Unlike pairwise comparisons, the MCS explicitly accounts for multiple testing and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

The left and right panels of Figure 8 summarize MCS inclusion patterns for CPI and PCE, respectively. At short horizons, RIS-based models (including LLaMA 70B, InflaBERT, and selected smaller LLaMA variants) enter the MCS frequently, and Aggr-RIS and Aggr-LLaMA70B remain in the MCS over several horizons. By contrast, Michigan expectations enter the MCS more consistently at longer horizons, and lexicon-based sentiment models tend to appear primarily at medium to longer

[27] Figures A.27–A.28 in Appendix E.3 report fluctuation-test charts across all forecast horizons for CPI and PCE, respectively, and confirm that performance differences are largely muted before 2019, while Aggr-RIS and Aggr-LLaMA70B become locally and statistically superior in 2020–2023 across horizons.

**Figure 7:** Fluctuation tests for forecast differences at $h = 1$.

**(a)** CPI (*headline*) **(b)** PCE (*core*)




*Notes*: The figure reports fluctuation tests for forecast differences for headline CPI and core PCE. The panels assess whether relative forecast performance is stable over time.

horizons. The composition plots in Figures A.29–A.30 in the Online Appendix 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`.

Overall, the evidence indicates that RIS-based models 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: the RIS delivers 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.

## 5.2 Nowcasting application

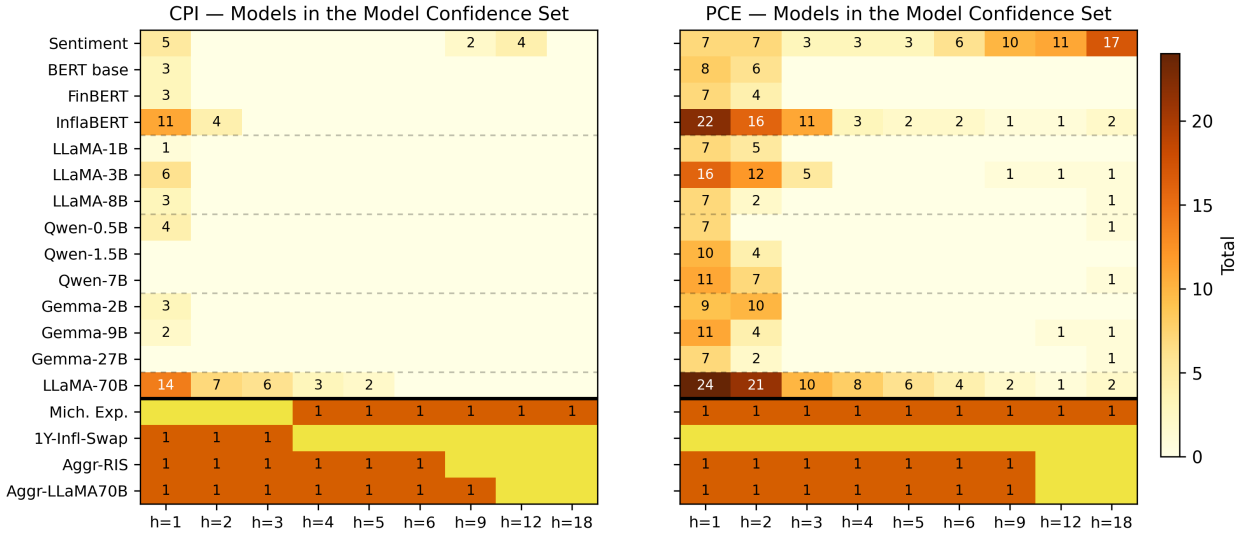
We next implement a nowcasting exercise that uses real-time vintages of CPI and PCE from ALFRED.[28] The sample split differs from the forecasting exercise to reflect both vintage availability and the construction of competitive benchmarks. The estimation window is 2009M1–2013M7 and the evaluation window is 2013M8–2025M5.[29]

Nowcasting introduces a ragged-edge setting: the RIS is observed daily, while CPI and PCE are

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

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

**Figure 8:** Model confidence set results for CPI and PCE.




*Notes*: For each horizon and model family, each heatmap reports which models enter the MCS (aggregates are coded as 1 if included and colour-coded accordingly).

released monthly with a delay. This feature allows us to quantify how much incremental information the RIS 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 $\pi_t$. 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 RIS information is always used strictly prior to the official release.[30]

Operationally, we estimate the following nowcasting regression

$$\pi_t = \alpha + \beta_1 \pi_{t-1} + \gamma X_{t+c} + \varepsilon_t, \tag{5}$$

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

[30]This constraint is respected at the level of individual comments as well: because each comment enters the daily series on its own date rather than being backdated to the parent submission, any comment posted on or after the release day of $\pi_t$ is automatically excluded from the information set used at each within-month cutoff. The cutoffs therefore use only Reddit content that was genuinely available before the official release.

**Table 5:** Nowcast results for CPI and PCE: RMSE and MAE ratios relative to the AR(1) benchmark.

| Panel A | | **CPI (headline)** | | | |
|---|---|---|---|---|---|
| Loss | Model | +5 days | +10 days | +14 days | +22 days |
| RMSE | Aggr-RIS | 0.949** | 0.943*** | 0.914*** | — |
| | Aggr-LLaMA70B | 0.869*** | 0.853*** | 0.825*** | — |
| MAE | Aggr-RIS | 0.957** | 0.948** | 0.918*** | — |
| | Aggr-LLaMA70B | 0.877*** | 0.854*** | 0.820*** | — |
| Panel B | | **PCE (core)** | | | |
| RMSE | Aggr-RIS | 0.984* | 0.997 | 0.963** | 0.880* |
| | Aggr-LLaMA70B | 0.968 | 0.962* | 0.945** | 0.940** |
| MAE | Aggr-RIS | 0.979* | 0.991 | 0.971* | 0.917** |
| | Aggr-LLaMA70B | 0.975 | 0.970 | 0.959 | 0.951 |

*Notes*: Entries report loss ratios relative to the AR(1) benchmark, computed as $loss(AR\text{-}X(1))/loss(AR(1))$. For the regressions with Reddit, the initial in-sample period spans 2009M1–2013M7, while the out-of-sample period covers 2013M8–2025M5. Nowcasts are generated recursively, with the model re-estimated each month. Significance levels are based on the Diebold–Mariano test $p$-values: $p < 0.01$ (***) $p < 0.05$ (**) $p < 0.1$ (*).

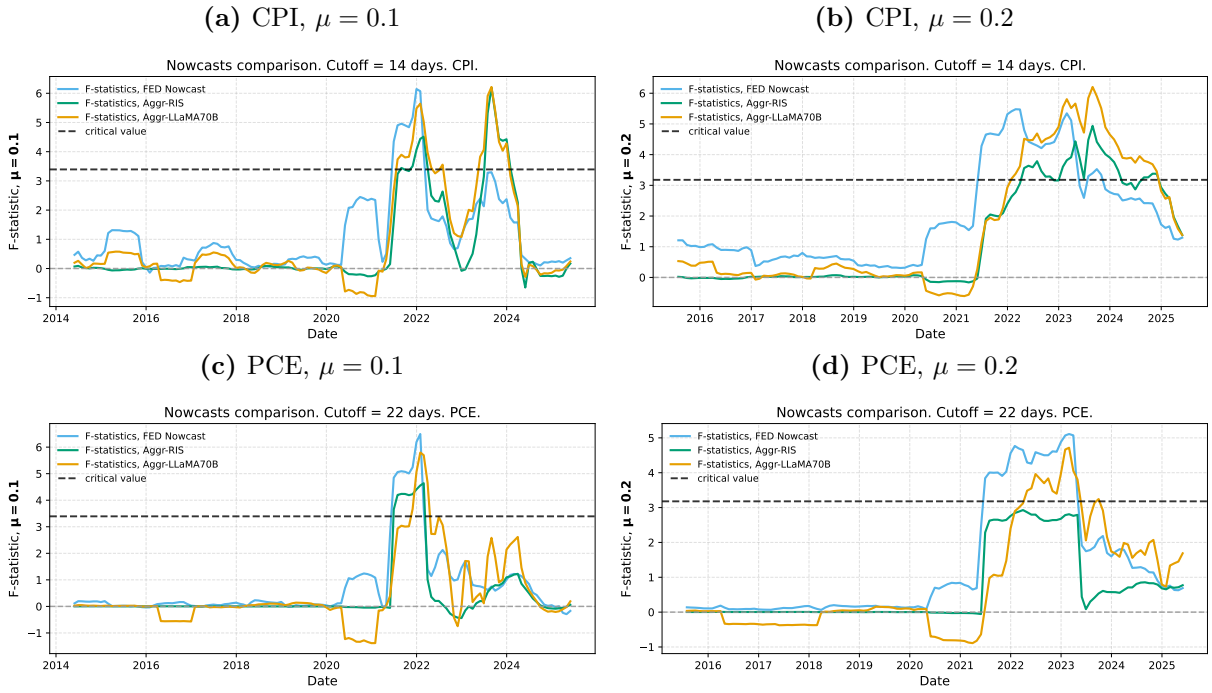
Panels A and B of Table 5 report nowcast accuracy in RMSE and MAE ratios relative to the $AR(1)$ benchmark for CPI and PCE, respectively. Two patterns stand out. First, for both CPI and PCE, Aggr-RIS and Aggr-LLaMA70B outperform $AR(1)$ across cutoffs. Second, using more within-month RIS information tends to improve accuracy, particularly for PCE, where the later cutoffs (+14 and +22 days) deliver the clearest gains.

### 5.2.1 Comparison with the Federal Reserve inflation nowcast

To benchmark the information content of the RIS, we compare our nowcasts with 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:

**Figure 9:** Fluctuation test statistics for nowcasts.




+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.[31]

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), while the RIS is available only from 2008 onward.[32] Moreover, the Fed nowcast explicitly exploits high-frequency commodity and energy-price information, which we deliberately exclude to isolate the predictive content of the RIS.

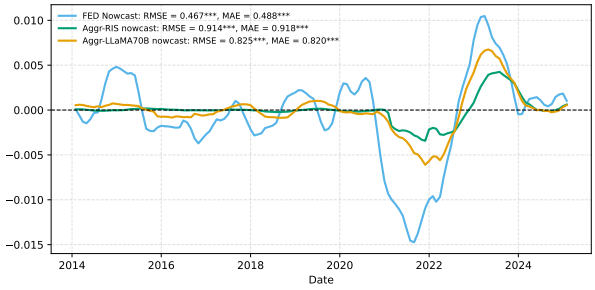
Figure 9 reports fluctuation-test statistics for nowcasts using two window lengths ($\mu = 0.1$ and $\mu = 0.2$). While pointwise significance depends on the window choice, the overall patterns are stable. For both CPI and PCE, the Fed nowcast outperforms $AR(1)$ primarily during the initial phase of

[31]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.

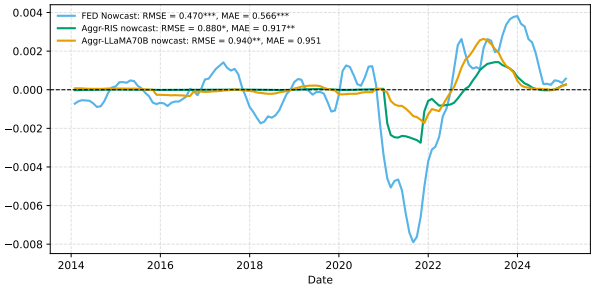
[32]This difference may affect comparability in ways that are not fully under our control.

**Figure 10:** Ten-period moving average of the differential loss of RIS- and Fed-based nowcasts relative to an AR(1) benchmark

**(a)** CPI




**(b)** PCE




*Notes*: The figure reports 10-period moving averages of the differential loss for Aggr-RIS, Aggr-LLaMA70B, the Fed nowcast, and the AR(1) benchmark. Values below zero indicate that the corresponding nowcast performs better than the AR(1) benchmark.

the post-pandemic inflation surge, when energy prices moved sharply. Aggr-LLaMA70B shows a similar profile, with gains concentrated in the high-inflation period. By contrast, Aggr-RIS tends to improve upon $AR(1)$ later in the sample: for CPI, this is particularly visible during the disinflation phase of 2023–2024, when its fluctuation statistic remains elevated even as the Fed nowcast recedes. For PCE, the RIS is less informative, with gains concentrated in Aggr-LLaMA70B, which exceeds the critical value during the 2022–2023 surge, whereas Aggr-RIS adds comparatively little beyond the $AR(1)$ benchmark for this target.

Figure 10 reports moving-average differential loss relative to $AR(1)$, confirming that the Fed nowcast achieves the lowest RMSE on average over the full sample. At the same time, for PCE the RIS-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. The RIS, by construction, captures inflation-related discourse—including perceived price pressures and expectations—that may be less tightly linked to commodity prices and appears particularly informative during the subsequent disinflationary phase.

## 5.3 Robustness and Economic Interpretation

### 5.3.1 Predictive regressions (in-sample evidence)

As a complementary check to the pseudo out-of-sample forecasting and nowcasting exercises, we estimate predictive regressions on the full sample to assess whether the RIS contains incremental information about future inflation beyond standard predictors. Specifically, for each horizon $h$, 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 RIS variant. We evaluate the marginal contribution of the RIS by testing whether its inclusion produces a statistically significant increase in adjusted $R^2$ (F-tests with HAC standard errors). Figure 11a illustrates the results for $h = 1$ using `r/Economics`: adding the RIS yields a positive and frequently statistically significant improvement in fit across a wide range of moving-average smoothing windows. More generally, across subreddits and horizons, RIS variants deliver systematic, statistically significant gains in explanatory power for both CPI and PCE, consistent with the interpretation that they capture information not contained in conventional survey- and market-based expectations measures.[33]

### 5.3.2 Density forecasts via quantile regression

To complement the point-forecast evidence, we also assess whether RIS variants improve *density* forecasts of inflation. We estimate quantile-regression versions of our baseline forecasting model, reconstruct predictive distributions from multiple conditional quantiles, and evaluate accuracy using the Continuous Ranked Probability Score (CRPS). Overall, RIS-based models improve distributional forecasts relative to the $AR(1)$ benchmark. The largest CRPS gains arise from LLaMA 70B-based RIS variants (especially from `r/Economics`) at short horizons ($h \leq 6$), while expectation-based measures become relatively more competitive at longer horizons (Tables A.17–A.18 in the Online Appendix). Beyond average performance, the quantile-regression results show that the partial association between the RIS and future inflation varies across the conditional distribution (Figure A.33 in the Online Appendix). In particular, the relationship tends to be stronger in high-inflation states, consistent with state-dependent attention and expectation formation. This pattern is also consistent with the mechanism emphasized in the recent literature: when inflation is elevated, *coverage intensity* and public attention increase, which can amplify the informativeness of text-based measures and strengthen their predictive content.[34]

[33] Full results and additional figures are reported in the Online Appendix (Appendix E.6).

[34] Full methodological details and additional results are provided in the Online Appendix (Appendix E.7).

### 5.3.3 Does the RIS anticipate inflation expectations?

To provide additional economic grounding for the RIS, we examine whether it helps predict survey-based inflation expectations. We estimate VARs including inflation, Michigan expectations, the one-year inflation swap rate, and a RIS variant, and we test for Granger causality between the RIS and expectations across model families, subreddits, and smoothing windows. The results in Table 6 show that RIS variants, especially those from thematically coherent communities such as `r/Economics` or `r/wallstreetbets`, as well as the cross-subreddit aggregate, frequently Granger-cause Michigan expectations, while the reverse direction is observed less often; bidirectional causality arises only in a limited number of specifications (see Table A.23 in the Online Appendix). These patterns are consistent with the interpretation that high-frequency online discussions both reflect and, in some cases, anticipate shifts in inflation expectations.[35]

**Table 6:** Summary of Granger-causality results between Michigan survey expectations and the RIS across subreddits and the cross-subreddit aggregate

| Model | r/Economics | r/economy | r/wallstreetbets | Aggregate |
|---|---|---|---|---|
| BERT base | $\oplus$ | $\star\ \nabla$ | $\square$ | – |
| FinBERT | $\circ\ \square\ \triangle\ \diamond\ \star^{\leftrightarrow}\ \oplus$ | – | $\diamond\ \star\ \nabla\ \oplus$ | $\circ\ \diamond\ \nabla\ \oplus$ |
| InflaBERT | $\circ\ \square\ \triangle\ \diamond^{\leftrightarrow}\ \star\ \nabla^{\leftrightarrow}\ \oplus^{\leftrightarrow}$ | $\nabla$ | $\star\ \nabla\ \oplus$ | $\circ\ \square\ \triangle\ \diamond^{\leftrightarrow}\ \star\ \nabla^{\leftrightarrow}\ \oplus$ |
| Gemma 2B | $\circ\ \square\ \triangle\ \diamond\ \star\ \nabla\ \oplus$ | $\nabla$ | $\star\ \oplus$ | $\star\ \nabla$ |
| Gemma 9B | – | – | $\star$ | $\star\ \oplus$ |
| Gemma 27B | $\otimes$ | – | – | – |
| LLaMA 1B | $\circ\ \diamond\ \star\ \nabla\ \oplus\ \otimes$ | – | – | $\star$ |
| LLaMA 3B | $\circ\ \square\ \triangle\ \diamond\ \star^{\leftrightarrow}\ \nabla^{\leftrightarrow}\ \oplus^{\leftrightarrow}$ | $\oplus$ | $\star^{\leftrightarrow}\ \nabla^{\leftrightarrow}\ \oplus^{\leftrightarrow}$ | $\circ\ \square\ \triangle\ \diamond\ \star\ \nabla^{\leftrightarrow}\ \oplus$ |
| LLaMA 8B | $\circ\ \square\ \triangle\ \diamond\ \nabla^{\leftrightarrow}\ \oplus$ | – | – | $\triangle\ \diamond$ |
| LLaMA 70B | $\circ\ \square\ \triangle\ \diamond\ \star\ \nabla^{\leftrightarrow}\ \oplus\ \otimes$ | $\nabla\ \oplus\ \otimes$ | $\star\ \nabla$ | $\circ\ \square\ \triangle\ \diamond^{\leftrightarrow}\ \star^{\leftrightarrow}\ \nabla^{\leftrightarrow}\ \oplus$ |
| Qwen 0.5B | $\triangle\ \diamond\ \star\ \nabla\ \oplus\ \otimes$ | $\oplus\ \otimes$ | $\star\ \nabla^{\leftrightarrow}\ \oplus^{\leftrightarrow}$ | $\circ\ \square\ \triangle\ \diamond^{\leftrightarrow}\ \star\ \oplus$ |
| Qwen 1.5B | $\diamond\ \star^{\leftrightarrow}$ | $\circ\ \square\ \triangle\ \diamond\ \star\ \nabla\ \oplus$ | $\star$ | $\circ\ \diamond\ \star\ \nabla^{\leftrightarrow}\ \oplus$ |
| Qwen 7B | – | $\nabla^{\leftrightarrow}$ | $\circ\ \square\ \triangle\ \nabla\ \oplus$ | – |

Notes: Symbols identify MA windows as follows: $\circ = 1$, $\square = 5$, $\triangle = 10$, $\diamond = 30$, $\star = 90$, $\nabla = 120$, $\oplus = 180$, $\otimes = 360$. A plain symbol indicates unidirectional Granger causality from the relevant RIS measure to survey expectations. A symbol with superscript $\leftrightarrow$ indicates bidirectional causality. Cells marked – indicate no evidence that the RIS predicts expectations at any MA window.

[35] Full details are provided in the Online Appendix (Appendix F).

### 5.3.4 RIS and news sentiment: overlap and differences

To clarify what the RIS measures, we examine its relationship with an established newspaper-based sentiment index from Barbaglia et al. (2022). That index uses economics-related articles from six major U.S. newspapers accessed through Dow Jones Factiva and applies the Fine-Grained Aspect-based Sentiment (FiGAS) framework of Consoli et al. (2022), which isolates text segments semantically linked to specific economic concepts (e.g., *inflation*) and assigns polarity scores using a domain-specific dictionary.

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[36]), whereas `r/wallstreetbets` is dominated by user-generated content. Second, despite these different content-generation mechanisms, the RIS closely co-moves with the news-sentiment series and exhibits a strong contemporaneous correlation, with limited evidence of systematic leads or lags (see Figure 11b for `r/Economics` and Figures A.36–A.37 in the Online Appendix for all subreddits). 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.[37] These patterns, together with the Granger-causality evidence in Section 5.3.3, are consistent with interpreting the RIS as a composite text-based measure in which reactions to news, perceptions of current prices, and forward-looking beliefs coexist, rather than as a pure measure of any one of them. Fully separating these channels—for instance, isolating the share attributable to news reactions versus genuinely forward-looking beliefs—is a research question that we leave to future research.

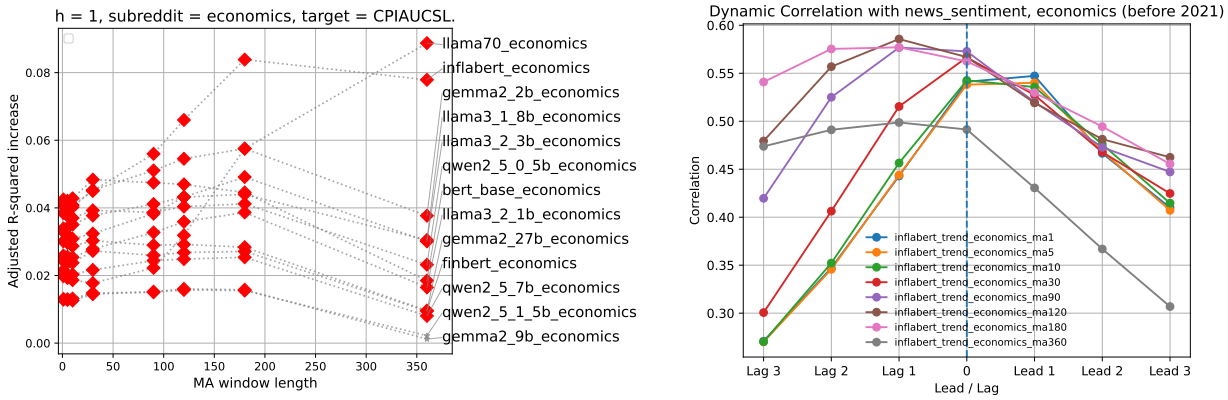
[36]Within Reddit, a "flair" denotes a granular label used to classify submissions (Post Flair) or identify user-specific credentials and affiliations (User Flair) within a given subreddit.

[37]Full evidence and figures are reported in the Online Appendix (Appendix G).

**Figure 11:** Robustness: Predictive regressions for CPI at $h = 1$ and lead–lag correlations between the RIS and newspaper-based sentiment (`r/Economics`)

**(a)** Predictive regression for CPI at $h = 1$ (`r/Economics`)

**(b)** Lead–lag correlations: RIS vs. newspaper sentiment (`r/Economics`)




*Notes*: **Left panel:** Incremental change in adjusted $R^2$ when adding a RIS variant $\text{RIS}_t$ (constructed from `r/Economics` and smoothed using the MA window on the horizontal axis) to a baseline predictive regression for one-month-ahead CPI inflation ($h = 1$) that includes lagged inflation, Michigan expectations, and the one-year inflation swap rate. The sample spans 2009M1–2025M8. Markers indicate rejection of the null hypothesis that the coefficient on $\text{RIS}_t$ is zero at the 5% level. **Right panel:** Lead–lag correlations between the `r/Economics` RIS variant constructed using InflaBERT and the newspaper-based sentiment index; the contemporaneous correlation (lag 0) is typically the largest.

# 6 Conclusion

This paper shows that large language models can transform Reddit discussions into timely and economically meaningful predictors of U.S. inflation. We treat LLMs as *measurement devices*: they map unstructured text into structured, interpretable RIS variants that can be embedded in standard econometric forecasting and nowcasting frameworks. Our empirical focus is on headline CPI and core PCE inflation, and our central question is whether the RIS adds predictive content relative to widely used benchmarks.

A first contribution is methodological. We develop a transparent, end-to-end pipeline that converts Reddit submissions and time-local comments into monthly RIS variants suitable for real-time monitoring. Each step—keyword filtering, LLM-based geographic screening, human-in-the-loop directional labeling (UP/DOWN/NEUTRAL), and backward-looking aggregation and smoothing—

is designed to avoid look-ahead bias. Forecasting and nowcasting are implemented in a pseudo-real-time setting with recursive expanding windows, a fixed train–test split (2009M1–2012M12 for estimation and 2013M1–2025M8 for evaluation in the forecasting exercise), and explicit data-availability constraints.

Empirically, RIS-based models deliver robust gains in point forecasting. In a recursive pseudo out-of-sample design with horizons up to 18 months, both the best single RIS variants and, especially, MSE-weighted forecast combinations improve accuracy relative to an $AR(1)$ benchmark. The gains are strongest at short horizons (up to six months ahead), where inflation is typically hardest to forecast due to transitory shocks. Relative to conventional predictors, RIS-based forecasts outperform autoregressive models augmented with survey-based expectations (Michigan) and market-based expectations (the one-year inflation swap) over short-to-medium horizons and remain competitive at longer horizons, where expectation-based measures tend to become relatively more informative. These results are reinforced by the MCS procedure, which shows that RIS-based models and forecast combinations frequently remain in the final set of superior models at short horizons.

The nowcasting results further highlight the operational value of the RIS. Using real-time vintages of CPI and PCE (ALFRED) and a ragged-edge design that exploits daily Reddit information, we show that the RIS improves nowcasts of the most recently completed month relative to $AR(1)$ when using information available only up to early-month cutoffs (e.g., day 5, 10, and 14; and up to day 22 for PCE). When benchmarked against the Cleveland Fed Inflation Nowcast, the two approaches appear complementary: the Fed nowcast tends to perform best during sharp commodity-driven movements (notably the post-pandemic inflation surge), while the RIS is comparatively more informative during the subsequent disinflation phase, consistent with Reddit capturing perceived price pressures and expectation-related discussion rather than high-frequency energy-price dynamics.

A broad set of robustness checks supports these findings. The results are not an artifact of the $AR(1)$ benchmark—a UCSV benchmark yields the same ranking—and are robust to alternative loss functions (MAE, MAD) and to a pre-COVID sample. Density forecasts (CRPS) show improved calibration for RIS-based specifications at short horizons, and full-sample predictive regressions confirm significant incremental explanatory power beyond lagged inflation and conventional expectations. Two external-validation exercises reinforce the economic interpretation: the RIS co-moves contemporaneously with newspaper sentiment, and Granger-causality tests indicate that it frequently anticipates Michigan survey expectations.

Finally, our results carry a practical implication for scalable deployment. While large models such as LLaMA 70B are strong performers, fine-tuned small language models often deliver comparable forecast accuracy, and parameter-efficient fine-tuning methods substantially reduce computational cost without materially degrading classification performance. This makes the approach feasible in resource-constrained environments and supports the integration of the RIS into applied inflation-monitoring toolkits.

