ChatMacro: Evaluating Inflation Forecasts of Generative AI[*]

M. Jahangir Alam   Shane Boyle   Huiyu Li   Tatevik Sekhposyan

January 27, 2026
Latest version

Abstract

Recent research suggests that generic large language models (LLMs) can match the accuracy of traditional methods when forecasting macroeconomic variables in pseudo out-of-sample settings generated via prompts. This paper assesses the out-of-sample forecasting accuracy of LLMs by eliciting real-time forecasts of U.S. inflation from ChatGPT. We find that out-of-sample predictions are largely inaccurate and stale, even though forecasts generated in pseudo out-of-sample environments are comparable to existing benchmarks. Our results underscore the importance of out-of-sample benchmarking for LLM predictions.

Keywords: large language models, generative AI, inflation forecasting
JEL Codes: C45, E31, E37

[*] Alam: Wilfrid Laurier Univ., moalam@wlu.ca; Boyle: Federal Reserve Bank of San Francisco, shane.boyle@sf.frb.org; Li: Federal Reserve Bank of San Francisco, huiyu.li@sf.frb.org; Sekhposyan: corresponding author, Texas A&M Univ., tsekhposyan@tamu.edu. We thank Regis Barnichon, Tom Cook, Domenico Giannone, Anne Hansen, Oscar Jorda, Sophia Kazinnik, Pablo Levi, Michael McCracken, Sergio Villalvazo Martin, Nitish Sinha, Ke Wang, and participants of the presentations at the 2025 Dolomiti Macro Meetings, Wilfrid Laurier Univ., American Univ., the Board of Governors, the FRB of San Francisco, and other presentations for their thoughtful comments. The views in this paper are solely the responsibility of the authors and should not be interpreted as reflecting the views of the Federal Reserve Bank of San Francisco or the Board of Governors of the Federal Reserve System. Chidambaram Ganesan provided excellent research assistance. The computational resources are funded through Microsoft Azure via the Microsoft for Startups program.

1 Introduction

The public release of generative AI tools (GenAI) marks a significant breakthrough, expected to revolutionize many aspects of modern life. In this article, we investigate the usefulness of these tools for macroeconomic forecasts or generating survey responses of a forecasting nature, such as surveys about expectations. More specifically, we examine whether an out-of-the-box general AI tool, such as OpenAI's ChatGPT-4 Turbo, which has been available for general public use, can accurately predict US inflation or generate usable survey responses on inflation expectations.

The predictive accuracy of GenAI is a relatively new and understudied topic. The few existing studies, such as Carriero, Pettenuzzo and Shekhar (2024) and Faria-e-Castro and Leibovici (2024) investigate the predictive performance of GenAI and find that they tend to produce more accurate forecasts than traditional macroeconomic surveys, can mimic professional forecasters with reasonable precision, while providing only mild and episodic gains relative to state-of-the-art time series forecasting models.[1]

In addition, several exercises have been conducted to simulate survey responses using GenAI. For example, Bybee (2025) compares the expectations generated by LLMs with existing surveys to assess whether they capture similar deviations from rational full-information expectations. Additionally, Wu, Xi and Xie (2025) and Zarifhonarvar (2026) propose a framework that leverages LLMs to generate consumer inflation expectations tailored to different demographic personas, while Hansen, Horton, Kazinnik, Puzzello and Zarifhonarvar (2025) simulate economic forecasts of professional forecasters using LLMs.

A commonality between these studies is that GenAI models are evaluated using historical data that may be in the estimation sample and 'training' of the model. That is, these studies generate forecasts or survey responses for outcomes that have already been realized and evaluate the success of the models by comparing model output with historical data. To avoid models simply re-

[1] More recent Time Series Foundational Models (TSFMs) and Time Series Language Models (TSLMs), such as Google's TimesFM by Das et al. (2024), Salesforce's Moirari by Liu et al. (2024), and Amazon's Chronos by Ansari et al. (2024) have gained traction.

calling historical data, most studies use prompts to instruct AI models not to use specific information. For example, a prompt may ask the model to forecast inflation for time t without using any information after the date $t - k$, where $k > 0$. The prompt is submitted after time t, and the AI models are trained on data covering period t.

We consider the above approaches as evaluating GenAI models in a *pseudo* out-of-sample environment. The traditional forecasting literature distinguishes between in-sample, out-of-sample, and pseudo out-of-sample environments. The in-sample approach utilizes all available data to estimate statistical relationships and characterizes a fit for various dates within the sample. The vast majority of the forecasting literature relies on out-of-sample environments designed to mimic a realistic scenario: the model is estimated using only the information available up to the forecast origin date and is used to predict a variable in the yet-to-be-observed future. As discussed in Clark and McCracken (2013), the out-of-sample prediction could be preferred to the in-sample one, as it guards against overfitting.[2]

A pseudo out-of-sample evaluation is a retrospective simulation that mimics the out-of-sample forecasting process using historical data. A key challenge in this setting is that the information set available to the forecaster in the prediction may differ from the information that was truly available in the past. This distinction, often referred to as using real-time data versus revised data, is crucial for reliable evaluation (Croushore, 2006).

For GenAI, the forecasts are inherently pseudo out-of-sample, since, when prompted to make a historical prediction, it is difficult to verify whether the "do not use information after $t - k$" period in the prompt successfully restricts the model from using that information. Lopez-Lira, Tang and Zhu (2025) conduct an experiment demonstrating that system and user prompt data restrictions are ineffective in preventing access to future data when querying LLMs.

Many pseudo out-of-sample LLM forecast evaluations demonstrate the absence of data leakage through date-restriction tests that query the model about significant events, such as presidential election outcomes or the COVID-19 pan-

[2] An exception is Inoue and Kilian (2005), which argues for an in-sample evaluation for predictive models due to more credible forecast evaluation tools (with higher power).

demic (see, for example, Wu et al., 2025), and compare its responses before and after the dates of these events. However, this method may have limitations, as highly significant events may be much easier for GenAI models to recognize and exclude (due to their relatively high coverage in the corpus) than more subtle financial or economic data points. For the latter, Crane, Karra and Soto (2025) and Lopez-Lira et al. (2025) document instances of perfect recall in the context of realized data.

Regardless of whether prompt engineering can control recall issues, this approach is also prone to forward-looking bias in model estimation — the parameters of the model are affected by the forward-looking information it predicts. As an alternative, in a limited number of studies, such as Bybee (2025) and Zarifhonarvar (2026), the knowledge cutoff date of LLMs is used to examine the effects of conditional information. This approach, too, has drawbacks and potential risks, since the deployment and update of LLMs may be staggered and automated (Microsoft, 2025). Therefore, it is not entirely clear whether relying on the knowledge cut-off dates is sufficient to establish complete information control.

