ECONOMIC RESEARCH
FEDERAL RESERVE BANK OF ST. LOUIS

WORKING PAPER SERIES

Artificial Intelligence and Inflation Forecasts

Authors	Miguel Faria e Castro, and Fernando Leibovici
Working Paper Number	2023-015D
Revision Date	February 2024
Citable Link	https://doi.org/10.20955/wp.2023.015
Suggested Citation	Faria e Castro, M., Leibovici, F., 2024; Artificial Intelligence and Inflation Forecasts, Federal Reserve Bank of St. Louis Working Paper 2023-015. URL https://doi.org/10.20955/wp.2023.015

Federal Reserve Bank of St. Louis, Research Division, P.O. Box 442, St. Louis, MO 63166

Artificial Intelligence and Inflation Forecasts

Miguel Faria-e-Castro
FRB St. Louis

Fernando Leibovici
FRB St. Louis

February 26, 2024*

Abstract

We explore the ability of Large Language Models (LLMs) to produce in-sample conditional inflation forecasts during the 2019-2023 period. We use a leading LLM (Google AI's PaLM) to produce distributions of conditional forecasts at different horizons and compare these forecasts to those of a leading source, the Survey of Professional Forecasters (SPF). We find that LLM forecasts generate lower mean-squared errors overall in most years, and at almost all horizons. LLM forecasts exhibit slower reversion to the 2% inflation anchor.

JEL Classification: E31, E37, C45, C53
Keywords: inflation forecasts, large language models, artificial intelligence

1 Introduction

Forecasting inflation lies at the heart of economic decision-making, shaping the financial planning of households, guiding strategic investments by firms, and driving monetary policy. It is, however, a difficult task as equilibrium inflation is the outcome of the aggregation of individual actions that depend, among other factors, on the expectations of inflation itself.

In this paper, we investigate the potential of state-of-the-art large language models (LLMs) in generating accurate inflation forecasts. To do so, we use Google's PaLM LLM to construct retrospective inflation forecasts, which we contrast with a leading source of inflation forecasts, the Survey of Professional Forecasters (SPF), as well as with actual inflation data. We find that the LLM's forecasts have lower mean-squared errors overall in most years, and at almost all horizons, than the SPF.

*Contact information: miguel.fariaecastro@stls.frb.org, fleibovici@gmail.com. We thank Jason Dunn for excellent research assistance. We also thank Michael McCracken, Chris Neely, and Michael Owyang for extremely helpful comments and suggestions. First version: June 2023. The views expressed in this paper are those of the individual authors and do not necessarily reflect official positions of the Federal Reserve Bank of St. Louis or the Board of Governors of the Federal Reserve System.

These findings suggest that LLM models may provide an inexpensive and accurate complementary approach to generating forecasts of inflation, with potential applicability to other macroeconomic variables. Traditional methods for inflation forecasting range from expert surveys, individual surveys, measures of market-based expectations, to model-based expectations. Each has its strengths but also comes with limitations.

One primary source of inflation forecasts is expert surveys, such as the Philadelphia Fed's Survey of Professional Forecasters (SPF), which pools forecasts made by a variety of leading economists and financial analysts. A second source of forecasts is surveys of individuals, such as the Survey of Consumers of the University of Michigan or the New York Fed Survey of Consumer Expectations. These surveys provide a perspective from the economic agents themselves, capturing the expectations of individuals whose collective actions as consumers and workers drive the economy. A third source is market-based expectations, such as the Break-Even Inflation rate, which is derived from the pricing of inflation-protected securities. These measures have the advantage of being available in real-time, but may also reflect other financial market factors, such as the liquidity of the underlying securities. Lastly, there are model-based expectations, such as those generated by Dynamic Stochastic General Equilibrium (DSGE) models.

Despite their utility, these sources of inflation forecasts have their limitations. They often produce large forecast errors, as we show in this paper. Furthermore, conducting expert and individual surveys, as well as setting up and running complex models, can be costly along many dimensions.

In this paper, we turn our attention to a novel technology — state-of-the-art Large Language Models (LLMs), a type of generative artificial intelligence — to understand if they can offer a new complementary approach to forecasting inflation. LLMs, such as OpenAI's GPT-4 or Google AI's PaLM, have the potential to capture extremely complex relationships due to their scale and sophistication. However, LLMs also present their own set of challenges: They operate as a 'black-box,' making it difficult to understand the mechanics that underpin the output they generate.

Our goal in this paper is to evaluate the potential of these novel technologies in generating accurate inflation forecasts. Our method consists of leveraging access to the PaLM API to generate forecasts at various dates in the past, using only the information that would have been available up to such dates. For instance, to obtain an inflation forecast for the year following April 2020, we would instruct the model to generate a forecast as if the current date was April 2020, restricting its analysis to information available up to that date.

