LLM-Powered Predictive Inference with Online Text Time Series

Yingying Fan[a,1,2], **Jinchi Lv**[a,1,2], **Ao Sun**[a,1], **and Yurou Wang**[b,1]

This manuscript was compiled on January 10, 2026

Time series predictive inference is an important yet challenging task in economics and business, where existing approaches are often designed for low-frequency, survey-based data. With the recent advances of large language models (LLMs), there is growing potential to leverage high-frequency online text data for improved time series prediction, an area still largely unexplored. This paper proposes the LLM-TS, an LLM-based approach for time series predictive inference incorporating online text data. The LLM-TS is based on a joint time series framework that combines survey-based low-frequency data with LLM-generated high-frequency surrogates. The framework relies only on an error correlation assumption, combining a text-embedding-augmented ARX model for the observed gold-standard measurements with a VARX model for the LLM-generated surrogates. LLM-TS employs LLMs such as ChatGPT and the trained BERT models to construct LLM surrogates. Online text embeddings are extracted via LDA and BERT. We establish the asymptotic properties of the method and provide two forms of constructed prediction intervals. We also extend LLM-TS to incorporate deep learning backbones. To demonstrate the practical power of LLM-TS, we apply it to a critical real-world application: inflation forecast. We construct two large high-frequency online text data sets from the U.S. and China, and use LLMs to extract inflation-related signals from texts that reflect price dynamics. The finite-sample performance and practical advantages of LLM-TS are illustrated through extensive simulations and two noisy real data examples, highlighting its potential to improve time series prediction in economic applications.

Large language models | Inflation prediction | Online texts | Asymptotic distributions | Time series | Deep learning

Significance Statement

Large language models (LLMs) are increasingly used to build time-series indices. By scanning online texts, they provide high-frequency signals and low-cost experimentation for applications such as market simulations and policy evaluations. However, compared to structural models using carefully collected field data, LLM-based indices are often less stable and provide limited insight into their underlying mechanisms. This paper introduces a general forecasting framework that merges traditional econometric models with LLM outputs, converting low-quality LLM signals into a statistically coherent framework that yields reliable forecasts and valid prediction intervals. Using a decade of Wall Street Journal news articles and a collection of 159.5 million social media posts, we show that our framework accurately forecasts inflation with robustness across different information environments.

Time series predictive inference is a longstanding problem with widespread applications in statistics, economics, business, health and medical sciences, genomics, biology, and other fields. Traditional time series prediction methods often rely on structural models (1–3) and laborious field-collected data. While these methods are fully interpretable, they face declining prediction accuracy and the high costs associated with large-scale data collection.

Emerging alternative data streams, particularly text data from news media and social platforms, demonstrate the viability of unstructured content as novel inputs for time series prediction (4–7). Such online data is cheaply collected, and the fact that large language models (LLMs) have capabilities in processing complex linguistic patterns suggests largely untapped potential for time series modeling and research (8, 9). However, the prediction advantages of LLMs are tempered by their black-box nature and limited interpretability regarding the sources of their predictive power.

The above dilemmas motivate a core research question: How can we effectively combine the prediction capabilities of LLMs using online text with the interpretability of established structural models using field-collected data? Our suggested framework addresses such challenge through an integrated forecasting system that harmoniously combines traditional econometric models with the LLM-powered prediction models. The method tackles two key obstacles: 1) the effective combination of limited high-accuracy, low-frequency official data and abundant high-frequency but less robust LLM-powered surrogates; and 2) the inherent conflict between complex machine learning and deep learning architectures, and the need for interpretable results. Through correlation modeling between traditional structure models and LLM-powered surrogates, our approach enhances the predictive inference accuracy while maintaining clear explanations for structural relationships. We name our new framework as the LLM-powered time series predictive inference (LLM-TS).

An illustrative application is inflation forecasting. Inflation is a key macroeconomic indicator that guides monetary policy and reflects overall economic conditions.

Author affiliations: [a]Data Sciences and Operations Department, Marshall School of Business, University of Southern California, Los Angeles, CA 90089; [b]Paula and Gregory Chow Institute for Studies in Economics, Xiamen University, China 361005

Author contributions: Y.F. and J.L. designed research; Y.F., J.L., A.S. and Y.W. performed research; A.S. and Y.W. collected data; Y.F., J.L., A.S. and Y.W. analyzed data; and Y.F., J.L., A.S and Y.W. wrote the paper.

The authors declare no competing interest.

[1]Four authors contributed equally to this work.

[2]To whom correspondence should be addressed. E-mail: fanyingy@usc.edu or jinchilv@usc.edu

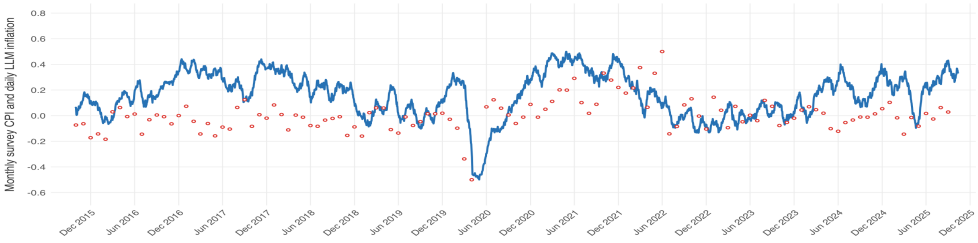


Fig. 1. A time-series comparison between the LLM-generated daily inflation index (Eq. (22)) and the CPI from September 2015 to August 2025 in the U.S. economy. The blue curve represents the LLM-generated daily inflation index capturing high-frequency variations through analysis of unstructured Wall Street Journal news articles, and the red circles depict the normalized monthly urban CPI. To facilitate the comparison of trend changes, we apply a 30-day moving average smoothing to LLM-generated daily inflation index in the figure. Additionally, both the monthly survey-based CPI and the LLM-generated daily inflation index are standardized and shifted by subtracting 0.5, ensuring that their fluctuations are centered around zero.

It is closely linked to the interest rates, labor-market dynamics (such as wage growth and employment), production costs, and financial-market sentiment (10–13). Traditional forecasting approaches rely on structural autoregressive models, which regress current inflation indices on their own lags and a small set of macro variables (14, 15). A new and complementary information source arises from economic narratives (16) embedded in newspapers, blogs, and social-media discussions about price changes. These online texts provide real-time signals that are often absent from official statistics. In this paper, we examine two distinct economic environments: the U.S. and China. We collect a ten-year text data set from the Wall Street Journal (https://www.wsj.com) and a corpus of 159.5 million posts from Sina Weibo (https://www.weibo.com), the largest social-media platform in China. From these unstructured texts we use an LLM pipeline to construct a daily, text-based inflation index (Eq. (22)). Fig. 1 and *Fig. S5 in SI Appendix* compare our high-frequency inflation indices derived from the Wall Street Journal (WSJ) news articles and Weibo posts, respectively, with the official monthly Consumer Price Index (CPI), a standard measurement of inflation. The LLM-derived series tracks the same broad inflationary patterns as the gold-standard CPI while providing richer, higher-frequency signals; the daily index exhibits fine-grained volatility that converges toward the monthly survey-based benchmark.

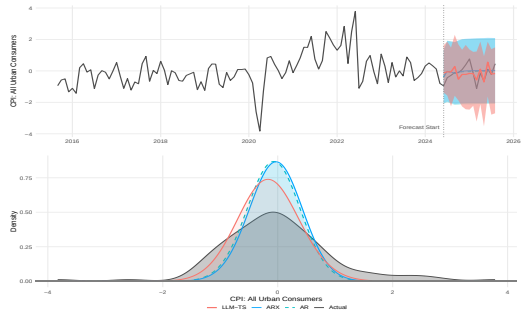


Fig. 2. Top panel: The CPI forecasts and corresponding prediction intervals across different dates given by the AR, ARX (unemployment rate as exogenous), and LLM-TS models. Bottom panel: The kernel density plots of predicted CPI values given by different models compared to the actual CPI values. The LLM-TS model exploits the LLM-generated synthetic surrogates constructed using ChatGPT and the trained BERT models, and the LDA embeddings for online text embeddings, as discussed in Section *LLM-CPI*. We adopt the period from January 2019 to May 2024 as the training sample, and the period from June 2024 to August 2025 to evaluate the out-of-sample forecasts.

The proposed LLM-TS leverages the predictive power of LLMs and the large-scale data sources of online text time series to enhance the effectiveness of predictive inference. Fig. 2 and *Fig. S10 in SI Appendix* illustrate the prediction power of the suggested LLM-TS method in forecasting high-frequency inflation dynamics using WSJ and Weibo text data, respectively. The black solid curve represents the standardized actual CPI values. Here, we consider three models: a classical autoregressive (AR) model relying on the historical CPI values (green dashed curve); an autoregressive model with the unemployment rate as an exogenous predictor (ARX), also referred to as Gordon's "triangle model" (17) (blue solid curve); and our suggested LLM-TS model that combines the prediction power of the structural ARX model and the LLM-powered surrogate model (red solid curve).

The results show clearly that the LLM-TS most closely tracks the actual CPI trajectory, successfully capturing turning points and underlying trends. More importantly, the prediction interval given by the LLM-TS model is substantially narrower than those given by the AR and ARX models, while still maintaining the desired nominal coverage rate (i.e., 95%). In contrast, the prediction intervals from traditional models are excessively wide and less informative. The advantages of the LLM-TS are more pronounced in the Chinese economic setting, which is considerably more challenging due to data scarcity. These findings highlight the effectiveness of the LLM-TS in delivering both accurate point forecasts and tight uncertainty quantification, especially in the presence of complex economic signals. While inflation forecasting serves as a concrete application in this work, we demonstrate later that it is broadly applicable and can be extended to more complex time-series prediction tasks, including those built upon state-of-the-art deep learning backbone models.

Problem setup

We observe $\{(y_t, \mathbf{z}_t), t = 1, \ldots, T\}$, where $y_t \in \mathbb{R}$ is a continuous measurement of interest, and $\mathbf{z}_t \in \mathbb{R}^d$ represents related field-collected data that may have predictive power for y_t. Additionally, we have a data set of online text time series $\{\mathcal{D}_t, t = 1, \ldots, T\}$, where $\mathcal{D}_t$ is an unstructured text data set collected during period t.