The novelty and the main contribution of this paper are in benchmarking the quality and accuracy of GenAI predictions in an out-of-sample environment. We prompt OpenAI's ChatGPT-4 Turbo to forecast *future* US Consumer Price Index (CPI) on an hourly basis from May 13, 2024, until June 30, 2025. By asking for future inflation, we obtain truly out-of-sample forecasts that are neither contaminated by memory, forward-looking bias of parameter estimates of the model, nor by direct look-up issues, since the future information is nonexistent and the model is trained only on past data.

We find that evaluating ChatGPT forecasts in a pseudo out-of-sample environment drastically overstates their performance in the out-of-sample environment. First, we compare ChatGPT Nowcasts with publicly available state-of-the-art model-based predictions provided by the Federal Reserve Bank of Cleveland. ChatGPT is inferior to the Cleveland Fed by mean-squared-error and mean-directional-error metrics in both the pseudo out-of-sample and out-of-sample periods. However, the performance gap is much smaller in the pseudo out-of-sample environment. Second, we analyze the long-horizon forecasts of the Consumer Price Index (CPI) and find that they are more stale in the out-of-sample environment than in the pseudo out-of-sample environment, as they vary little with fluctuations in the CPI. Moreover, the forecast distributions are more dispersed and less informative in the out-of-sample environment. We also isolate episodes of leakage from direct look-ups of future information when providing multi-period predictions in pseudo out-of-sample.

Our study reveals a critical divergence: LLM out-of-sample predictions may be significantly inferior to their in-sample or pseudo out-of-sample performance. This underscores the importance of out-of-sample benchmarking of current LLMs when applied to forecasting. This concern is echoed in a related paper by Dunn et al. (2025), which uses newly released FOMC minutes to evaluate the extent to which data leakage contributes to the capabilities of LLMs models. Similarly, Kazinnik and Sinclair (2025) uses newly released FOMC minutes to evaluate the performance of ChatGPT simulated FOMC responses.

We are related to the broader and more long-standing literature on the promises and pitfalls of prediction with machine learning and big data. Earlier studies raised concerns about transparency, reproducibility, and overfitting (e.g. see Lazer, Kennedy, King and Vespignani, 2014 on Google Flu Trends). We add to this literature by highlighting the difficulty of using prompts to restrict information. These are new issues that are perhaps unique to forecasting with LLMs.

In what follows, Section 2 describes our methodology and evaluation metrics. Section 3 compares the pseudo out-of-sample and out-of-sample performances, while Section 4 provides evidence on leakage. Section 5 concludes and discusses future directions. Auxiliary results are delegated to the Supplemental Appendix.

2 Data and methodology

This section describes how we construct pseudo out-of-sample and out-of-sample inflation forecasts by LLM. We construct the sample by querying OpenAI GPT-4 Turbo, which was the latest model available in April 2024, when we configured the infrastructure to collect out-of-sample predictions.[3]

To make our study comparable to the literature on forecasting with LLMs, we use the following prompt to elicit forecasts for the US CPI inflation:

> Assume that you are in {`current_date`}. Please give me your best forecast of year-over-year seasonally adjusted Consumer Price Index (CPI) inflation in the United States (US) for the current month and the next 12 months. Please give me numeric values for these forecasts. Do not use any information that was not available to you as of {`current_date`} to formulate these forecasts. Give me a single estimate for each month in following format and also provide the information sources for the forecast.
> Example:
> Jan2024: 3.2%
> ...
> Information Sources:

To generate out-of-sample forecasts, we submit the above query every hour from May 13th, 2024, until June 30th, 2025, with the placeholder 'current_date' replaced by the actual date when the query is submitted. For example, we started collecting out-of-sample forecasts on May 13, 2024. The first prompt on that day has '2024-05-13 07:17:02' for the 'current_date' and ChatGPT is asked to forecast inflation for May 2024 and the next 12 months without using information that became available after '2024-05-13 07:17:02'.[4] The forecast is out-of-sample because inflation data for the current month (May) and the next 12 months were unavailable at the time of the query. This design ensures that ChatGPT is not using forward-looking information to formulate a prediction.

In the following, we use $f(t, t+h, \tau)$ to denote a forecast made on origin month-year t for inflation in target month-year $t+h$. The forecast horizon h takes values from 0 to 12, corresponding to a nowcast ($h = 0$) through to a twelve-month-ahead forecast ($h = 12$). Since we elicit multiple forecasts

[3] We deploy GPT-4 Turbo on Azure OpenAI Service through the Representational State Transfer (RESTful) API, which is suitable for large-scale retrievals.

[4] The system dates for the out-of-sample queries are in GMT.

throughout the same month-year t, we use τ to track the date and time (for example, 3 pm on the 1st of the month) of the query. For example, the response for our first out-of-sample query on May 13, 2024 at 07:17:02 reads:

> As of 2024-05-13 07:17:02, here are the CPI inflation forecasts for the United States from May 2024 to May 2025 based on available economic data, trends and economic policy analysis up to this date:
>
> May2024: 2.9%
> Jun2024: 3.0%
> Jul2024: 3.1%
> ...
> Mar2025: 2.4%
> Apr2025: 2.3%
> May2025: 2.3%
>
> Information Sources:
> 1. Historical CPI data and trends from the U.S. Bureau of Labor Statistics (BLS).
> 2. Current and forecasted economic policies from the Federal Reserve and U.S. government reports.
> 3. Global economic conditions and their impacts on the U.S. economy, as reported by the International Monetary Fund (IMF) and World Bank.
> 4. Economic analysis and forecasts from major financial institutions and think tanks.
>
> Please note that these estimates are based on the current understanding of economic indicators and are subject to change with new economic data or policy adjustments.

Using the aforementioned notation, the forecast in this response is

$$\begin{aligned} f(\text{May2024}, \text{May2024}, \text{7am of the 13th}) &= 2.9\% \\ f(\text{May2024}, \text{Jun2024}, \text{7am of the 13th}) &= 3.0\% \\ &\vdots \end{aligned}$$

For pseudo out-of-sample, we use the same prompt but generate forecasts using 'current_date' that ranges from January 1, 2019, to April 30th, 2024.[5] Additionally, instead of formulating the example as 'Jan2024: 3.2%', we give an example of 'Dec2018: 2.0%' to avoid providing forward looking information.[6] The forecasts obtained from these queries are pseudo out-of-sample because the data for the target dates are realized at the time we submitted the query and mimics the pseudo out-of-sample construction used in prior studies.

2.1 Performance metrics

We evaluate ChatGPT's performance by comparing its forecasts to realized inflation and a frontier forecast. We obtain realized inflation from the Federal Reserve Economic Data (FRED), using the monthly, seasonally adjusted, year-over-year Consumer Price Index for All Urban Consumers: All Items in U.S. City Average (CPIAUCSL). This series covers the period from January 2019 to June 2025 and we use the August 2025 vintage of the data. For the frontier forecast, we use CPI Nowcasts from the Federal Reserve Bank of Cleveland, which provide daily nowcasts of monthly year-over-year CPI inflation.