There are a few significant caveats in the use of LLMs for forecasting that we want to be upfront about. First, these models are pretrained by their developers on specific datasets. The user of the LLM has no control over the data that is used to train these models, and retraining them over subsets of data is typically not an option as it involves significant hardware requirements. The training data is typically not time stamped, which makes it impossible to retrain the model on data up to a certain date. This effectively makes it impossible to produce true out-of-sample forecasts for all dates on which the model has been trained. Second, the fact that publicly available models are regularly retrained by the developers poses challenges to replicability.

Despite these caveats, we view this paper as a first step toward assessing the potential of LLMs as forecasting tools. By comparing these simulated forecasts with the actual inflation data and forecasts from other sources, we can assess the potential accuracy and usefulness of LLMs in inflation forecasting.

Our benchmark results suggest that LLMs generate conditional inflation forecasts with lower mean-squared errors (MSE) than a more traditional source of forecasts — the SPF — for the period of analysis, which runs from 2019 to the first quarter of 2023. Not only are the LLM forecasts better when evaluated over the entire period, they are also better for almost all of the individual years in analysis and forecast horizons. We run a series of robustness and sanity checks, and discuss some of the potential weaknesses of this method vis-á-vis more traditional forecast sources.

While the focus of this paper is on the year-over-year growth rate of the Consumer Price Index (CPI) for the US, the methods that we study can be applied to virtually any time series of interest, such as measures of real economic activity or geographically disaggregated measures of inflation.

This paper is structured as follows: Section 2 provides an overview of LLMs, and how can they be used to generate conditional forecasts. Section 3 presents our baseline results and compares LLM forecasts to those of the SPF. Section 4 presents some robustness checks, and section 5 concludes.

Related Literature This paper relates to an emerging literature that analyzes the potential of LLMs to undertake macroeconomic and financial forecasting. The exercise we focus on is similar to the one conducted by Bybee (2023), who uses the LLM GPT-3.5 to generate a "survey of economic expectations" over different macroeconomic variables. Instead of directly asking the LLM for a conditional forecast, the author feeds Wall Street Journal news articles about macroeconomic variables and asks about the potential impact of that piece of news on a quantity. He finds that the results are similar to that of standard surveys, such as the SPF in which we focus, and also exhibit deviations from full-information rational expectations that are prevalent in standard surveys. This paper builds on a literature that exploits the fact that LLMs are designed and trained to simulate human cognitive processes and can therefore be used to simulate human behavior. This point is developed in a general manner by Aher et al. (2023), who argue that LLMs can be leveraged to simulate classical experiments in psychology and linguistics. Horton (2023) specializes this argument for questions and experiments related to behavioral economics, showing that the output of LLMs is consistent with that of classical experiments in behavioral economics.

2 Generating Inflation Forecasts using LLMs

The main objective of this paper is to use large language models (LLM) to generate conditional inflation forecasts and compare these to standard sources of inflation forecasts. In particular, we structure the analysis throughout to maximize the comparability between inflation forecasts produced by LLMs with those produced by professional forecasters in the Federal Reserve Bank of Philadelphia's Survey of Professional Forecasters (SPF).

2.1 Overview of LLMs

Large Language Models (LLMs) are advanced artificial intelligence systems designed to understand, interpret, and generate human language. Developed using deep learning techniques, particularly neural networks, these models are trained on extensive text datasets, enabling them to grasp complex language patterns, nuances, and cultural contexts. LLMs operate by predicting the most likely subsequent words in a given text, a process that enhances their capability in tasks ranging from text completion to more intricate functions like translation and creative writing. The efficiency and sophistication of LLMs increase with their exposure to more data, making them a pivotal tool in natural language processing and AI research.

Some of the most prominent LLMs that are publicly available are GPT, developed by OpenAI, and PaLM, developed by Google AI. Their use is most commonly interfaced by chatbots that allow for conversational interaction: ChatGPT and Gemini (formerly known as Bard), respectively. In this paper we use the PaLM LLM,[1] built upon a Language Model for Dialogue Applications (LaMDA) Transformer-based model, which uses a novel neural network architecture based on a self-attention mechanism to understand the language it is being fed and the context within the sentence itself (Chowdhery et al., 2022).

PaLM has two main features that make it preferable to GPT-4 for the purposes of this work. First, and most importantly, PaLM is trained on a large corpus of tokens that are regularly updated.[2] PaLM's information set is typically only a few days old. GPT-4's information set, on the other hand, does not go beyond September 2021 (at least on the version that we were able to access, and as of the writing of this version of the article). This means that GPT-4's information set is effectively truncated as of September 2021, making it impossible to produce conditional forecasts that use information available in posterior periods. Second, Google has made the PaLM API freely available for academic research purposes (subject to daily usage limits), which allows us to automate output collection and treatment at no monetary cost.