Overall, the evidence suggests that the RIS—constructed with disciplined, backward-looking procedures and evaluated in a pseudo-real-time setting—can complement standard survey-, market-, and news-based sources in inflation forecasting and nowcasting. More broadly, the paper illustrates how LLMs can be used to operationalize high-frequency public discourse as a measurable economic predictor for macroeconomic measurement and policy analysis.

# References


Alam, M. Jahangir, Shane Boyle, Huiyu Li, and Tatevik Sekhposyan (2026). *ChatMacro: Evaluating Inflation Forecasts of Generative AI*. Tech. rep. (cit. on p. 2).

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. 2, 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. 6).

Audrino, Francesco, Jessica Gentner, and Simon Stalder (2024). "Quantifying uncertainty: a new era of measurement through large language models". In: (cit. on p. 6).

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 pp. 5, 32).

Bybee, Leland (2023). "Surveying Generative AI's Economic Expectations". In: *arXiv preprint arXiv:2305.02823* (cit. on pp. 2, 6, 18).

Carriero, Andrea, Davide Pettenuzzo, and Shubhranshu Shekhar (2024). "Macroeconomic Forecasting with Large Language Models". In: *arXiv preprint arXiv:2407.00890* (cit. on pp. 5, 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. 3, 19).

Consoli, Sergio, Luca Barbaglia, and Sebastiano Manzan (2022). "Fine-grained, aspect-based sentiment analysis on economic and financial lexicon". In: *Knowledge-Based Systems* 247, p. 108781. ISSN: 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2022.108781 (cit. on p. 32).

Dettmers, Tim, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer (2023). *QLoRA: Efficient Finetuning of Quantized LLMs*. arXiv: 2305.14314 (cit. on p. 12).

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 pp. 2, 6).

Faust, Jon and Jonathan Wright (2013). "Forecasting Inflation". In: vol. 2. Elsevier. Chap. Chapter 1, pp. 2–56 (cit. on p. 20).

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

Google AI Team (2024). *Gemma 2 Technical Report* (cit. on p. 12).

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. 5).

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

Hayou, Soufiane, Nikhil Ghosh, and Bin Yu (2024). *LoRA+: Efficient Low Rank Adaptation of Large Models*. arXiv: 2402.12354 (cit. on p. 12).

Hong, Yongmiao, Fuwei Jiang, Lingchao Meng, and Bowen Xue (2025). "Forecasting Inflation Using Economic Narratives". In: *Journal of Business & Economic Statistics* 43.1, pp. 216–231 (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).

Jiang, Fuwei, Kunpeng Li, Lingchao Meng, and Bowen Xue (2025). "Measuring Real-Time Economic Condition with Economic Narratives". In: *Journal of Futures Markets* 45.11, pp. 2186–2207 (cit. on p. 13).

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. 5).

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. 3, 27).

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. 12).

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. 12).

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. 5).

Sarkar, Suproteem K and Keyon Vafa (2024). "Lookahead Bias in Pretrained Language Models". In: *Available at SSRN 4754678* (cit. on p. 2).

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. 6).

Stock, James H. and Mark W. Watson (2007). "Why Has U.S. Inflation Become Harder to Forecast?" In: *Journal of Money, Credit and Banking* 39.s1, pp. 3–33. DOI: https://doi.org/10.1111/j.1538-4616.2007.00014.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1538-4616.2007.00014.x (cit. on p. 20).

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. 12).

# 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)

October 8, 2026

# A Reddit data: Further details

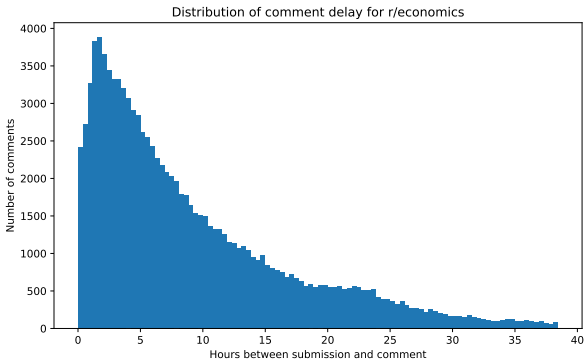
## A.1 Comment delay

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.

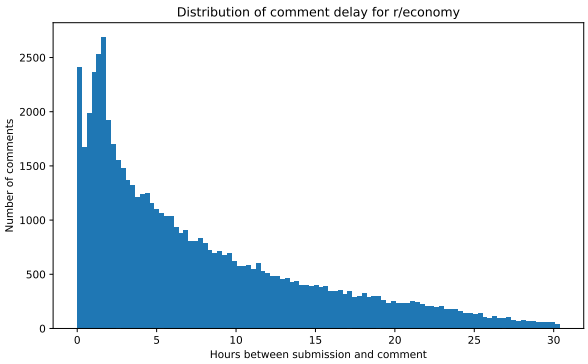
Figure A.1 plots the delay distribution for the central 95% of comments, trimming the upper tail to remove outliers that are not informative for constructing the daily and monthly RIS. Engagement on Reddit is highly front-loaded: most comments are posted shortly after the submission appears, with the 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 RIS.

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

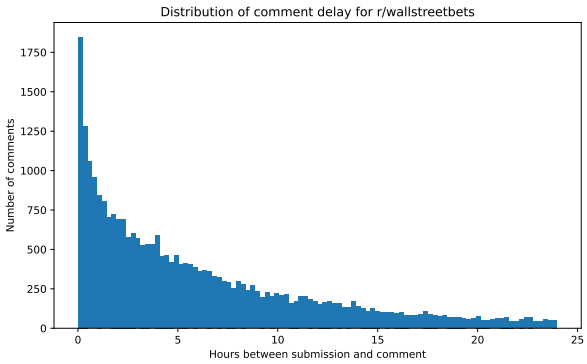
**(a)** `r/Economics`




**(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 LLaMA 70B (zero-shot) prompts for country identification and price direction

This appendix reports the zero-shot prompts used with the unfine-tuned LLaMA 70B model. We employ two prompts. The first prompt is used to classify the geographic reference of each Reddit submission (U.S., non-U.S., or no clear geographic attribution) and is shown in Figure A.2. The second prompt is applied to the geo-filtered corpus (submissions and comments) to extract directional labels (UP, DOWN, NEUTRAL) and is reported in Figure A.3.

**Figure A.2:** LLM prompt for geo-reference classification using LLaMA 70B-Instruct.

```
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.
```

*Notes*: The title of the submission is given in the user message.

**Figure A.3:** LLM prompt for directional classification using LLaMA 70B-Instruct.

```
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.
```

*Notes*: `df['title'][i]` contains all the submissions or comments referred to the U.S. and containing at least an inflation-related keyword. The prompt elicits a directional classification; as discussed in the main text, the resulting label need not map one-to-one into a strictly forward-looking expectation.

# 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-related content*. Each submission is assigned to one of three mutually exclusive categories: UP (the text conveys rising or increasing inflation/prices), DOWN (the text conveys falling or decreasing inflation/prices), 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.

**Figure A.4:** Prompt given to LLaMA 70B for labeling 700 Reddit submissions as UP, DOWN, or NEUTRAL.

```
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
```

*Notes*: This prompt was used to add labeled posts for fine-tuning the LLMs. 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 using the prompt reported in Figure A.5, 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 directional classifications used to construct the RIS variants in the main analysis.

**Figure A.5:** ChatGPT prompt for labeling as UP, DOWN or NEUTRAL Reddit submissions.

```
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?

Post:  [Text of the submission...]
```

# D A Primer on Large Language Models

Large language models 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 directional classifications[38] 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.

[38] In previous versions we called them "educated" signals.

**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.[39] 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.[40] We consider the 2B, 9B, and 27B versions; Gemma 27B is the largest model we fine-tune.[41]

**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.[42] 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

[39] `https://ai.meta.com/blog/meta-llama-3/`
[40] `Italsofeaturesavocabularyof256,128tokens.Link:https://huggingface.co/blog/gemma`
[41] `https://storage.googleapis.com/deepmind-media/gemma/gemma-2-report.pdf`
[42] `https://qwenlm.github.io/blog/qwen2.5/`

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 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 $r \ll \min(d, k)$.[43]

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,[44] 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 `LoRA` by decomposing the weight matrix into *magnitude* and *direction* components and fine-tuning both.

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

[44] 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.

Instead of applying all fine-tuning adjustments in one place, DoRA spreads them out, improving the model's adaptability while maintaining efficiency. DoRA decomposes 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

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

By separating directional and magnitude updates, DoRA can 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, quantization yields QLoRA (Dettmers et al., 2023); when combined with DoRA updates, it 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).[45] 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 linguistic patterns and narrative cues that are informative for inflation forecasting using Reddit discussions.

**Fine-Tuning LLMs to Extract Directional Inflation Classifications from Reddit Posts** To construct the RIS variants 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, while that for `r/wallstreetbets` contains around 4,800 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.[46] The preferred specification is selected based on the (weighted)

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

[46] 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` and for the 4,801 titles from `r/wallstreetbets`.

F1 score.

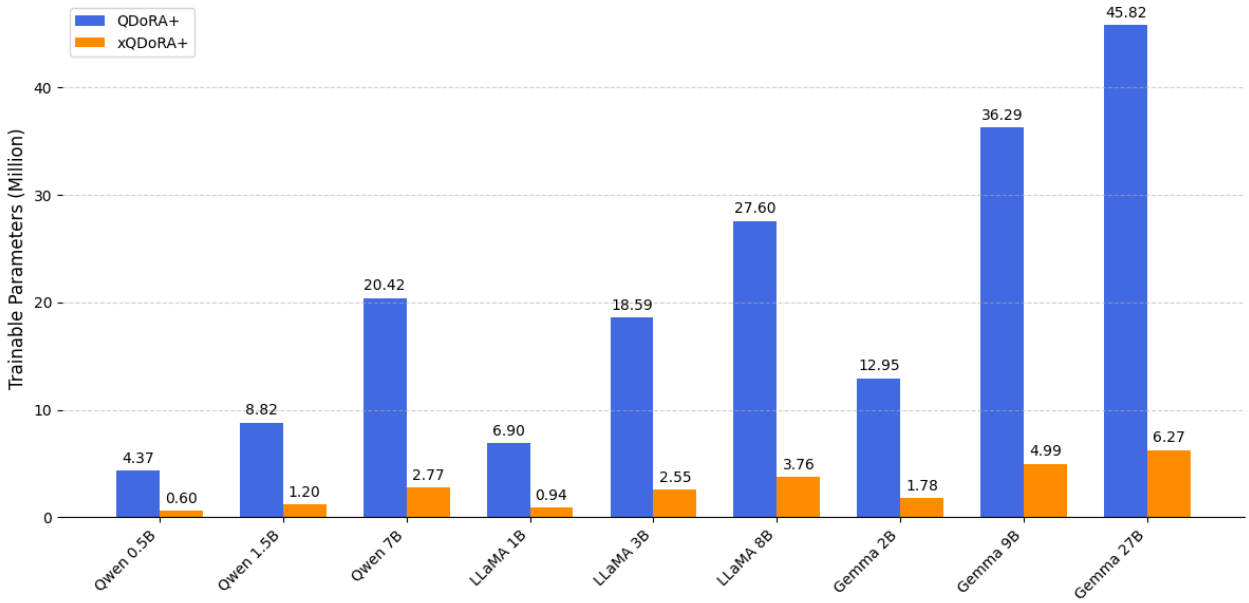
To make fine-tuning feasible on a single GPU (NVIDIA A100), we combine 4-bit quantization with PEFT. In particular, we use `DoRA` (Liu et al., 2024), a weight-decomposed variant of LoRA, which introduces low-rank adapters and updates only a small fraction of model parameters while keeping the base weights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

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, in some cases, an even higher weighted F1 score than 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 RIS variants.

**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

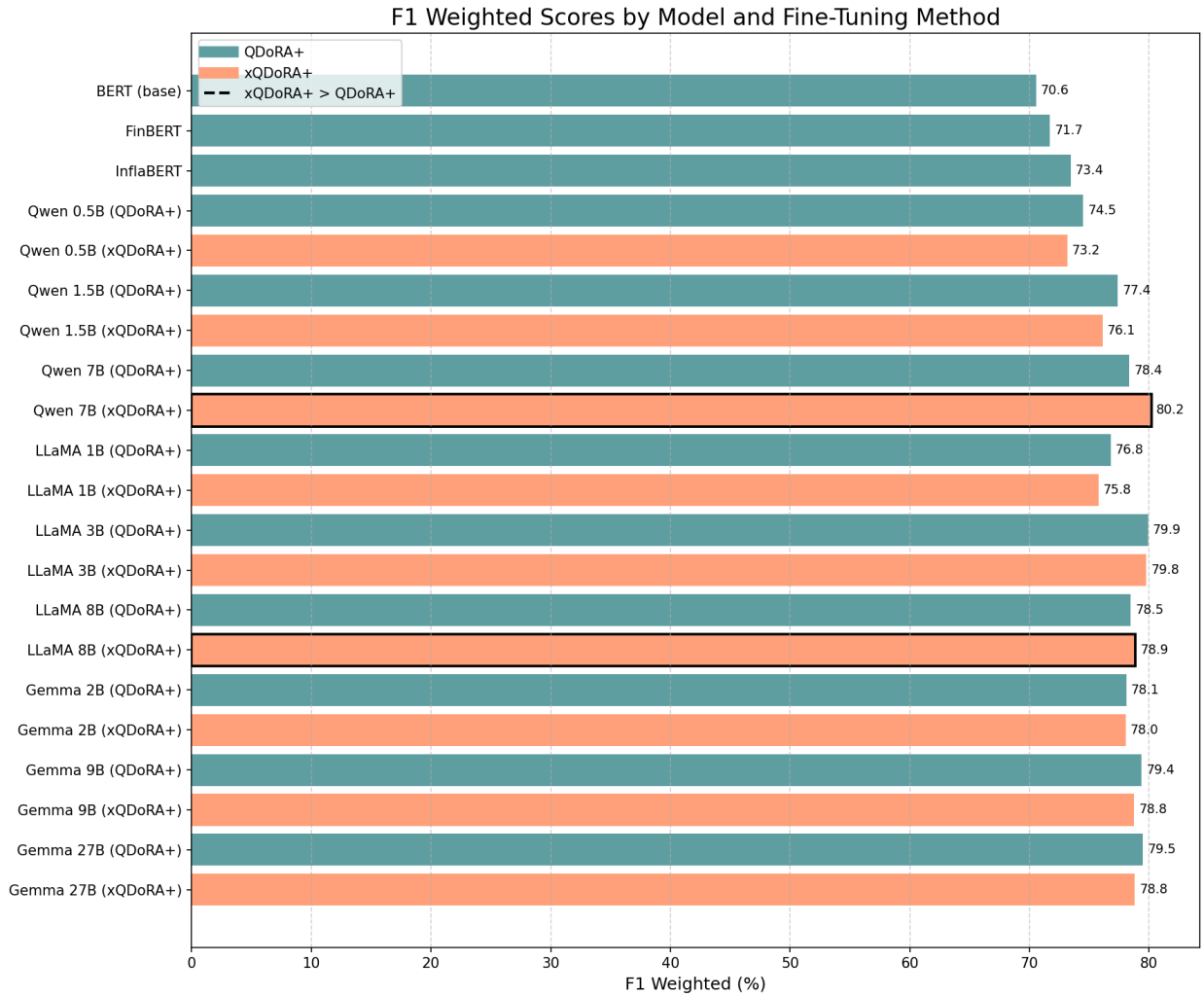
| **Model** | **Size in billion (B) of parameters** | **Percentage of trainable parameters** |
|---|---|---|
| BERT base | 0.109 B | 100.0% |
| FinBERT | 0.109 B | 100.0% |
| InflaBERT | 0.082 B | 100.0% |
| Qwen 2.5 0.5B | 0.498 B | 0.8782% (QDoRA+)<br>0.1198% (xQDoRA+) |
| Qwen 2.5 1.5B | 1.553 B | 0.5682% (QDoRA+)<br>0.0773% (xQDoRA+) |
| Qwen 2.5 7B | 7.091 B | 0.2880% (QDoRA+)<br>0.0391% (xQDoRA+) |
| LLaMA 3.2 1B | 1.243 B | 0.5555% (QDoRA+)<br>0.0760% (xQDoRA+) |
| LLaMA 3.2 3B | 3.231 B | 0.5753% (QDoRA+)<br>0.0788% (xQDoRA+) |
| LLaMA 3.1 8B | 7.533 B | 0.3664% (QDoRA+)<br>0.0499% (xQDoRA+) |
| Gemma 2 2B | 2.627 B | 0.4930% (QDoRA+)<br>0.0677% (xQDoRA+) |
| Gemma 2 9B | 9.278 B | 0.3911% (QDoRA+)<br>0.0538% (xQDoRA+) |
| Gemma 2 27B | 27.273 B | 0.1680% (QDoRA+)<br>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+)<br>71.223% (xQDoRA+) | 77.519% (QDoRA+)<br>75.000% (xQDoRA+) | 74.627% (QDoRA+)<br>73.063% (xQDoRA+) | 74.509% (QDoRA+)<br>73.167% (xQDoRA+) | 68.345% (QDoRA+)<br>69.784% (xQDoRA+) |
| Qwen 2.5 1.5B | 76.259% (QDoRA+)<br>73.381% (xQDoRA+) | 78.519% (QDoRA+)<br>79.688% (xQDoRA+) | 77.372% (QDoRA+)<br>76.404% (xQDoRA+) | 77.379% (QDoRA+)<br>76.116% (xQDoRA+) | 73.381% (QDoRA+)<br>70.504% (xQDoRA+) |
| Qwen 2.5 7B | 79.137% (QDoRA+)<br>76.978% (xQDoRA+) | 78.014% (QDoRA+)<br>83.594% (xQDoRA+) | 78.571% (QDoRA+)<br>80.150% (xQDoRA+) | 78.353% (QDoRA+)<br>80.201% (xQDoRA+) | 73.381% (QDoRA+)<br>74.820% (xQDoRA+) |
| LLaMA 3.2 1B | 74.820% (QDoRA+)<br>76.259% (xQDoRA+) | 78.195% (QDoRA+)<br>75.177% (xQDoRA+) | 76.471% (QDoRA+)<br>75.714% (xQDoRA+) | 76.785% (QDoRA+)<br>75.765% (xQDoRA+) | 71.942% (QDoRA+)<br>73.381% (xQDoRA+) |
| LLaMA 3.2 3B | 78.417% (QDoRA+)<br>76.978% (xQDoRA+) | 80.741% (QDoRA+)<br>82.946% (xQDoRA+) | 79.562% (QDoRA+)<br>79.851% (xQDoRA+) | 79.914% (QDoRA+)<br>79.784% (xQDoRA+) | 76.978% (QDoRA+)<br>76.978% (xQDoRA+) |
| LLaMA 3.1 8B | 77.698% (QDoRA+)<br>78.417% (xQDoRA+) | 79.412% (QDoRA+)<br>79.562% (xQDoRA+) | 78.545% (QDoRA+)<br>78.986% (xQDoRA+) | 78.453% (QDoRA+)<br>78.878% (xQDoRA+) | 74.820% (QDoRA+)<br>76.978% (xQDoRA+) |
| Gemma 2 2B | 76.978% (QDoRA+)<br>77.698% (xQDoRA+) | 79.259% (QDoRA+)<br>78.261% (xQDoRA+) | 78.102% (QDoRA+)<br>77.978% (xQDoRA+) | 78.096% (QDoRA+)<br>78.042% (xQDoRA+) | 74.820% (QDoRA+)<br>76.978% (xQDoRA+) |
| Gemma 2 9B | 79.137% (QDoRA+)<br>77.698% (xQDoRA+) | 79.710% (QDoRA+)<br>80.000% (xQDoRA+) | 79.422% (QDoRA+)<br>78.832% (xQDoRA+) | 79.360% (QDoRA+)<br>78.762% (xQDoRA+) | 75.540% (QDoRA+)<br>70.504% (xQDoRA+) |
| Gemma 2 27B | 79.856% (QDoRA+)<br>77.698% (xQDoRA+) | 79.286% (QDoRA+)<br>80.000% (xQDoRA+) | 79.570% (QDoRA+)<br>78.832% (xQDoRA+) | 79.486% (QDoRA+)<br>78.816% (xQDoRA+) | 76.978% (QDoRA+)<br>76.978% (xQDoRA+) |

**Figure A.7:** Weighted F1 scores for each LLM under different fine-tuning strategies.

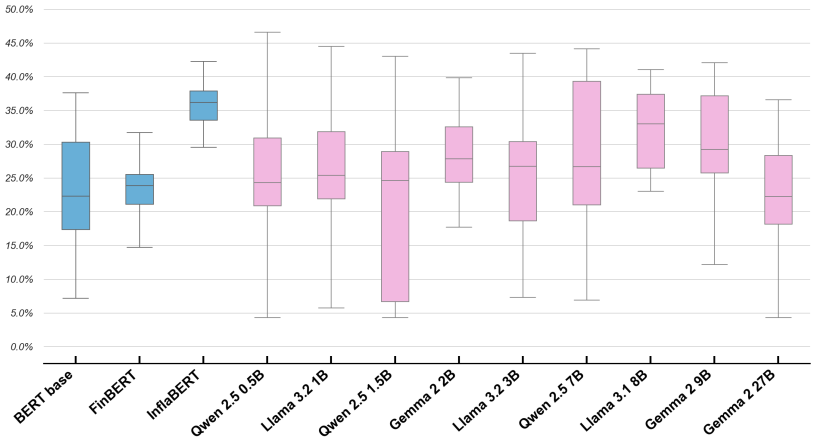



*Notes*: The figure shows the weighted F1 score for each LLM using the two parameter-efficient fine-tuning techniques, QDoRA+ and xQDoRA+.

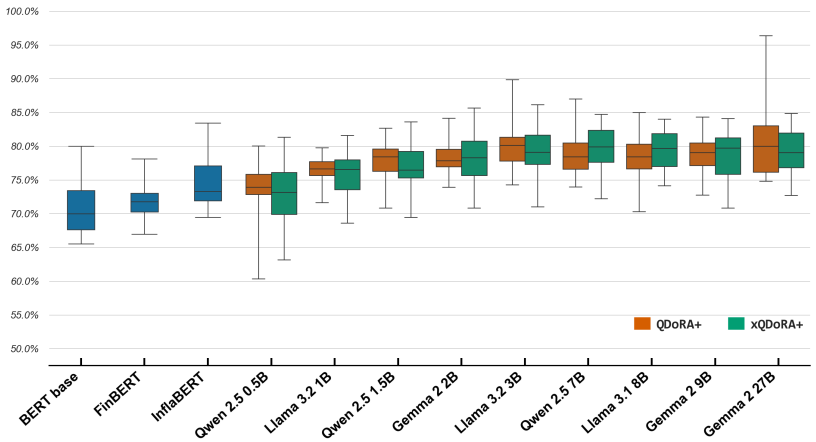
Figure A.8 shows box plots of weighted F1 scores for unfine-tuned and fine-tuned models. The fine-tuned models display a clear advantage in classification accuracy.

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

**(a)** 'Raw' models




**(b)** Fine-tuned models

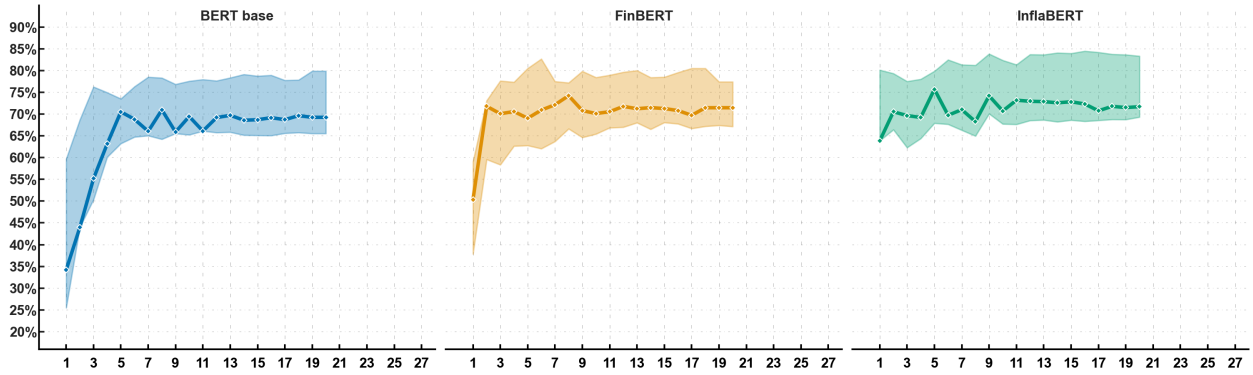



*Notes*: The figure shows box plots of the F1 scores for unfine-tuned (i.e., "raw") models and for models fine-tuned over 20 independent runs with different random seeds

Figures A.9, A.10, A.11, and A.12 show the evolution of the weighted F1 score on the validation set over training epochs until the process terminates after five consecutive epochs without improvement (early stopping).

The charts show that weighted F1 scores for the BERT-based models evolve more smoothly

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




across epochs than those for the `LLaMA`, `Gemma`, and `Qwen` LLM families. Fine-tuning also terminates earlier for the Qwen family than for 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.