Since inflation is a monthly variable, we evaluate the accuracy of the now-

[5] The queries were submitted in batch on October 19th (for 2019 data) and 20th (for all other years), 2025. For example, our first pseudo out-of-sample query asks ChatGPT to provide forecasts of inflation for January 1st, 2019, and the next 12 months, without using information after January 1st, 2019. More specifically, the first pseudo out-of-sample query has '2019-01-01 09:00:00 EST' for 'current_date' and repeats this for every subsequent date.

[6] Neither of these examples matches the exact values for the respective reference months, but they are close in magnitude. When the out-of-sample prompts were configured, the example was intended to serve the purpose of getting and storing data in a format we prefer. However, we can not rule out that GenAI is treating this as an informative data point. Hence, we moved the example to a date before January 1st, 2019 in the pseudo out-of-sample exercise. The performance of pseudo out-of-sample does not materially change if we use the same Jan2024 example.

casts and forecasts by averaging the high-frequency forecasts within each month. We first compute monthly averages of ChatGPT-4's hourly forecasts and the Cleveland Fed's daily forecasts for each forecast origin month-year t to predict inflation for that month-year, that is, we take a simple average of $f(t, t, \tau)$ over τ.[7] These averaged forecasts are then compared with the realized inflation for month-year t using two standard evaluation metrics: mean squared error (MSE) and mean directional error (MDE).[8] To assess multi-horizon performance, we compare the monthly averages of ChatGPT-4's hourly forecasts from time t through $t + 12$ with the corresponding realized inflation values over the same horizon. [9]

2.2 Knowledge Cutoff

In traditional forecast evaluation exercises, the information set available to the model is an essential component for credible evaluation, since any evidence that models or forecasts with more information dominate those with less or inferior information content would be statements about the information sets, not necessarily about the models or their implied forecasts. Given that surveys are usually run at different frequencies and different times within a calendar year, the alignment of information sets becomes an important component of the analysis and can even change how we perceive policy and its effects (see Bauer and Swanson, 2023, among others).

Thus, understanding ChatGPT 4 Turbo's knowledge cutoff is an important characteristic to control for when evaluating pseudo out-of-sample and out-of-

[7] While the Cleveland Fed's daily forecasts often extend into time $t + 1$ before actual inflation data are released, we trim the data to include only forecasts made at time t, for consistency with ChatGPT-4. Cleveland Fed nowcasts are publically available at https://www.clevelandfed.org/indicators-and-data/inflation-nowcasting.

[8] MSE $= \frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2$, MDE $= \frac{1}{T}\sum_{t=1}^{T}\mathbf{1}\left(\text{sign}(y_t - y_{t-1}) \neq \text{sign}(\hat{y}_t - \hat{y}_{t-1})\right)$; where T is effective monthly sample size used for evaluation, t is the specific month-year, y_t is the realized value of the target variable at time t, and $\hat{y}_t$ is the aggregated predicted value for the target date t, calculated as $\hat{y}_t = \frac{1}{T_\tau}\sum_{\tau=1}^{T_\tau} f(t - h, t, \tau)$, where T_τ denotes the number of high-frequency (either daily or hourly) forecast observations available throughout a month.

[9] We have also experimented with weekly as opposed to monthly averaging of the high-frequency forecasts. While the resulting series are somewhat noisier, the results remain qualitatively the same as those reported for the monthly averages.

sample results. The challenge with GenAI models is that their exact knowledge cut-off is often difficult to determine. For instance, the GPT-4 Turbo overview provided by OpenAI Platform suggests December 1, 2023 as a knowledge cutoff date.[10] Instead, when GPT-4 Turbo was announced on November 6, 2023, it was mentioned that it has knowledge up to April 2023.[11] In the following, we treat April 2023 as the knowledge cut-off date and use it to assess whether the knowledge cut-off can be used to credibly evaluate forecast performance. As we demonstrate in Section 3, this cutoff date leads to a stark divergence between pseudo out-of-sample accuracy and actual out-of-sample performance.

3 Can ChatGPT really forecast inflation?

In this section, we show a drastic difference in ChatGPT's performance between pseudo out-of-sample and out-of-sample environments. In addition, we find deterioration in predictions after the April 2023 knowledge cut-off date.

3.1 Nowcast Performance

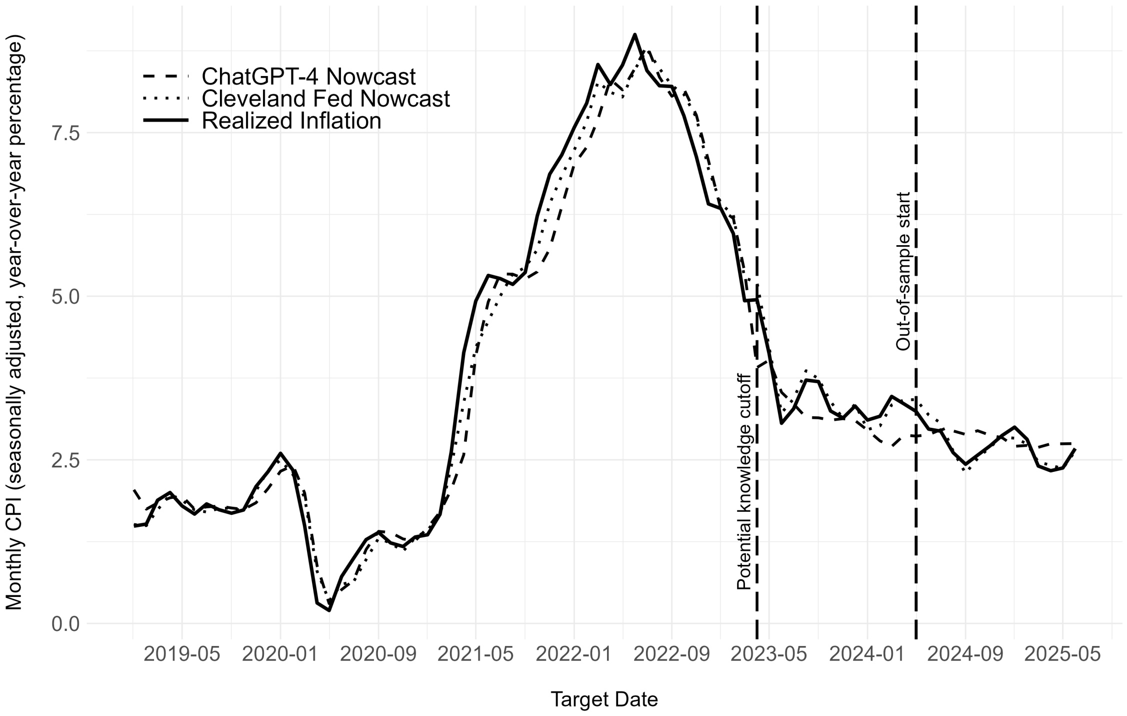
We first assess the performance of the nowcasts, i.e., predictions for the month associated with the month-year in the 'current_date' in the prompt.