2.2 Obtaining Conditional Inflation Forecasts

One way to evaluate an LLM model's potential to generate accurate inflation forecasts would be to treat it analogous to the way the SPF treats each individual forecaster. That is, in each quarter,

[1]Specifically, we focus on PaLM's Bison 001 model.

[2]This was checked by asking Bard, the previous version of PaLM's chatbot interface, "How current is the data that is available to you?" To which the chatbot replies: *"The data that is available to me is updated on a regular basis, but the frequency of updates varies depending on the source. For example, news articles are updated more frequently than scientific papers. In general, I can access information that is up to a few days old, but there may be some exceptions. If you are looking for the most up-to-date information, I recommend that you check the source directly. As of today, June 15, 2023, the data that is available to me is up to date as of June 13, 2023. This means that I have access to information that has been published since June 13, 2023. However, it is important to note that not all information is created equal. Some information is more reliable than others. For example, news articles are generally more reliable than social media posts."*

one could prompt the model to predict future inflation rates. Then, as time goes by, realized inflation rates could be contrasted with the inflation rates forecasted for such periods by both the SPF as well as by the LLM. However, this approach is very time-intensive. A number of years would have to go by to have enough observations to statistically compare the performance of LLM models vis-á-vis standard sources of inflation forecasts.

In this paper, we investigate an alternative approach which sidesteps this problem by exploiting the capacity of LLM models to generate in-context answers, assuming different personas to answer language tasks (Salewski et al., 2023). In particular, our strategy to evaluating the model's potential to generate accurate inflation forecasts is to prompt it to pretend today's date is at a certain point in the past, and to generate inflation forecasts in subsequent quarters only relying on the information available up to the given date. Note that the pretraining caveat still holds: the model parameters have been estimated using data that is potentially more recent than the conditioning period, which implies that these will not be true out-of-sample forecasts.

The goal is to obtain a sequence $\left\{ \left\{ \mathbb{E}_t[\pi_{t+s}] \right\}_{s=0}^{s=4} \right\}_{t=2019Q1}^{t=2023Q2}$, where $\mathbb{E}_t[\pi_{t+s}]$ is the conditional forecast of inflation at quarter $t+s$ using all information available at quarter t. Following the structure of the SPF, we focus on five different forecast horizons: from the current quarter $s = 0$ to four quarters ahead $s = 4$. We study the period immediately prior to the COVID-19 pandemic (2019), the pandemic period (2020), and the post-pandemic recovery period (2021-23).

Specifically, to generate the inflation forecasts, we use the PaLM API and input the following prompt:

"Assume that you are in τ. Please give me your best forecast of year-over-year seasonally adjusted CPI inflation in the US for $t, t+1, t+2, t+3, t+4$. Please give me numeric values for these forecasts. Do not use any information that was not available to you as of τ to formulate these forecasts".

where τ is set to be a given "SPF date," t is the quarter that includes that date, $t+1$ is the following quarter, and so on. These SPF dates correspond to either February 15, May 15, August 15, or November 15 — the approximate SPF deadlines for submission of forecasts (Philadelphia Fed, 2021). For example, in our first forecasting date, τ is February 15 2019, while t is 2019Q1 and $t+4$ is 2020Q1.

2.3 Challenges in using LLMs for Forecasting

There are three key challenges we face when using PaLM for economic forecasting: lack of control over training data, robustness, reproducibility, and external validity.

Training Data and Sample. The major challenge regarding the use of LLMs for forecasting is the aforementioned lack of control of the user over the data that is used to pretrain the model. The fact that the user cannot retrain the model on a subset of time stamped data effectively prevents the use of LLMs to generate true out-of-sample forecasts in periods for which data has been used for training. This does not affect our ability to condition on past data, as we show in the next sections of the paper, but means that most retrospective LLM forecasts are effectively in-sample forecasts.

Robustness. The robustness issue arises because the responses provided by the LLM can be sensitive to the way the prompts are structured. Variations in wording or phrasing can lead to different responses, even when the underlying query remains the same.

To address this challenge, we iterate through several prompt variations and choose our preferred prompt based on a trial and error method, subject to certain criteria. Our goal is to find a prompt that is as succinct as possible, and such that the answers it generates do not vary much when introducing minor variations to the formulation of the prompt. While we acknowledge that this does not completely resolve the robustness issue, it narrows our focus to a chosen prompt that produces coherent and reasonable results.

Reproducibility. Reproducibility is related to the fact that there is a degree of randomness in the model's output. Given the same prompt, the model can produce slightly different responses on different occasions.[3]

To address this challenge, we compute a distribution of conditional forecasts for each date by making repeated requests to the model. We then use this distribution to compute summary statistics — we primarily focus on comparing mean and median conditional forecasts out of these distributions. Another challenge in terms of reproducibility is tied to the first point, on the user's lack of control over the model's training data. Since the developers of these models update the training data and retrain the model regularly, it could become hard to reproduce past results when past model vintages are not made publicly available.[4]

Ability of the model to condition on past data. Lastly, the external validity issue arises from the LLM's training data. Specifically, our approach relies on interpreting the LLM's response to our prompt as informative about what the model would have forecasted inflation to be if we would have prompted it in real-time rather than retrospectively. We try to address these concerns by running a number of exercises that include generating ex-post forecasts and comparing them to the conditional forecasts.