Our goal is to construct a $(1-\alpha)100\%$ prediction interval PI_{T+h} for h-step-ahead prediction, such that

$$\liminf_{T\to\infty} \mathbb{P}\{y_{T+h} \in \mathrm{PI}_{T+h}\} \geq 1-\alpha.$$

Traditional approaches rely on structural models, such as the AR or ARX models, where $\mathbf{z}_t$ serves as exogenous variables. However, these methods often generate overly wide and noninformative prediction intervals PI_{T+h}.

The key question we address is: How can the online text time series $\{\mathcal{D}_t, t = 1, \ldots, T\}$ be effectively leveraged in predictive inference to achieve both valid and informative predictions under mild assumptions? To tackle this problem, we adopt a two-step procedure. First, we use LLMs to form surrogate data $\{(\mathbf{y}_t^S, \mathbf{x}_t), t = 1, \ldots, T\}$ using the online text time series, where $\mathbf{y}_t^S \in \mathbb{R}^K$ is an LLM-powered surrogate vector correlated with the target response y_t, and $\mathbf{x}_t \in \mathbb{R}^p$ represents text embeddings from period t. This step is nontrivial, requiring the extraction of meaningful signals from highly noisy text data. We provide an information extraction process in this work (i.e., Algorithm 1).

Based on $\{(\mathbf{y}_t^S, \mathbf{x}_t), t = 1, \ldots, T\}$, we propose the LLM-powered time series predictive inference (LLM-TS) framework, which integrates LLM-generated surrogates with conventional field-collected measurements. The framework combines a text-embedding-augmented autoregressive model for the observed gold-standard measurements with a vector autoregressive model for the LLM-generated surrogates. These two models are connected through their cross-sectional error correlation structure. Extending the methodology of McCaw et al. (18) to time-series settings, LLM-TS achieves improved time series prediction by reducing the model error of the text-embedding-augmented autoregressive model using the surrogate data.

The proposed LLM-TS framework provides tight prediction intervals with theoretical guarantees, as demonstrated through simulations and real-world data examples.

Related work

Prediction-powered inference and synthetic surrogate joint modeling. Our study contributes to recent methodological developments in *prediction-powered inference* and *synthesized surrogate joint modeling*, which aim to incorporate LLM-generated predictions into formal statistical procedures. For example, Angelopoulos et al. (19) and Zrnic and Candès (20) proposed combining experimental and synthetic data through cross-validation to improve statistical efficiency, while McCaw et al. (18) stabilized raw LLM outputs by treating them as synthetic proxies within a joint-likelihood framework. These works focus primarily on estimation and hypothesis testing rather than time-series forecasting. A recent work of Bashari et al. (21) introduced the synthetic-powered conformal predictive inference. Their framework is highly general and yields strong performance when the surrogate and target nonconformity score distributions are well aligned. Our approach provides a distinct direction based on the joint residual modeling, which does not require such distributional alignment. A more detailed comparison between the two methods is provided in *Section 6 of SI Appendix*.

Time series predictive inference. Our work is also related to time series predictive inference, a long-standing problem in statistics and econometrics. Phillips (1) rigorously discussed the sampling distribution of forecasts for a first-order autoregressive model. Fuller and Hasza (2) and Stine (22) extended this discussion to general-order autoregressive models. Bootstrap methods for time series prediction were explored in (23, 24). For an overview of traditional time series predictive inference, readers can refer to the works of (3, 25). However, for modern complex time series predictive inference, traditional structural approaches may exhibit reduced predictive power and provide less informative prediction intervals. A substantial body of work has focused on leveraging the predictive power of machine learning and deep learning in the time series field; see, e.g., (26–29). For a comprehensive review, refer to Petropoulos et al. (30). Our LLM-TS framework provides a general approach for combining the strengths of classical time-series methodologies with the predictive power of modern LLMs. It preserves the interpretability and provides tighter prediction intervals. Moreover, the core idea extends naturally to deep learning architectures.

LLMs empowering economic and business research. Our study contributes to recent advances in LLMs empowering economic and business research. Recent advancements in LLMs have shown potential to support economic and business research through their ability to generate synthetic data that approximates real-world patterns (31, 32). These models enable cost-effective experimentation in applications such as market simulations and policy evaluations, offering researchers a flexible tool for preliminary analysis. However, the reliability of LLM-driven insights remains uncertain due to challenges such as inherent biases, reliance on outdated training data, and limited adaptability to real-time events (33–35). These limitations highlight the importance of developing complementary methods to evaluate and refine the LLM outputs. Our work addresses these challenges by providing an effective approach that leverages LLM outputs, even when they are of low quality or unreliable, to develop a rigorous statistical framework for time series predictive inference. We further demonstrate the power of our methodology through simulations and complex real-world data examples.

Inflation forecasting. Our work also contributes to the literature on modern inflation forecasting, a long-studied challenge in economics (14, 17, 36). Recent advances in data availability have enabled new approaches to this problem. For instance, Medeiros et al. (37) demonstrated improved inflation predictions by applying machine learning methods to a broad set of macroeconomic indicators from Mccracken and Ng (38). Many theoretical and empirical studies have also examined the important role of economic narratives in economic fluctuations (5, 6, 39). More recently, Hong et al. (7) incorporated text data from Wall Street Journal articles to enhance forecasting accuracy. Building upon these developments, we explore an alternative approach that integrates both text and macroeconomic data sources while leveraging the power of LLMs for prediction and inference. Our results suggest that this combined method can offer significant improvements over existing approaches.

LLM-TS

We proceed with assuming that the LLM-powered surrogates data $\{(\mathbf{y}_t^S, \mathbf{x}_t), t = 1, \ldots, T\}$ has already been obtained. The detailed process for obtaining it is deferred to Section *LLM-CPI*. A key ingredient of the suggested LLM-TS method is a joint time-series model integrating a target structural model and an LLM-powered surrogate model on the surrogate data.

We begin with introducing the target model on the observed data set $\{y_t, t = 1, \ldots, T\}$. For concreteness, we showcase the idea using an autoregressive model with exogenous variables of order q_1, referred to as ARX(q_1) model, as the target structural model

$$y_t = \sum_{l=1}^{q_1} \alpha_l y_{t-l} + \mathbf{z}_t^\top \boldsymbol{\theta} + \mathbf{x}_t^\top \boldsymbol{\beta} + \epsilon_t \quad [1]$$

for $t = 1, \ldots, T$. Here, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{q_1})^\top \in \mathbb{R}^{q_1}$ denotes the autoregressive coefficient vector, $\boldsymbol{\theta} \in \mathbb{R}^d$ represents the regression coefficient vector for the field-collected covariates, $\boldsymbol{\beta} \in \mathbb{R}^p$ stands for the regression coefficient vector for the text embedding features based on online time series, and ϵ_t is the scalar model error.

While the LLM-generated surrogates $\mathbf{y}_t^S$ may be related to the target response, they should not be used directly as predictors of the target response due to some intrinsic limitations of theirs. First, the LLM-based predictions are inherently stochastic, such that repeated executions or slight prompt perturbations can yield noticeable different results. When such high-variance outputs are used directly as covariates, this variability can propagate into the forecasting model, leading to unstable estimates and reduced reproducibility. Second, because LLM surrogates and embeddings are derived from the same textual sources, directly including $\mathbf{y}_t^S$ could inflate variance through redundant information, exacerbating collinearity and further inflating variance of predictive inference.

Instead of the direct prediction approach, we incorporate the LLM-generated surrogates into the target structural model to reduce its error, thereby enhancing the target model's predictive and inference performance. This idea is inspired by the recent work of (18), who introduced a general framework for integrating synthetic surrogates to empower the testing procedure for genome-wide association studies. Building upon this, we adapt and extend their approach to the time series prediction framework, enabling the incorporation of LLM-generated synthetic surrogates into the joint time-series modeling. To formalize this idea, we choose the LLM-powered surrogate model as the vector autoregressive model with exogenous variables of order q_2 (referred to as VARX(q_2) model hereafter)

$$\mathbf{y}_t^S = \sum_{l=1}^{q_2} \mathbf{A}_l^S \mathbf{y}_{t-l}^S + \mathbf{B}^S \mathbf{x}_t + \boldsymbol{\epsilon}_t^S, \quad [2]$$

where $\mathbf{y}_t^S = (y_{t,1}^S, \ldots, y_{t,K}^S)^\top \in \mathbb{R}^K$ contains the K-dimensional LLM-generated surrogates at time t, $\mathbf{A}_l^S \in \mathbb{R}^{K \times K}$ with $l = 1, \ldots, q_2$ denote the autoregressive coefficient matrices, $\mathbf{B}^S \in \mathbb{R}^{K \times p}$ represents the regression coefficient matrix for the exogenous text embedding features, and $\boldsymbol{\epsilon}_t^S = (\epsilon_{t,1}^S, \cdots, \epsilon_{t,K}^S)^\top$ is the model error vector. Here, we assume $q_2 \leq q_1$ without loss of generality. The LLM-powered surrogate model (2) is not intended to capture the true data-generating process of the LLM surrogates, but serves as a working model for their temporal dynamics and relationships with the text data. Indeed, the surrogate model can be a mis-specified model for surrogate data $\{(\mathbf{y}_t^S, \mathbf{x}_t), t = 1, \cdots, T\}$, while still achieving the goal of enhancing target model's performance.