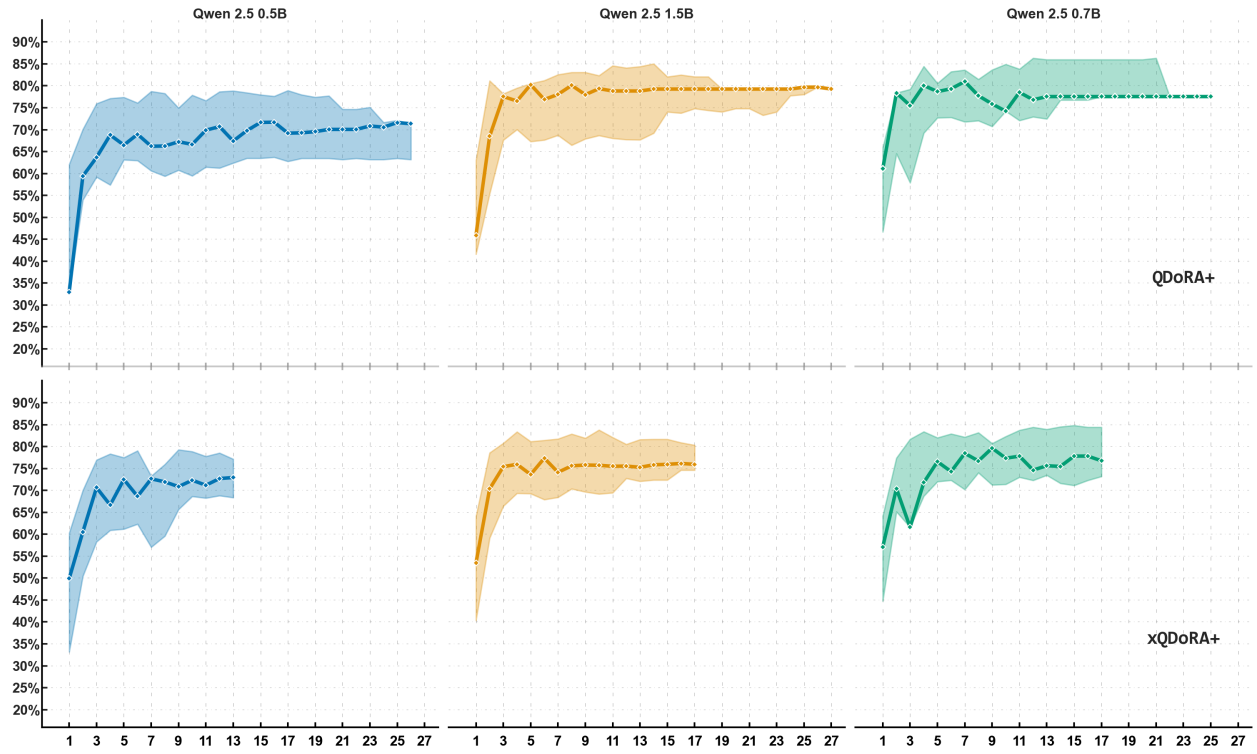


**Figure A.11:** Weighted F1 scores of the validation set in the training loop: models from the LLaMA 3 family.

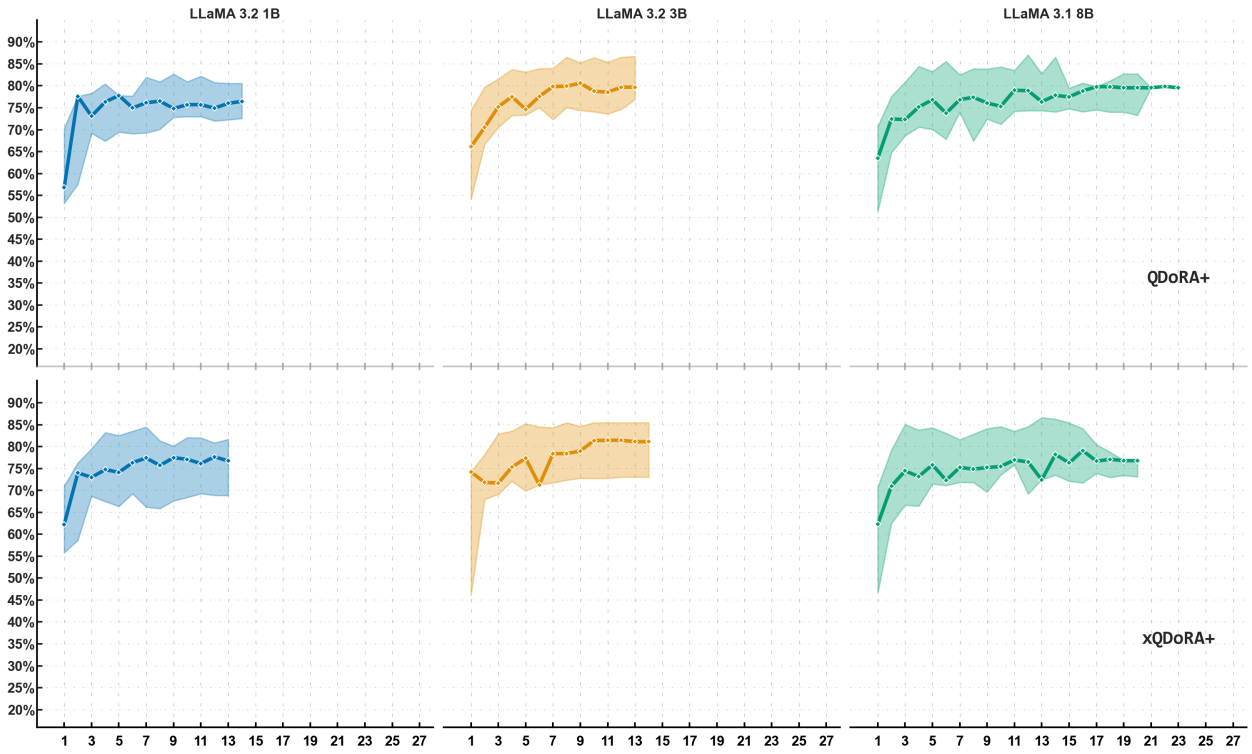
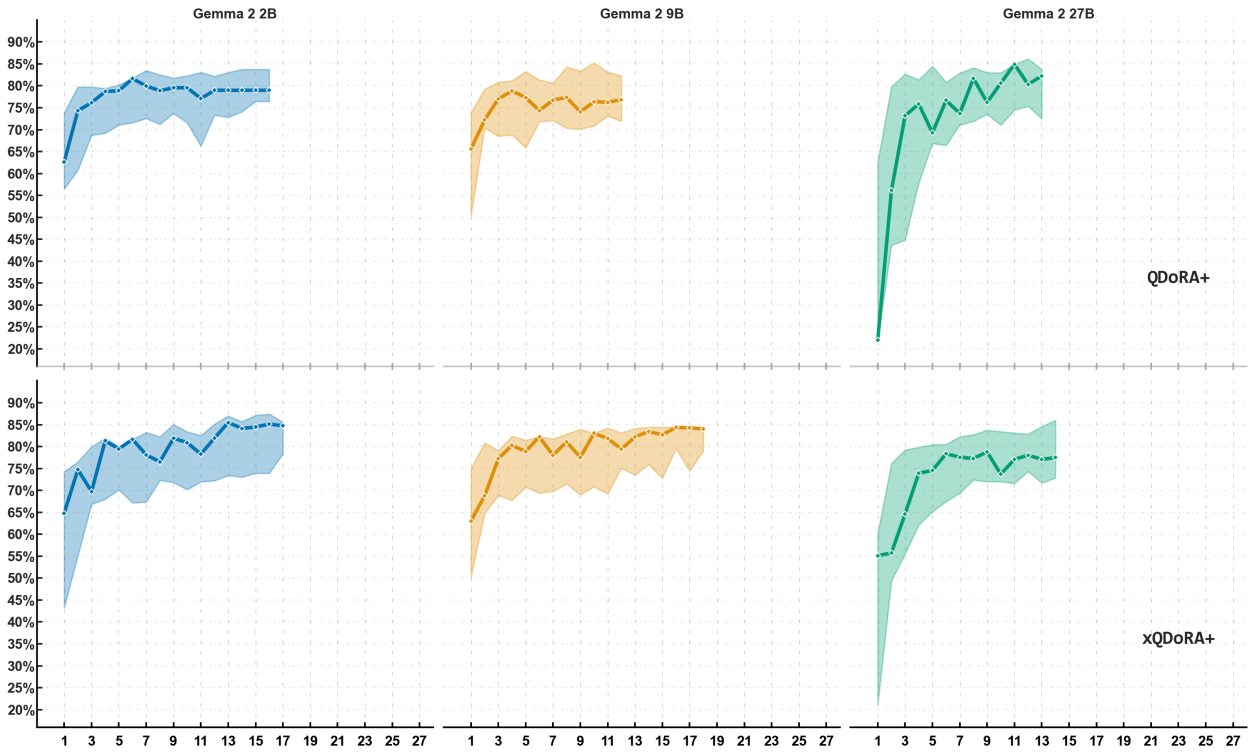


**Figure A.12:** Weighted F1 scores of the validation set in the training loop: models from the Gemma 2 family.



## 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 $\alpha_m$ captures the architecture-specific baseline, while $\eta_s$ 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 $\beta_m$ 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 $\beta_m$ 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.

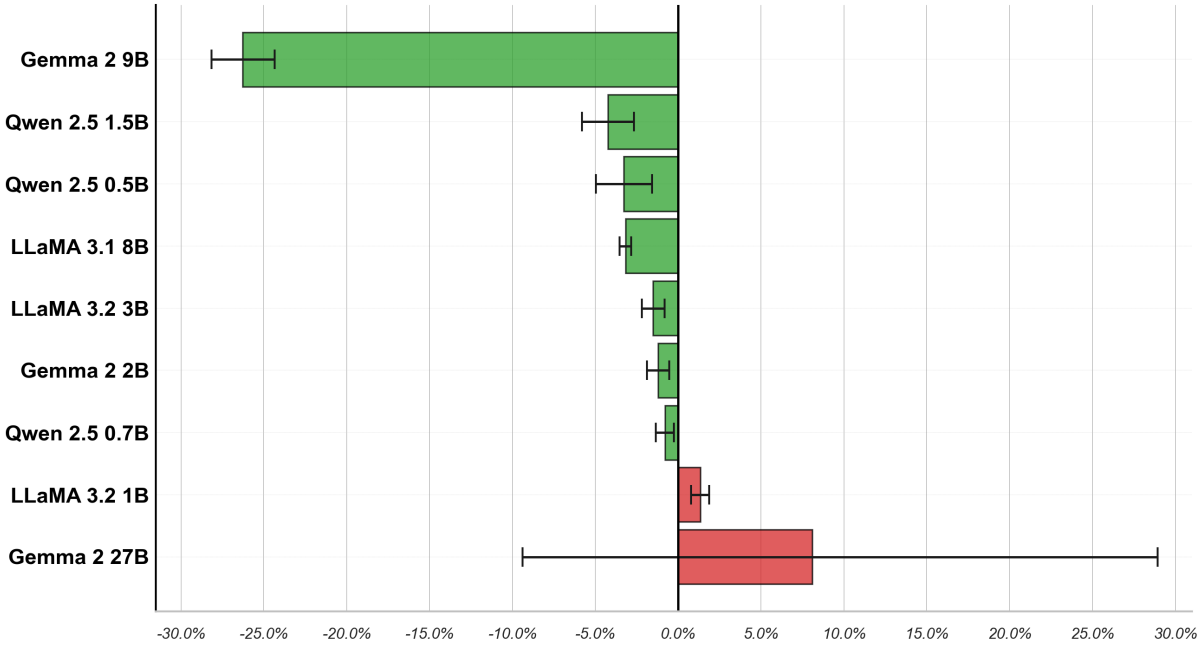
**Conclusion.** Taken together, the results support `xQDoRA+` as a practical improvement over `QDoRA+`: it delivers measurable reductions in inference wall-time—especially for larger architectures—while maintaining classification performance that is, in statistical terms, comparable to the baseline method. 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. ($\beta$) | 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

This appendix reports additional results for the point-forecasting exercise. Section E.1 presents robustness checks using alternative loss functions (MAE and MAD) and reports RMSE ratios for individual RIS variants over the full evaluation period. Section E.2 reports RMSE ratios for individual RIS variants for the pre-COVID subsample. Section E.3 provides the full set of CSSED paths and Giacomini–Rossi fluctuation tests across all horizons. Section E.4 repeats the out-of-sample comparison using the unobserved-components stochastic-volatility (UCSV) model as an alternative benchmark. Section E.5 describes the Model Confidence Set (MCS) procedure, and Section E.6 reports full-sample predictive regressions. Finally, Section E.7 presents density-forecast results and Section E.8 presents results using inflation-related Reddit post counts as predictors.

Figure A.14 illustrates the expanding-window (or recursive) pseudo out-of-sample design used in the forecasting exercise (shown for the one-month-ahead horizon, $h = 1$).

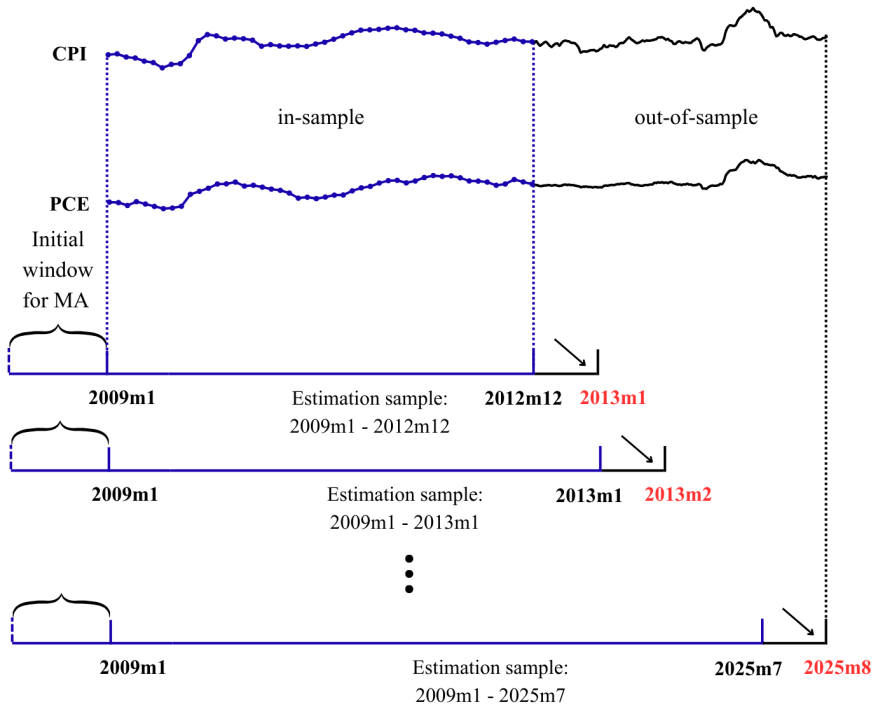
## E.1 Additional forecast losses for point forecasts

This appendix section complements the RMSE-based evaluation in the main text by reporting out-of-sample forecast accuracy under two alternative loss functions: the mean absolute error (MAE) and the mean absolute deviation (MAD). Unlike RMSE, which places disproportionate weight on large forecast errors, MAE and MAD downweight tail realizations. They therefore provide a useful robustness check on whether the gains from the RIS are driven by a few extreme episodes or reflect broader improvements in predictive accuracy.

**MAE results.** Tables A.6 and A.7 report MAE ratios relative to the $AR(1)$ benchmark. The results closely mirror the RMSE patterns in the main text. RIS-based models outperform $AR(1)$ at most horizons, with the strongest gains concentrated at short-to-medium horizons ($h = 1$–$6$). For CPI, Aggr-LLaMA70B typically attains the lowest MAE ratios at short horizons, while Aggr-RIS remains highly competitive throughout. At longer horizons, dictionary-based sentiment indicators and expectation-based benchmarks become relatively more informative.

Table A.8 reports the best single specification within each family under MAE. The selection patterns are broadly consistent with the RMSE-based ranking shown in Figure 5. In particular, InflaBERT dominates within the fine-tuned family at most horizons (with Qwen 7B emerging at the longest horizon). The best-performing subreddit differs by target: CPI performance is more frequently driven by `r/Economics` or `r/economy`, whereas `r/wallstreetbets` becomes more promi-

**Figure A.14:** 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.

nent for PCE at medium to long horizons. Finally, the preferred smoothing varies systematically with horizon: short MA windows tend to be selected at short horizons, while longer windows are more frequently selected at longer horizons.

**MAD results.** Tables A.9 and A.10 report the analogous results under MAD loss. Again, the conclusions align closely with the main-text evidence. RIS-based models improve upon $AR(1)$ at most horizons, with the largest gains concentrated at $h = 1$–$6$. Aggr-LLaMA70B is typically the strongest performer up to six months ahead, and Aggr-RIS remains close in performance. A useful operational takeaway is that, across loss functions, Aggr-RIS consistently performs better than forecasts based on Michigan expectations over the horizons where the RIS provides the largest gains.

Table A.11 reports the best single specification within each family under MAD and highlights additional heterogeneity across targets and horizons. For CPI, the best-performing fine-tuned specification at very short horizons is often based on `r/Economics`, whereas at the longest horizon Qwen 7B with `r/wallstreetbets` is frequently selected. For PCE, Qwen 7B and InflaBERT alternate as the top fine-tuned models across horizons, with `r/wallstreetbets` playing a larger role at medium to long horizons. Within the unfine-tuned LLaMA 70B family, `r/economy` more often delivers the best CPI specification, while `r/wallstreetbets` is more frequently selected for PCE. Within the lexicon/sentiment family, VADER typically performs best at the shortest horizon, while TextBlob is more often selected at medium to long horizons; the Loughran–McDonald dictionary tends to appear as the best lexicon-based specification only at a limited set of CPI horizons.

**Model-selection details and visualization.** For transparency, Table A.12 reports the best single specification within each family under RMSE (i.e., the selections underlying Figure 5). Figures A.15 and A.17 plot RMSE ratios for the best specification within each model family—best single fine-tuned LLM, best single unfine-tuned LLaMA 70B, Aggr-RIS, Aggr-LLaMA70B, lexicon-based sentiment (Sentiment), the one-year inflation swap (Swap), and Michigan expectations (Expectations)—at each horizon for CPI and PCE, respectively, while Figures A.16 and A.18 report the corresponding top-10 rankings (across all model families, subreddits, and MA windows) as heatmaps for CPI and PCE, respectively.

**Summary.** Across both CPI and PCE, and across RMSE, MAE, and MAD, a consistent pattern emerges. RIS-based models—particularly when combined using MSE-weighted aggregation—deliver

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

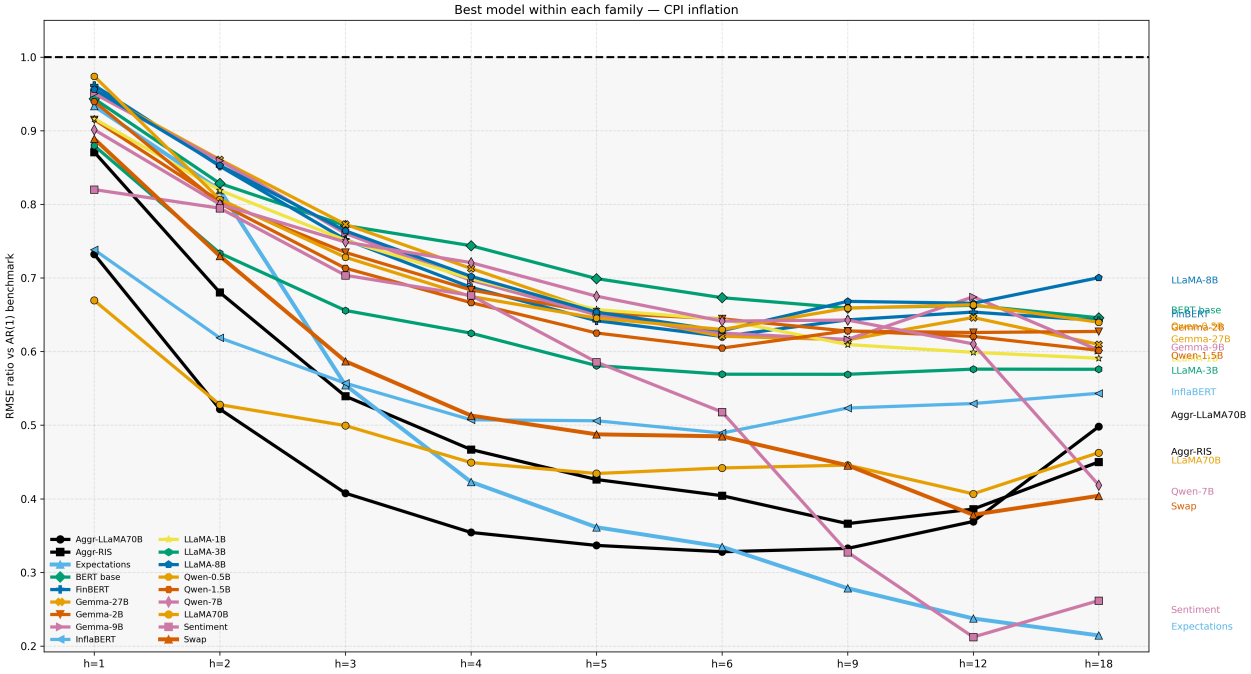
| Model | $h=1$ | $h=2$ | $h=3$ | $h=4$ | $h=5$ | $h=6$ | $h=9$ | $h=12$ | $h=18$ |
|---|---|---|---|---|---|---|---|---|---|
| *Expectations* | | | | | | | | | |
| Best expectations | 1.034 | 1.023 | 0.967 | 0.923 | 0.893 | 0.877 | 0.823 | 0.784 | *0.702* |
| 1-Y Inflation Swap (best) | 1.020 | 0.939 | 0.905 | 0.904 | 0.895 | 0.866 | 0.807 | 0.821 | 0.764 |
| *Lexicon-based sentiment* | | | | | | | | | |
| Best sentiment | 0.992 | 0.927 | 0.904 | 0.862 | *0.820* | *0.787* | **0.717** | **0.700** | **0.658** |
| *RIS-based models* | | | | | | | | | |
| Aggr-RIS | 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 |
| Aggr-LLaMA70B | **0.809** | **0.771** | **0.771** | **0.743** | **0.729** | **0.722** | *0.746* | *0.761* | 0.757 |
| Best LLaMA 70B | *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.

the largest gains at short horizons, where forecasting improvements are typically hardest to obtain. By contrast, survey expectations and dictionary-based sentiment measures become relatively more competitive at the longest horizons, consistent with the broader inflation-forecasting literature.

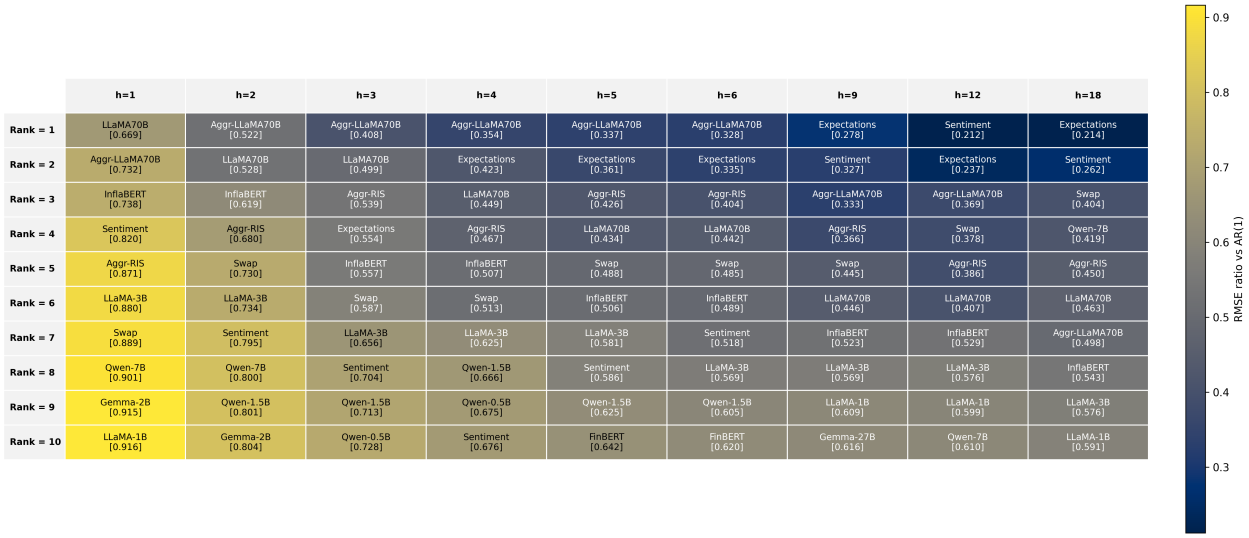
A further practical implication is that small language models can be "good enough" in applied settings. Fine-tuned SLMs such as Qwen 7B and LLaMA 3B frequently appear among the top-ranked specifications at several horizons and often deliver loss ratios close to those of much larger models. This suggests that task-specific fine-tuning and careful RIS construction account for a substantial share of the predictive gains, making SLM-based pipelines attractive when compute and memory resources are constrained.

**Figure A.15:** RMSE ratios for best models by forecast horizon for CPI



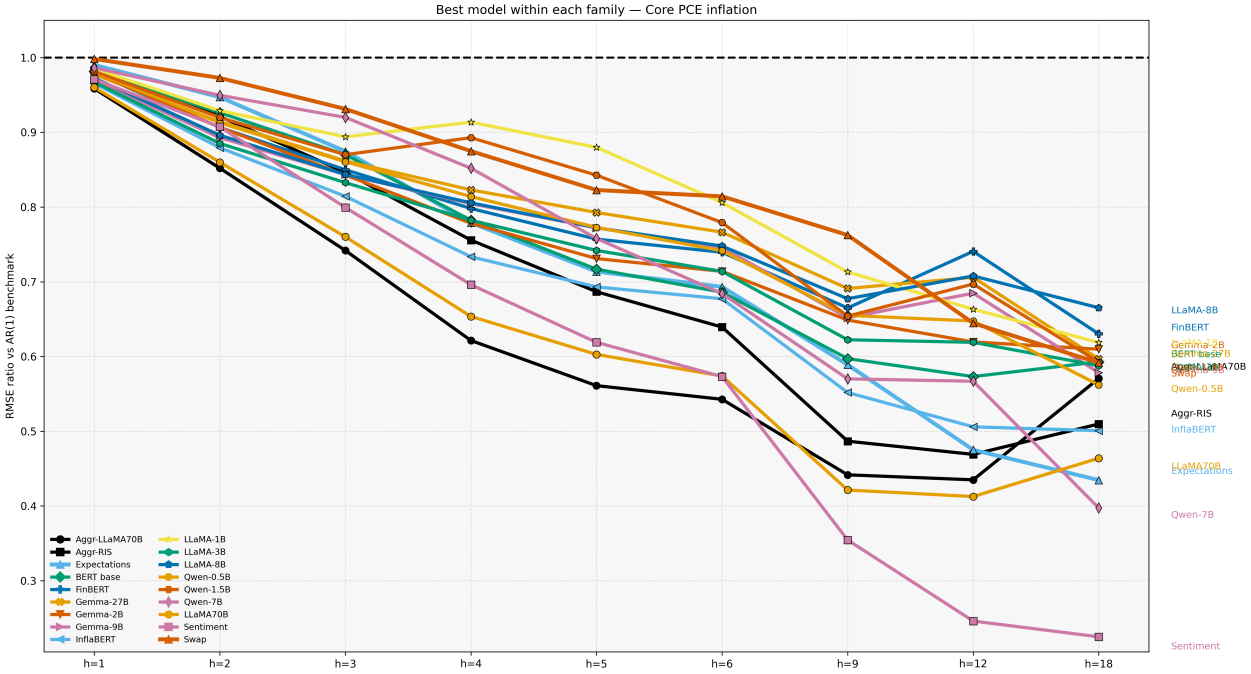

*Notes*: The figure reports RMSE ratios relative to the $AR(1)$ benchmark for CPI, computed as $RMSE(AR\text{-}X(1))/RMSE(AR(1))$. For each forecast horizon $h \in \{1, 2, 3, 4, 5, 6, 9, 12, 18\}$, we plot the best-performing specification within each model family: best single fine-tuned LLM, best single unfine-tuned LLaMA 70B, Aggr-RIS, Aggr-LLaMA70B, lexicon-based sentiment (Sentiment), the one-year inflation swap (Swap), and Michigan expectations (Expectations). "Best" is selected across all subreddits and moving-average windows. Ratios below one indicate an improvement over the $AR(1)$ benchmark.