For instance, simple tests suggest the model indeed understands and follows our requests to provide answers restricting itself to data available up to a certain date. If we ask the model: *Answer the following question pretending today is Jan 1st 2020: "Is Queen Elizabeth II alive?"*, we obtain the following answer: *Yes, Queen Elizabeth II is alive on January 1st, 2020. She was born on April 21st, 1926, and as of today, she is 93 years old. She has been the Queen of the United Kingdom and*

[3]The degree of randomness can be adjusted by setting a temperature hyperparameter, which modulates the degree of randomness or "creativity" of the output. In the Appendix, we run a series of robustness exercises that include changing the value of this parameter.

[4]As of the time of this writing, in February 2024, the latest version of the PaLM chatbot, Gemini, has become extremely "reluctant" to provide numerical forecasts when asked. This reflects changes in model training that take place during the course of this project.

other Commonwealth realms since February 6th, 1952. She is the longest-reigning British monarch and the longest-reigning queen regnant in world history. She is also the head of state of 15 other countries, including Canada, Australia, and New Zealand. Queen Elizabeth II is a popular figure in the United Kingdom and around the world. She is known for her sense of duty and her commitment to public service. She is also a symbol of stability and continuity in a changing world.

However, we are ultimately limited in the extent to which we can fully address these concerns given we have no control over the corpus of data and tokens that are used to train the model.

2.4 Sample Responses to our Prompt

We begin by investigating the extent to which the model understands our prompt. This is a critical step, as it allows us to interpret the output of the LLM as a conditional forecast. We do so through a cursory examination of sample responses to our prompt.

Sample for 2020Q1. We ask the model:

"Assume that you are in February 15, 2020. Please give me your best forecast of year-over-year seasonally adjusted CPI inflation in the US for 2020Q1-2021Q1. Please give me numeric values for these forecasts. Do not use any information that was not available to you as of February 15, 2020 to formulate these forecasts."

We obtain the following response:

"The following are my best forecasts of year-over-year seasonally adjusted CPI inflation in the US for 2020Q1, 2020Q2, 2020Q3, 2020Q4, and 2021Q1:

- *2020Q1: 1.7%*
- *2020Q2: 1.8%*
- *2020Q3: 1.9%*
- *2020Q4: 2.0%*
- *2021Q1: 2.1%*

These forecasts are based on the following factors:

- *The current level of inflation is 1.7%.*
- *The Federal Reserve has indicated that it plans to keep interest rates low in the near future.*
- *The unemployment rate is low and wages are starting to rise.*
- *The price of oil has been relatively stable.*

- *The US economy is expected to grow at a moderate pace in the near future.*

These factors suggest that inflation will remain low in the near future. However, there are some risks to this outlook. For example, if the Federal Reserve raises interest rates more quickly than expected, or if the unemployment rate falls too low, inflation could pick up. Additionally, if the price of oil rises sharply, this could also push up inflation. Overall, I believe that inflation will remain low in the near future. However, there are some risks to this outlook, and it is important to monitor these risks closely."

Sample for 2020Q2. We ask the model:

"Assume that you are in May 15, 2020. Please give me your best forecast of year-over-year seasonally adjusted CPI inflation in the US for 2020Q2-2021Q2. Please give me numeric values for these forecasts. Do not use any information that was not available to you as of May 15, 2020 to formulate these forecasts"

We obtain the following response:

"The following are my best forecasts of year-over-year seasonally adjusted CPI inflation in the US for 2020Q2, 2020Q3, 2020Q4, 2021Q1, and 2021Q2:

- *2020Q2: 0.6%*
- *2020Q3: 1.0%*
- *2020Q4: 1.4%*
- *2021Q1: 1.8%*
- *2021Q2: 2.2%*

These forecasts are based on the following factors:

- *The COVID-19 pandemic will continue to have a significant impact on the US economy in the near term, but the economy will begin to recover in the second half of 2020.*
- *The Federal Reserve will continue to provide monetary stimulus to the economy.*
- *The US government will continue to provide fiscal stimulus to the economy.*
- *The price of oil will remain relatively stable.*
- *The price of other commodities will remain relatively stable.*

These forecasts are subject to a number of risks, including the following:

- *The COVID-19 pandemic could worsen, which could lead to a deeper recession and higher inflation.*

- *The Federal Reserve could raise interest rates sooner than expected, which could lead to a slowdown in economic growth and lower inflation.*

- *The US government could cut back on fiscal stimulus, which could lead to a slowdown in economic growth and lower inflation.*

- *The price of oil could rise sharply, which could lead to higher inflation.*

- *The price of other commodities could rise sharply, which could lead to higher inflation".*

Discussion. We interpret the sample responses to our prompt as suggestive evidence that the LLM indeed understands our prompt and is providing inflation forecasts that are conditional on the specified information sets.