We are now ready to introduce our joint LLM-TS model. Following (18), we assume that the errors of both models jointly follow a multivariate normal distribution

$$(\epsilon_t, (\boldsymbol{\epsilon}_t^S)^\top)^\top \sim N(\mathbf{0}^\top, \boldsymbol{\Sigma}), \quad [3]$$

with the error covariance matrix given by

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{TT}^2 & \boldsymbol{\Sigma}_{TS} \\ \boldsymbol{\Sigma}_{ST} & \boldsymbol{\Sigma}_{SS} \end{bmatrix}.$$

We further assume that the errors are independent across different periods, meaning that the autoregressive terms in the models account for all cross-temporal correlations across months. Here, σ_{TT}^2 represents the variance of the target model error, $\boldsymbol{\Sigma}_{SS}$ denotes the covariance matrix of the LLM-powered surrogate model errors, and $\boldsymbol{\Sigma}_{TS}$ captures the cross-covariance between the target model and the LLM-powered surrogate model errors. Such formulation allows us to model the correlation structures between the target model and LLM-powered surrogate predictions. It is important to emphasize that the target model and the LLM-powered surrogate model do *not* share any model parameters; their only connection is through the correlations of their model errors.

Estimation based on the joint modeling. In view of the joint LLM-TS model (3), we can rewrite the random error of the target model Eq. (1) using the properties of conditional normal distribution as

$$\epsilon_t = \boldsymbol{\gamma}^\top \boldsymbol{\epsilon}_t^S + e_t, \quad [4]$$

where $\boldsymbol{\gamma} = \boldsymbol{\Sigma}_{SS}^{-1} \boldsymbol{\Sigma}_{ST}$ represents the regression coefficient vector linking the LLM-generated surrogate model errors to the target model error, and random error $e_t \sim N(0, \sigma_e^2)$ with $\sigma_e^2 = \sigma_{TT} - \boldsymbol{\Sigma}_{TS} \boldsymbol{\Sigma}_{SS}^{-1} \boldsymbol{\Sigma}_{ST}$. The decomposition in (4) allows us to isolate the portion of variation in the target model error that can be explained by the surrogate model errors, thereby reducing the overall variance of the target model error and ensuring more accurate prediction and inference.

By substituting Eq. (2) into decomposition Eq. (4), it holds that

$$\begin{aligned} \epsilon_t =& \boldsymbol{\gamma}^\top \left(\mathbf{y}_t^S - \sum_{l=1}^{q_2} \mathbf{A}_l^S \mathbf{y}_{t-l}^S \right) - \boldsymbol{\gamma}^\top \mathbf{B}^S \mathbf{x}_t + e_t \\ :=& \boldsymbol{\gamma}^\top \mathbf{D}(\mathbf{y}_t^S) - \boldsymbol{\gamma}^\top \mathbf{B}^S \mathbf{x}_t + e_t, \end{aligned} \quad [5]$$

where $\mathbf{D}(\mathbf{y}_t^S) := \mathbf{y}_t^S - \sum_{l=1}^{q_2} \mathbf{A}_l^S \mathbf{y}_{t-l}^S$ captures the error component of the LLM predictions after accounting for their autoregressive structure. Plugging expression (5) into the target model Eq. (1), we can equivalently write the joint LLM-TS model (3) as a joint LLM-powered ARX model given by

$$\begin{aligned} y_t =& \sum_{l=1}^{q_1} \alpha_l y_{t-l} + \mathbf{z}_t^\top \boldsymbol{\theta} + \mathbf{x}_t^\top \left(\boldsymbol{\beta} - \boldsymbol{\gamma}^\top \mathbf{B}^S \right) + \boldsymbol{\gamma}^\top \mathbf{D}(\mathbf{y}_t^S) + e_t \\ =& \sum_{l=1}^{q_1} \alpha_l y_{t-l} + \mathbf{z}_t^\top \boldsymbol{\theta} + \mathbf{x}_t^\top \boldsymbol{\delta} + \boldsymbol{\gamma}^\top \mathbf{D}(\mathbf{y}_t^S) + e_t, \end{aligned} \quad [6]$$

where $\boldsymbol{\delta} = \boldsymbol{\beta} - \boldsymbol{\gamma}^\top \mathbf{B}^S$. It is seen that the model error e_t in (6) has a reduced variance compared to that in the target model Eq. (1). The stronger the model error correlations, the

greater gains in the variance reduction, and thereby, better prediction and inference accuracy using Eq. (6). Throughout the rest of the paper, the joint LLM-TS model is implicitly referred to as model Eq. (6).

To estimate the parameters of the joint LLM-powered ARX model Eq. (6), we exploit a two-step approach. We first construct estimators $(\widehat{\mathbf{A}}_1^S, \ldots, \widehat{\mathbf{A}}_{q_2}^S, \widehat{\mathbf{B}}^S)$ of the parameters in Eq. (2) by solving the optimization problem

$$\min_{\mathbf{A}_1^S,\ldots,\mathbf{A}_{q_2}^S,\mathbf{B}^S} \frac{1}{T}\sum_{t=q_2+1}^{T}\left\|\mathbf{y}_t^S - \sum_{l=1}^{q_2}\mathbf{A}_l^S\mathbf{y}_{t-l}^S - \mathbf{B}^S\mathbf{x}_t\right\|^2. \quad [7]$$

Given such estimates, we further estimate the error component $\mathbf{D}(\mathbf{y}_t^S)$ using the plug-in estimator

$$\widehat{\mathbf{D}}(\mathbf{y}_t^S) = \mathbf{y}_t^S - \sum_{l=1}^{q_2}\widehat{\mathbf{A}}_l^S\mathbf{y}_{t-l}^S. \quad [8]$$

We then form estimates $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\gamma}})$ of the parameters in the joint LLM-powered ARX model Eq. (6) by solving the optimization problem

$$\min_{\boldsymbol{\alpha},\boldsymbol{\theta},\boldsymbol{\delta},\boldsymbol{\gamma}} \frac{1}{T}\sum_{t=q_q+1}^{T}\left(y_t - \sum_{l=1}^{q_1}\alpha_l y_{t-l} - \mathbf{z}_t^\top\boldsymbol{\theta} - \mathbf{x}_t^\top\boldsymbol{\delta} - \boldsymbol{\gamma}^\top\widehat{\mathbf{D}}(\mathbf{y}_t^S)\right)^2. \quad [9]$$

With the estimates given in (7)–(9) above, we can construct the one-step-ahead forecast $\hat{y}_{T+1}$ using the joint LLM-powered ARX model Eq. (6) as

$$\hat{y}_{T+1} = \sum_{l=1}^{q_1}\hat{\alpha}_l y_{T+1-l} + \mathbf{z}_{T+1}^\top\hat{\boldsymbol{\theta}} + \mathbf{x}_{T+1}^\top\hat{\boldsymbol{\delta}} + \hat{\boldsymbol{\gamma}}^\top\widehat{\mathbf{D}}(\mathbf{y}_{T+1}^S), \quad [10]$$

where $\widehat{\mathbf{D}}(\mathbf{y}_{T+1}^S)$ is calculated using Eq. (8) with $t = T+1$. For the multi-step-ahead forecasts, we employ a rolling horizon approach. Specifically, the h-step-ahead forecast $\hat{y}_{T+h}$ can be constructed as

$$\hat{y}_{T+h} = \sum_{l=1}^{q_1}\hat{\alpha}_l\hat{y}_{T+h-l} + \mathbf{z}_{T+h}^\top\hat{\boldsymbol{\theta}} + \mathbf{x}_{T+h}^\top\hat{\boldsymbol{\delta}} + \hat{\boldsymbol{\gamma}}^\top\widehat{\mathbf{D}}(\mathbf{y}_{T+h}^S), \quad [11]$$

where $\hat{y}_{T+h-1}, \hat{y}_{T+h-2}, \ldots$ are iteratively computed based on the forecasts $\hat{y}_t$ from previous time stamps with $\hat{y}_t = y_t$ for $t \le T$, and covariates $\mathbf{x}_{T+h}$, $\mathbf{z}_{T+h}$, and $\widehat{\mathbf{D}}(\mathbf{y}_{T+h}^S)$ are used as the inputs.

Under mild regularity conditions, we can show that for each fixed $h \ge 1$,

$$\hat{y}_{T+h} - y_{T+h} = \sum_{r=0}^{h-1}(\mathbf{A}^r)_{11}e_{T+h-r} + o_p(1), \quad [12]$$

where matrix $\mathbf{A}$ is defined in *Section 1 of SI Appendix* and $\mathbf{A}^0$ is defined as the identity matrix $\mathbf{I}$. The proof details are deferred to *Section 1 of SI Appendix*.

In contrast, if we ignore the LLM-generated surrogates and rely solely on the traditional ARX model, the h-step-ahead forecast is then given by

$$\hat{y}_{T+h}^a = \sum_{l=1}^{q_1}\hat{\alpha}_l\hat{y}_{T+h-l}^a + \mathbf{z}_{T+h}^\top\hat{\boldsymbol{\theta}} + \mathbf{x}_{T+h}^\top\hat{\boldsymbol{\beta}}, \quad [13]$$

where $\hat{y}_{T+h-1}^a, \hat{y}_{T+h-2}^a, \ldots$ are iteratively computed via the ARX model prediction with $\hat{y}_t^a = y_t$ for each $t \le T$. The prediction error for this benchmark model is

$$\hat{y}_{T+h}^a - y_{T+h}^a = \sum_{r=0}^{h-1}(\mathbf{A}^r)_{11}\epsilon_{T+h-r} + o_p(1).$$

The efficiency gain of LLM-TS compared to the traditional ARX model can be quantified by the ratio of prediction error variances

$$\text{Efficiency} = \frac{\text{Var}\left(\sum_{r=0}^{h-1}(\mathbf{A}^r)_{11}\epsilon_{T+h-r}\right)}{\text{Var}\left(\sum_{r=0}^{h-1}(\mathbf{A}^r)_{11}e_{T+h-r}\right)} = \frac{\sigma_{TT}^2}{\sigma_{TT}^2 - \boldsymbol{\Sigma}_{TS}\boldsymbol{\Sigma}_{SS}^{-1}\boldsymbol{\Sigma}_{ST}}. \quad [14]$$

Under the simplified assumptions of $\boldsymbol{\Sigma}_{SS} = \mathbf{I}$ and $\boldsymbol{\Sigma}_{ST} = \rho\mathbf{1}$ with $\mathbf{I}$ and $\mathbf{1}$ the identity matrix and the vector of ones, respectively, the efficiency gain in (14) reduces to

$$\text{Efficiency} = \frac{1}{1 - |S|\rho^2}. \quad [15]$$

Here, we require $|S|\rho^2 < 1$ to guarantee the positive definiteness of the joint covariance matrix $\boldsymbol{\Sigma}$. In light of (15), we see the practical benefits of incorporating the LLM-generated synthetic surrogates in the LLM-TS, and that the stronger the correlations between the target and the surrogates, the greater the efficiency gain. Moreover, under the residual correlation framework in Eq. (3), our joint modeling approach theoretically achieves uniformly lower prediction mean squared error (PMSE) than approaches that use directly the surrogate response $\mathbf{y}_t^S$ as a predictor; see *Section 7 of SI Appendix* for a detailed theoretical comparison.