**Figure A.16:** Top-10 best models by forecast horizon for CPI



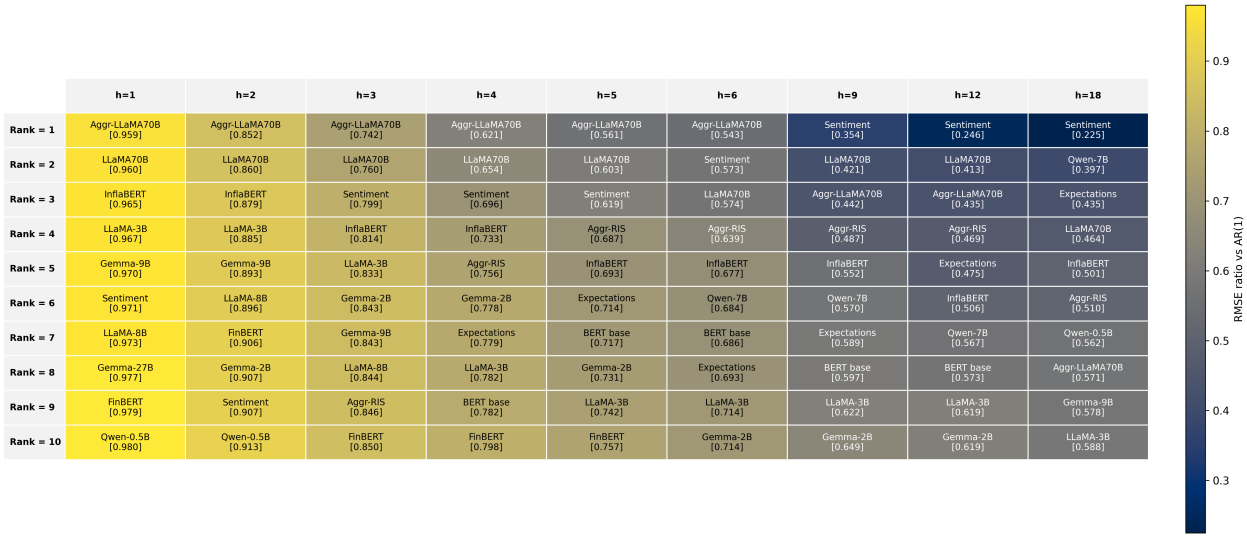

*Notes*: The heatmap summarizes the top-10 forecasting specifications for CPI at each horizon, ranked by RMSE ratio relative to the $AR(1)$ benchmark, $RMSE(AR\text{-}X(1))/RMSE(AR(1))$. Columns correspond to forecast horizons $h \in \{1, 2, 3, 4, 5, 6, 9, 12, 18\}$; each tile reports the model identifier and its RMSE ratio. The ranking is computed over all candidate specifications obtained by combining (i) model family (single fine-tuned LLM, single unfine-tuned LLaMA 70B, Aggr-RIS, Aggr-LLaMA70B, lexicon-based sentiment, one-year inflation swap, and Michigan expectations), (ii) subreddit, and (iii) moving-average window. Color encodes performance: warmer colors indicate ratios closer to one (little or no gain), while cooler colors indicate lower ratios (larger improvements). Ratios below one imply that the corresponding specification outperforms $AR(1)$.

**Figure A.17:** RMSE ratios for best models by forecast horizon for PCE




*Notes*: The figure reports RMSE ratios relative to the $AR(1)$ benchmark for PCE, computed as $RMSE(AR\text{-}X(1))/RMSE(AR(1))$. For each forecast horizon $h \in \{1, 2, 3, 4, 5, 6, 9, 12, 18\}$, we plot the best-performing specification within each model family: best single fine-tuned LLM, best single unfine-tuned LLaMA 70B, Aggr-RIS, Aggr-LLaMA70B, lexicon-based sentiment (Sentiment), the one-year inflation swap (Swap), and Michigan expectations (Expectations). "Best" is selected across all subreddits and moving-average windows. Ratios below one indicate an improvement over the $AR(1)$ benchmark.

**Figure A.18:** Top-10 best models by forecast horizon for PCE




*Notes*: The heatmap summarizes the top-10 forecasting specifications for PCE at each horizon, ranked by RMSE ratio relative to the $AR(1)$ benchmark, $RMSE(AR\text{-}X(1))/RMSE(AR(1))$. Columns correspond to forecast horizons $h \in \{1, 2, 3, 4, 5, 6, 9, 12, 18\}$; each tile reports the model identifier and its RMSE ratio. The ranking is computed over all candidate specifications obtained by combining (i) model family (single fine-tuned LLM, single unfine-tuned LLaMA 70B, Aggr-RIS, Aggr-LLaMA70B, lexicon-based sentiment, one-year inflation swap, and Michigan expectations), (ii) subreddit, and (iii) moving-average window. Color encodes performance: warmer colors indicate ratios closer to one (little or no gain), while cooler colors indicate lower ratios (larger improvements). Ratios below one imply that the corresponding specification outperforms $AR(1)$.

**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 |
| *Lexicon-based sentiment* | | | | | | | | | |
| Best sentiment | 0.978 | 0.962 | 0.915 | 0.857 | *0.793* | **0.745** | **0.621** | **0.536** | **0.465** |
| *RIS-based models* | | | | | | | | | |
| Aggr-RIS | 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* |
| Aggr-LLaMA70B | **0.878** | **0.838** | **0.823** | **0.809** | **0.784** | *0.785* | 0.813 | 0.840 | 0.758 |
| Best LLaMA 70B | 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-based 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, "LM" corresponds to the financial dictionary of Loughran and McDonald (2011).

**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 |
| *Lexicon-based sentiment* | | | | | | | | | |
| Best sentiment | 0.982 | 0.927 | 0.899 | 0.849 | *0.796* | *0.753* | **0.665** | **0.647** | *0.600* |
| *RIS-based models* | | | | | | | | | |
| Aggr-RIS | 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 |
| Aggr-LLaMA70B | **0.799** | **0.727** | **0.713** | **0.688** | **0.675** | **0.671** | *0.681* | 0.696 | 0.699 |
| Best LLaMA 70B | *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 |
| *Lexicon-based sentiment* | | | | | | | | | |
| Best sentiment | 0.985 | 0.970 | 0.918 | 0.862 | 0.788 | **0.722** | **0.573** | **0.471** | **0.395** |
| *RIS-based models* | | | | | | | | | |
| Aggr-RIS | *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* |
| Aggr-LLaMA70B | **0.880** | **0.826** | **0.796** | **0.776** | **0.762** | *0.767* | 0.770 | 0.752 | 0.682 |
| Best LLaMA 70B | 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-based 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 | Aggr-RIS | – | – | LLaMA 70B | r/economy | 30 | TextBlob | r/economy | 10 |
| 3 | Aggr-RIS | – | – | LLaMA 70B | r/economy | 30 | TextBlob | r/economy | 10 |
| 4 | Aggr-RIS | – | – | LLaMA 70B | r/economy | 5 | TextBlob | r/economy | 5 |
| 5 | Aggr-RIS | – | – | LLaMA 70B | r/economy | 90 | LM | r/Economics | 90 |
| 6 | Aggr-RIS | – | – | LLaMA 70B | r/economy | 90 | LM | r/Economics | 90 |
| 9 | Aggr-RIS | – | – | LLaMA 70B | r/economy | 120 | LM | r/Economics | 180 |
| 12 | Aggr-RIS | – | – | LLaMA 70B | r/economy | 360 | LM | r/Economics | 180 |
| 18 | Qwen 7B | r/wallstreetbets | 360 | Aggr-LLaMA70B | – | – | 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. Aggr-RIS and Aggr-LLaMA70B denote the aggregated forecast models in the fine-tuned and LLaMA 70B families, respectively; for these rows, subreddit and MA are not applicable. In the lexicon/sentiment family, "LM" corresponds to the financial dictionary of Loughran and McDonald (2011).

**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-based 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. In the lexicon/sentiment family, "LM" corresponds to the financial dictionary of Loughran and McDonald (2011).

## E.2 Point forecasts in the pre-COVID sample

This appendix section repeats the point-forecast evaluation on a sample that ends just before the COVID-19 pandemic, thereby reducing the influence of the pandemic shock and the subsequent energy-price episode on forecast comparisons. We report accuracy as RMSE ratios relative to the $AR(1)$ benchmark, $RMSE(AR\text{-}X)/RMSE(AR(1))$, so values below one indicate an improvement over $AR(1)$. As in the main analysis, for each model family we select the best specification across subreddits and backward-looking moving-average (MA) windows.

Tables A.13 and A.14 report RMSE ratios for CPI and PCE, respectively, in the pre-COVID sample. The main message is that RIS variants processed with LLMs continue to outperform $AR(1)$ across horizons and targets, confirming that the gains are not driven solely by the extraordinary volatility of the post-2020 period.

**CPI.** For CPI, RIS-based models dominate at short and medium horizons. At $h = 1$–$2$, the best-performing specifications are the best single fine-tuned LLM and the best single LLaMA 70B model. For longer horizons, Aggr-LLaMA70B typically delivers the lowest (or second-lowest) RMSE ratios, closely followed by the best single LLaMA 70B specification. Aggr-RIS remains highly competitive throughout and outperforms Michigan expectations at all horizons. By contrast, survey- and market-based benchmarks do not systematically dominate: Michigan expectations ranks near the top only at isolated horizons (e.g., second at $h = 12$), and the inflation swap reaches second place only at $h = 9$. Lexicon-based sentiment models are generally not competitive in this pre-COVID CPI setting.

Figures A.19 and A.20 provide complementary visual summaries. In particular, the unfine-tuned LLaMA 70B-based RIS variant is the top performer at most horizons, with large gains at $h = 1$–$9$, while the fine-tuned InflaBERT specification becomes the best (or among the best) at the longest horizons. Smaller fine-tuned models (e.g., Gemma 2B, Qwen 7B, LLaMA 3B, and Qwen 0.5B) also appear frequently among the best-performing specifications, typically with RMSE ratios modestly below one.

**PCE.** For PCE, the ranking differs somewhat, but the overall picture remains favorable for RIS variants. At horizons $h = 1$–$9$, Aggr-LLaMA70B is typically the best-performing model family, indicating that pooling LLaMA 70B-based RIS variants is particularly effective for core inflation in the pre-COVID period. At longer horizons, Aggr-RIS becomes relatively more competitive and often attains the best performance. Survey expectations and lexicon-based sentiment measures

**Table A.13:** Forecast results (RMSE ratios) for *CPI* (pre-COVID sample).

| Model | $h=1$ | $h=2$ | $h=3$ | $h=4$ | $h=5$ | $h=6$ | $h=9$ | $h=12$ | $h=18$ |
|---|---|---|---|---|---|---|---|---|---|
| *Expectations* | | | | | | | | | |
| Expectations | 1.133 | 1.191 | 1.198 | 1.196 | 1.173 | 1.153 | 1.030 | *0.931* | 0.947 |
| 1-Y Inflation Swap | 1.062 | 0.991 | 0.963 | 0.986 | 0.982 | 0.969 | *0.935* | 0.987 | 1.051 |
| *Lexicon-based sentiment* | | | | | | | | | |
| Best sentiment | 0.947 | 0.967 | 0.881 | 0.860 | 0.853 | 0.862 | 1.050 | 0.982 | 1.083 |
| *RIS-based models* | | | | | | | | | |
| Aggr-RIS | 0.983 | 0.966 | 0.965 | 0.960 | 0.956 | 0.953 | 0.958 | 0.965 | 0.946 |
| Best fine-tuned LLM | **0.832** | **0.799** | *0.821* | 0.830 | 0.858 | 0.820 | 0.946 | 1.026 | 1.047 |
| Aggr-LLaMA70B | 0.931 | 0.835 | **0.817** | *0.801* | **0.798** | *0.808* | **0.884** | **0.903** | *0.897* |
| Best LLaMA 70B | *0.849* | *0.804* | 0.831 | **0.777** | *0.800* | **0.799** | 0.952 | 1.015 | **0.746** |

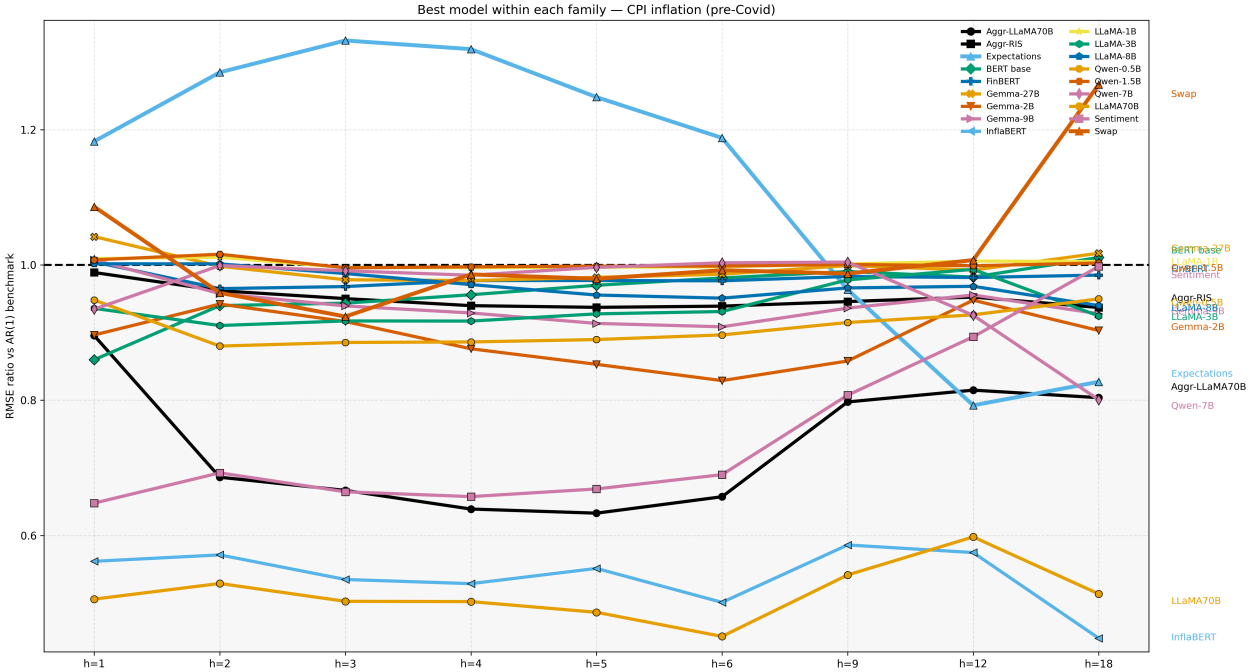
*Notes*: RMSE is computed as a ratio: $RMSE(AR\text{-}X(1))/RMSE(AR(1))$, where the benchmark is $AR(1)$. The evaluation is conducted on a pre-COVID out-of-sample period. Forecasts are generated using a recursive window, with the model re-estimated at each step. Bold (italic) indicates the best (second-best) model for each horizon.

generally do not outperform $AR(1)$ in this sample, while the one-year inflation swap is sometimes competitive at medium horizons (often ranking second between $h = 4$ and $h = 9$).

Figures A.21 and A.22 show that many fine-tuned small language models–including Qwen and LLaMA variants–regularly place among the top-10 specifications, and that InflaBERT is frequently among the best models up to $h = 12$. At the longest horizon ($h = 18$), the best-performing specification is a fine-tuned SLM (Qwen 0.5B), with an RMSE ratio well below one.

**Summary.** Three conclusions emerge from the pre-COVID evaluation. First, the outperformance of RIS-based models relative to $AR(1)$ persists when the post-2020 period is excluded, indicating that the main results are not driven mechanically by the COVID and energy shocks. Second, LLaMA 70B-based RIS variants and Aggr-LLaMA70B remain particularly strong at short and medium horizons, while fine-tuned models—most notably InflaBERT—become more prominent at longer horizons. Third, fine-tuned SLMs are often competitive and sometimes best-in-class, reinforcing the practical implication that high-quality RIS variants can be built even under constrained computational resources.

**Figure A.19:** RMSE ratios for best models by forecast horizon for CPI - (pre-COVID)



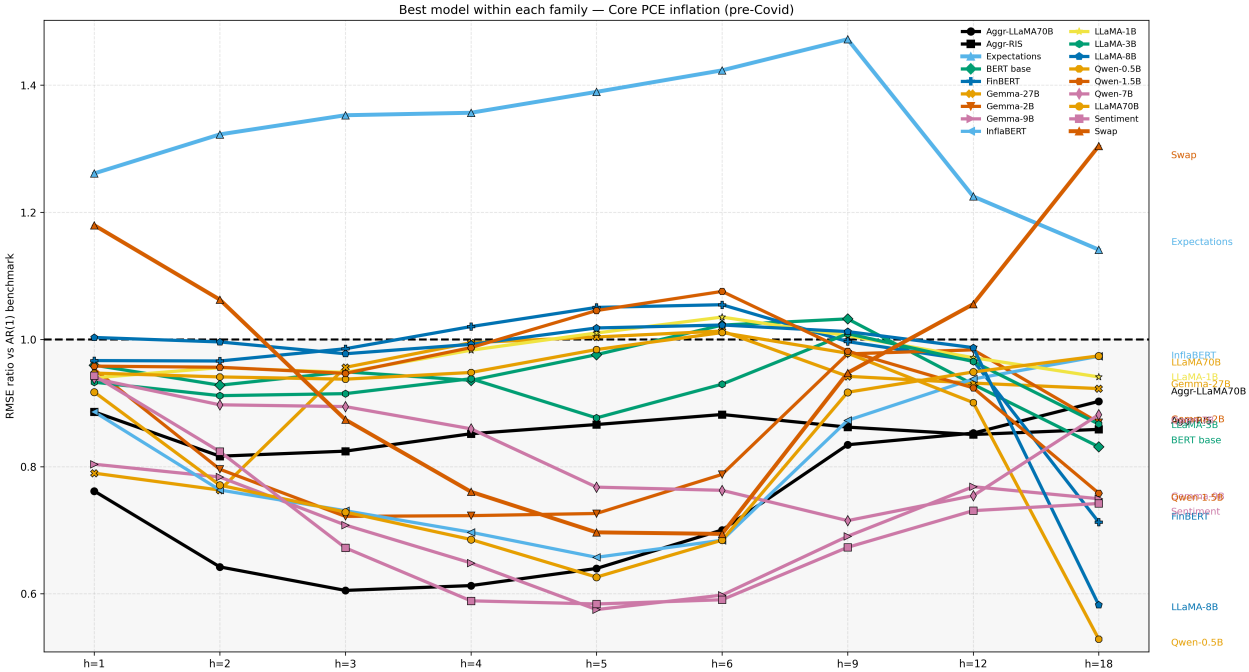

*Notes*: The figure reports RMSE ratios relative to the $AR(1)$ benchmark for CPI in the pre-COVID period, computed as $RMSE(AR\text{-}X(1))/RMSE(AR(1))$. For each forecast horizon $h \in \{1, 2, 3, 4, 5, 6, 9, 12, 18\}$, we plot the best-performing specification within each model family: best single fine-tuned LLM, best single unfine-tuned LLaMA 70B, Aggr-RIS, Aggr-LLaMA70B, lexicon-based sentiment (Sentiment), the one-year inflation swap (Swap), and Michigan expectations (Expectations). "Best" is selected across all subreddits and moving-average windows. Ratios below one indicate an improvement over the $AR(1)$ benchmark.

**Figure A.20:** Top-10 best models by forecast horizon for CPI - (pre-COVID)




*Notes*: The heatmap summarizes the top-10 forecasting specifications for CPI in the pre-COVID period at each horizon, ranked by RMSE ratio relative to the $AR(1)$ benchmark, $RMSE(AR\text{-}X(1))/RMSE(AR(1))$. Columns correspond to forecast horizons $h \in \{1, 2, 3, 4, 5, 6, 9, 12, 18\}$; each tile reports the model identifier and its RMSE ratio. The ranking is computed over all candidate specifications obtained by combining (i) model family (single fine-tuned LLM, single unfine-tuned LLaMA 70B, Aggr-RIS, Aggr-LLaMA70B, lexicon-based sentiment, one-year inflation swap, and Michigan expectations), (ii) subreddit, and (iii) moving-average window. Color encodes performance: warmer colors indicate ratios closer to one (little or no gain), while cooler colors indicate lower ratios (larger improvements). Ratios below one imply that the corresponding specification outperforms $AR(1)$.

**Figure A.21:** RMSE ratios for best models by forecast horizon for PCE - (pre-COVID)




*Notes*: The figure reports RMSE ratios relative to the $AR(1)$ benchmark for PCE in the pre-COVID period, computed as $RMSE(AR\text{-}X(1))/RMSE(AR(1))$. For each forecast horizon $h \in \{1, 2, 3, 4, 5, 6, 9, 12, 18\}$, we plot the best-performing specification within each model family: best single fine-tuned LLM, best single unfine-tuned LLaMA 70B, Aggr-RIS, Aggr-LLaMA70B, lexicon-based sentiment (Sentiment), the one-year inflation swap (Swap), and Michigan expectations (Expectations). "Best" is selected across all subreddits and moving-average windows. Ratios below one indicate an improvement over the $AR(1)$ benchmark.

**Figure A.22:** Top-10 best models by forecast horizon for PCE - (pre-COVID)




*Notes*: The heatmap summarizes the top-10 forecasting specifications for PCE in the pre-COVID period at each horizon, ranked by RMSE ratio relative to the $AR(1)$ benchmark, $RMSE(AR\text{-}X(1))/RMSE(AR(1))$. Columns correspond to forecast horizons $h \in \{1, 2, 3, 4, 5, 6, 9, 12, 18\}$; each tile reports the model identifier and its RMSE ratio. The ranking is computed over all candidate specifications obtained by combining (i) model family (single fine-tuned LLM, single unfine-tuned LLaMA 70B, Aggr-RIS, Aggr-LLaMA70B, lexicon-based sentiment, one-year inflation swap, and Michigan expectations), (ii) subreddit, and (iii) moving-average window. Color encodes performance: warmer colors indicate ratios closer to one (little or no gain), while cooler colors indicate lower ratios (larger improvements). Ratios below one imply that the corresponding specification outperforms $AR(1)$.

**Table A.14:** Forecast results (RMSE ratios) for *PCE* (pre-COVID sample).

| Model | $h=1$ | $h=2$ | $h=3$ | $h=4$ | $h=5$ | $h=6$ | $h=9$ | $h=12$ | $h=18$ |
|---|---|---|---|---|---|---|---|---|---|
| *Expectations* | | | | | | | | | |
| Expectations | 1.107 | 1.134 | 1.143 | 1.148 | 1.164 | 1.187 | 1.217 | 1.107 | 1.040 |
| 1-Y Inflation Swap | 1.070 | 1.002 | *0.910* | *0.873* | *0.854* | *0.856* | *0.932* | 0.983 | 1.063 |
| *Lexicon-based sentiment* | | | | | | | | | |
| Best sentiment | 1.126 | 1.182 | 1.304 | 1.387 | 1.969 | 1.946 | 1.749 | 1.611 | 1.385 |
| *RIS-based models* | | | | | | | | | |
| Aggr-RIS | *0.953* | *0.921* | 0.921 | 0.932 | 0.934 | 0.939 | **0.920** | **0.917** | **0.927** |
| Best fine-tuned LLM | 0.962 | 0.926 | 0.945 | 0.971 | 0.979 | 0.994 | 1.044 | 1.093 | 0.966 |
| Aggr-LLaMA70B | **0.883** | **0.829** | **0.813** | **0.815** | **0.815** | **0.848** | **0.920** | *0.930* | *0.959* |
| Best LLaMA 70B | 0.958 | 0.936 | 0.971 | 0.991 | 0.918 | 0.943 | 1.023 | 1.046 | 1.112 |

*Notes*: RMSE is computed as a ratio: $RMSE(AR\text{-}X(1))/RMSE(AR(1))$, where the benchmark is $AR(1)$. The evaluation is conducted on a pre-COVID out-of-sample period. Forecasts are generated using a recursive window, with the model re-estimated at each step. Bold (italic) indicates the best (second-best) model for each horizon.

## E.3 CSSED and fluctuation tests

This appendix provides additional graphical diagnostics for forecast performance across *all* horizons. We report (i) CSSED paths, which summarize how relative accuracy evolves, and (ii) the Giacomini and Rossi (2010) fluctuation tests, which assess whether forecast differences relative to the $AR(1)$ benchmark are locally significant within the out-of-sample period.

**CSSED.** Figures A.23 and A.24 plot CSSED paths for headline CPI and core PCE, respectively, for four competing forecasting models: Michigan expectations, the one-year inflation swap, Aggr-RIS, and Aggr-LLaMA70B. CSSED cumulates differences in squared forecast errors over the evaluation period and therefore provides a time-varying measure of relative performance.

For CPI, both Aggr-RIS and Aggr-LLaMA70B outperform the Michigan- and swap-augmented specifications at short horizons, with Aggr-LLaMA70B generally delivering the strongest performance up to $h = 6$. At longer horizons, Aggr-LLaMA70B remains the best-performing specification for most of the out-of-sample period, with its advantage particularly pronounced until early 2024.

For core PCE, the ranking is even clearer. Across horizons from $h = 1$ through $h = 18$, Aggr-LLaMA70B typically dominates, followed by Aggr-RIS; the Michigan- and swap-based specifications generally perform worse. Two exceptions are worth noting: at $h = 18$, Aggr-RIS briefly overtakes Aggr-LLaMA70B in the final part of the sample, and in the last months of 2025 the Michigan-based specification shows a relative improvement.

Figures A.25 and A.26 replicate the CSSED analysis for the pre-COVID subsample. For CPI, the ordering at short horizons remains broadly unchanged: up to $h = 6$, Aggr-RIS and Aggr-LLaMA70B generally outperform Michigan- and swap-based models. In this period, Michigan expectations occasionally perform relatively well—most notably between 2015 and 2017 for horizons $h = 4$–$6$—but otherwise tend to underperform $AR(1)$. At longer horizons ($h \geq 9$), the RIS-based forecasts continue to outperform the benchmark in most subperiods, while Michigan expectations improve mainly between mid-2015 and 2018.

For PCE in the pre-COVID sample, Aggr-RIS and Aggr-LLaMA70B outperform Michigan expectations up to $h = 9$. Michigan-based forecasts are typically weaker than $AR(1)$ up to $h = 12$, with only a short-lived improvement around late 2015; they improve in 2016–2017 and then weaken again in 2018. Inflation-swap forecasts perform poorly at the shortest horizons but become more competitive around 2018 up to $h = 9$, while they tend to underperform at longer horizons ($h = 12$ and $h = 18$). For $h \geq 9$, Aggr-RIS is often stronger than Aggr-LLaMA70B in the pre-COVID

period.