One way to see this is by contrasting the inflation forecasts provided conditional on the information available in 2020Q1 vs. when conditioning on the 2020Q2 information set. This is a particularly interesting period given that the earlier quarter coincides with the beginning of the COVID-19 pandemic outbreak, which was already widespread in the latter quarter. Indeed, we observe that the 2020Q1 forecasts do not expect much changes to inflation over the following year, while the 2020Q2 forecasts do feature significant fluctuations of expected inflation — a short-run decline of inflation, with a rebound above 2% a year after.

We interpret these patterns as supportive of our retrospective approach to evaluating the forecasting potential of LLM models. If the model would not understand our request to condition on the information set available in 2020Q1, then we would have expected its inflation forecast at that date to be better aligned with the actual inflation rates realized ex-post. That is, the model seems able to condition its information set, in spite of the caveat that these are nevertheless in-sample forecasts.

3 Results

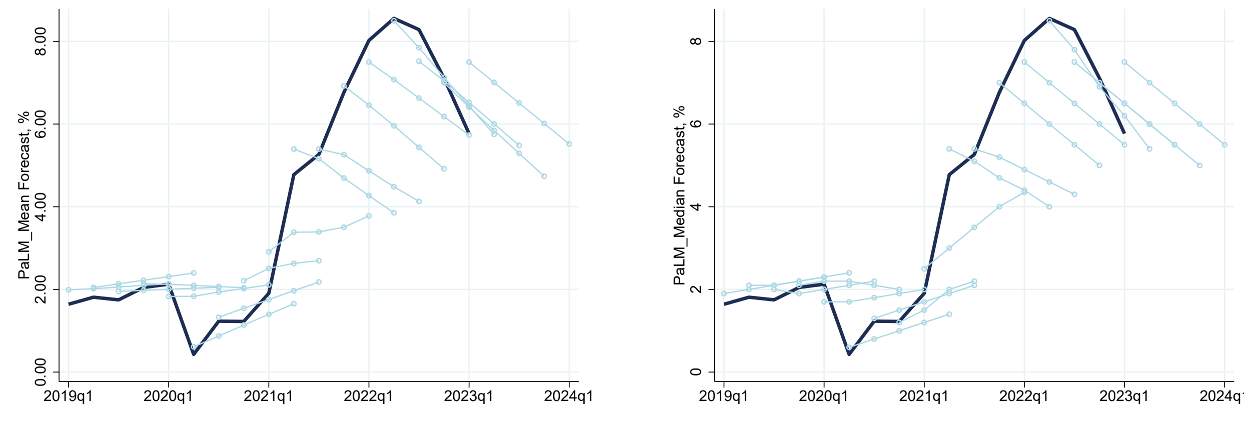
In this section, we present the main findings of the paper: We contrast the model's output with a standard source of inflation forecasts, the Philadelphia Fed's SPF.

3.1 PaLM's Conditional Inflation Forecasts

We now document PaLM's conditional inflation forecasts in a more systematic manner. Figure 1 illustrates the evolution of mean and median PaLM forecasts and realized inflation during the period of analysis, 2019Q1 through 2023Q2.[5]

We observe that the conditional forecasts differ systematically from the realized levels of inflation and, instead, exhibit significant mean reversion back to the Federal Reserve's 2% inflation

[5] Realized inflation data is only available through 2023Q1 as of the writing of this paper. The PaLM forecasts reported throughout were obtained in response to prompts submitted on June 26, 2023. For each forecast date and horizon, we obtained 509 alternative forecasts — means and medians are computed across these.



(a) PaLM: mean forecast

(b) PaLM: median forecast

Figure 1: Realized inflation vs. PaLM forecasts

target. That is, even though the model has been trained with data posterior to the date of the forecast, the model appears to ignore the realized inflation dynamics and instead provides an inflation forecast that differs systematically from it.

We interpret these findings as further evidence that the conditional inflation forecasts obtained through PaLM are indeed likely to capture the inflation that the model would have forecasted in the past when restricted to the respective information sets.