Predictive inference via LLM-TS

We now introduce two ways of constructing the LLM-TS prediction intervals for time series predictive inference.

Box–Jenkins prediction interval. Based on Eq. (12), we can construct an asymptotic prediction interval for the h-step-ahead prediction $\hat{y}_{T+h}$ once we obtain a consistent estimator of the error variance. One common approach to estimating such variance is through the sum of squared residuals (40)

$$\hat{\sigma}_e^2 := \frac{1}{T - q_1}\sum_{t=q_1+1}^{T}\hat{e}_t^2 \quad [16]$$

with $\hat{e}_t = y_t - \sum_{l=1}^{q_1}\hat{\alpha}_l y_{t+1-l} - \mathbf{z}_t^\top\hat{\boldsymbol{\theta}} - \mathbf{x}_t^\top\hat{\boldsymbol{\delta}} - \hat{\boldsymbol{\gamma}}^\top\widehat{\mathbf{D}}(\mathbf{y}_t^S)$. Using the above error variance estimator, we can construct the Box–Jenkins (BJ) prediction interval (3) with confidence level $1 - \alpha$ as

$$\text{PI}^{BJ} = (\hat{y}_{T+h})\left[\hat{y}_{T+h} - \hat{z}_{\alpha/2}^h, \hat{y}_{T+h} + \hat{z}_{\alpha/2}^h\right], \quad [17]$$

where $\hat{z}_{\alpha/2}^h = |z_{\alpha/2}|\sqrt{\sum_{r=0}^{h-1}(\widehat{\mathbf{A}}^r)_{11}^2}\hat{\sigma}_e$, $z_{\alpha/2}$ is the $\alpha/2$ quantile of the standard normal distribution, and $\alpha \in (0, 1)$. The BJ prediction interval asymptotically covers the true value y_{T+h} if $\hat{\sigma}_e^2$ is a consistent estimator of σ_e^2.

Bootstrap prediction interval. We also suggest a residual-based bootstrap prediction interval for LLM-TS (41, 42). Given the estimated parameters $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\gamma}})$, we compute the residuals as

$$\hat{e}_t = y_t - \sum_{l=1}^{q_1} \hat{\alpha}_l y_{t-l} - \mathbf{z}_t^\top \hat{\boldsymbol{\theta}} - \mathbf{x}_t^\top \hat{\boldsymbol{\delta}} - \hat{\boldsymbol{\gamma}}^\top \widehat{\mathbf{D}}(\mathbf{y}_t^S) \qquad [18]$$

for $t = q_1 + 1, \ldots, T$. We generate the bootstrap residuals by drawing $T + h$ samples with replacement from $\{\hat{e}_t - \hat{\mu},\ t = q_1 + 1, \ldots, T\}$, where $\hat{\mu} = \sum_{t=q_1+1}^{T} \hat{e}_t/(T - q_1)$. Denote by $\{e_t^*,\ t = 1, \ldots, T + h\}$ the bootstrap residuals. We then recursively calculate

$$y_t^* = \sum_{l=1}^{q_1} \hat{\alpha}_l y_{t-l}^* + \mathbf{z}_t^\top \hat{\boldsymbol{\theta}} + \mathbf{x}_t^\top \hat{\boldsymbol{\delta}} + \hat{\boldsymbol{\gamma}}^\top \widehat{\mathbf{D}}(\mathbf{y}_t^S) + e_t^* \qquad [19]$$

for $t = q_1 + 1, \ldots, T + h$, with initial points $\{y_t^* = e_t^*,\ t \le q_1\}$. We next refit the joint LLM-powered ARX model Eq. (6) using the bootstrap sample $\{y_t^*,\ t = q_1 + 1, \ldots, T\}$ and denote the refitted parameters as $\{\hat{\boldsymbol{\alpha}}^*, \hat{\boldsymbol{\theta}}^*, \hat{\boldsymbol{\delta}}^*, \hat{\boldsymbol{\gamma}}^*\}$. The h-step-ahead forecast for the bootstrap sample is calculated as

$$\hat{y}_{T+h}^* = \sum_{l=1}^{q_1} \hat{\alpha}_l^* \hat{y}_{T+h-l}^* + \mathbf{z}_{T+h}^\top \hat{\boldsymbol{\theta}}^* + \mathbf{x}_{T+h}^\top \hat{\boldsymbol{\delta}}^* + (\hat{\boldsymbol{\gamma}}^*)^\top \widehat{\mathbf{D}}(\mathbf{y}_{T+h}^S), \qquad [20]$$

where $\hat{y}_{T+h-1}^*, \hat{y}_{T+h-2}^*, \ldots$ are computed similarly with $\hat{y}_t^*$ for each $t \le T$ being y_t^* . Using $\hat{y}_{T+h}^*$ introduced above, the bootstrap residual is calculated as $\hat{e}_{T+h}^* = y_{T+h}^* - \hat{y}_{T+h}^*$.

We repeat the bootstrap procedure (i.e., Eq. (19) and Eq. (20)) $B \ge 1$ times to obtain a sequence of bootstrap residuals $\{\hat{e}_{T+h}^{*,(b)},\ b = 1, \ldots, B\}$. For each $\alpha \in (0, 1)$, denote the $\alpha/2$ quantile and $1 - \alpha/2$ quantile of the bootstrap residuals as $\hat{q}_{\alpha/2}^h$ and $\hat{q}_{1-\alpha/2}^h$, respectively. Then we can construct the bootstrap prediction interval with confidence level $1 - \alpha$ as

$$\mathrm{PI}^{BOOT}(\hat{y}_{T+h}) = \left[\hat{y}_{T+h} + \hat{q}_{\alpha/2}^h,\ \hat{y}_{T+h} + \hat{q}_{1-\alpha/2}^h\right]. \qquad [21]$$

We provide rigorous theoretical guarantees for the asymptotic coverage rates of both the BJ and bootstrap prediction intervals in Theorems 2 and 3 of *Section 1 of SI Appendix*, respectively. These results ensure that our two proposed prediction intervals achieve the desired coverage level asymptotically under some mild regularity conditions.

Extension of LLM-TS to deep learning models. We emphasize that the linear structures in Eq. (1) and Eq. (2) are chosen primarily to align with the macroeconomic case studies considered in this paper. Macroeconomic indices are typically observed at a monthly frequency, resulting in short samples with relatively stable trend components. In such settings, linear specifications are effective and highly interpretable for short-horizon forecasting, and often perform competitively relative to more complex deep learning models; see *Tables S18 and S19 of SI Appendix* for empirical evidence.

Our joint modeling framework is, however, not restricted to linear time series models. Its two core components: i) LLM-derived text embedding features and ii) the joint residual modeling extend naturally to modern deep learning architectures. Recent studies have shown that LLM-derived representations can serve as informative inputs for time-series forecasting (29, 43); our case analysis further confirms that such representations remain effective even when embedded within classical structural models. We now introduce a deep-learning model extension of LLM-TS.

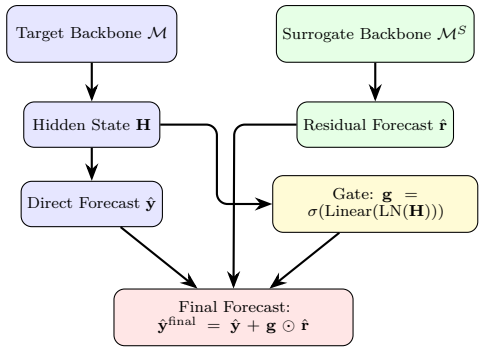


Fig. 3. LLM-TS with deep learning backbones. The target backbone $\mathcal{M}$ produces a direct forecast and a hidden representation $\mathbf{H}$. The surrogate backbone $\mathcal{M}^S$ predicts a residual correction $\hat{\mathbf{r}}$. A gating mechanism computes horizon-wise weights $\mathbf{g} \in [0, 1]^H$ and adaptively combines the two paths.

The central innovation of our framework lies in the joint residual modeling strategy. In many applications, some surrogate series related to the target variable are available, but directly incorporating them as symmetric inputs can distort the intrinsic temporal lag structure of the target series and degrade forecasting accuracy (27, 28). We therefore impose strict *channel independence* in the target model: the target series is forecast *solely* from its own lagged history. At the same time, discarding surrogate information entirely may forfeit valuable leading or cross-series signals.

Our joint residual modeling principle resolves such tension by allowing surrogate series to contribute *only* through residual corrections rather than as direct inputs. To generalize beyond the linear models, we replace the target and surrogate components with deep time-series backbones $\mathcal{M}$ and $\mathcal{M}^S$, respectively. The target backbone produces an H-step direct forecast $\hat{\mathbf{y}}$ along with a hidden representation $\mathbf{H}$, while the surrogate backbone predicts the corresponding residual correction $\hat{\mathbf{r}}$. Because deep learning models typically perform direct multi-step forecasting, surrogate information is most informative at short horizons, whereas long-horizon behavior is governed primarily by the intrinsic dynamics of the target series. To accommodate such heterogeneity, we introduce a *gating mechanism* (44) that regulates adaptively the contribution of surrogate residuals across different forecast horizons, as shown in Figure 3. The gate incorporates selectively the surrogate information only when it improves prediction accuracy, thereby ensuring stability and robustness while preserving the dominant role of the target backbone in long-horizon forecasting. A full deep learning extension of the LLM-TS framework, together with comparisons against state-of-the-art forecasting models on standard benchmarks, is provided in *Section 5 of SI Appendix*.