**Fluctuation tests.** Figures A.27 and A.28 report Giacomini and Rossi (2010) fluctuation statistics for CPI and PCE across all horizons ($h = 1$ to $h = 18$). Forecasts are compared to the $AR(1)$ benchmark using a rolling window of 10 observations, and critical values follow Giacomini and Rossi (2010). Values above the threshold indicate that the model's forecast performance differs significantly from $AR(1)$ over that local window.

Consistent with the main-text evidence for $h = 1$, fluctuation statistics are largely flat prior to 2019, reflecting both stable inflation dynamics and lower RIS density in earlier Reddit years. In contrast, Aggr-RIS and Aggr-LLaMA70B display statistically significant gains immediately after the COVID shock and during the 2022 energy-price episode, with the strongest and most persistent improvements concentrated between 2020 and end-2023. This pattern holds for both CPI and core PCE across horizons. At the longest horizon ($h = 18$), significant differences are more localized, clustering around late 2023 and early 2024.

**Summary.** Two conclusions emerge. First, CSSED paths confirm that RIS-based models deliver sustained accuracy gains relative to traditional expectation- and swap-based predictors, with particularly strong performance for Aggr-LLaMA70B in the full sample. Second, fluctuation tests show that these gains are state-dependent: they become statistically salient during high-volatility episodes (COVID and the energy-price shock), while differences are muted in the low-inflation, low-variance pre-2019 period. Together, these diagnostics corroborate the main results and clarify when RIS variants are most informative.

**Figure A.23:** CSSED plots for CPI (headline) at different forecast horizons.

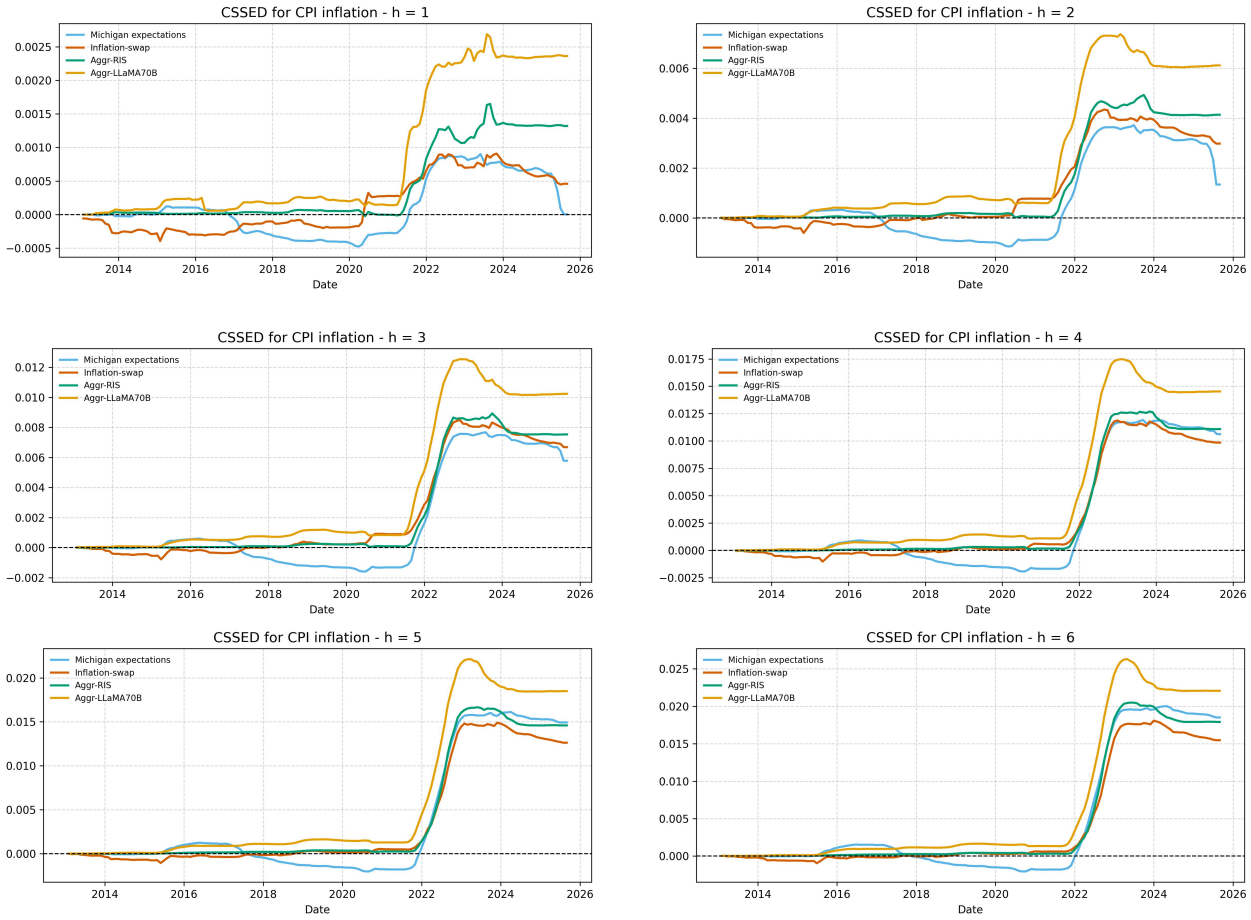
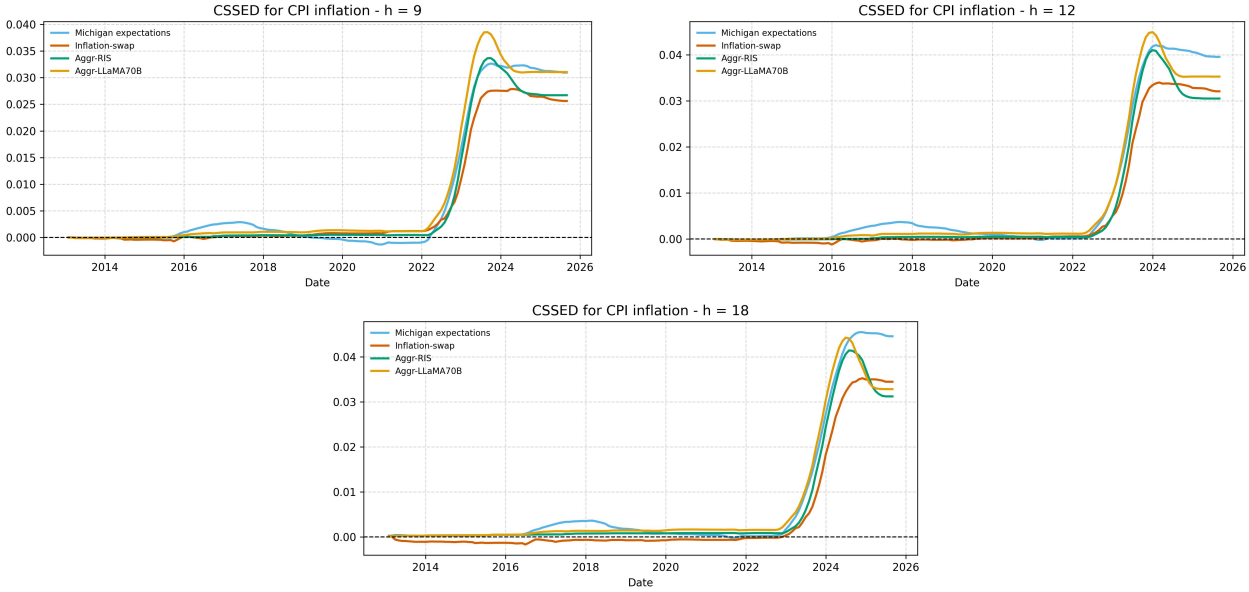


**Figure A.23:** CSSED plots for CPI (headline) at different forecast horizons (cont.).




*Notes*: This figure reports the CSSED for CPI (headline) forecasts across alternative horizons. Positive values indicate that the alternative model outperforms the autoregressive benchmark.

**Figure A.24:** CSSED plots for PCE (core) at different forecast horizons.

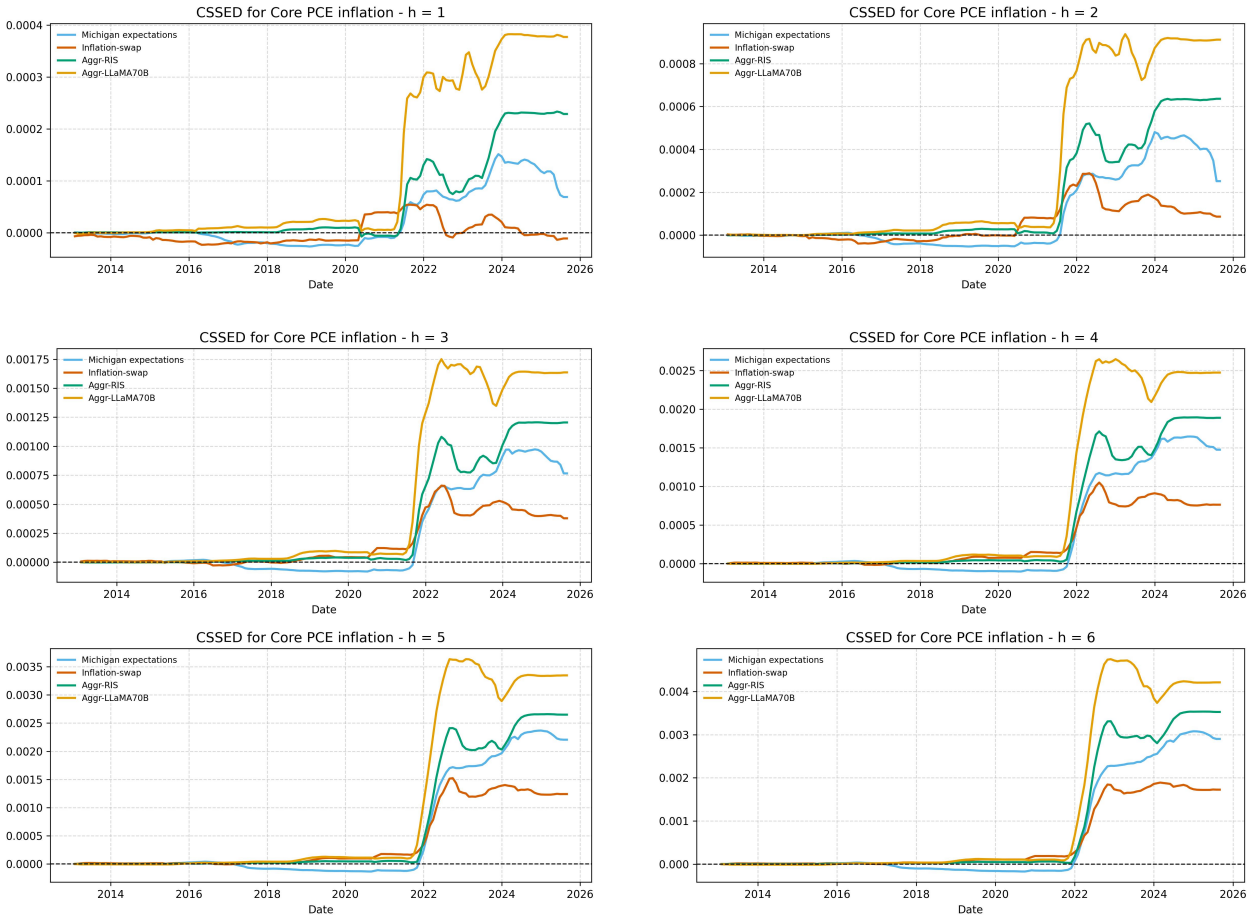
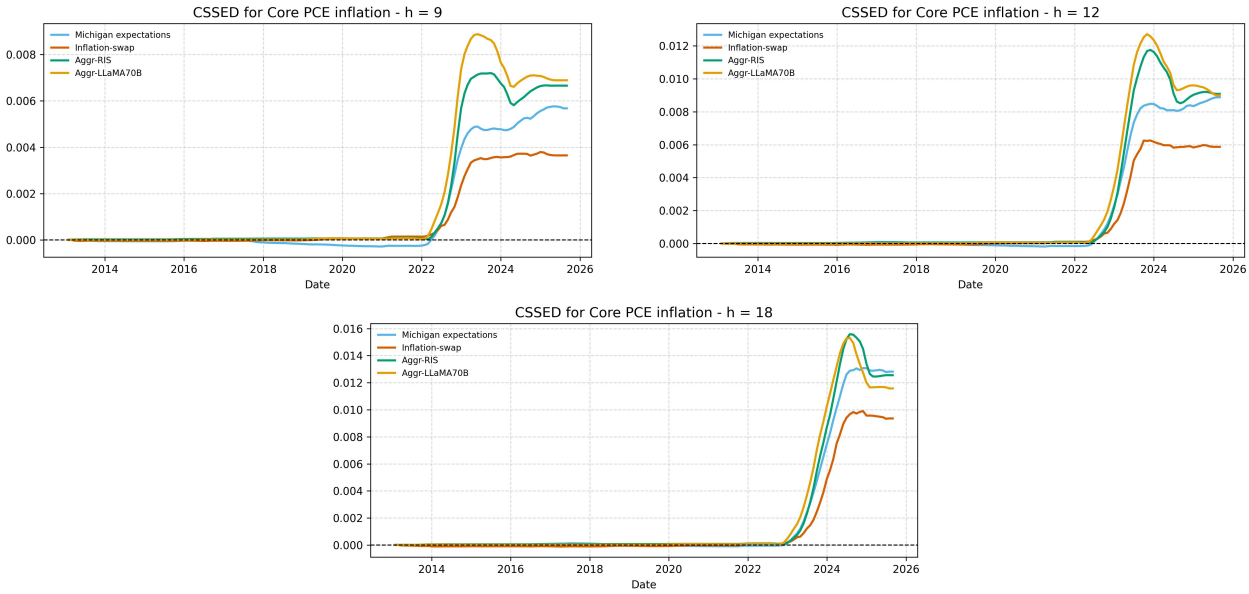


**Figure A.24:** CSSED plots for PCE (core) at different forecast horizons (cont.).

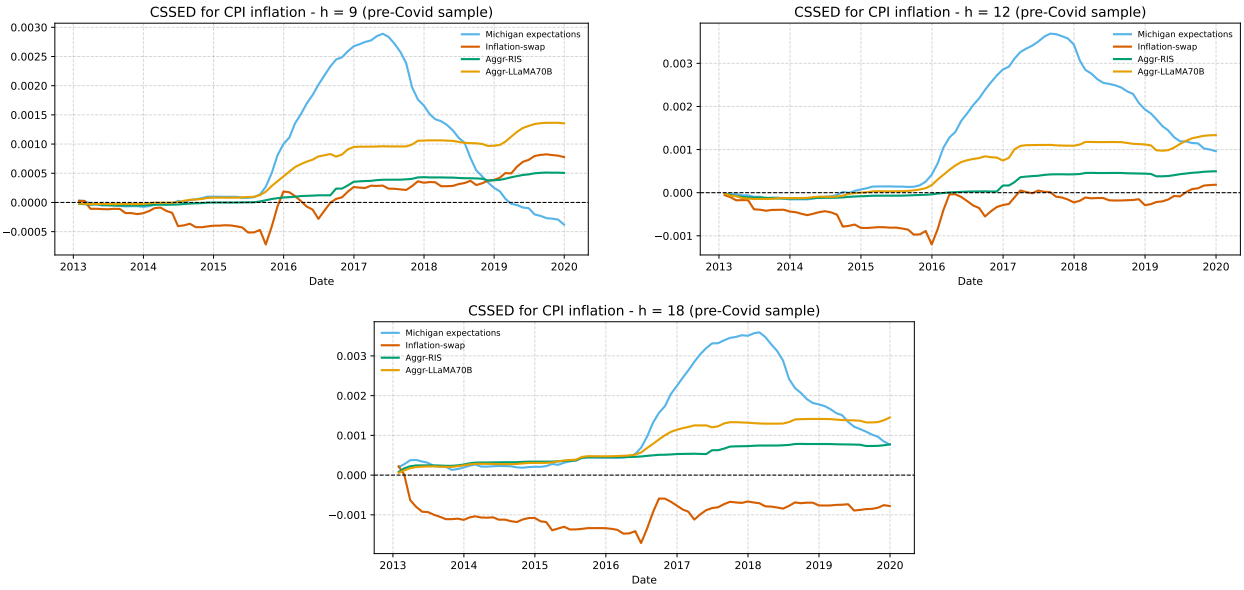



*Notes*: This figure reports the CSSED for PCE (core) forecasts across alternative horizons. Positive values indicate that the alternative model outperforms the autoregressive benchmark.

**Figure A.25:** CSSED plots for CPI (headline) at different forecast horizons (pre-COVID).



**Figure A.25:** CSSED plots for CPI (headline) at different forecast horizons (pre-COVID, cont.).




*Notes*: This figure reports the CSSED for CPI (headline) forecasts across alternative horizons in the pre-COVID period. Positive values indicate that the alternative model outperforms the autoregressive benchmark.

**Figure A.26:** CSSED plots for PCE (core) at different forecast horizons (pre-COVID).

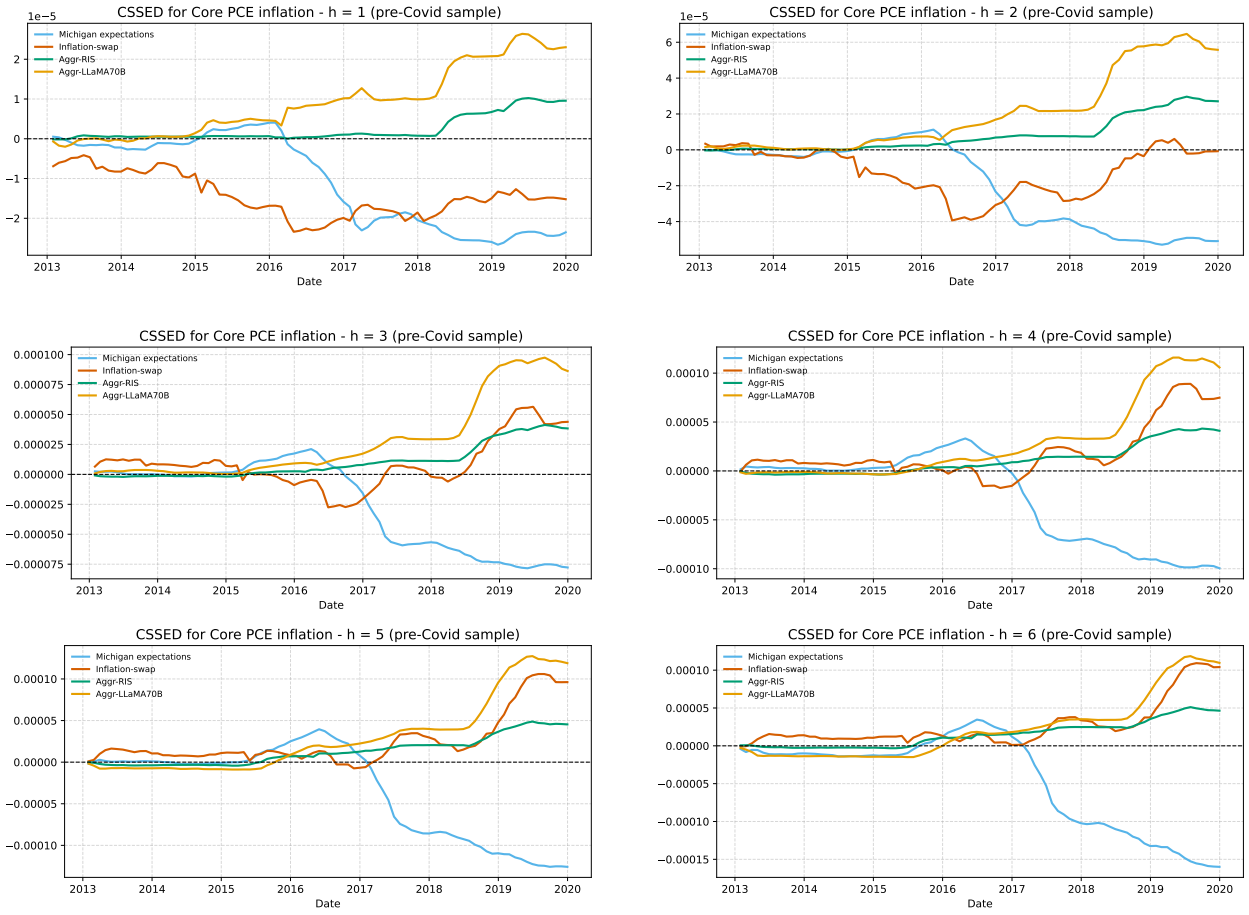
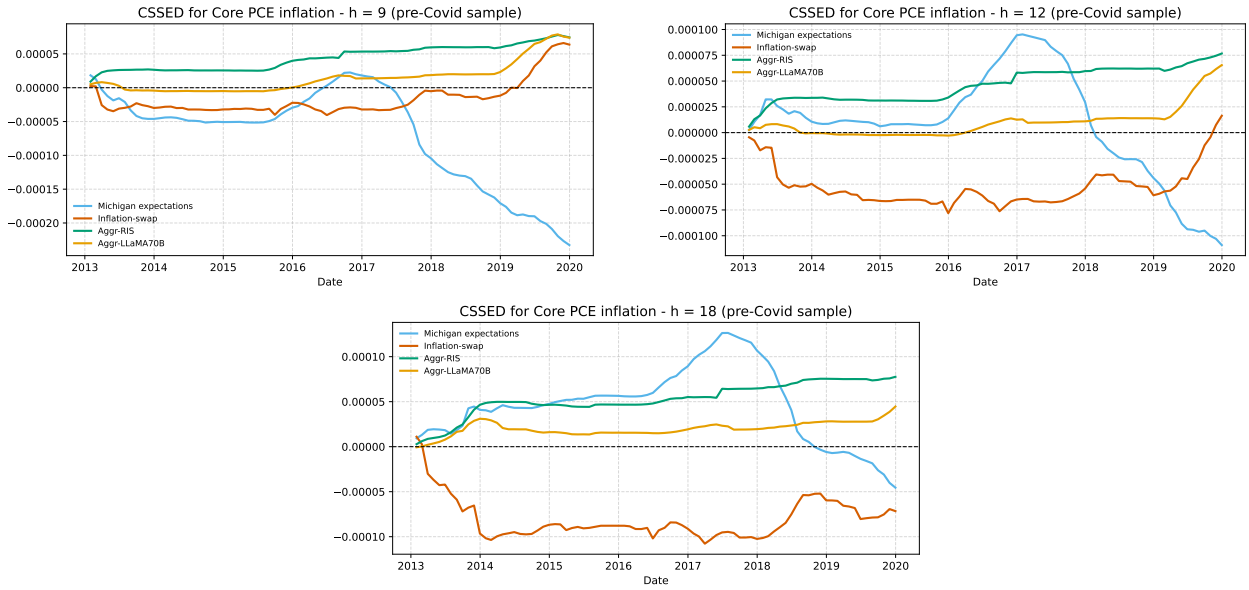


**Figure A.26:** CSSED plots for PCE (core) at different forecast horizons (pre-COVID, cont.).




*Notes*: This figure reports the CSSED for PCE (core) forecasts across alternative horizons in the pre-COVID period. Positive values indicate that the alternative model outperforms the autoregressive benchmark.