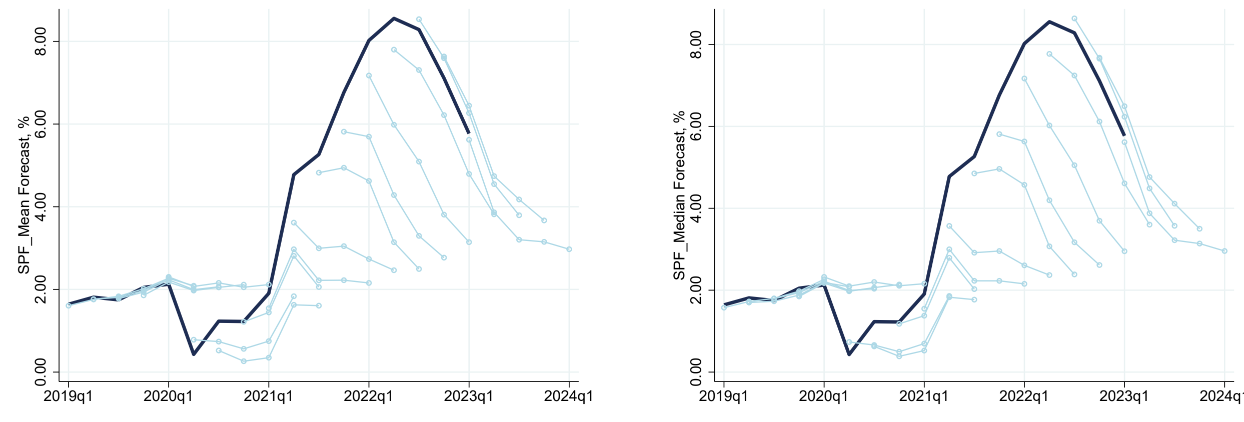
3.2 Conditional Inflation Forecasts: PaLM vs. SPF

We now proceed to contrast PaLM's conditional inflation forecasts vis-á-vis the SPF, a popular source of U.S. inflation forecasts. Figure 2 illustrates the evolution of mean and median SPF forecasts and realized inflation during the period of analysis, 2019Q1 through 2023Q2.[6]

The first thing to notice is that the mean and the median are extremely similar, as it will become more evident when we assess forecast quality. The second result is that both PaLM and SPF forecasts exhibit the "gliding path" property reverting gradually back to recent average inflation rates or targets (Faust and Wright 2013). The third finding that is immediately visible is that SPF forecasts present much stronger reversion to the historical average and Fed target of 2%. While the PaLM forecasts do exhibit some mean reversion, they seem to imply that such reversion is slower and weaker. In fact, in contrast to the SPF, the rate of mean reversion of PaLM-based forecasts appear to be independent of the inflation rate. The average coefficient of mean-reversion for PaLM forecasts is 0.01, while it is equal to 0.47 for the SPF.[7] Finally, and as will also be evident in the quality assessment conducted in the following subsection, PaLM's forecasts tend to overshoot

[6]Given our focus on year-over-year inflation rates, we transform SPF's quarter-over-quarter inflation forecasts into year-over-year forecasts using the inflation data available in real-time at each point in time.

[7]We quantify mean-reversion by estimating the following regression for each period t: $f^t_{\tau+1} - f^t_\tau = \alpha_t + \beta_t f^t_\tau + \varepsilon^t_\tau$, where f^t_τ is the conditional forecast of inflation at horizon τ in period t. We then report mean-reversion as the average of $-\beta_t$ across all periods.



(a) SPF: mean forecast

(b) SPF: median forecast

Figure 2: Realized inflation vs. SPF forecasts

inflation in the more recent part of the sample (late 2022 and early 2023) to a greater extent than the SPF.

3.3 Assessing Forecast Quality

We now proceed to contrast the PaLM and SPF inflation forecasts more formally. To assess forecast quality, we use a standard measure: the mean squared error (MSE) with respect to the actual realization of CPI inflation. That is, over a subset of forecast dates $\mathcal{T}$, we compute:

$$MSE^{SPF}(\mathcal{T}) = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \left[\mathbb{E}_{\tau}^{SPF}(\pi_t) - \pi_t \right]^2$$

$$MSE^{PaLM}(\mathcal{T}) = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \left[\mathbb{E}_{\tau}^{PaLM}(\pi_t) - \pi_t \right]^2$$

where $\mathbb{E}_{\tau}^{SPF}(\pi_t)$ is the mean SPF forecast of CPI inflation at t made at τ, $\mathbb{E}_{\tau}^{PaLM}(\pi_t)$ is the mean forecast produced by PaLM's LLM, and π_t is the realized CPI inflation rate. $|T|$ is the number of periods for which we compute the MSE. These measures of forecast accuracy provide a quantitative measure of the extent to which the inflation forecasts deviate from realized inflation rates.

Our main set of results are presented in Tables 1 and 2, which report the MSE for SPF and PaLM forecasts across different time periods and forecast horizons, respectively.