Simulation examples

In this section, we present several simulation examples to examine the finite-sample performance of the LLM-TS method. The synthetic data sets used in these simulations

are derived from the *Wall Street Journal* (WSJ) corpus; see *Section 2 of SI Appendix* for simulation setting details.

As seen in *Table S1 in SI Appendix*, the LLM-TS model consistently achieves lower both relative root prediction mean squared error (rPMSE) and relative sign prediction error (rSign) than the benchmark models. Such performance advantage becomes even more pronounced as the model error correlation level ρ increases. Importantly, the improvement holds consistently across both short and long term horizons, demonstrating the advantages and robustness of the LLM-TS method in forecasting tasks compared to classical approaches.

Table S2 in SI Appendix reports the average coverage of the prediction intervals and their mean interval lengths (in parentheses) for each forecast horizon H and correlation level ρ. The AR model maintains high coverage across all cases. The two LLM-TS variants—based on the BJ and bootstrap intervals—match the nominal coverage in the short run and stay close in the long run, with only occasional slight undercoverage. Both LLM-TS variants produce much narrower prediction intervals than the AR model, demonstrating their ability to produce more accurate forecasts without losing coverage.

We also conduct additional simulations to test LLM-TS's robustness to different types of model misspecifications: the omitted key predictor, overfitting, and non-Gaussian error distribution (i.e., t-distribution). In all these cases, the LLM-TS remains robust by providing reliable predictive inference results; see *Section 2 of SI Appendix* for details.

LLM-CPI

In this section, we validate the LLM-TS framework for inflation forecasting across two distinct economic environments: the U.S. and China. Our approach to constructing high-frequency, LLM-powered surrogate signals is readily generalizable to a broad class of time-series prediction problems whenever unstructured data is available as auxiliary information. When applied specifically to inflation forecasting, we refer to this framework as the LLM-CPI.

LLM-based high-frequency inflation index construction.

Online text time-series data collection. To capture real-time inflation narratives across the distinct economic environments of the U.S. and China, we leverage two complementary information platforms that reflect both professional and grassroots perspectives. In the U.S. context, news coverage in *The Wall Street Journal* (WSJ) tracks systematically the monetary policy developments, price pressures, and macro-financial conditions, serving as a forward-looking signal of dynamics in a mature market economy. In contrast, reflecting China's unique information ecosystem, Weibo provides a critical real-time channel for public discourse, with 588 million monthly active users (45). As a platform studied widely for its role in shaping and amplifying public sentiment (46, 47), Weibo offers a granular view of consumer expectations and perceptions that differs fundamentally from traditional Western media outlets.

To operationalize this design, we collect the full corpus of WSJ news articles published between September 1, 2015 and August 31, 2025. For the Chinese market, we aggregate all Weibo posts containing predefined inflation-related keywords following (6) from January 1, 2019 to September 30, 2025. See *Sections 3 and 4 of SI Appendix* for the data details. The resulting WSJ corpus comprises 422,444 news articles, while the Weibo data set contains approximately 159.5 million short posts. We should emphasize that the collected text data sets not only pertain to inflation-related contents, but also include background noise, including advertisements, e-commerce contents, and sales promotions.

LLM-based inflation index construction. Our text analysis begins with standard preprocessing steps to remove duplicate samples and non-textual elements such as emojis and special symbols. The primary challenge, however, lies in identifying inflation-related texts from a large volume of unrelated contents. Extracting such signals from large text collections is difficult for three major reasons. First, *contextual ambiguity* arises because commercial or promotional texts often resemble genuine economic discussions, making it difficult to distinguish informative content from noise. Second, *linguistic differences* across platforms add complexity, ranging from the formal financial language used in the *Wall Street Journal* to the informal and highly varied expressions found in user-generated contents on Weibo. Third, *semantic diversity* implies that inflation and price changes are described in many different ways across professional and grassroots discussions, which limits the effectiveness of simple filtering rules. As a result, directly applying unsupervised learning methods to identify inflation-related content has limited effectiveness.

To address these practical challenges, we develop a hierarchical LLM-based learning framework that is summarized in Algorithm 1. We first employ a chain-of-thought (CoT; 48) prompting strategy using GPT-4 (specifically `gpt-4-turbo-2024-04-09`) to annotate two random samples of 8,000 WSJ news articles and 30,000 Weibo posts, respectively. Leveraging the few-shot learning capabilities of GPT-4 (49), we implement a hierarchical annotation procedure in which documents are first classified into broad categories (e.g., Inflation, Lifestyle, and Entertainment), and those identified as inflation-related are subsequently assigned a continuous severity score; see *Sections 3 and 4 of SI Appendix* for prompt details.

Using these high-quality annotations, we further fine-tune a set of domain-specific BERT models. Specifically, we use *Category-BERT* for thematic classification and *CPI-BERT* for sentiment quantification across both text data sets. In addition, for corpora where commercial content is prevalent, we fine-tune an *Advertisement-BERT* model to filter promotional noise. Given the linguistic characteristics of the data, we adopt the `bert-base-cased` architecture for the WSJ corpus to preserve case-sensitive information in English text, and the `bert-base-chinese` architecture for processing Weibo posts to ensure accurate character-level tokenization. The fine-tuned BERT models exhibit strong performances on the held-out test sets, which are reported in *SI Appendix*; see *Figs. S13–S14* for the WSJ data and *Figs. S2–S4* for the Weibo data.

In the final deployment stage, we apply such cascaded framework sequentially to the full text data sets. The WSJ news articles are processed directly for thematic relevance, while for Weibo, the LLM pipeline first removes commercial content using *Advertisement-BERT* and then applies *Category-BERT* to identify inflation-related texts. Only texts classified as inflation-related are passed to the final scoring

Algorithm 1 Constructing the LLM-generated high-frequency index

Input: The preprocessed text data set.
Output: The LLM-generated high-frequency index.
Step 1: Sampling
Randomly sample a subset of text documents.
Step 2: Annotation
for each document in the sampled subset **do**
    Use *chain-of-thought prompting* with GPT-4 to annotate the document:
        (a) If commercial content exists, determine if the document is an advertisement;
        (b) Determine whether the document is relevant to inflation;
        (c) If related to inflation, assess the continuous degree of inflation it represents.
Step 3: Fine-tuning
Fine-tune the following BERT models using the annotated data set:
    (a) **Advertisement-BERT:** Identify advertisement content (when applicable);
    (b) **Category-BERT:** Identify whether non-advertisement posts are related to inflation;
    (c) **CPI-BERT:** Estimate the continuous degree for target-related documents.
Step 4: Sequential prediction
if commercial content exists in the corpus **then**
    Apply **Advertisement-BERT** to remove advertisement documents.
Apply **Category-BERT** to filter out non-relevant documents.
Apply **CPI-BERT** *only* to the remaining target-related documents.
Compute the LLM-generated daily high-frequency index using Eq. (22).

stage. This filtering procedure yields 256,747 relevant WSJ news articles (a 60.8% retention rate), and reduces the original Weibo data set to 7.35 million inflation-relevant posts (a 4.6% retention rate) from 1.49 million unique users. The retained documents are then processed by the *CPI-BERT* regression model to generate fine-grained sentiment scores, $\text{Score}_i \in [0, 1]$. *Figs. S6 and S15 in SI Appendix* depict the daily volumes of inflation-up and inflation-down discussions at this final stage for both WSJ and Weibo data sets, respectively. See *Sections 3 and 4 of SI Appendix* for more fine-tuning and prediction details.

To construct the LLM-generated daily inflation index, we pair each document's continuous inflation score with its publication date, forming a data set $\{(\text{Score}_i, \text{Date}_i),\ i = 1, \ldots, N\}$, where Date_i denotes the posting date of the ith document entry. The *LLM-generated daily inflation index* for day d is defined as

$$\text{Inflation}_d = \frac{\sum_{i=1}^{N} \text{Score}_i\, \mathbb{I}(\text{Date}_i = d)}{\sum_{i=1}^{N} \mathbb{I}(\text{Date}_i = d)}, \qquad [22]$$

where $\mathbb{I}(\cdot)$ is the indicator function. Fig. 1 plots the LLM-generated daily inflation index Inflation_d in (22) and the monthly survey CPI for the WSJ data. See *Fig. S5 in SI Appendix* for the corresponding plot for the Weibo data. Since the LLM-generated daily inflation index Inflation_d highly fluctuates, we smooth it into three ten-day periods within each month in practical application. We denote the resulting LLM-generated inflation index as $\{y^S_{t,k} \in \mathbb{R}, t = 1, \ldots, T, k = 1, \ldots, K\}$, where $K = 3$ represents the three periods within each month, and each $y^S_{t,k}$ corresponds to a surrogate of the CPI generated by LLMs (i.e., ChatGPT and the trained BERT models) for the kth period of the tth month.

Online text embeddings. Our suggested LLM-CPI framework incorporates two text embedding methods: the topic probability embeddings from the latent Dirichlet allocation (LDA) (50), and the BERT embeddings extracted from the fine-tuned CPI-BERT model architecture. Specifically, we implement the LDA model to derive topic probability distributions from the text data. Each document is represented as a K-dimensional vector, where elements correspond to posterior probabilities of memberships in K latent thematic clusters. These document-topic distributions are temporally aggregated monthly through averaging; see *Figs. S7 and S16 in SI Appendix* for the LDA topic results, as well as *Tables S22 and S23 in SI Appendix* for related hashtages from the Weibo data. For the BERT embeddings, we extract 768-dimensional vectors through mean pooling of the final hidden layer right before the output layer of the fine-tuned CPI-BERT model (i.e., a deep neural network), capturing semantic patterns in individual documents (i.e., news articles or posts). These document-level embeddings are averaged within each month to create monthly LLM-based economic text features; see *Sections 3 and 4 of SI Appendix* for the text embedding details.

LLM-CPI forecasting and inference results.