**Figure A.27:** Fluctuation test plots for CPI (headline) at different forecast horizons.

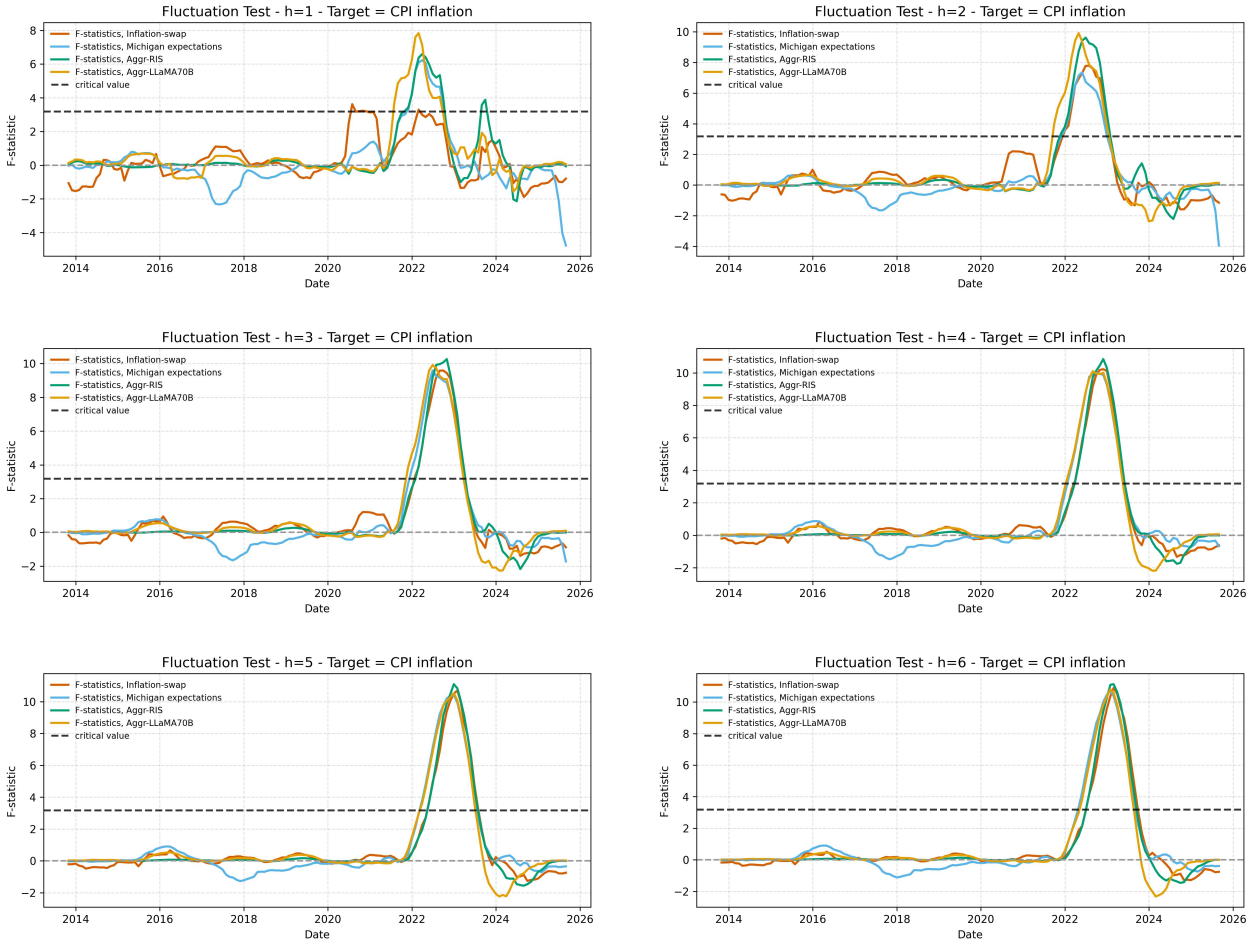
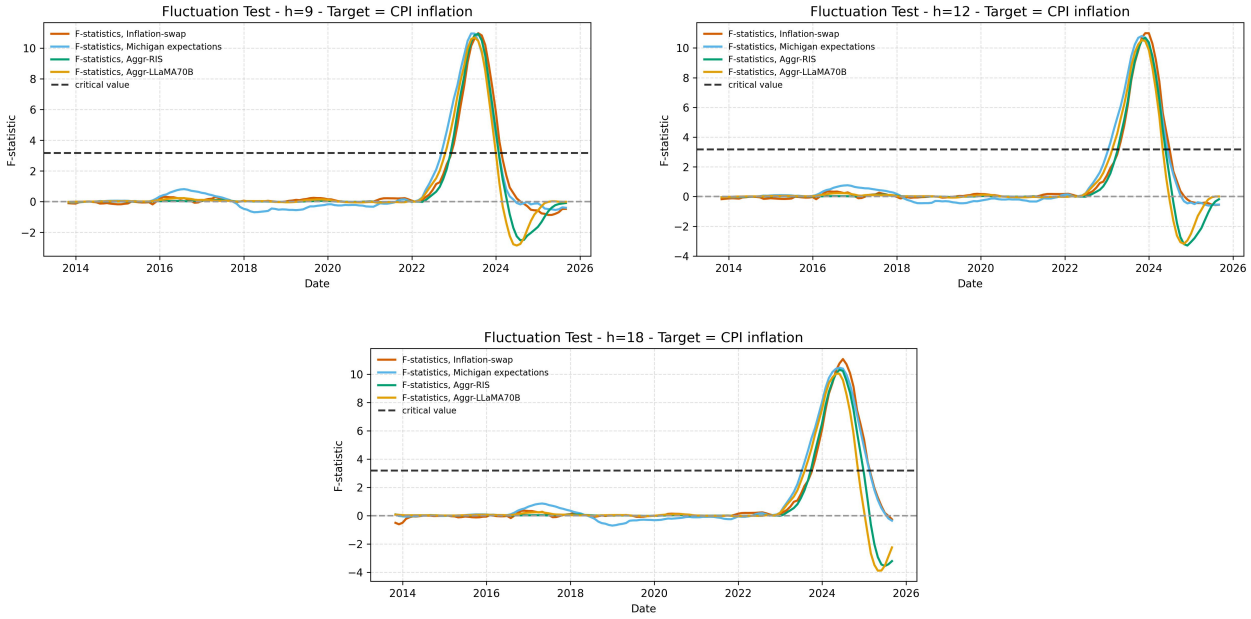


**Figure A.27:** Fluctuation test plots for CPI (headline) at different forecast horizons (cont.).




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

**Figure A.28:** Fluctuation test plots for PCE (core) at different forecast horizons.

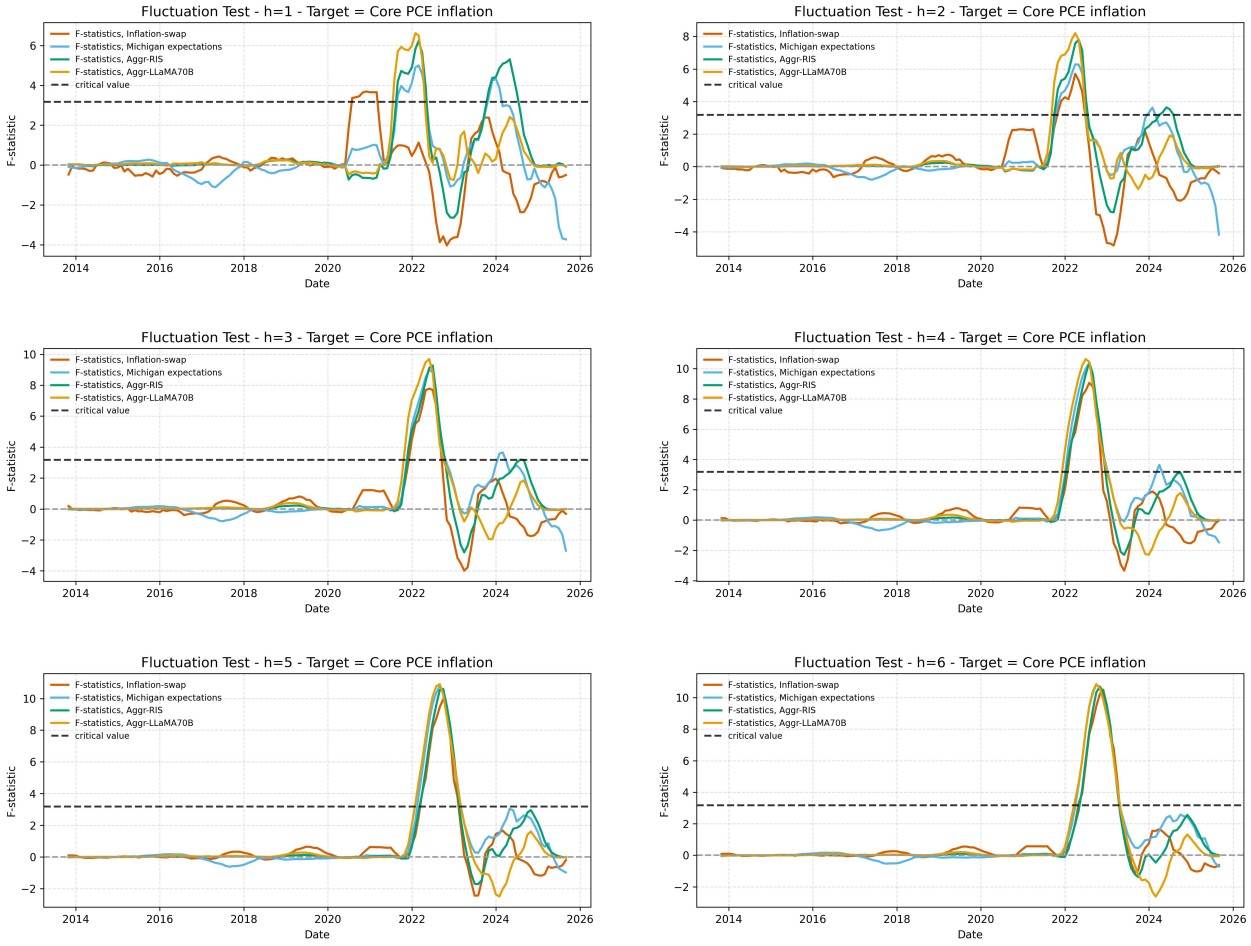
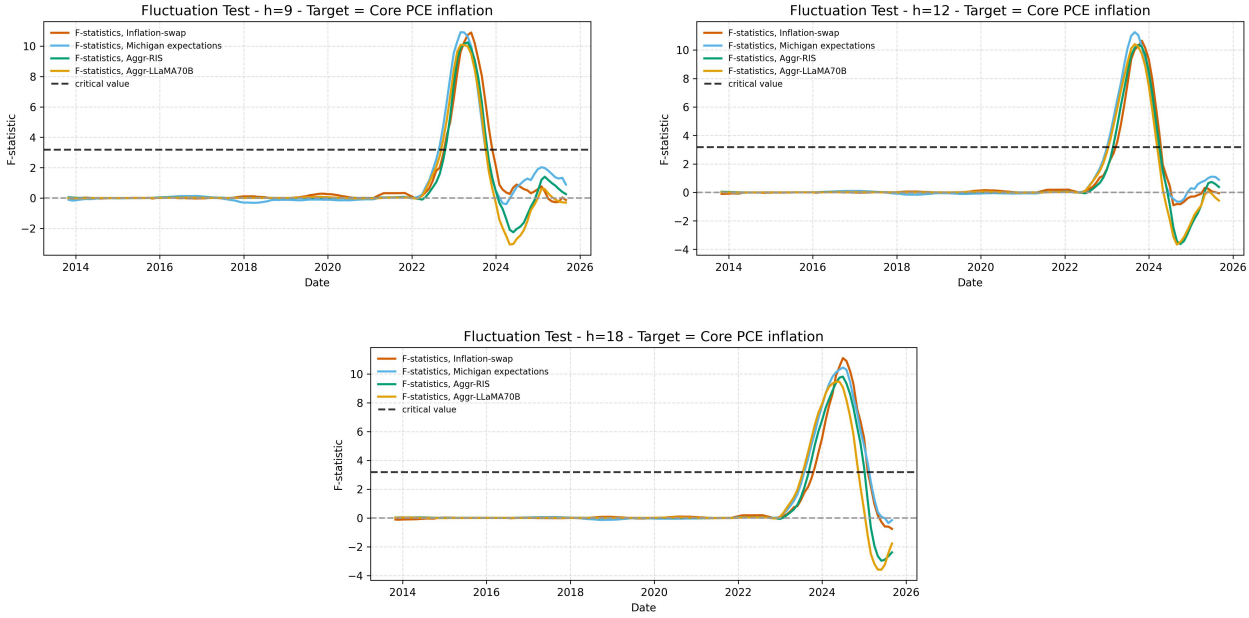


**Figure A.28:** Fluctuation test plots for PCE (core) at different forecast horizons (cont.).




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

## E.4 Alternative benchmark: UCSV

As a robustness check, we evaluate forecast accuracy relative to the unobserved-components stochastic-volatility (UCSV) model, a widely used univariate benchmark for U.S. inflation forecasting (Stock and Watson, 2007; Faust and Wright, 2013). The model decomposes inflation into a time-varying trend and a transitory component:

$$\pi_t = \tau_t + \eta_t^T, \tag{A.3}$$

$$\tau_t = \tau_{t-1} + \eta_t^P, \tag{A.4}$$

where $\eta_t^T \overset{iid}{\sim} \mathcal{N}(0, \sigma_{T,t}^2)$ and $\eta_t^P \overset{iid}{\sim} \mathcal{N}(0, \sigma_{P,t}^2)$. The volatilities evolve as random walks in logs:

$$\log(\sigma_{T,t}^2) = \log(\sigma_{T,t-1}^2) + \psi_{1,t}, \tag{A.5}$$

$$\log(\sigma_{P,t}^2) = \log(\sigma_{P,t-1}^2) + \psi_{2,t}, \tag{A.6}$$

with $(\psi_{1,t}, \psi_{2,t})' \overset{iid}{\sim} \mathcal{N}(0, 0.2\, I_2)$. Forecasts are obtained from the filtered estimate of the trend component $\tau_t$; in particular, the $h$-step-ahead forecast of $\pi_{t+h}$ is given by the filtered trend estimate at time $t$.

Tables A.15 and A.16 report RMSE ratios for the main forecast specifications relative to the UCSV benchmark for headline CPI and core PCE, respectively. For reference, we also report the RMSE of the $AR(1)$ benchmark used in the main text. Two patterns emerge. First, the ranking of benchmarks depends on horizon: $AR(1)$ tends to outperform UCSV at very short horizons (up to three months for CPI and up to two months for PCE), whereas UCSV becomes more competitive—and typically outperforms $AR(1)$—at longer horizons. Second, and most importantly, the relative performance of the RIS-based models is qualitatively unchanged. For CPI, Aggr-LLaMA70B remains the strongest specification up to six months ahead, followed by the best single LLaMA 70B model, mirroring the main-text results. For PCE, Aggr-LLaMA70B performs best at short horizons (up to five months), while at longer horizons the dictionary-based sentiment benchmark becomes more competitive, followed by the best LLaMA 70B or the best fine-tuned LLM specification.

Overall, replacing the $AR(1)$ benchmark with the standard UCSV benchmark does not alter the central conclusion of the paper: RIS variants—especially those derived from LLMs and aggregated using MSE-based weights—deliver robust forecasting gains relative to strong univariate inflation benchmarks.

**Table A.15:** RMSE ratios relative to the UCSV benchmark for CPI forecasts.

| | Horizon $h$ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Model | 1 | 2 | 3 | 4 | 5 | 6 | 9 | 12 | 18 |
| *Expectations* | | | | | | | | | |
| Expectations | 0.786 | 0.875 | 0.836 | 0.807 | *0.799* | *0.804* | 0.744 | *0.625* | **0.471** |
| Swap | 0.760 | 0.812 | 0.808 | 0.828 | 0.855 | 0.871 | 0.830 | 0.722 | 0.580 |
| *Lexicon-based sentiment* | | | | | | | | | |
| Sentiment | 0.743 | 0.828 | 0.852 | 0.855 | 0.856 | 0.851 | **0.736** | **0.578** | *0.481* |
| *RIS-based models* | | | | | | | | | |
| Aggr-RIS | 0.708 | 0.764 | 0.781 | 0.794 | 0.808 | 0.818 | 0.813 | 0.741 | 0.611 |
| Best fine-tuned LLM | 0.686 | 0.771 | 0.810 | 0.832 | 0.864 | 0.889 | 0.935 | 0.843 | 0.567 |
| Aggr-LLaMA70B | *0.639* | **0.676** | **0.687** | **0.691** | **0.704** | **0.718** | *0.742* | 0.682 | 0.596 |
| Best LLaMA 70B | **0.621** | *0.705* | *0.758* | *0.771* | 0.802 | 0.829 | 0.899 | 0.789 | 0.637 |
| *Benchmarks* | | | | | | | | | |
| UCSV | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| AR(1) | 0.786 | 0.922 | 0.997 | 1.059 | 1.113 | 1.153 | 1.158 | 1.041 | 0.853 |

*Notes*: Entries report RMSE ratios relative to the UCSV benchmark for CPI forecasts. Values below one indicate an improvement over UCSV. Bold (italic) indicates the lowest (second-lowest) RMSE ratio in each column.

**Table A.16:** RMSE ratios relative to the UCSV benchmark for PCE forecasts.

| | Horizon $h$ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Model | 1 | 2 | 3 | 4 | 5 | 6 | 9 | 12 | 18 |
| *Expectations* | | | | | | | | | |
| Expectations | 0.761 | 0.878 | 0.919 | 0.944 | 0.982 | 1.026 | 1.077 | 1.047 | 0.767 |
| Swap | 0.777 | 0.904 | 0.967 | 1.021 | 1.074 | 1.123 | 1.197 | 1.196 | 0.879 |
| *Lexicon-based sentiment* | | | | | | | | | |
| Sentiment | 0.751 | 0.872 | 0.890 | 0.905 | *0.901* | **0.880** | **0.761** | **0.651** | **0.443** |
| *RIS-based models* | | | | | | | | | |
| Aggr-RIS | 0.729 | 0.816 | 0.862 | 0.896 | 0.936 | 0.971 | 1.015 | 1.036 | 0.777 |
| Best fine-tuned LLM | 0.716 | 0.819 | 0.881 | 0.935 | 0.977 | 1.018 | 1.093 | 1.064 | *0.660* |
| Aggr-LLaMA70B | **0.698** | **0.769** | **0.801** | **0.823** | **0.860** | *0.906* | 0.999 | 1.040 | 0.809 |
| Best LLaMA 70B | *0.710* | *0.801* | *0.842* | *0.870* | 0.902 | 0.926 | *0.954* | *0.955* | 0.754 |
| *Benchmarks* | | | | | | | | | |
| UCSV | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| AR(1) | 0.775 | 0.917 | 1.012 | 1.098 | 1.182 | 1.253 | 1.387 | 1.443 | 1.126 |

*Notes*: Entries report RMSE ratios relative to the UCSV benchmark for PCE forecasts. Values below one indicate an improvement over UCSV. Bold (italic) indicates the lowest (second-lowest) RMSE ratio in each column.

## E.5 Model Confidence Set

In this section we comment on the results of the MCS procedure of Hansen et al. (2011), which 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 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

$$d_{mn,t+h} \;=\; L(m, t+h) - L(n, t+h), \qquad m, n = 1, \ldots, M, \;\; t = R, \ldots, T-h. \tag{A.7}$$

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

$$d_{m\cdot,t+h} \;=\; \frac{1}{|\mathcal{M}| - 1} \sum_{n \in \mathcal{M}} d_{mn,t+h}, \qquad m \in \mathcal{M}. \tag{A.8}$$

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

$$H_{0,\mathcal{M}} : \; c_{mn} = 0, \quad \forall\, m, n \in \mathcal{M}, \tag{A.9}$$

$$H_{A,\mathcal{M}} : \; c_{mn} \neq 0, \quad \text{for some } m, n \in \mathcal{M}, \tag{A.10}$$

while the "against-the-average" formulation is

$$H_{0,\mathcal{M}} : \; c_{m\cdot} = 0, \quad \forall\, m \in \mathcal{M}, \tag{A.11}$$

$$H_{A,\mathcal{M}} : \; c_{m\cdot} \neq 0, \quad \text{for some } m \in \mathcal{M}, \tag{A.12}$$

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

$$\bar{d}_{mn} := \frac{1}{P_h} \sum_{t=R}^{T-h} d_{mn,t+h}, \qquad \bar{d}_{m\cdot} := \frac{1}{|\mathcal{M}| - 1} \sum_{n \in \mathcal{M}} \bar{d}_{mn}. \tag{A.13}$$

The corresponding studentized statistics are

$$t_{mn} = \frac{\bar{d}_{mn}}{\widehat{\sigma}(\bar{d}_{mn})}, \qquad t_{m\cdot} = \frac{\bar{d}_{m\cdot}}{\widehat{\sigma}(\bar{d}_{m\cdot})}, \tag{A.14}$$

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.9) naturally leads to the range statistic

$$T_{R,\mathcal{M}} = \max_{m,n \in \mathcal{M}} |t_{mn}|, \tag{A.15}$$

while the "against-the-average" EPA hypothesis in (A.11) maps to

$$T_{\max,\mathcal{M}} = \max_{m \in \mathcal{M}} t_{m\cdot}. \tag{A.16}$$

**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.15), the eliminated model is

$$e_{R,\mathcal{M}} = \arg\max_{m \in \mathcal{M}} \left( \sup_{n \in \mathcal{M}} t_{mn} \right), \tag{A.17}$$

whereas under the max statistic in (A.16) one removes

$$e_{\max,\mathcal{M}} = \arg\max_{m \in \mathcal{M}} t_{m\cdot}. \tag{A.18}$$

**Algorithm (MCS procedure).**

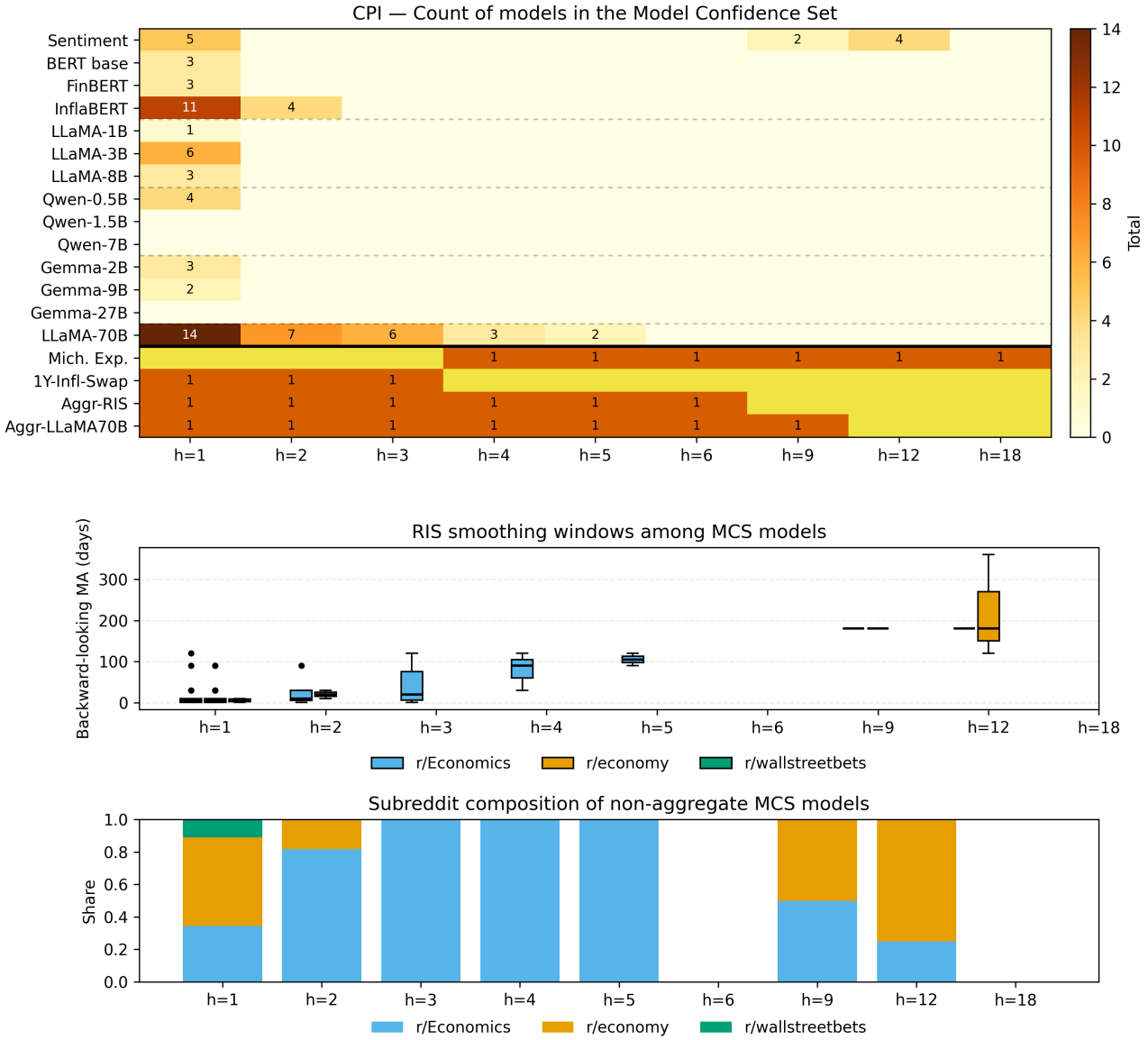
1. Initialize $\mathcal{M} \leftarrow \mathcal{M}_0$.
2. Test the EPA null $H_{0,\mathcal{M}}$ using either $T_{R,\mathcal{M}}$ in (A.15) or $T_{\max,\mathcal{M}}$ in (A.16). If EPA is *not* rejected at level $\alpha$, stop and set $\mathcal{M}^*_{1-\alpha} = \mathcal{M}$.
3. If EPA is rejected, determine the worst model using (A.17) or (A.18), 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.29 shows the MCS results for the CPI forecasting exercise. The top chart shows the number of RIS-based forecasting models entering the MCS at horizons from one to 18 months. At one month ahead, the models entering most frequently are those based on the unfine-tuned LLaMA 70B (14 models), fine-tuned InflaBERT (11), fine-tuned LLaMA 3B (6), a standard lexicon (5), and fine-tuned Qwen 0.5B (4). Forecast aggregations based on LLaMA 70B variants (Aggr-LLaMA70B) or smaller fine-tuned LLM variants (Aggr-RIS) enter the final MCS at nearly all horizons up to nine months. Thus, forecasts based on aggregates of fine-tuned SLMs or the unfine-tuned LLaMA 70B generally outperform forecasts based on Michigan expectations (which become more competitive from four to 18 months ahead) and the one-year inflation swap (which enters the MCS mainly at horizons up to three months). At nine and 12 months ahead, a few lexicon-based sentiment models also enter the MCS. The middle chart shows the distribution of MA windows across subreddits. At the one-month horizon, models in the MCS use all three subreddits with MA windows up to 100 days. At longer horizons up to nine months, selected models rely primarily on `r/Economics` with MA windows below 200 days; at 12 months ahead, they rely mainly on `r/economy` and MA windows longer than 100 days. The bottom chart reports each subreddit's contribution to MCS inclusion by horizon. At one month ahead, all three subreddits contribute; at longer horizons, `r/Economics` and `r/economy` account for most RIS variants that enter the MCS.

Figure A.30 shows the corresponding MCS results for PCE. The top chart reports the number of RIS-based forecasting models entering the MCS at horizons from one to 18 months. At one month ahead, the models entering most frequently are those based on the unfine-tuned LLaMA 70B (14 models), fine-tuned InflaBERT (11), fine-tuned LLaMA 3B (6), a standard lexicon (5), and fine-tuned Qwen 0.5B (4). Forecast aggregations based on LLaMA 70B variants (Aggr-LLaMA70B) or smaller fine-tuned LLM variants (Aggr-RIS) enter the final MCS at nearly all horizons up to nine months. Thus, forecasts based on aggregates of fine-tuned SLMs or the unfine-tuned LLaMA 70B

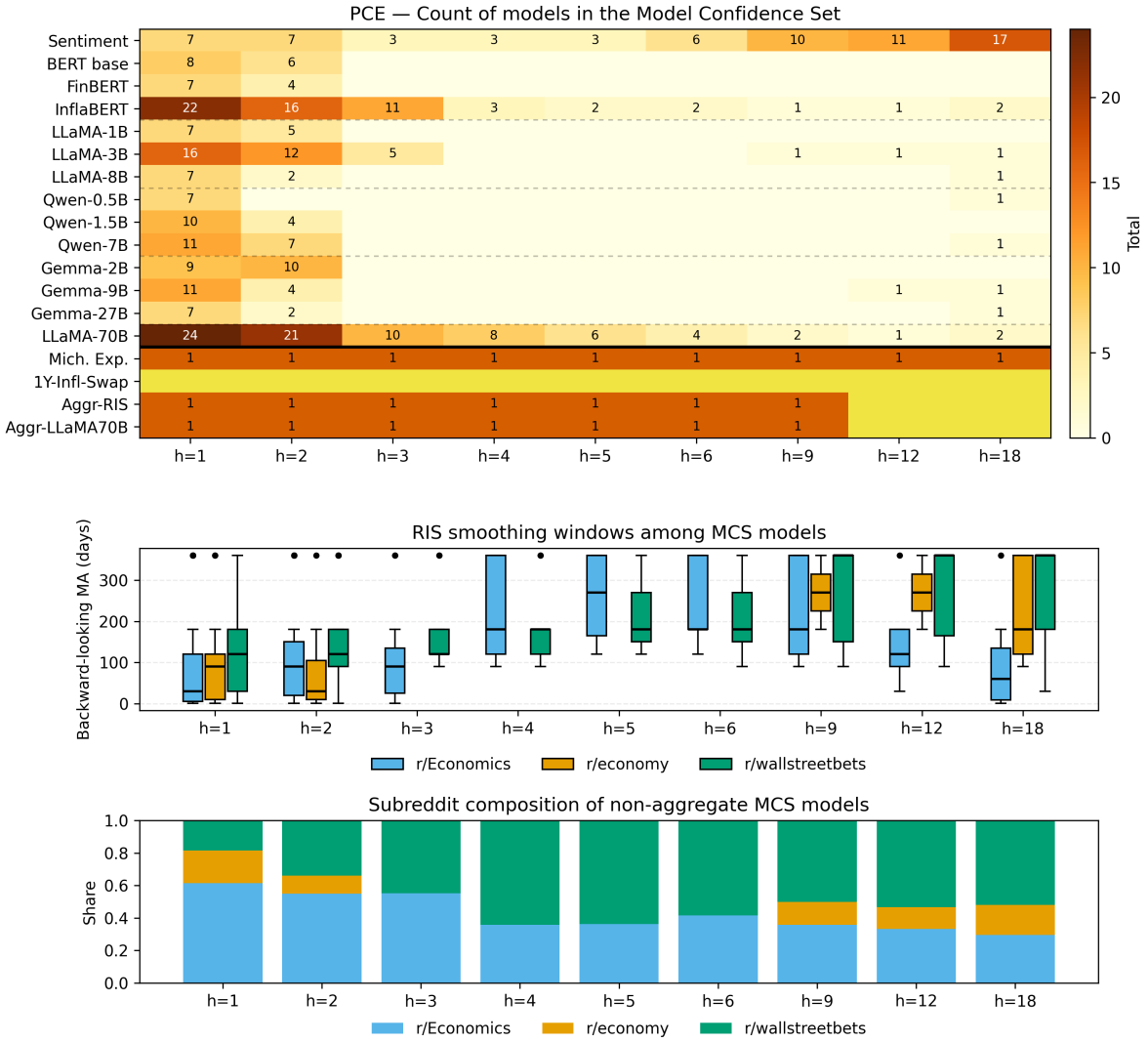
generally outperform forecasts based on Michigan expectations and the one-year inflation swap. At nine and 12 months ahead, some lexicon-based sentiment models also enter the MCS. The middle chart shows the distribution of MA windows across subreddits, while the bottom chart reports each subreddit's contribution to MCS inclusion by horizon.