Figure 1 compares the ChatGPT nowcasts (dashed line) with the realized inflation (solid line) and the Cleveland Fed's nowcast (dotted line) for forecasting the monthly year-over-year inflation rate (in percentage points) from January 2019 to June 2025. January 2019 to April 2024 nowcasts are generated using the pseudo out-of-sample query, while those from May 2024 onward are generated using out-of-sample queries. We also mark the knowledge cutoff date of April 2023 consistent with the discussion in the previous section, though there could be other cutoff dates — December 2023 is the one frequently referenced. The "Out-of-sample" vertical line marks the start of when we began cumulating hourly inflation predictions from ChatGPT.

[10] See "GPT-4 Turbo Model," OpenAI API, accessed December 13, 2025, .

[11] See "New models and developer products announced at DevDay," OpenAI Blog, November 6, 2023.

Figure 1: Nowcast Evaluation, Pseudo Out-of-Sample vs. Out-of-Sample

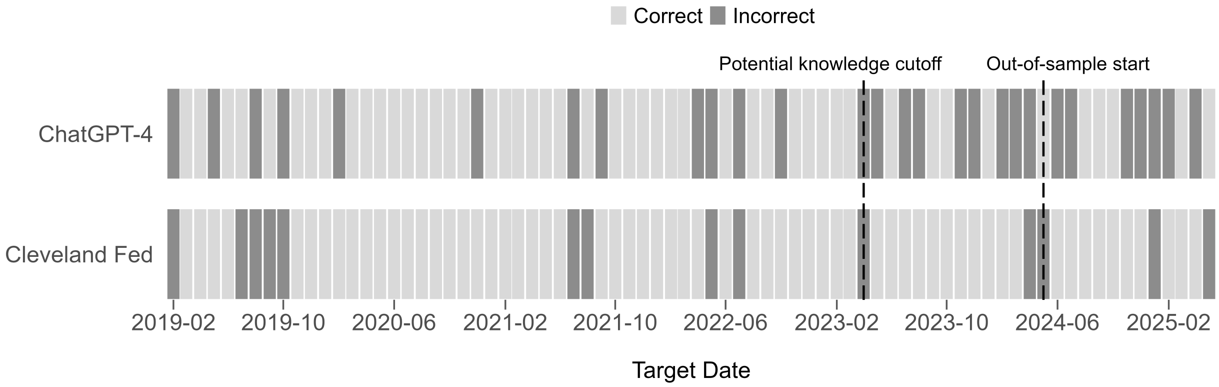


Note: Nowcasts reflect monthly averages of ChatGPT-4's hourly forecasts and the Cleveland Fed's daily forecasts. "Potential knowledge cutoff" vertical line corresponds to April 2023, and "Out-of-sample start" vertical line corresponds to May 2024. See the text on the discussion of these dates. Realized inflation is calculated with August 2025 vintage data.

Visually, ChatGPT-4 forecasts track realized very well before the knowledge cutoff date, April 2023, capturing the level and the directional changes during a period of large swings in inflation. This is similar to the findings of Faria-e-Castro and Leibovici (2024), which used Google AI's PaLM to generate pseudo out-of-sample inflation forecasts. When benchmarked against standard models such as the Cleveland Fed nowcast, ChatGPT-4 remains competitive, although with a slight lag.

In contrast, Figure 1 shows a more marked deviation of the ChatGPT forecast from the realized inflation and Cleveland Fed forecasts post April 2023. This covers the pseudo out-of-sample period between May 2023 and April 2024 and the out-of-sample period from May 2024 to June 2025. The performance of ChatGPT appears to have deteriorated even though the actual inflation rate is much more stable during the out-of-sample period than before April 2023.

Figure 2: Directional Accuracy of Nowcasts



Note: Nowcasts reflect monthly averages of ChatGPT-4's hourly forecasts and the Cleveland Fed's daily forecasts. The light shade indicates when the model correctly captures the directional change in inflation, while a dark shade indicates when it misses. "Potential knowledge cutoff" vertical line corresponds to April 2024, and "Out-of-sample start" vertical line corresponds to May 2024. See the text on the discussion of these dates. Realized inflation is calculated with August 2025 vintage data.

Figure 2 highlights the deterioration of the forecasts after April 2023 and for the out-of-sample period by showing the directional accuracy of the ChatGPT (top panel) and Cleveland Fed (bottom panel) forecasts. Directional accuracy mitigates the influence of drastically different realized inflation rates between the two periods for our analysis. For each target month-year, a light shade indicates when the model correctly captures the directional change in inflation, while a dark shade indicates when it misses.[12] Visually, it is striking that the ChatGPT panel shows a marked increase in dark shades post-April 2023. Since the Cleveland Fed nowcast does not show such a change, the deterioration in ChatGPT nowcast around the knowledge cutoff date is likely due to a decline in ChatGPT's forecast quality rather than changes in underlying inflation dynamics, which arguably would also affect the Cleveland Fed nowcast.

[12] Directional Error$_t$ $= \mathbf{1}\,(\text{sign}(y_t - y_{t-1}) \neq \text{sign}(\hat{y}_t - \hat{y}_{t-1}))$; where t is the specific time period, y_t is the observed value and $\hat{y}_t$ is the predicted value at time t, as defined in footnote 8.

Table 1: Forecast Accuracy Metrics

Metrics	Pseudo Out-of-Sample (Pre-Knowledge Cutoff)	Pseudo Out-of-Sample (Post-Knowledge Cutoff)	Out-of-Sample
$\text{MSE}_{\text{ChatGPT}}$	0.217	0.230	0.080
$\text{MSE}_{\text{CLE Fed}}$	0.098	0.018	0.011
$\frac{\text{MSE}_{\text{ChatGPT}}}{\text{MSE}_{\text{CLE Fed}}}$	**2.214**	**12.778**	**7.273**
$\text{MDE}_{\text{ChatGPT}}$	0.240	0.667	0.538
$\text{MDE}_{\text{CLE Fed}}$	0.180	0.083	0.154
$\frac{\text{MDE}_{\text{ChatGPT}}}{\text{MDE}_{\text{CLE Fed}}}$	**1.333**	**8.036**	**3.494**

MSE = Mean Squared Error; MDE = Mean Directional Error. For the relative performance metrics in rows 3 and 6 of columns 1, 2,and 3, values less than 1 indicate that ChatGPT-4 performs worse than the benchmark. 'Pre-Knowledge Cutoff' refers to a pseudo out-of-sample from January 2019 to April 2023, 'Post-Knowledge Cutoff' refers to a pseudo out-of-sample from May 2023 to April 2024, while the 'Out-of-Sample' pertains to the May 2024 to June 2025 period. See the text on the discussion of these dates.