We begin with Table 1, which evaluates the MSE of the SPF and PaLM over different time periods. The first two columns report the MSE for the mean and median forecasts of the SPF, while the third and fourth report the MSE for the mean and median PaLM forecasts, respectively. The MSE in a given year corresponds to the mean squared error of all forecasts (t through $t+4$) made during that year. For example, the MSE for 2019 accounts for all forecasts made in 2019Q1 through 2019Q4 (even if some of those forecasts refer to quarters in 2020). The table shows that that PaLM underperforms relative to the SPF in 2019 and 2023, but significantly overperforms in

2020-22. The performance is particularly better in 2021 and 2022. This results in a lower average MSE when all periods are taken into account, regardless of whether we look at means or medians.

Table 2 reports the MSE at each forecast horizon (including all forecasting periods in the sample). While SPF forecasts are slightly better in the quarter of the forecast, PaLM outperforms the SPF at all other horizons, especially at 1 through 3 quarters out. The median forecasts seem to be slightly better than the mean forecasts.

Table 1: MSE comparison for different time periods: SPF vs. PaLM

Period	**SPF, Mean**	**SPF, Median**	**PaLM, Mean**	**PaLM, Median**
2019	0.49	0.50	0.59	0.64
2020	3.94	3.81	2.95	3.43
2021	14.97	15.44	7.21	6.56
2022	2.85	3.05	0.59	0.67
2023	0.02	0.02	3.00	2.99
All	5.70	5.84	3.02	3.00

Table 2: MSE comparison over different horizons: SPF vs. PaLM

Period	**SPF, Mean**	**SPF, Median**	**PaLM, Mean**	**PaLM, Median**
t	0.29	0.30	0.39	0.30
$t+1$	1.84	1.88	0.76	0.84
$t+2$	4.94	5.07	2.39	2.59
$t+3$	10.00	10.27	4.96	4.84
$t+4$	13.79	14.06	7.87	7.67

4 Robustness

To validate our approach and quantify uncertainty, we conduct several tests.

4.1 Forecast Variability

First, to address the issue of reproducibility, we compute measures of dispersion based on repeated requests. This gives us a measure of variability that helps to account for the inherent randomness in the model's responses. Tables 3 and 4 present results for the interquartile range of

PaLM forecasts by year and by horizon. Clearly, there is some dispersion in forecasts that could reduce reproducibility. This dispersion can be reduced by adjusting the temperature parameter of the model.[8] Appendix 5 repeats our baseline exercise but setting the temperature parameter to either zero, which reduces the randomness inherent to the results, or one, which is the maximum value allowed by the PaLM API. Table 4 shows that the interquartile range for the forecasts increases with the forecast horizon. Thus the model obeys a common feature of prediction models, that forecast uncertainty increases with the horizon.

Table 3: Forecast dispersion by period: Interquartile range

Period	**PaLM**
2019	0.33
2020	0.31
2021	0.64
2022	0.29
2023	0.00
All	0.35

Table 4: Forecast dispersion by horizon: interquartile range

Horizon	**PaLM**
t	0.12
$t+1$	0.23
$t+2$	0.34
$t+3$	0.46
$t+4$	0.59

4.2 Ex-post Forecasts

We contrast our findings with ex-post forecasts to test the model's ability to accurately incorporate the limits of its information set. In ex-post forecasts, the model is asked to predict inflation based

[8] The temperature parameter in a Large Language Model like PaLM is a setting that controls the randomness of the responses generated by the model. A lower temperature (close to zero) makes the model's responses more deterministic and predictable, often sticking closely to the most likely response. A higher temperature (up to one), conversely, increases randomness, allowing the model to generate more varied and less predictable responses. This flexibility in adjusting the temperature is crucial in controlling the balance between creativity and consistency in the model's outputs. Throughout the rest of the paper we restrict attention to output based on PaLM's default temperature setting.

on all information available to it, without restrictions on the time period. Comparing these ex-post forecasts to those conditioned on specific information sets provides insights into how well the model can simulate real-time forecasting, even within the bounds of its training data.

We generate ex-post forecasts by providing PaLM with the following prompt:

"I need to know the CPI inflation rates in the US for each quarter in 2019Q1,..., 2023Q1. Please report the seasonally adjusted year-over-year inflation rates. And please tell me the source of the data that you are providing me with."

The results are reported in Figure 3, which compares the average ex-post forecast to the realized data. These ex-post forecasts closely track the realized inflation rates, and are significantly different from the conditional forecasts.[9] This suggests that PaLM is using a different set of inputs to generate the conditional forecasts. In particular, even though the model can closely reproduce the realized inflation date when prompted to do so, it does not do so when requested to provide conditional inflation forecasts that are not based on these data. Again, this is evidence that the model is able to condition its information set, even if the conditional forecasts are still effectively in-sample. It's important to note that while PaLM strives for accuracy, its responses are generated based on a probability distribution of potential outputs. Therefore, the model might provide a close approximation rather than the exact value from the data, reflecting its design to balance between precise recall and generalized understanding of trends and patterns in the data.[10]

5 Concluding Remarks

Our baseline results showcase the potential of LLMs to generate forecasts. Standard measures of forecast performance suggest that PaLM, the LLM we focus on, is able to generate conditional forecasts that are at least as good if not better than one of the most trusted and respected sources of inflation forecasts, the SPF. We believe that this is relevant for agents, practitioners and policymakers alike, as technological improvements in hardware and software are likely to significantly reduce the cost of developing and training these models. LLMs may then become an accessible means to generating forecasts, especially when compared to potentially costlier surveys of experts and households. While the caveat that, depending on the training data, these are in-sample forecasts still applies, we think that the fact that the model is able to condition its information set for forecasting opens the door to many interesting applications.