Low-frequency survey CPI and other indicators. The target variable y_t in Eq. (1) is the observed monthly inflation rate. For both the U.S. and Chinese economic analyses, we use the urban CPI as the measure of inflation. Following standard normalization, we obtain the series y_t for $t = 1, \ldots, T$, where the sign of y_t indicates whether inflation has increased or decreased relative to the previous month. As an additional macroeconomic control, we also collect the national urban surveyed unemployment rate reported monthly by the NBSC and standardize it in the same manner. See *Sections 3 and 4 of SI Appendix* for details on data collection and preprocessing. The predictors in the LLM-CPI model consist of two components. The first component comprises text embeddings extracted from the online texts, as described above. The second component is a widely used inflation-related macroeconomic control variable, standardized and denoted as z_t for $t = 1, \ldots, T$.

Out-of-sample forecasting. Based on the proposed LLM-TS framework in Eq. (1), Eq. (2), and Eq. (6), we have the high-frequency LLM-based inflation index $y^S_{t,k}$ as our surrogates in Eq. (2), the text embeddings $\mathbf{x}_t$, and the target variable y_t in Eq. (1). For easy reference, we will name the LLM-CPI method with the LDA and BERT embeddings as LLM+LDA and LLM+BERT, respectively.

We now assess the out-of-sample forecasting performance of these two methods. To isolate the effect of the text embedding features, we first exclude the macroeconomic predictor z_t (i.e., the unemployment rate). We compare

the LLM+LDA and LLM+BERT models against three well-established inflation forecasting benchmarks: the AR model, RW model, and AVE model, as well as two deep learning models: the PatchTST (28) and Time-LLM (29). In addition, we compare to the direct text-based model using the LDA and BERT embeddings as covariates (without LLM-CPI for model error variance reduction). For simplicity, we refer to the text-based prediction model with the LDA embeddings as the LDA model, and that with the BERT embeddings as the BERT model (with slight abuse of terminology). Detailed model specifications are provided in *Section 3 of SI Appendix*.

The out-of-sample forecasting period spans the H months immediately preceding the end of the online text time series up to its endpoint, yielding H forecast steps. Specifically, the online text series ends on August 31, 2025 for the WSJ corpus and on September 30, 2025 for the Weibo corpus. This evaluation window is strictly independent of the samples used for model fitting and model selection, and occurs strictly after the release dates of the LLMs used for annotation. These design choices ensure the integrity of the out-of-sample assessment. Additional discussions on potential information leakage considerations are provided in *Sections 3 and 4 of SI Appendix*. We choose the AR model as the baseline and evaluate the performance using the relative root prediction mean squared error ($\text{rPMSE}^{\text{AR}}(H)$) and the relative sign prediction error ($\text{rSign}^{\text{AR}}(H)$) defined in Eq. (23) and Eq. (24) of *SI Appendix*. For both performance measures, we now have $Q = 1$ due to a single observation for each month.

Table 1 summarizes the results across different forecast horizons H for the WSJ data; see *Table S10 of SI Appendix* for the case of Weibo data. We outline the key findings below. 1) *Text embeddings enhance prediction accuracy.* Incorporating textual signals (without the LLM-CPI framework) reduces consistently the forecast errors compared to the baselines. Specifically, the LDA model achieves an average relative prediction mean squared error ($\text{rPMSE}^{\text{AR}}(H)$) of 0.996, outperforming the RW (1.325), AVE (1.096), Time-LLM (1.811), and PatchTST (1.157) benchmarks. The BERT-based model performs even better, attaining an average $\text{rPMSE}^{\text{AR}}(H)$ of 0.950. Both embedding-based models also exhibit strong directional accuracy. The relative sign prediction error ($\text{rSign}^{\text{AR}}(H)$) is 0.889 for the LDA model and 0.806 for the BERT model, both substantially lower than those of all classical time series benchmark models and comparable to the performance of deep learning–based approaches. 2) *The LLM-CPI method improves substantially the prediction performance.* Incorporating LLM-based surrogate signals systematically improves performance across both point and directional forecast metrics. For the LDA-based specification, the LLM-CPI framework reduces the $\text{rPMSE}^{\text{AR}}(H)$ from 0.996 to 0.916. The improvement is most pronounced for the BERT-based specification. The LLM+BERT model achieves the best overall performance, lowering $\text{rPMSE}^{\text{AR}}(H)$ to 0.882. More importantly, it delivers a marked improvement in directional accuracy: the relative sign prediction error drops from 0.806 for the standalone BERT model to 0.639 under the LLM-CPI framework.

Inflation predictive inference. We further evaluate the predictive inference performance of LLM-CPI. We employ the Box–Jenkins (BJ) procedure for constructing the prediction intervals in LLM-CPI. We compare with both AR and SPI prediction intervals (21). The corresponding results for LLM-CPI with the bootstrap prediction interval are presented in *Sections 3 and 4 of SI Appendix*.

Table 1 reports the prediction interval coverage rates and average interval lengths across different forecast horizons $H = 8$ to 15; see *Table S11 of SI Appendix* for the case of Weibo data. The major findings are summarized below. 1) *Text embeddings improve prediction interval efficiency.* Models that incorporate text embeddings produce shorter prediction intervals than the autoregressive benchmark while maintaining nominal coverage. As shown in the bottom panel of Table 1, the LDA and BERT models (without the LLM-CPI structure) achieve average interval lengths of 3.902 and 3.806, respectively, both of which are tighter than the AR benchmark (3.932). Importantly, both models retain a 100% coverage rate across all horizons H. These results reinforce the conclusion that textual signals contain valuable predictive information that reduces uncertainty, with BERT embeddings yielding the sharpest intervals among the standalone text-based models. 2) *LLM-CPI yields further gains in inference precision.* The LLM-augmented models deliver substantial additional reductions in interval length, reflecting markedly improved inference efficiency. In particular, the LLM+LDA model reduces the average interval length to 2.920, corresponding to a reduction of more than 25% relative to the AR benchmark, while maintaining a reliable average coverage rate of 0.912. Similarly, the LLM+BERT model achieves an average interval length of 3.731 with a coverage rate of 0.930. The SPI method yields wider prediction intervals and exhibits lower empirical coverage. Together, these results demonstrate that the LLM-CPI framework produces tighter and more informative prediction intervals by leveraging LLM-based surrogate signals to denoise textual information and reduce prediction variance.

Additional empirical results that incorporate the unemployment rate as a predictor are reported in *Tables S14 and S15 of SI Appendix* for the case of WSJ data, and *Tables S10 and S11 of SI Appendix* for the case of Weibo data. These results continue to show that the LLM-CPI model outperforms consistently the benchmark methods. Moreover, the corresponding analyses for the WSJ corpus and Weibo corpus exhibit the same qualitative pattern, indicating that the performance gains of the LLM-CPI framework are robust and broadly applicable across different economic environments. Together, these findings suggest a degree of *universality* in the effectiveness of online text signals and LLM-based surrogate information within the LLM-TS framework.

The impact of COVID-19 on inflation. A fundamental question in the evaluation of machine learning-based economic forecasting models is whether observed performance gains reflect genuine predictive content or instead arise from overfitting or spurious correlations. To address such concern, we exploit the COVID-19 pandemic, which began in early 2020 and constituted a major global economic shock affecting both the U.S. and Chinese economies, as a natural exogenous event to conduct a structural robustness check of the LLM-CPI framework. To address this concern, we partition the data into three regimes for the U.S. analysis: the *pre-pandemic* period (October 2016 to January 2020), the *during-pandemic* period (February 2020 to May 2023), and the *post-pandemic* period (June 2023

Table 1. The $\mathrm{rPMSE}^{\mathbf{AR}}(H)$, $\mathrm{rSign}^{\mathbf{AR}}(H)$, $\mathrm{Coverage}_m(H)$, and $\mathrm{Length}_m(H)$ results across different horizons H.

Type	Method	8	9	10	11	12	13	14	15	Ave.
$\mathrm{rPMSE}^{\mathrm{AR}}(H)$	RW	1.172	0.987	0.985	1.081	1.955	1.852	1.026	1.541	1.325
	AVE	1.102	1.095	1.076	1.033	1.194	0.964	1.057	1.246	1.096
	LDA	0.977	0.978	0.981	1.001	0.993	1.003	1.029	1.009	0.996
	BERT	0.919	0.916	0.919	0.954	0.968	0.984	0.969	0.972	0.950
	LLM+LDA	0.882	0.887	0.888	0.898	0.926	0.927	0.964	0.955	0.916
	LLM+BERT	0.827	0.826	0.828	0.877	0.929	0.934	0.925	0.912	0.882
	Time-LLM	1.202	2.638	1.904	1.507	1.237	2.334	1.380	2.284	1.811
	PatchTST	1.124	0.911	0.934	0.895	1.638	1.400	1.002	1.354	1.157
$\mathrm{rSign}^{\mathrm{AR}}(H)$	RW	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	AVE	0.833	2.000	2.000	1.667	0.556	0.556	0.444	1.000	1.132
	LDA	0.667	1.000	1.000	1.333	0.444	0.556	0.778	1.333	0.889
	BERT	0.500	1.000	1.000	1.333	0.444	0.444	0.556	1.167	0.806
	LLM+LDA	0.667	1.000	1.000	1.333	0.667	0.444	0.556	1.167	0.854
	LLM+BERT	0.500	1.000	1.000	0.667	0.333	0.333	0.444	0.833	0.639
	Time-LLM	0.500	2.667	1.667	2.000	0.556	0.667	0.667	1.167	1.236
	PatchTST	0.500	1.000	1.000	1.667	0.333	0.333	0.333	2.000	0.896
$\mathrm{Coverage}_m(H)$ ($\mathrm{Length}_m(H)$)	AR	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		(3.873)	(3.894)	(3.917)	(3.941)	(3.959)	(3.954)	(3.948)	(3.973)	(3.932)
	LDA	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		(3.853)	(3.875)	(3.897)	(3.916)	(3.928)	(3.916)	(3.901)	(3.926)	(3.902)
	BERT	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		(3.754)	(3.774)	(3.796)	(3.817)	(3.832)	(3.816)	(3.818)	(3.844)	(3.806)
	LLM+LDA (BJ)	0.875	0.889	0.900	0.909	1.000	0.923	0.929	0.867	0.912
		(2.864)	(2.960)	(3.052)	(3.151)	(3.189)	(2.852)	(2.614)	(2.681)	(2.920)
	LLM+BERT (BJ)	0.875	0.889	0.900	1.000	0.917	1.000	0.929	0.933	0.930
		(3.420)	(3.560)	(3.750)	(3.985)	(4.031)	(3.765)	(3.411)	(3.927)	(3.731)
	SPI (LDA)	0.875	0.889	0.900	0.909	0.917	0.923	0.929	0.933	0.909
		(3.318)	(3.318)	(3.318)	(3.320)	(3.311)	(3.320)	(3.324)	(3.319)	(3.319)
	SPI (BERT)	0.875	0.889	0.900	0.909	0.917	0.923	0.929	0.933	0.909
		(3.414)	(3.416)	(3.413)	(3.389)	(3.337)	(3.328)	(3.347)	(3.335)	(3.372)

Top panel: the relative cumulative PMSE values compared to the AR benchmark; smaller values indicate better performance. Middle panel: the relative cumulative sign prediction error values compared to the AR benchmark. Bottom panel: the values in the parentheses are interval length. All methods are without unemployment rate, and the BJ prediction interval is used.

to August 2025)*. This time segmentation enables us to examine changes in text content, topic structures, and model performance across markedly different economic regimes, thereby providing a stringent test of the robustness of the proposed approach.