In addition to visual inspection, Table 1 compares ChatGPT nowcasts with Cleveland Fed nowcasts using standard metrics of mean square error and mean direction error. Column (1) is for the pseudo out-of-sample before April 2023, (2) is for the pseudo out-of-sample from April 2023 to April 2024, and (3) the out-of-sample period from May 2024 to June 2025. The level of MSEs are not comparable across samples because the level of realized inflation changed. For this reason, our preferred metric is the MSE of ChatGPT forecasts relative to the Cleveland Fed's highlighted in bold. For all three periods, the relative MSE is above 1, meaning that Cleveland Fed was more accurate on average than ChatGPT. However, Cleveland Fed had a smaller advantage in the pseudo out-of-sample in the pre-knowledge cutoff period. The relative MSE in the pseudo out-of-sample before the knowledge cutoff is much smaller than that in the post-knowledge cutoff and out-of-sample (2.214 vs 12.778 or 7.273). The MDE metrics tell the same story. Cleveland Fed nowcast are more reliable at predicting the direction of change, but its advantage is smaller in the pseudo out-of-sample period before the knowledge cutoff (1.333 vs 8.036 or 3.494).

3.2 Long-Horizon Forecasts and Responsiveness

Figure 3 plots ChatGPT-4's multi-period forecasts in pseudo out-of-sample and out-of-sample with the dashed lines showing the 12-month-ahead forecasts. In panel (a), ChatGPT-4 performs reasonably well in the pseudo out-of-sample before the knowledge cutoff date. The 12-month-ahead forecasts are largely flat when inflation, depicted with the solid line, is stable at the beginning of the sample period. When inflation is below its mean, as in 2020, ChatGPT predicts it will go up, while between 2021 and the end of 2023, ChatGPT predicts it will go down, consistent with the view of mean reversion.

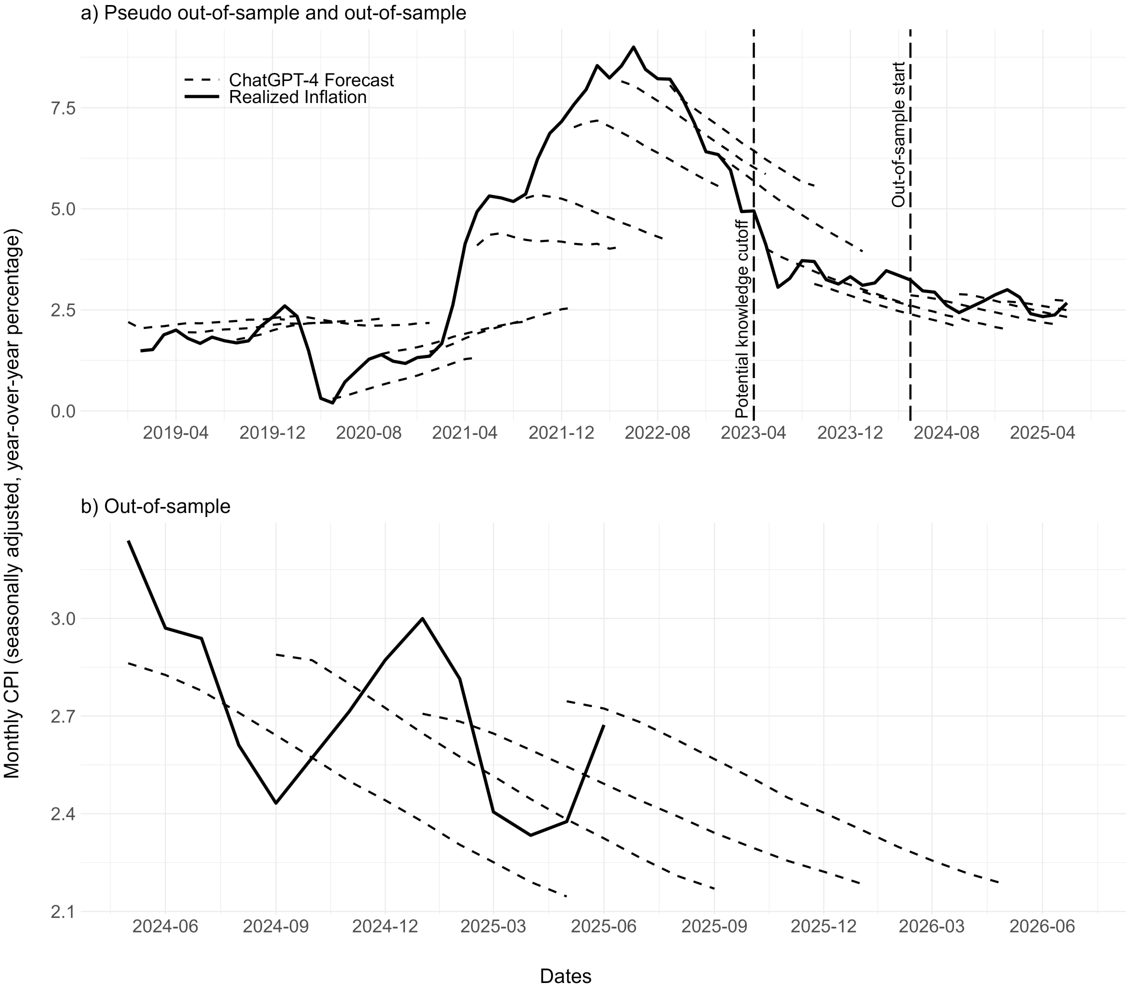
In contrast, the prediction of ChatGPT after the knowledge cutoff became rigid and unresponsive. The inflation predictions are flat which, at first glance, might be taken as an indication that ChatGPT is doing a good job of pinning down the overall trend of inflation. However, Panel (b) zooms in on the out-of-sample period and shows that the dashed lines start around the same level and are largely parallel to each other. That is, the 1 to 12-month-ahead forecasts remain essentially unchanged over time, regardless of when they are produced.

3.3 Further Discussions

What data sources are used for the forecast? As the prompt in Section 2 shows, we query ChatGPT for the information sources used to produce the forecasts. The information sources appear reasonable and along the same lines as those listed in the sample response in Section 2. However, we do not find a significant correlation between the sources cited and the forecasts. Furthermore, we have prompted ChatGPT to provide pseudo out-of-sample predictions that exclude the forecasts of various agencies. The results are somewhat inferior to those reported here, but the difference is minor. In general, it is not clear how Chat-

GPT uses the source it cites to produce the forecast. This finding highlights the black-box nature of the LLM-produced forecasts.

Figure 3: Multi-Horizon Forecasts Evaluation, Pseudo Out-of-Sample vs. Out-of-Sample



Note: Shows the monthly averages of hourly forecasts for t to $t+12$ months ahead at select forecast origin dates with four-month intervals between origin dates. "Potential knowledge cutoff" vertical line corresponds to April 2024, and "Out-of-sample start" vertical line corresponds to May 2024. See the text on the discussion of these dates. Realized inflation is calculated with August 2025 vintage data.

Retrieval augmentation (RAG) The ChatGPT-4 Turbo we used was configured to not access the internet. We tried retrieval augmentation (RAG) by providing the top 10 Bing search results of our prompt as context. The unreported results show that RAG generates more movements in the forecasts, but these movements do not necessarily improve the accuracy of the predictions.