**Figure A.29:** Results for MCS test for CPI ($h = 1, \ldots, 18$)




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

**Figure A.30:** Results for MCS test for PCE ($h = 1, \ldots, 18$)




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

## E.6 Predictive Regressions

This section reports an *in-sample* predictive-regression exercise designed to quantify the incremental explanatory power of RIS variants 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 the RIS adds statistically significant predictive content for inflation beyond conventional expectation measures and lagged inflation.

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

$$\pi_{t+h} = \alpha + \beta_1 \pi_t + \beta_2 \pi_t^e + \beta_3 \text{1YSwap}_t + \beta_4 \text{RIS}_t + \varepsilon_t, \tag{A.19}$$

where $\pi_t$ denotes inflation, $\pi_t^e$ is survey-based inflation expectations (Michigan), $\text{1YSwap}_t$ is the one-year inflation swap rate, and $\text{RIS}_t$ denotes the RIS variant under consideration. The regression is estimated separately for each RIS variant (fine-tuned LLM and LLaMA 70B variants), subreddit, and moving-average (MA) specification.

**Hypothesis and inference.** While individual coefficient estimates for $\text{RIS}_t$ are often significant, our main focus is whether the RIS improves 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 $\text{RIS}_t$ 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).[47]

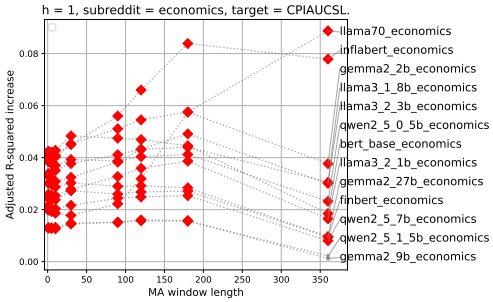
[47] The shorter sample reflects the later emergence of the subreddit and implies that results for `r/wallstreetbets` are not directly comparable in precision to those for the other subreddits. This limitation is less consequential in the forecasting/nowcasting applications because the forecast-combination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

**Graphical summary.** Figures A.31 and A.32 summarize results for CPI and PCE, respectively, focusing on horizons $h = 1$ and $h = 6$ for readability.[48] In each panel, the vertical axis reports the incremental change in adjusted $R^2$ from adding $\text{RIS}_t$ to the baseline specification, while the horizontal axis indexes the MA window used to smooth the RIS variant. Markers indicate whether the F-test rejects $H_0 : \beta_4 = 0$ at the 5% level.

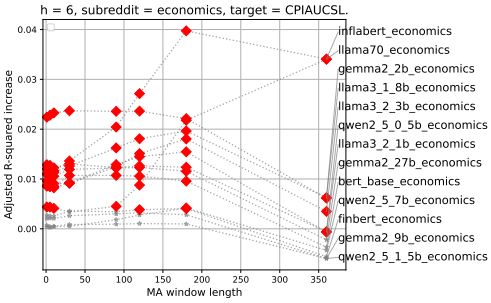
**Results.** Figures A.31 and A.32 show that RIS variants provide statistically significant incremental explanatory power across model families, subreddits, and horizons, and for both CPI and PCE. The improvements in fit are broadly robust to the choice of MA window: while the magnitude of the adjusted $R^2$ 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 the RIS is distributed across heterogeneous sources rather than concentrated in a single community 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, RIS variants often remain significant. Overall, the predictive-regression evidence is consistent with the out-of-sample results: the RIS contains incremental information about inflation dynamics beyond lagged inflation and standard survey- and market-based expectations measures.

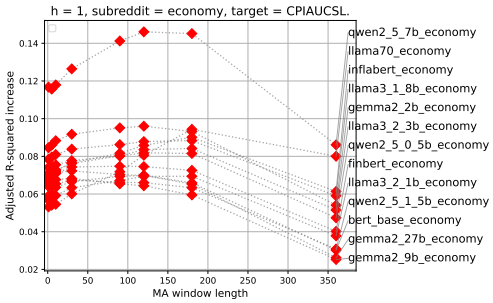
[48] We omit intermediate horizons to preserve clarity; patterns are similar across the remaining horizons.



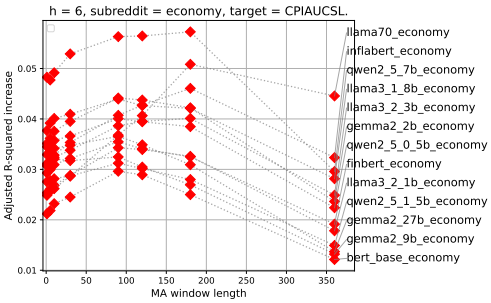

**(a)** `r/Economics` ($h = 1$), CPI.



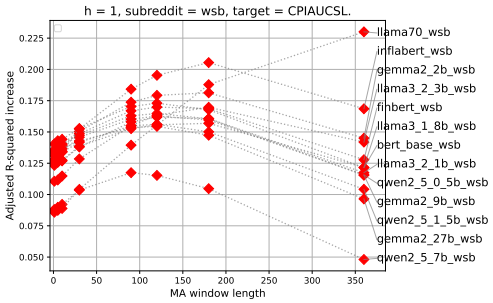

**(b)** `r/Economics` ($h = 6$), CPI.



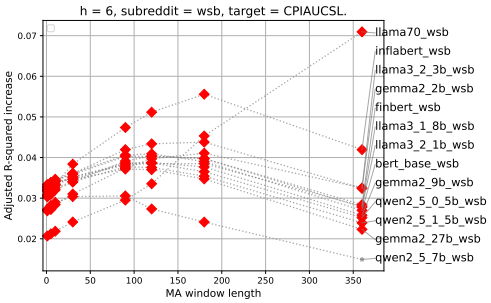

**(c)** `r/economy` ($h = 1$), CPI.




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

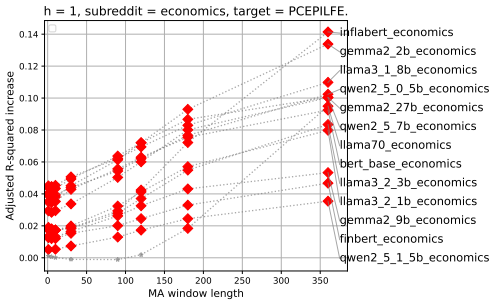



**(e)** `r/wallstreetbets` ($h = 1$), CPI.

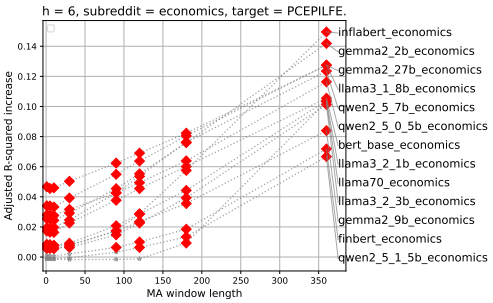



**(f)** `r/wallstreetbets` ($h = 6$), CPI.

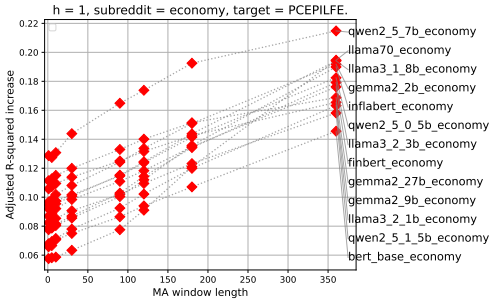
**Figure A.31:** CPI predictive regressions: incremental adjusted $R^2$ from adding a RIS variant $\text{RIS}_t$ to a baseline regression including lagged inflation, Michigan expectations, and the one-year inflation swap rate. Markers denote rejection of $H_0 : \beta_4 = 0$ at the 5% level (HAC inference). Samples: 2009M1–2025M8 for `r/Economics` and `r/economy`; 2016M1–2025M8 for `r/wallstreetbets`.



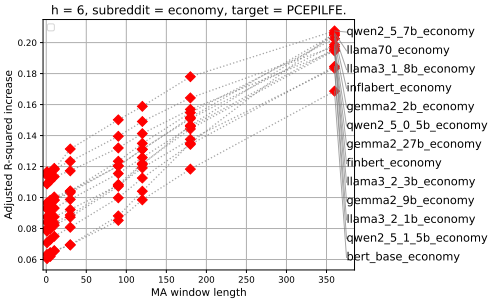

**(a)** `r/Economics` ($h = 1$), PCE.



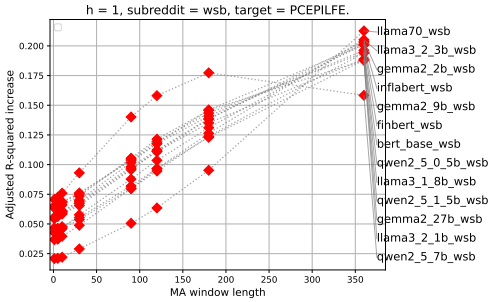

**(b)** `r/Economics` ($h = 6$), PCE.



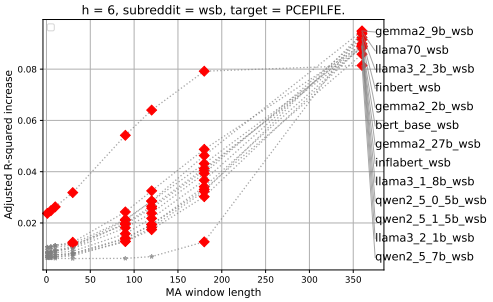

**(c)** `r/economy` ($h = 1$), PCE.




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




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




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

**Figure A.32:** PCE predictive regressions: incremental adjusted $R^2$ from adding a RIS variant $\text{RIS}_t$ 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.7 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 the RIS improves 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

$$\mathcal{Y}_{t|\tau} = x_t'\boldsymbol{\beta}(\tau) + \epsilon_{t|\tau}, \qquad \tau \in (0,1), \tag{A.20}$$

so that

$$\Pr\big(\mathcal{Y}_t \leq x_t'\boldsymbol{\beta}(\tau) \mid x_t\big) = \tau. \tag{A.21}$$

The estimator minimizes the quantile ("check") loss

$$\rho_\tau(u) = \begin{cases} \tau u, & u \geq 0, \\ (\tau - 1)u, & u < 0, \end{cases} \tag{A.22}$$

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

$$\widehat{\boldsymbol{\beta}}(\tau) = \arg\min_{\boldsymbol{\beta}} \frac{1}{T}\sum_{t=1}^{T} \rho_\tau(u_t(\tau))\,. \tag{A.23}$$

Because the objective is not differentiable at $u = 0$, estimation is performed numerically.[49]

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

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

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

$$\text{CRPS}(F, y) = \int \left(F(x) - \mathbf{1}_{\{x \geq y\}}\right)^2 dx, \tag{A.24}$$

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.17 and A.18 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 RIS variants but focus on variants smoothed with a 90-day backward moving average. For computational reasons, we do not compute forecast combinations for density forecasts.[50] 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-based RIS variants 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-based RIS variants remain the best performers up to $h = 6$, while Michigan expectations become relatively more competitive at longer horizons. Among the fine-tuned models, BERT-based specifications—especially InflaBERT—are consistently among the strongest performers.

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

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

[50] 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.

**Table A.17:** Forecast results (CRPS ratios) for *CPI* (90-day MA RIS variants).

| 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 |
| *RIS variants (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: $CRPS(AR\text{-}X(1))/CRPS(AR(1))$. 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 at each step. The **bold** entry in each column marks the lowest relative CRPS among the reported models for that horizon. Stars $(*, **, ***)$ are the reported significance indicators. LLM rows use a **90-day moving-average** smoothing of the RIS.

where $\text{RIS}_t$ denotes either the InflaBERT-based variant or the LLaMA 70B-based variant. 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).

**Table A.18:** Forecast results (CRPS ratios) for *PCE* (90-day MA RIS variants).

| 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 |
| *RIS variants (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: $CRPS(AR\text{-}X(1))/CRPS(AR(1))$. 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 the reported significance levels. LLM rows use a **90-day moving-average** smoothing of the RIS.

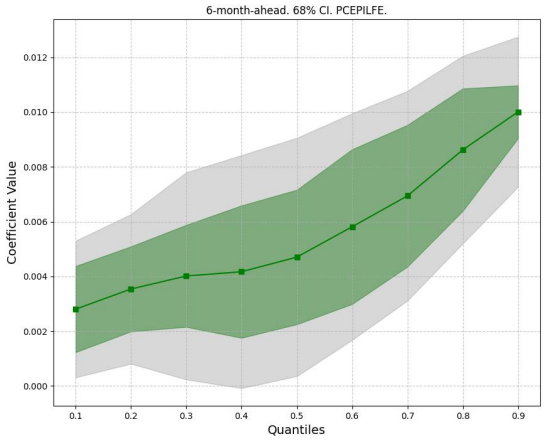
Our approach is inspired by Lopez-Salido and Loria (2024), who study inflation dynamics through quantile-dependent coefficients. Our objective here is not structural identification,[51] but rather to document whether the RIS is more informative in particular regions of the inflation distribution.

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

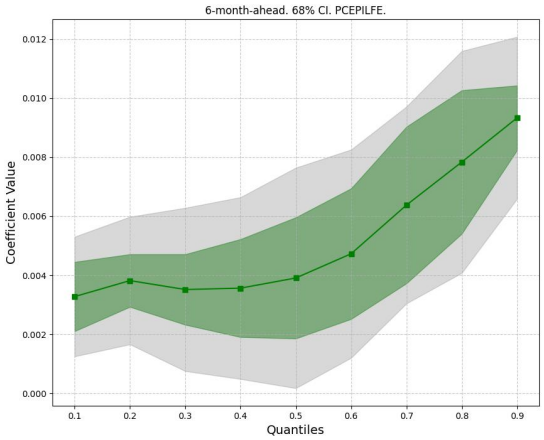
Figure A.33 plots the estimated coefficients across quantiles. The RIS variant displays a profile similar to conventional expectation measures: its predictive contribution tends to be stronger in parts of the distribution associated with elevated inflation outcomes, consistent with state-dependent attention 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.

**Figure A.33:** Estimated coefficients across quantiles of the dependent variable (PCE or CPI).

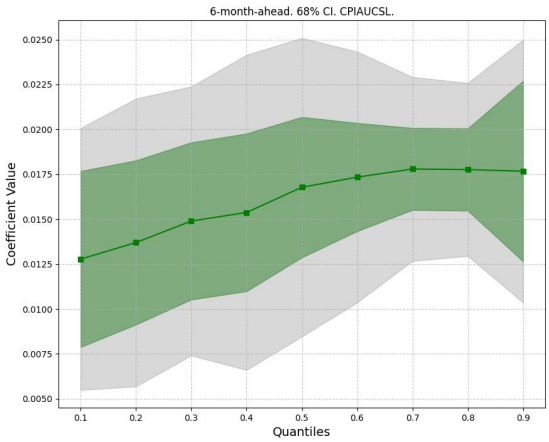
**(a)** PCE / `LLaMA 3.1 70B`



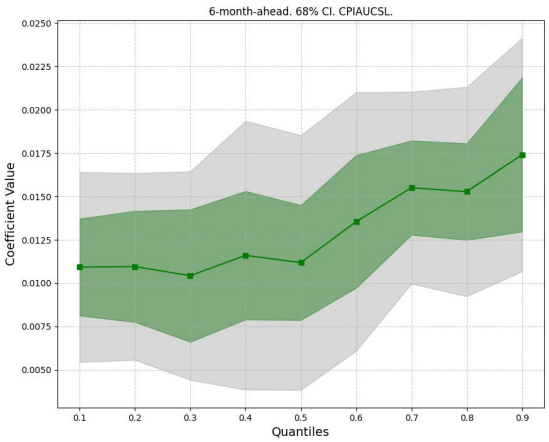

**(b)** PCE / `InflaBERT`




**(c)** CPI / `LLaMA 3.1 70B`




**(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 RIS variant $\text{RIS}_t$ (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.8 Inflation-related post counts as leading indicators

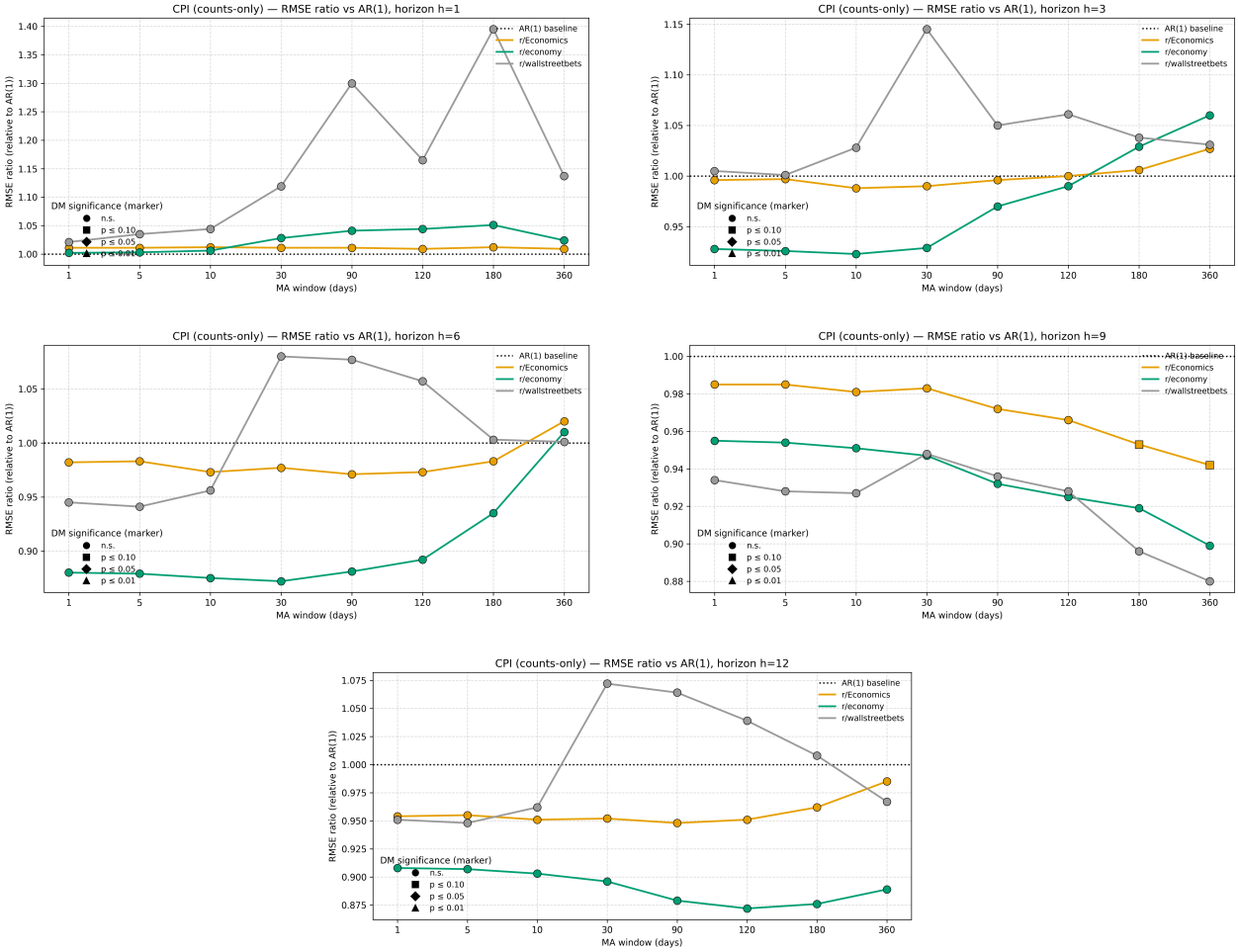
In this section, we check whether a simpler measure of Reddit activity—namely, the *volume* of inflation-related posts (as in Figure 2, left panel)—can itself serve as a useful leading indicator for inflation. To address this question, we replicate the point-forecasting exercise by replacing the LLM-based directional component of the RIS with the raw count of inflation-related submissions. Specifically, for each subreddit, we construct a daily series of inflation-related submission counts and apply the same backward-looking moving-average (MA) smoothers used elsewhere in the paper. We then aggregate the smoothed daily counts to monthly frequency and use them as the predictor in the $AR$-$X(1)$ forecasting regression. Figures A.34 and A.35 report RMSE ratios relative to the $AR(1)$ benchmark for headline CPI and core PCE at horizons $h \in \{1, 3, 6, 9, 12\}$, respectively.

**CPI.** For headline CPI, post-count indicators have limited value at the shortest horizon. At $h = 1$, none of the count-based specifications improves upon $AR(1)$. At intermediate horizons the evidence is mixed: at $h = 3$, counts from `r/Economics` and `r/economy` yield modest improvements for MA windows up to roughly 90–120 days; at $h = 6$, `r/economy` counts outperform for most MA windows, while `r/Economics` improves only slightly and `r/wallstreetbets` outperforms mainly with very short MA windows. At $h = 9$, counts from all subreddits outperform $AR(1)$, with the strongest performance typically coming from `r/wallstreetbets`, followed by `r/economy` and `r/Economics`. At $h = 12$, `r/economy` counts remain clearly stronger than the benchmark, while `r/Economics` improves only marginally and `r/wallstreetbets` does so primarily for very short MA windows (and the longest window). In terms of statistical evidence, only a small subset of these improvements is significant under the Diebold–Mariano test (notably, some long MA windows for `r/Economics` at the 10% level).

**PCE.** For core PCE, count-based indicators perform somewhat better than for CPI, but the improvements are still modest. At $h = 1$, `r/Economics` counts slightly outperform $AR(1)$ for most MA windows, although only the 10-day MA is significant at the 10% level; `r/economy` counts yield small gains up to short MA windows but are not significant. At $h = 3$ and $h = 6$, many count-based specifications outperform $AR(1)$, but none is statistically significant. At $h = 9$, all counts outperform the benchmark, with statistical significance appearing only for `r/Economics` at long MA windows (180 and 360 days) at the 10% level. At $h = 12$, all three subreddit counts improve on $AR(1)$, and `r/Economics` counts are frequently significant (at 5% for short MA windows and at 10% for longer windows).

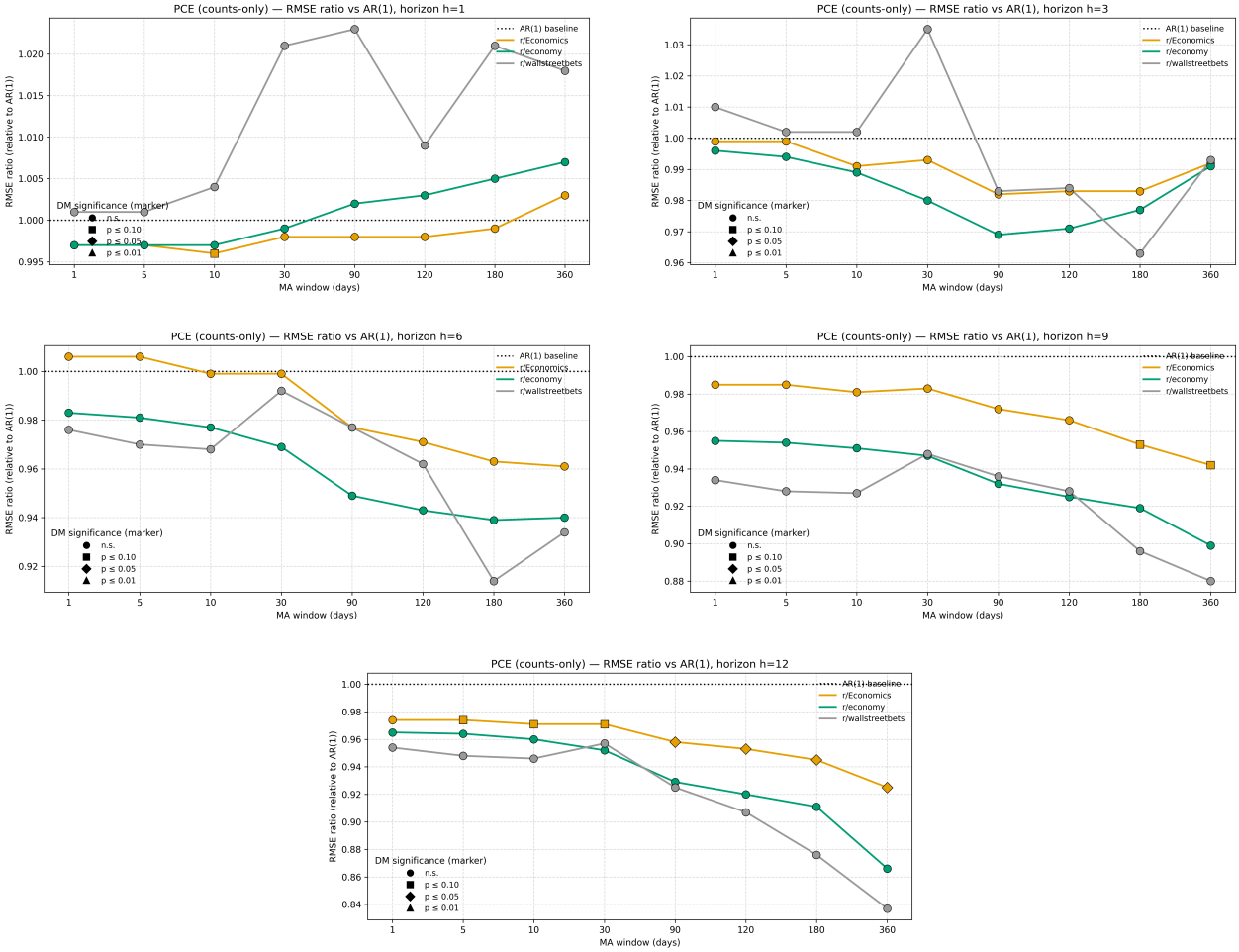
**Takeaway.** Overall, raw post counts—a proxy for *attention intensity*—can contain some predictive information, primarily at medium horizons and more clearly for core PCE than for headline CPI. However, these gains are generally small and rarely statistically significant, and they fall well short of the improvements delivered by the RIS, particularly at short horizons where inflation is dominated by shocks. The comparison suggests that the forecasting gains in the paper are not driven by posting volume alone, but by the *semantic content* extracted by LLMs.