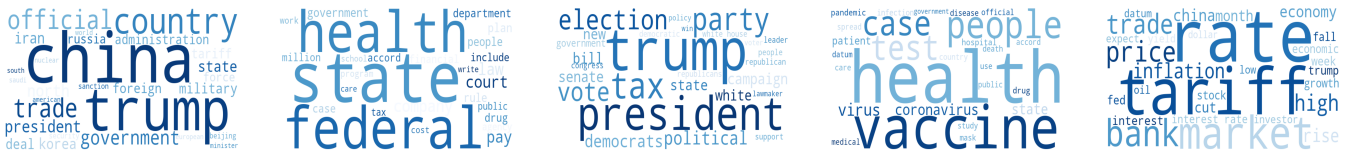
For each period, we independently train the LDA model to extract period-specific latent topics. To evaluate the prediction performance, we designate the final six months of each period as the testing sample, with the remaining (i.e., earlier) months used for training. We also apply the model selection procedure on the training sample as discussed in *Section 3 of SI Appendix* to reduce the dimensionality of text embeddings. *Tables S12 and S16 in SI Appendix* report the root prediction mean squared error (PMSE) and the sign prediction error (Sign) for the corresponding periods in the U.S. and Chinese economies, respectively†. These results unveil that the LLM-CPI model commonly outperforms the benchmark AR model in terms of both PMSE and Sign over different periods across both U.S. and Chinese economic settings.

Fig. 4 displays topic visualizations for the WSJ corpus across the three periods. The changes in dominant topics and keywords reveal a clear shift in textual narratives across these regimes. The pre-pandemic topics capture policy- and geopolitics-driven cost pressures, regulatory and fiscal influences on prices, and political uncertainty affecting expectations, all of which play a central role in shaping inflation dynamics and help explain their strong predictive power for the U.S. inflation. During the pandemic, the topic reflects the evolution of the COVID-19 public health crisis, emphasizing infection spread, medical capacity, vaccination efforts, and government responses that shaped economic activity during the pandemic. The post-pandemic macroeconomic environment—characterized by interest rate adjustments, inflation concerns, trade and tariff dynamics, and financial market responses involving banks, investors, and asset prices—captures policy-driven and market-mediated inflation dynamics in the post-pandemic period, making it highly informative for inflation forecasting.

For the Weibo data, which reflects grassroots public opinion, the text signals provide more diverse information. The selected topics are illustrated in *Figs. S11 and S12 in SI Appendix*. Beyond topic modeling, the Weibo platform also allows for deeper analysis by exploiting user-generated hashtags. We select the 10 most frequently occurring hashtags in posts to represent the human-readable meaning

*The timeline of the COVID-19 pandemic in the U.S. is available at https://en.wikipedia.org/wiki/COVID-19_pandemic_in_the_United_States. For China, given the presence of nationwide lockdowns, we instead define the *pre- and during-lockdown* period from January 1, 2019 to December 31, 2021, and the *post-lockdown* period from January 1, 2022 to December 31, 2023; see https://en.wikipedia.org/wiki/COVID-19_pandemic_in_mainland_China.

†We do not use the relative errors here since the baseline AR model occasionally yields zero error when the forecast horizon H is small.



(a) Pre-pandemic: Trade tensions and geopolitics **(b)** Pre-pandemic: Domestic regulation and public spending **(c)** Pre-pandemic: Electoral politics and fiscal policy **(d)** During-pandemic: Public health crisis **(e)** Post-pandemic: Monetary policy and inflation

Fig. 4. The word cloud visualizations of the LDA topics extracted from the WSJ corpus.

of each topic word, see *Section 7 of SI Appendix* for more details. *Table S24 in SI Appendix* presents the most-discussed original hashtags corresponding to our selected topic during the lockdown period. It is seen that the lockdown-period topics emphasize crisis-related themes such as debt defaults, bankruptcies, and city lockdowns, reflecting the immediate economic fallout. *Tables S25–S28 in SI Appendix* list the four selected topics and related hashtags during the post-lockdown period. These LDA-identified topics highlight how inflationary pressures in China arise from the interaction of input cost shocks, financial market adjustments, household behavior, and external risks.

Robustness analysis of LLM-TS in the U.S. economy. The preceding analysis demonstrates that the LLM-TS performs well in predicting urban CPI in both the U.S. and Chinese economies. A natural next question is whether LLM-TS remains effective when applied to other related macroeconomic variables. We address this question by extending LLM-TS to a broad set of macroeconomic price and financial indices. To this end, we analyze a comprehensive panel of U.S. macroeconomic indicators from the FRED-MD database (38). Detailed descriptions of the selected indices are provided in *Table S13 of SI Appendix*.

As shown in *Table S17 of SI Appendix*, the LLM-TS delivers substantial accuracy gains across the vast majority of series. Relative to the AR benchmark, LLM-TS improves the relative prediction mean squared error (rPMSE) in 17 out of 23 cases, and improves the sign accuracy in 16 out of 23 cases.

Table S19 of SI Appendix further compares LLM-TS with deep learning-based forecasting methods. The LLM-TS outperforms substantially these alternatives. This result is consistent with the nature of the data: deep forecasting models typically require large-scale time-series corpora to fully exploit their expressive capacity, whereas monthly macroeconomic series contain only a few dozen observations, placing these methods in a few-shot learning regime. Moreover, the CPI and related macroeconomic indices exhibit relatively stable and low-complexity temporal dynamics, for which classical AR-type models and the joint residual modeling employed by LLM-TS are already well suited. In such settings, prediction difficulty arises primarily from data scarcity rather than model flexibility, limiting the advantages of high-capacity neural network models.

Table 2 summarizes the predictive inference results across different U.S. macroeconomic price indices. Overall, LLM-TS produces shorter prediction intervals than both the AR benchmark and SPI while maintaining nominal coverage, achieving the best overall performance for 20 out of 23 series. Relative to the AR benchmark, these results show that

Table 2. The $\mathrm{Coverage}_m(H)$ and interval length results across different price indices.

Target	AR	LLM-TS (BJ)	SPI
WPSFD49207	1.00 (4.031)	1.00 (3.105)	1.00 (4.361)
WPSFD49502	1.00 (4.031)	1.00 (3.085)	1.00 (4.302)
PPICMM	0.92 (3.751)	0.58 (2.546)	0.75 (3.330)
CPIAUCSL	1.00 (3.997)	1.00 (2.865)	1.00 (3.268)
CPIAPPSL	0.92 (3.949)	0.83 (2.659)	0.92 (3.428)
CPITRNSL	1.00 (4.024)	1.00 (2.455)	1.00 (3.369)
CPIMEDSL	1.00 (3.959)	1.00 (3.189)	0.92 (3.311)
CUSR0000SAC	1.00 (4.033)	1.00 (2.335)	1.00 (3.975)
CUSR0000SAD	1.00 (3.928)	0.83 (1.474)	1.00 (2.485)
CUSR0000SAS	1.00 (3.914)	1.00 (3.056)	1.00 (2.858)
CPIULFSL	1.00 (4.061)	1.00 (2.739)	1.00 (3.210)
CUSR0000SA0L2	1.00 (4.033)	1.00 (2.889)	1.00 (3.384)
CUSR0000SA0L5	1.00 (3.996)	1.00 (2.823)	1.00 (3.057)
PCEPI	1.00 (3.988)	1.00 (2.704)	1.00 (3.347)
DNDGRG3M086SBEA	1.00 (4.032)	1.00 (2.319)	1.00 (3.577)
DSERRG3M086SBEA	1.00 (3.806)	1.00 (3.882)	0.92 (3.048)
DDURRG3M086SBEA	1.00 (3.937)	0.75 (2.469)	1.00 (3.223)
OILPRICEx	1.00 (4.071)	1.00 (2.277)	1.00 (3.351)
WPSID61	1.00 (3.996)	1.00 (3.818)	1.00 (3.545)
WPSID62	1.00 (4.015)	0.83 (2.381)	1.00 (4.104)
S&P 500	0.83 (3.889)	1.00 (2.950)	0.75 (3.357)
S&P div yield	1.00 (3.927)	1.00 (2.850)	0.75 (2.899)
VIXCLSx	1.00 (3.993)	0.92 (2.794)	0.83 (2.862)
Count (smallest length)	0	20	3

The values in the parentheses are interval length, and coverage near the nominal level of 0.95 with smaller interval length is preferred.

incorporating text embeddings and joint residual modeling in the LLM-TS framework yields robust and systematic gains in predictive inference. The gains are particularly pronounced for the headline CPI, core inflation components, PCE inflation, and apparel and transportation prices, as well as key financial market variables, including the S&P 500, dividend yield, and the VIX. These series are especially sensitive to public narratives, market sentiment, and forward-looking economic expectations that are reflected prominently in financial news coverage.