4 Leakage

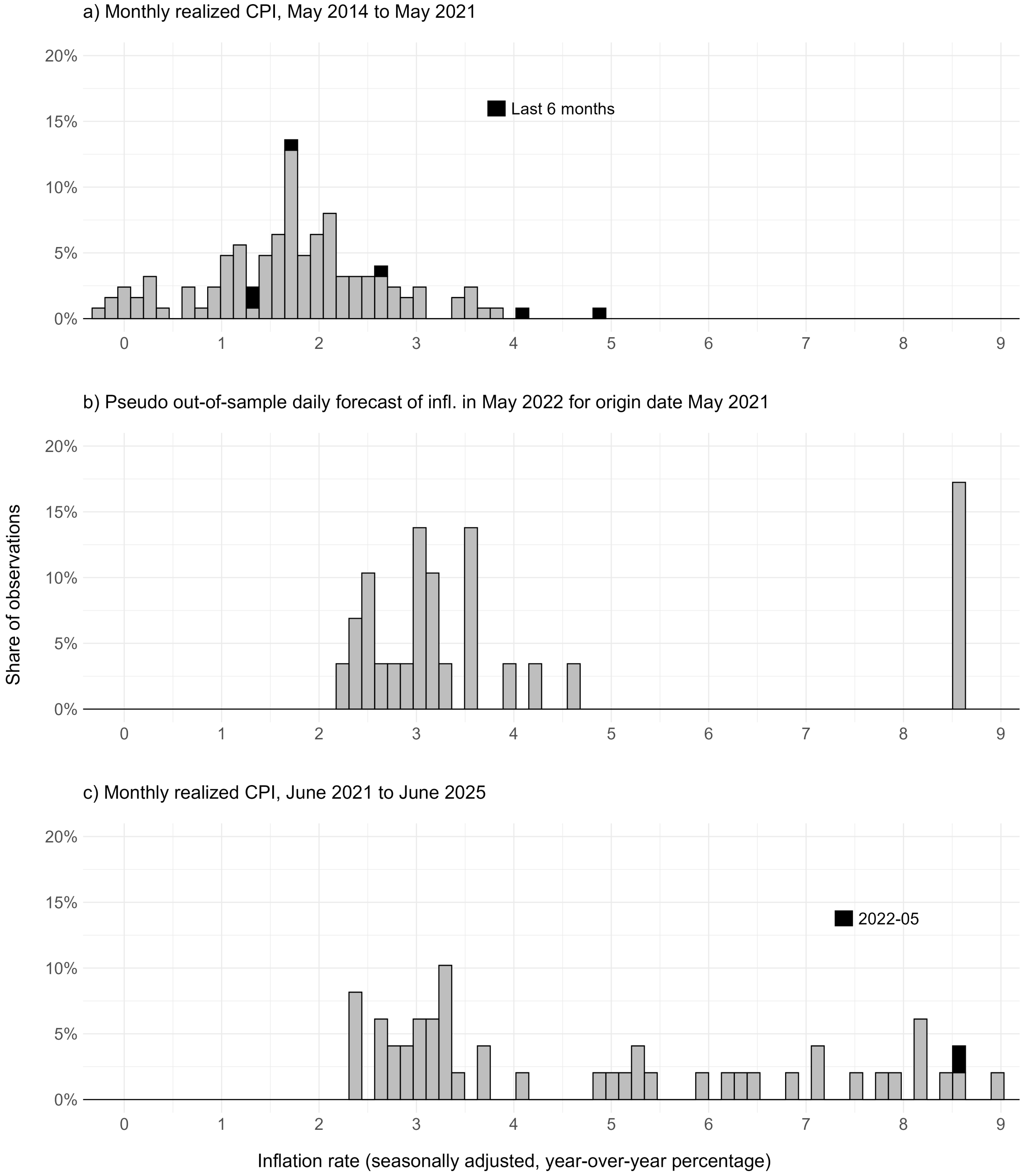
The relatively strong performance of ChatGPT-4's pseudo out-of-sample forecasts prior to the knowledge cutoff date is consistent with ChatGPT using forwarding looking information despite being asked not to. Since these forecasts were generated after the corresponding inflation data had already been released and are included in ChatGPT's training data, it is possible that ChatGPT-4 inadvertently accessed or internalized future information when forecasting.

Figure 4 provides more direct evidence of such leakage using the distribution of daily pseudo out-of-sample forecasts made for an origin date of May 2021 for the inflation rate in May 2022 ($h = 12$-steps-ahead). The top Panel (a) displays the distribution of realized inflation rates from January 2011 to May 2021, capturing realizations that should be in the information set of ChatGPT when making predictions. The black areas of the histogram separately mark the inflation rate in the 6 months leading up to May 2021. Some of the realizations over the 6 months are typical, falling near the center of the distribution, and others are in the tails. Overall, historical inflation rates never exceeded 5% at any point in the 10 years before May 2021.

The middle Panel (b) of Figure 4 shows the distribution of daily forecasts (these are not aggregated forecasts) in the pseudo out-of-sample. The forecast distribution is shifted to the right of the top panel. In particular, it includes a value of 8.6% that occurs 17.5% of the time in May 2021, which appears highly unlikely given the distribution in the top panel.

The distribution in the middle Panel (b) is closer to the distribution of the inflation rate after May 2021 plotted in the bottom panel of Figure 4. Since May 2021, realized inflation has shifted up to the 7%, 8%, and 9% range. In fact, the actual inflation rate for May 2022—the target of the 12-month-ahead forecast—was 8.53%, almost identical to the forecast 'outlier'. This alignment raises

Figure 4: Pseudo Out-of-Sample Leakage Exploration



Note: We construct pseudo out-of-sample forecasts using the real-time prompt, but generated them retrospectively, after inflation outcomes were known. Realized CPI data correspond to August 2025 vintage. Realized CPI data correspond to August 2025 vintage.

the possibility that some of the model's responses for that forecast origin month were influenced by information it should not have had access to, suggesting potential leakage.

These findings underscore the need for caution when using prompt conditioning to create pseudo out-of-sample environments and to evaluate the performance of GenAI forecasts. Our results show that a more reliable approach is to assess predictions in out-of-sample or pseudo out-of-sample periods *after* the knowledge cutoff dates, if these can be credibly established (as we have shown, this could be difficult ex ante).

5 Conclusion

In this paper, we collect out-of-sample inflation forecasts by a generic LLM to evaluate its performance. We find that the forecasts are significantly worse than those made in a pseudo out-of-sample environment created by informational conditioning of the prompt. Our findings suggest that the benchmarks used in the emerging literature may be misleading about the ability of these tools to forecast or generate survey responses for macroeconomic variables. A better alternative is to benchmark using out-of-sample forecasts or pseudo out-of-sample forecasts for dates after the knowledge cutoff, if such a cutoff can be established.