**Figure A.34:** RMSE ratios using inflation-related Reddit post counts as predictors for CPI headline




*Notes*: The panels report RMSE ratios relative to the $AR(1)$ benchmark for headline CPI forecasts that use the count of inflation-related submissions from each subreddit as the predictor. Horizons are $h \in \{1, 3, 6, 9, 12\}$, ordered from the top-left panel to the bottom panel.

**Figure A.35:** RMSE ratios using inflation-related Reddit post counts as predictors for PCE core




*Notes*: The panels report RMSE ratios relative to the $AR(1)$ benchmark for core PCE forecasts that use the count of inflation-related submissions from each subreddit as the predictor. Horizons are $h \in \{1, 3, 6, 9, 12\}$, ordered from the top-left panel to the bottom panel.

# F RIS and inflation expectations

Social-media measures 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 discussions are available at high frequency and without publication lags, offering policymakers timely and potentially lower-cost information about inflation-related views.

This appendix provides additional evidence supporting our interpretation of the RIS as an expectation-related measure. The central question is whether Reddit discussions merely *co-move* with expectations or whether the RIS also *anticipates* 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 RIS variant extracted using different LLMs:

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

where $\pi_t$ is inflation (PCE), $\pi_t^e$ denotes Michigan inflation expectations, $\text{1YSwap}_t$ is the one-year inflation swap rate, and $\text{RIS}_t$ is a RIS variant.[52] 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 the RIS and survey expectations, focusing on two directions: (i) whether the RIS Granger-causes Michigan expectations ($\text{RIS}_t \rightarrow \pi_t^e$), and (ii) whether Michigan expectations Granger-cause the RIS ($\pi_t^e \rightarrow \text{RIS}_t$). Bidirectional causality is recorded when both tests reject.

Tables A.19, A.20, and A.21 report the results by subreddit; Table A.22 repeats the exercise for an aggregate RIS variant, denoted by $\text{RIS}_t^{\text{Agg}}$ constructed by averaging the three subreddit variants within each model family.

[52] 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 the RIS more frequently Granger-causes 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 $\text{RIS}_t \to \pi_t^e$ (Reddit anticipates expectations), 10 indicate bidirectional causality, and 7 indicate $\pi_t^e \to \text{RIS}_t$ (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 $\text{RIS}_t \to \pi_t^e$, 1 indicates bidirectional causality, and 7 correspond to $\pi_t^e \to \text{RIS}_t$.

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

**Aggregate across subreddits.** Table A.22 reports results for the aggregate RIS variant $\text{RIS}_t^{\text{Agg}}$ (the average across the three subreddits, computed within each model family). The conclusions are consistent with the subreddit-level evidence: in most specifications, $\text{RIS}_t^{\text{Agg}}$ Granger-causes $\pi_t^e$; the reverse direction occurs less frequently; and bidirectional relations are comparatively rare.

Table 6 shows that predictive content from the RIS to survey expectations is concentrated in a subset of LLMs and medium-to-long MA windows, with the cross-subreddit aggregate RIS and `r/Economics` delivering the broadest evidence. Bidirectional causality is comparatively rare and tends to appear for stronger models at intermediate horizons, while reverse-only causality is limited and reported separately in Table A.23.

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 the RIS can anticipate shifts in survey-based inflation expectations. These results provide additional support for interpreting the RIS as expectation-related and help rationalize its forecasting performance in the main analysis.

**Table A.19:** Granger causality between survey expectations $\pi_t^e$ and the RIS $\text{RIS}_t$ for subreddit `r/Economics`. Yellow cells indicate cases where the RIS Granger-causes expectations. Green cells indicate bidirectional causality; – indicates no causality in either direction.

| Model | 1 | 5 | 10 | 30 | 90 | 120 | 180 | 360 |
|---|---|---|---|---|---|---|---|---|
| BERT base | – | – | – | – | $\pi_t^e \to \text{RIS}_t$ | $\pi_t^e \to \text{RIS}_t$ | $\text{RIS}_t \to \pi_t^e$ | – |
| FinBERT | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \leftrightarrow \pi_t^e$ | $\pi_t^e \to \text{RIS}_t$ | $\text{RIS}_t \to \pi_t^e$ | – |
| InflaBERT | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \leftrightarrow \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \leftrightarrow \pi_t^e$ | $\text{RIS}_t \leftrightarrow \pi_t^e$ | – |
| Gemma 2B | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ | – |
| Gemma 9B | – | – | – | – | – | $\pi_t^e \to \text{RIS}_t$ | – | – |
| Gemma 27B | – | – | – | – | – | $\pi_t^e \to \text{RIS}_t$ | – | $\text{RIS}_t \to \pi_t^e$ |
| LLaMA 1B | $\text{RIS}_t \to \pi_t^e$ | – | – | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ |
| LLaMA 3B | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \leftrightarrow \pi_t^e$ | $\text{RIS}_t \leftrightarrow \pi_t^e$ | $\text{RIS}_t \leftrightarrow \pi_t^e$ | – |
| LLaMA 8B | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ | $\pi_t^e \to \text{RIS}_t$ | $\text{RIS}_t \leftrightarrow \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ | – |
| LLaMA 70B | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \leftrightarrow \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ |
| Qwen 0.5B | – | – | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ |
| Qwen 1.5B | – | – | – | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \leftrightarrow \pi_t^e$ | $\pi_t^e \to \text{RIS}_t$ | – | – |
| Qwen 7B | – | – | – | – | – | – | – | – |

**Table A.20:** Granger causality between survey expectations $\pi_t^e$ and the RIS $\text{RIS}_t$ for subreddit `r/economy`. Yellow cells indicate cases where the RIS Granger-causes expectations. Green cells indicate bidirectional causality; – indicates no causality in either direction.

| Model | 1 | 5 | 10 | 30 | 90 | 120 | 180 | 360 |
|---|---|---|---|---|---|---|---|---|
| BERT base | – | – | – | – | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ | – | – |
| FinBERT | – | – | – | – | – | – | – | – |
| InflaBERT | $\pi_t^e \to \text{RIS}_t$ | – | – | – | – | $\text{RIS}_t \to \pi_t^e$ | – | – |
| Gemma 2B | – | – | – | – | – | $\text{RIS}_t \to \pi_t^e$ | – | – |
| Gemma 9B | – | – | – | – | – | – | – | $\pi_t^e \to \text{RIS}_t$ |
| Gemma 27B | – | – | – | – | – | – | – | – |
| LLaMA 1B | – | – | – | – | – | – | – | – |
| LLaMA 3B | – | – | – | – | – | – | $\text{RIS}_t \to \pi_t^e$ | – |
| LLaMA 8B | – | – | – | – | – | – | – | – |
| LLaMA 70B | $\pi_t^e \to \text{RIS}_t$ | $\pi_t^e \to \text{RIS}_t$ | $\pi_t^e \to \text{RIS}_t$ | $\pi_t^e \to \text{RIS}_t$ | $\pi_t^e \to \text{RIS}_t$ | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ |
| Qwen 0.5B | – | – | – | – | – | – | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ |
| Qwen 1.5B | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ | – |
| Qwen 7B | – | – | – | – | – | $\text{RIS}_t \leftrightarrow \pi_t^e$ | – | – |

**Table A.21:** Granger causality between survey expectations $\pi_t^e$ and the RIS $\text{RIS}_t$ for subreddit `r/wallstreetbets`. Yellow cells indicate cases where the RIS Granger-causes expectations. Green cells indicate bidirectional causality; – indicates no causality in either direction.

| Model | 1 | 5 | 10 | 30 | 90 | 120 | 180 | 360 |
|---|---|---|---|---|---|---|---|---|
| BERT base | – | $\text{RIS}_t \to \pi_t^e$ | – | – | – | – | – | – |
| FinBERT | – | – | – | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ | – |
| InflaBERT | – | – | – | – | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ | $\pi_t^e \to \text{RIS}_t$ |
| Gemma 2B | – | – | – | – | $\text{RIS}_t \to \pi_t^e$ | – | $\text{RIS}_t \to \pi_t^e$ | – |
| Gemma 9B | – | – | – | – | $\text{RIS}_t \to \pi_t^e$ | – | – | – |
| Gemma 27B | – | – | – | – | – | – | – | – |
| LLaMA 1B | – | – | – | – | – | – | – | – |
| LLaMA 3B | – | – | – | – | $\text{RIS}_t \leftrightarrow \pi_t^e$ | $\text{RIS}_t \leftrightarrow \pi_t^e$ | $\text{RIS}_t \leftrightarrow \pi_t^e$ | $\pi_t^e \to \text{RIS}_t$ |
| LLaMA 8B | – | – | – | – | $\pi_t^e \to \text{RIS}_t$ | – | $\pi_t^e \to \text{RIS}_t$ | – |
| LLaMA 70B | – | – | – | – | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ | – | $\pi_t^e \to \text{RIS}_t$ |
| Qwen 0.5B | – | – | – | $\pi_t^e \to \text{RIS}_t$ | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \leftrightarrow \pi_t^e$ | $\text{RIS}_t \leftrightarrow \pi_t^e$ | – |
| Qwen 1.5B | – | – | – | – | $\text{RIS}_t \to \pi_t^e$ | – | $\pi_t^e \to \text{RIS}_t$ | – |
| Qwen 7B | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ | – | – | $\text{RIS}_t \to \pi_t^e$ | $\text{RIS}_t \to \pi_t^e$ | – |

**Table A.22:** Granger causality between survey expectations $\pi_t^e$ and the aggregate RIS $\text{RIS}_t^{\text{Agg}}$. Yellow cells indicate cases where the RIS Granger-causes expectations. Green cells indicate bidirectional causality; – indicates no causality in either direction.

| Model | 1 | 5 | 10 | 30 | 90 | 120 | 180 | 360 |
|---|---|---|---|---|---|---|---|---|
| BERT base | – | – | – | – | – | – | – | $\pi_t^e \to \text{RIS}_t^{\text{Agg}}$ |
| FinBERT | $\text{RIS}_t^{\text{Agg}} \to \pi_t^e$ | – | – | $\text{RIS}_t^{\text{Agg}} \to \pi_t^e$ | – | $\text{RIS}_t^{\text{Agg}} \to \pi_t^e$ | $\text{RIS}_t^{\text{Agg}} \to \pi_t^e$ | $\pi_t^e \to \text{RIS}_t^{\text{Agg}}$ |
| InflaBERT | $\text{RIS}_t^{\text{Agg}} \to \pi_t^e$ | $\text{RIS}_t^{\text{Agg}} \to \pi_t^e$ | $\text{RIS}_t^{\text{Agg}} \to \pi_t^e$ | $\text{RIS}_t^{\text{Agg}} \leftrightarrow \pi_t^e$ | $\text{RIS}_t^{\text{Agg}} \to \pi_t^e$ | $\text{RIS}_t^{\text{Agg}} \leftrightarrow \pi_t^e$ | $\text{RIS}_t^{\text{Agg}} \to \pi_t^e$ | $\pi_t^e \to \text{RIS}_t^{\text{Agg}}$ |
| Gemma 2B | – | – | – | – | $\text{RIS}_t^{\text{Agg}} \to \pi_t^e$ | $\text{RIS}_t^{\text{Agg}} \to \pi_t^e$ | – | $\pi_t^e \to \text{RIS}_t^{\text{Agg}}$ |
| Gemma 9B | – | – | – | – | $\text{RIS}_t^{\text{Agg}} \to \pi_t^e$ | $\pi_t^e \to \text{RIS}_t^{\text{Agg}}$ | $\text{RIS}_t^{\text{Agg}} \to \pi_t^e$ | – |
| Gemma 27B | – | – | – | – | – | – | – | $\pi_t^e \to \text{RIS}_t^{\text{Agg}}$ |
| LLaMA 1B | – | – | – | – | $\text{RIS}_t^{\text{Agg}} \to \pi_t^e$ | $\pi_t^e \to \text{RIS}_t^{\text{Agg}}$ | – | $\pi_t^e \to \text{RIS}_t^{\text{Agg}}$ |
| LLaMA 3B | $\text{RIS}_t^{\text{Agg}} \to \pi_t^e$ | $\text{RIS}_t^{\text{Agg}} \to \pi_t^e$ | $\text{RIS}_t^{\text{Agg}} \to \pi_t^e$ | $\text{RIS}_t^{\text{Agg}} \to \pi_t^e$ | $\text{RIS}_t^{\text{Agg}} \to \pi_t^e$ | $\text{RIS}_t^{\text{Agg}} \leftrightarrow \pi_t^e$ | $\text{RIS}_t^{\text{Agg}} \to \pi_t^e$ | $\pi_t^e \to \text{RIS}_t^{\text{Agg}}$ |
| LLaMA 8B | – | – | $\text{RIS}_t^{\text{Agg}} \to \pi_t^e$ | $\text{RIS}_t^{\text{Agg}} \to \pi_t^e$ | – | – | – | – |
| LLaMA 70B | $\text{RIS}_t^{\text{Agg}} \to \pi_t^e$ | $\text{RIS}_t^{\text{Agg}} \to \pi_t^e$ | $\text{RIS}_t^{\text{Agg}} \to \pi_t^e$ | $\text{RIS}_t^{\text{Agg}} \leftrightarrow \pi_t^e$ | $\text{RIS}_t^{\text{Agg}} \leftrightarrow \pi_t^e$ | $\text{RIS}_t^{\text{Agg}} \leftrightarrow \pi_t^e$ | $\text{RIS}_t^{\text{Agg}} \to \pi_t^e$ | $\pi_t^e \to \text{RIS}_t^{\text{Agg}}$ |
| Qwen 0.5B | $\text{RIS}_t^{\text{Agg}} \to \pi_t^e$ | $\text{RIS}_t^{\text{Agg}} \to \pi_t^e$ | $\text{RIS}_t^{\text{Agg}} \to \pi_t^e$ | $\text{RIS}_t^{\text{Agg}} \leftrightarrow \pi_t^e$ | $\text{RIS}_t^{\text{Agg}} \to \pi_t^e$ | $\pi_t^e \to \text{RIS}_t^{\text{Agg}}$ | $\text{RIS}_t^{\text{Agg}} \to \pi_t^e$ | – |
| Qwen 1.5B | $\text{RIS}_t^{\text{Agg}} \to \pi_t^e$ | – | – | $\text{RIS}_t^{\text{Agg}} \to \pi_t^e$ | $\text{RIS}_t^{\text{Agg}} \to \pi_t^e$ | $\text{RIS}_t^{\text{Agg}} \leftrightarrow \pi_t^e$ | $\text{RIS}_t^{\text{Agg}} \to \pi_t^e$ | $\pi_t^e \to \text{RIS}_t^{\text{Agg}}$ |
| Qwen 7B | – | – | – | – | – | – | – | – |

**Table A.23:** MA windows with 'reverse-only' Granger causality from survey expectations to the RIS, by subreddit and for the cross-subreddit aggregate

| Model | r/Economics | r/economy | r/wallstreetbets | Aggregate |
|---|---|---|---|---|
| BERT base | 90, 120 | – | – | 360 |
| FinBERT | 120 | – | – | 360 |
| InflaBERT | – | 1 | 360 | 360 |
| Gemma 2B | – | – | – | 360 |
| Gemma 9B | 120 | 360 | – | 120 |
| Gemma 27B | 120 | – | – | 360 |
| LLaMA 1B | – | – | – | 120, 360 |
| LLaMA 3B | – | – | 360 | 360 |
| LLaMA 8B | 90 | – | 90, 180 | – |
| LLaMA 70B | – | 1, 5, 10, 30, 90 | 360 | 360 |
| Qwen 0.5B | – | – | 30 | 120 |
| Qwen 1.5B | 120 | – | 180 | 360 |
| Qwen 7B | – | – | – | – |

Notes: Entries report MA windows for which survey expectations Granger-cause the RIS, but not vice versa.

# G The relationship between Reddit and news

This appendix investigates how the RIS relates to traditional news sentiment. Establishing this link is important for interpretation: if Reddit mainly echoes news, the RIS may reflect a socially filtered version of the news cycle; if it diverges, it may capture additional content generated by users.

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

Table A.24 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 consists of user-generated text, 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.24:** 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. The RIS therefore captures 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 the RIS from traditional news-sentiment measures in two key ways. First, Reddit acts as a filtering mechanism: only content that attracts attention is posted and can therefore enter the RIS. Second, discussion adds a social layer that can amplify, contest, or regularize the sentiment associated with a given news item. Deviations between the RIS and news sentiment can therefore arise precisely because of selection and discussion dynamics.

## G.3 Empirical relationship between Reddit and news sentiment

We compare the RIS with a widely used newspaper-based sentiment measure from Barbaglia et al. (2022). Their news corpus draws on six major U.S. newspapers (The *New York Times*, *Wall Street Journal*, *Washington Post*, *Dallas Morning News*, *San Francisco Chronicle*, and the *Chicago Sun-Times*) accessed via Dow Jones Factiva. Articles are selected if Factiva assigns them to at least one of four topical categories (excluding sports): (i) economic news (ECAT), (ii) monetary/financial news (MCAT), (iii) corporate news (CCAT), and (iv) general news (GCAT).

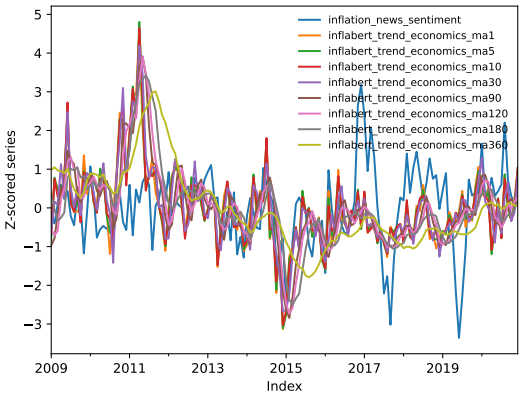
Barbaglia et al. (2022) compute sentiment using the Fine-Grained Aspect-based Sentiment (FiGAS) framework of Consoli et al. (2022). FiGAS isolates text segments that are semantically linked to specific economic concepts (e.g., *inflation*, *unemployment*, *monetary policy*) and assigns polarity scores in $[-1, 1]$ using a domain-specific dictionary. These scores are then aggregated into daily sentiment indices that reflect both the tone and the intensity of coverage across macroeconomic aspects.

Figure A.36 plots the newspaper-based sentiment series alongside the RIS (z-scored for comparability), using the InflaBERT-based RIS variant for each subreddit. Across all three subreddits, the RIS series closely co-moves with the news-sentiment index, indicating a strong informational connection between the two sources.

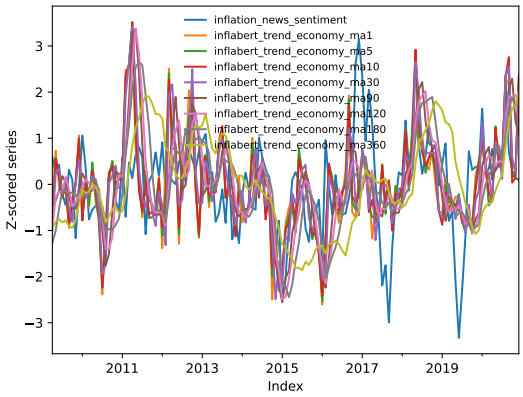
To examine timing, Figure A.37 reports lead–lag correlations between the RIS and the news-sentiment series. In all three panels, the peak correlation occurs at lag zero, indicating that the strongest association is contemporaneous. Correlations at positive and negative lags are gener-

**Figure A.36:** Time series (z-scored) of newspaper-based news sentiment and the InflaBERT-based RIS by subreddit.

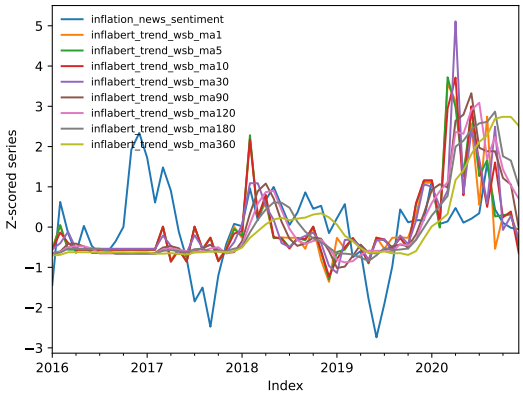
**(a)** `r/Economics`




**(b)** `r/economy`




**(c)** `r/wallstreetbets`



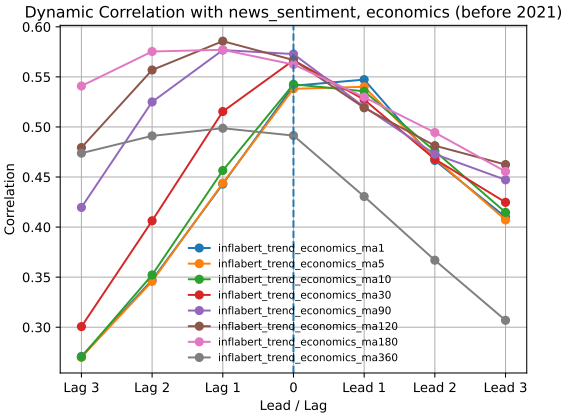

*Notes*: Each panel reports the standardized time series of newspaper-based news sentiment and the InflaBERT-based RIS for the corresponding subreddit.

ally smaller, providing limited evidence that the RIS systematically leads or lags traditional news sentiment.

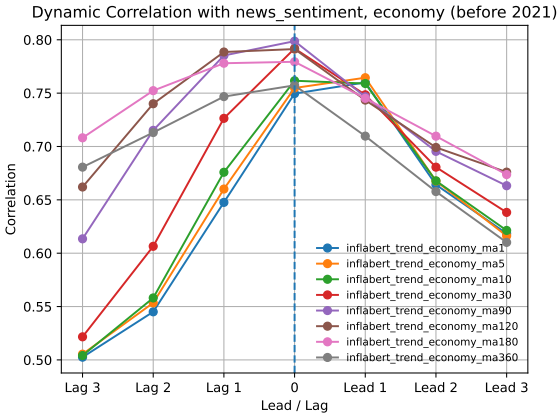
Taken together, the evidence indicates that the RIS strongly co-moves with traditional news sentiment and is primarily contemporaneous with it. This suggests that Reddit operates largely as a real-time *amplifier* of the news cycle rather than as a systematic leading or lagging indicator of newspaper 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 RIS reflects both the underlying news and the community's collective framing. In this sense, the RIS can be viewed as a socially filtered counterpart to standard news-sentiment measures.

**Figure A.37:** Lead–lag correlations between the RIS and newspaper-based sentiment.

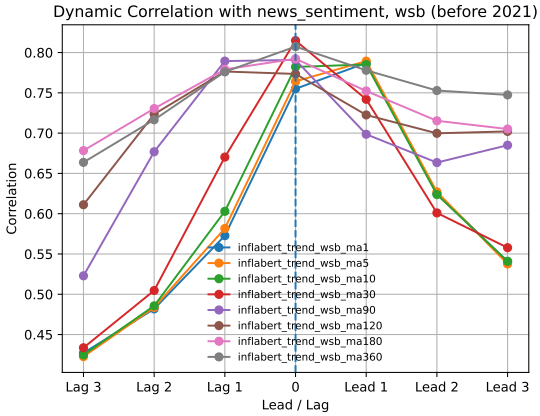
**(a)** `r/Economics`




**(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. A84).

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. A90).

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. A84).

Bracha, Anat and Jenny Tang (2025). "Inflation levels and (in) attention". In: *Review of Economic Studies* 92.3, pp. 1564–1594 (cit. on p. A78).

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 p. A78).

Consoli, Sergio, Luca Barbaglia, and Sebastiano Manzan (2022). "Fine-grained, aspect-based sentiment analysis on economic and financial lexicon". In: *Knowledge-Based Systems* 247, p. 108781. ISSN: 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2022.108781 (cit. on p. A90).

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. A84).

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).

Faust, Jon and Jonathan Wright (2013). "Forecasting Inflation". In: vol. 2. Elsevier. Chap. Chapter 1, pp. 2–56 (cit. on p. A61).

Giacomini, Raffaella and Barbara Rossi (2010). "Forecast comparisons in unstable environments". In: *Journal of Applied Econometrics* 25.4, pp. 595–620 (cit. on pp. A47, A48).

Goldstein, Nathan (2023). "Tracking inattention". In: *Journal of the European Economic Association* 21.6, pp. 2682–2725 (cit. on p. A78).

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. A84).

Grattafiori, Aaron et al. (2024). *The Llama 3 Herd of Models*. arXiv: 2407.21783 [cs.AI] (cit. on p. A9).

Hansen, Peter R, Asger Lunde, and James M Nason (2011). "The Model Confidence Set". In: *Econometrica* 79.2, pp. 453–497 (cit. on p. A64).

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. A10).

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, A13).

Lopez-Salido, David and Francesca Loria (2024). "Inflation at risk". In: *Journal of Monetary Economics*, p. 103570 (cit. on pp. A76, A77).

Loshchilov, Ilya and Frank Hutter (2017). *Decoupled Weight Decay Regularization*. arXiv: 1711.05101 (cit. on p. A12).

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. A36, A38, A39).

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 p. A74).

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).

Stock, James H. and Mark W. Watson (2007). "Why Has U.S. Inflation Become Harder to Forecast?" In: *Journal of Money, Credit and Banking* 39.s1, pp. 3–33. DOI: `https://doi.org/10.1111/j.1538-4616.2007.00014.x`. eprint: `https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1538-4616.2007.00014.x` (cit. on p. A61).

Team, Gemma et al. (2024). *Gemma 2: Improving Open Language Models at a Practical Size*. arXiv: `2408.00118 [cs.CL]` (cit. on p. A9).

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 p. A78).