Compared to SPI, the improvements can be partly attributed to the Box–Jenkins framework, which exploits effectively the approximate normality of macroeconomic index innovations while leveraging LLM-based surrogate information.

Discussions

We have investigated in this paper the problem of time series forecasting. Motivated by the recent developments in LLMs, our suggested framework of LLM-TS predictive inference is rooted on the joint time series model of both observed low-frequency field data and LLM-generated high-

frequency text surrogates. We have applied the proposed framework to inflation forecasting and inference, exploiting the correlations between low-frequency, survey-based inflation measurements and high-frequency, LLM-generated inflation signals constructed from two large-scale online text data sets. The model conditions on lagged monthly inflation indices, lagged LLM-generated daily inflation surrogates, macroeconomic covariates, and online text embeddings. Supported by theoretical guarantees, LLM-TS is shown to deliver accurate point forecasts and tight inflation predictive inference results in both U.S. and Chinese economic environments, thanks to the power of LLMs such as ChatGPT and the trained BERT models as well as text embeddings via LDA and BERT. We have further generalized the LLM-TS framework to deep learning-based time-series models, demonstrating that the underlying joint residual modeling principle is broadly applicable and not restricted to linear specifications.

It would be of interest to incorporate time-series surrogates generated by different LLM tools within the joint modeling framework. Exploring more advanced text-embedding architectures to extract informative textual features is another promising direction. In addition, extending conformal predictive inference to deep learning-based LLM-TS models would further enhance uncertainty quantification in nonlinear forecasting settings. These directions lie beyond the scope of the current paper and are left for future research.

Data availability

The analytic code is fully documented in executable `R Markdown` files and publicly available at https://github.com/suntiansheng/LLM-CPI-prediction-and-inference. The underlying data sets include the proprietary Wall Street Journal content and Weibo posts accessed under the platform terms. Because the WSJ terms of use may treat certain derived data sets (e.g., embeddings) as reproductions, creating potential legal uncertainty, we do not share the WSJ data. We release only non-expressive derived data sets from Weibo to support reproducibility.

1. PC Phillips, The sampling distribution of forecasts from a first-order autoregression. *J. Econom.* **9**, 241–261 (1979).
2. WA Fuller, DP Hasza, Properties of predictors for autoregressive time series. *J. Am. Stat. Assoc.* **76**, 155–161 (1981).
3. GE Box, GM Jenkins, GC Reinsel, GM Ljung, *Time Series Analysis: Forecasting and Control.* (John Wiley & Sons), (2015).
4. LA Thorsrud, Words are the new numbers: a newsy coincident index of the business cycle. *J. Bus. & Econ. Stat.* **38**, 393–409 (2020).
5. VH Larsen, LA Thorsrud, J Zhulanova, News-driven inflation expectations and information rigidities. *J. Monet. Econ.* **117**, 507–520 (2021).
6. C Angelico, J Marcucci, M Miccoli, F Quarta, Can we measure inflation expectations using twitter? *J. Econom.* **228**, 259–277 (2022).
7. Y Hong, F Jiang, L Meng, B Xue, Forecasting inflation using economic narratives. *J. Bus. & Econ. Stat.* **43**, 216–231 (2025).
8. A Agrawal, J Gans, A Goldfarb, *Prediction Machines, Updated and Expanded: The Simple Economics of Artificial Intelligence.* (Harvard Business Press), (2022).
9. E Brynjolfsson, D Li, L Raymond, Generative AI at work. *The Q. J. Econ.* **140**, 889–942 (2025).
10. JB Taylor, Discretion versus policy rules in practice in *Carnegie-Rochester Conference Series on Public Policy.* (Elsevier), Vol. 39, pp. 195–214 (1993).
11. RS Gürkaynak, B Sack, E Swanson, The sensitivity of long-term interest rates to economic news: evidence and implications for macroeconomic models. *Am. Econ. Rev.* **95**, 425–436 (2005).
12. CE Borio, AJ Filardo, Globalisation and inflation: new cross-country evidence on the global determinants of domestic inflation. *BIS Work. Pap.* (2007).
13. O Blanchard, The Phillips curve: back to the 60's? *Am. Econ. Rev.* **106**, 31–34 (2016).
14. A Atkeson, LE Ohanian, Are Phillips curves useful for forecasting inflation? *Fed. Reserv. Bank Minneapolis Q. Rev.* **25**, 2–11 (2001).
15. JH Stock, MW Watson, Why has US inflation become harder to forecast? *J. Money, Credit. Bank.* **39**, 3–33 (2007).
16. M Weber, F D'Acunto, Y Gorodnichenko, O Coibion, The subjective inflation expectations of households and firms: Measurement, determinants, and implications. *J. Econ. Perspectives* **36**, 157–184 (2022).
17. RJ Gordon, US inflation, labor's share, and the natural rate of unemployment. *NBER Work. Pap. Ser.* (1988).
18. ZR McCaw, J Gao, X Lin, J Gronsbell, Synthetic surrogates improve power for genome-wide association studies of partially missing phenotypes in population biobanks. *Nat. Genet.* **56**, 1527–1536 (2024).
19. AN Angelopoulos, S Bates, C Fannjiang, MI Jordan, T Zrnic, Prediction-powered inference. *Science* **382**, 669–674 (2023).
20. T Zrnic, EJ Candès, Cross-prediction-powered inference. *Proc. Natl. Acad. Sci.* **121**, e2322083121 (2024).
21. M Bashari, RM Lotan, Y Lee, E Dobriban, Y Romano, Synthetic-powered predictive inference in *Conference on Neural Information Processing Systems.* (2025).
22. RA Stine, Estimating properties of autoregressive forecasts. *J. Am. statistical association* **82**, 1072–1078 (1987).
23. LA Thombs, WR Schucany, Bootstrap prediction intervals for autoregression. *J. Am. Stat. Assoc.* **85**, 486–492 (1990).
24. DN Politis, JP Romano, The stationary bootstrap. *J. Am. Stat. association* **89**, 1303–1313 (1994).
25. C Chatfield, Calculating interval forecasts. *J. Bus. & Econ. Stat.* **11**, 121–135 (1993).
26. H Wu, J Xu, J Wang, M Long, Autoformer: Decomposition Transformers with auto-correlation for long-term series forecasting. *Adv. Neural Inf. Process. Syst.* **34**, 22419–22430 (2021).
27. A Zeng, M Chen, L Zhang, Q Xu, Are Transformers effective for time series forecasting? in *Proceedings of the AAAI Conference on Artificial Intelligence.* (2023).
28. Y Nie, N H. Nguyen, P Sinthong, J Kalagnanam, A time series is worth 64 words: Long-term forecasting with Transformers in *International Conference on Learning Representations.* (2023).
29. M Jin, et al., Time-LLM: Time series forecasting by reprogramming large language models in *International Conference on Learning Representations.* (2024).
30. F Petropoulos, et al., Forecasting: theory and practice. *Int. J. forecasting* **38**, 705–871 (2022).
31. J Brand, A Israeli, D Ngwe, Using LLMs for market research. *Harv. Bus. Sch. Mark. Unit Work. Pap.* (2023).
32. JJ Horton, Large language models as simulated economic agents: what can we learn from homo silicus? *NBER Work. Pap. Ser.* (2023).
33. A Goli, A Singh, Frontiers: can large language models capture human preferences? *Mark. Sci.* **43**, 709–722 (2024).
34. T De Kok, ChatGPT for textual analysis? How to use generative LLMs in accounting research. *Manag. Sci.* **71**, 7223–8095 (2025).
35. Z Ye, H Yoganarasimhan, Y Zheng, LOLA: LLM-assisted online learning algorithm for content experiments. *Mark. Sci.* **44**, 975–1215 (2025).
36. JH Stock, MW Watson, Forecasting inflation. *J. Monet. Econ.* **44**, 293–335 (1999).
37. MC Medeiros, GF Vasconcelos, Á Veiga, E Zilberman, Forecasting inflation in a data-rich environment: the benefits of machine learning methods. *J. Bus. & Econ. Stat.* **39**, 98–119 (2021).
38. MW McCracken, S Ng, FRED-MD: a monthly database for macroeconomic research. *J. Bus. & Econ. Stat.* **34**, 574–589 (2016).
39. R Chahrour, K Nimark, S Pitschner, Sectoral media focus and aggregate fluctuations. *Am. Econ. Rev.* **111**, 3872–3922 (2021).
40. TL Lai, CZ Wei, Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals Stat.* **10**, 154–166 (1982).
41. PJ Bickel, DA Freedman, Some asymptotic theory for the bootstrap. *The Annals Stat.* **9**, 1196–1217 (1981).
42. DA Freedman, Bootstrapping regression models. *The Annals Stat.* **9**, 1218–1228 (1981).
43. Z Pan, et al., S^2IP-LLM: Semantic space informed prompt learning with LLM for time series forecasting in *International Conference on Machine Learning.* (2024).
44. YN Dauphin, A Fan, M Auli, D Grangier, Language modeling with gated convolutional networks in *International Conference on Machine Learning.* (PMLR), pp. 933–941 (2017).
45. Weibo Corporation, Weibo announces first quarter 2024 unaudited financial results. *http://ir.weibo.com/news-releases/news-release-details/weibo-announces-first-quarter-2024-unaudited-financial-results* (2024).
46. X Feng, AC Johansson, Top executives on social media and information in the capital market: evidence from China. *J. Corp. Finance* **58**, 824–857 (2019).
47. B Qin, D Strömberg, Y Wu, Social media and collective action in China. *Econometrica* **92**, 1993–2026 (2024).
48. J Wei, et al., Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* **35**, 24824–24837 (2022).
49. F Gilardi, M Alizadeh, M Kubli, ChatGPT outperforms crowd workers for text-annotation tasks. *Proc. Natl. Acad. Sci.* **120**, e2305016120 (2023).
50. DM Blei, AY Ng, MI Jordan, Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022 (2003).