Our paper focuses on a general use LLM, yet it may have similar implications for models designed specifically for time series forecasting. Time Series Foundational Models (TSFMs) and Time Series Language Models (TSLMs), such as Google's TimesFM (Das et al., 2024), Salesforce's Moirari (Liu et al., 2024), and Amazon's Chronos (Ansari et al., 2024), have gained traction and are built on pre-trained models similar to those used in large language models (LLMs). These models are marketed as capable of zero-shot forecasting, meaning they can predict future events or trends without relying on specific prior data or examples directly related to those events. They are typically evaluated in an out-of-sample setting, where predictions are made for future time periods that do not overlap with the data used during pre-training, ensuring the model forecasts truly unseen scenarios. However, if users are unaware of the data and time periods used in the pre-training, they may inadvertently produce a pseudo out-of-sample forecast.

The rapidly developing literature has proposed methods to avoid look-ahead bias. In the context of forecasting stock returns, He et al. (2025) proposes using chronologically consistent LLMs, and Engelberg et al. (2025) proposes masking user-provided data to look-ahead bias. We leave it to future research to investigate the success of these methods for forecasting macroeconomic variables, where data is much more sparsely spread over time and space.

We cast doubt on the ability of a generic LLM to provide accurate forecasts or mimic professional forecasters in real time. However, such tools can still be helpful to households to better anchor their expectations in real-time, among others. We leave it to future studies to compare LLM forecasts with household surveys such as the Michigan Survey of Consumers. We also provide all out-of-sample GenAI prompts and respective responses to the public through the website AI Inflation Expectations so that researchers can explore other aspects of the forecasting with LLMs.

References

Ansari, Abdul Fatir, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang, "Chronos: Learning the Language of Time Series," *arXiv preprint arXiv:2403.07815*, 2024.

Bauer, Michael D. and Eric T. Swanson, "An Alternative Explanation for the "Fed Information Effect"," *American Economic Review*, 2023, *113* (3), 664–700.

Bybee, J Leland, "The ghost in the machine: Generating beliefs with large language models," *arXiv preprint arXiv:2305.02823*, 2025.

Carriero, Andrea, Davide Pettenuzzo, and Shubhranshu Shekhar, "Macroeconomic forecasting with large language models," *arXiv preprint arXiv:2407.00890*, 2024.

Clark, Todd and Michael McCracken, "Chapter 20 - Advances in Forecast Evaluation," in Graham Elliott and Allan Timmermann, eds., *Handbook of Economic Forecasting*, Vol. 2 of *Handbook of Economic Forecasting*, Elsevier, 2013, pp. 1107–1201.

Crane, Leland D., Akhil Karra, and Paul E. Soto, "Total Recall? Evaluating the Macroeconomic Knowledge of Large Language Models," *Finance and Economics Discussion Series (FEDS)*, 2025, *2025-44*.

Croushore, Dean, "Forecasting with Real-Time Macroeconomic Data," in G. Elliott, C.W.J. Granger, and A. Timmermann, eds., *Handbook of Economic Forecasting*, Vol. 1, Elsevier, 2006, pp. 961–982.

Das, Abhimanyu, Weihao Kong, Rajat Sen, and Yichen Zhou, "A decoder-only foundation model for time-series forecasting," *arXiv preprint arXiv:2310.10688*, 2024.

Dunn, Wendy, Ellen Meade, Nitish Sinha, and Raakin Kabir, “Teaching to the Test: Evaluating the Performance of Generative AI Models for Economic Analysis is Harder than you Think,” *Mimeo.*, 2025.

Engelberg, Joseph, Asaf Manela, William Mullins, and Luka Vulicevic, “Entity neutering,” *Available at SSRN*, 2025.

Faria-e-Castro, Miguel and Fernando Leibovici, “Artificial Intelligence and Inflation Forecasts,” *Federal Reserve Bank of St. Louis Review*, Fourth Quarter 2024, *106* (12), 1–14.

Hansen, Anne Lundgaard, John J Horton, Sophia Kazinnik, Daniela Puzzello, and Ali Zarifhonarvar, “Simulating the survey of professional forecasters,” *Available at SSRN*, 2025.

He, Songrun, Linying Lv, Asaf Manela, and Jimmy Wu, “Chronologically Consistent Large Language Models,” *arXiv preprint arXiv:2502.21206*, 2025.

Inoue, Atsushi and Lutz Kilian, “In-Sample or Out-of-Sample Tests of Predictability: Which One Should We Use?,” *Econometric Reviews*, 2005, *23* (4), 371–402.

Kazinnik, Sophia and Tara M Sinclair, “Fomc in silico: A multi-agent system for monetary policy decision modeling,” *Available at SSRN 5424097*, 2025.

Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani, “The Parable of Google Flu: Traps in Big Data Analysis,” *Science*, 2014, *343* (6176), 1203–1205.

Liu, Xu, Juncheng Liu, Taha Aksu, Yuxuan Liang, Roger Zimmermann, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo, “Moirai-MoE: Empowering Time Series Foundation Models with Sparse Mixture of Experts,” *arXiv preprint arXiv:2410.10469*, 2024.

Lopez-Lira, Alejandro, Yuehua Tang, and Mingyin Zhu, “The Memorization Problem: Can We Trust LLMs’ Economic Forecasts?,” *arXiv preprint arXiv:2504.14765*, 2025.

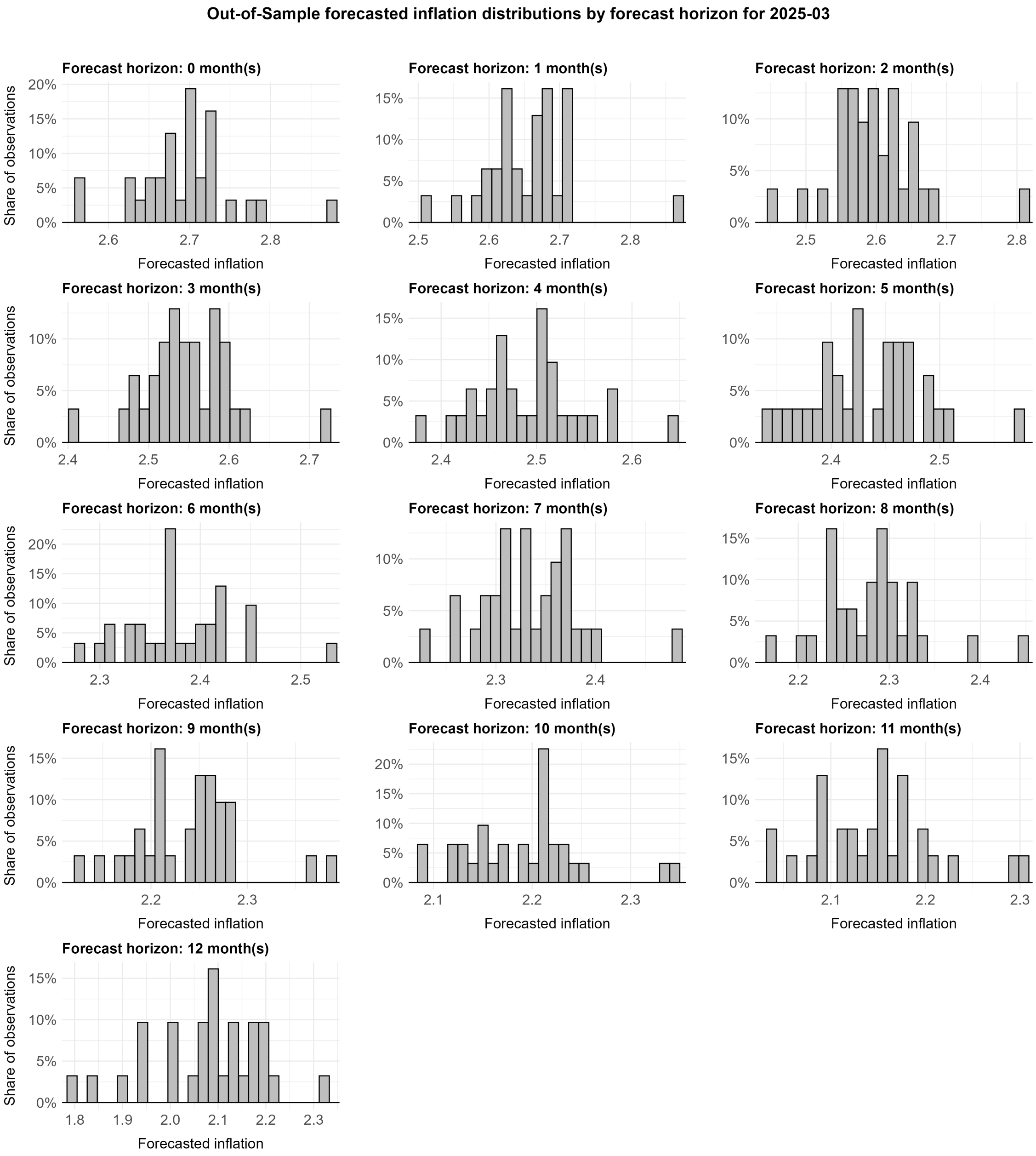
Microsoft, “Azure OpenAI Model Retirements and Replacements,” https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/model-retirements 2025. Microsoft Learn. Accessed January 16, 2026.

Wu, Jing Cynthia, Jin Xi, and Shihan Xie, “LLM Survey Framework: Coverage, Reasoning, Dynamics, Identification,” *National Bureau of Economic Research Working Paper*, 2025.

Zarifhonarvar, Ali, “Generating inflation expectations with large language models,” *Journal of Monetary Economics*, 2026, *157*, 103859.

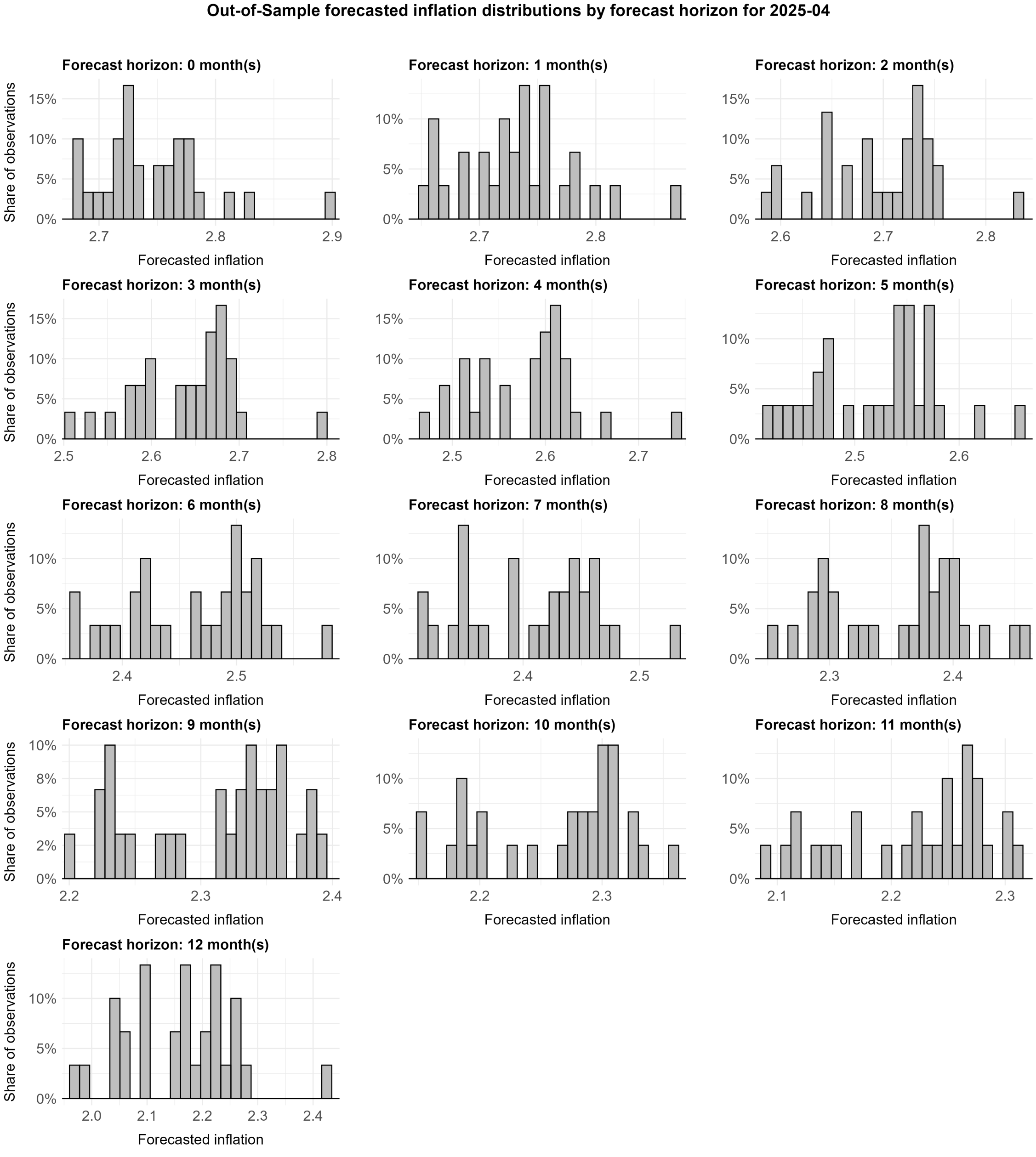
A Appendix, not for publication

Figure A1: Out-of-sample Distributions



Note: Out-of-sample forecast distribution for March 2025.

Figure A2: Out-of-sample Distributions



Note: Out-of-sample forecast distribution for April 2025.