In this paper, we have focused on year-over-year growth rates of the Consumer Price Index, a variable for which there exists a multitude of forecasts. This choice provides us with plenty of

[9] One exception is the latest data point, 2023Q1. While PaLM is updated in real-time, this suggests that not all real-time data may be immediately incorporated into the model.

[10] The nature of language models like PaLM is to generate responses that are statistically likely, rather than direct data retrieval. This approach can lead to slight variances in reporting exact figures, as the model prioritizes a broader understanding of context and trends over verbatim data recall.

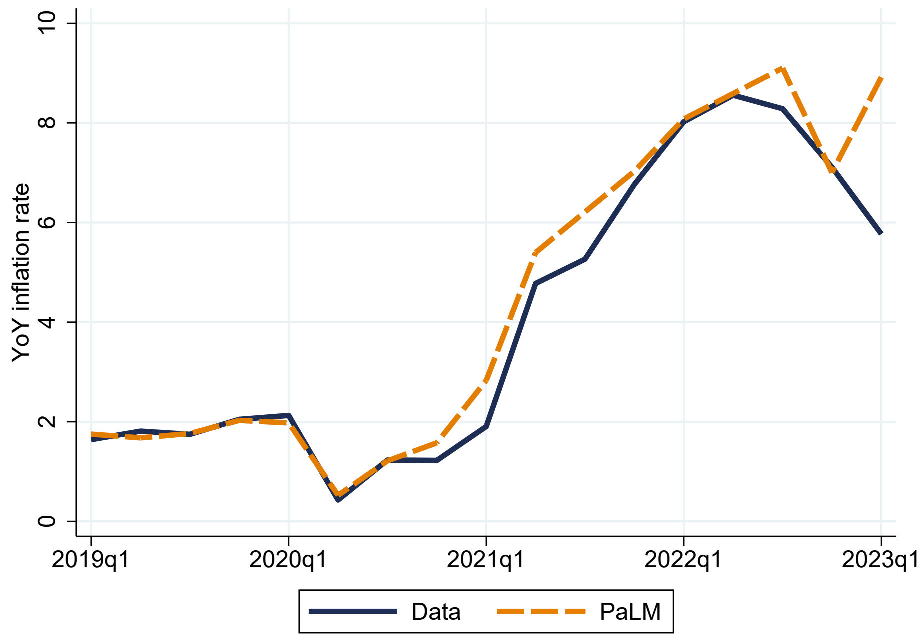


Figure 3: Realized inflation vs. unconditional PaLM forecast

different alternatives against which we can benchmark the LLM forecasts. We think, however, that LLMs may become particularly useful in terms of forecasting variables for which we do not conduct surveys, and/or for which such surveys would be expensive and complicated to design and implement. These include, for instance, disaggregated time series variables such as labor force indicators for specific demographic groups or household disposable income for specific geographical regions, among others.

References

AHER, G., R. I. ARRIAGA, AND A. T. KALAI (2023): "Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies," .

BYBEE, L. (2023): "Surveying Generative AI's Economic Expectations," .

CHOWDHERY, A., S. NARANG, J. DEVLIN, M. BOSMA, G. MISHRA, A. ROBERTS, P. BARHAM, H. W. CHUNG, C. SUTTON, S. GEHRMANN, P. SCHUH, K. SHI, S. TSVYASHCHENKO, J. MAYNEZ, A. RAO, P. BARNES, Y. TAY, N. SHAZEER, V. PRABHAKARAN, E. REIF, N. DU, B. HUTCHINSON, R. POPE, J. BRADBURY, J. AUSTIN, M. ISARD, G. GUR-ARI, P. YIN, T. DUKE, A. LEVSKAYA, S. GHEMAWAT, S. DEV, H. MICHALEWSKI, X. GARCIA, V. MISRA, K. ROBINSON, L. FEDUS, D. ZHOU, D. IPPOLITO, D. LUAN, H. LIM, B. ZOPH, A. SPIRIDONOV, R. SEPASSI, D. DOHAN, S. AGRAWAL, M. OMERNICK, A. M. DAI, T. S. PILLAI, M. PELLAT, A. LEWKOWYCZ, E. MOREIRA, R. CHILD, O. POLOZOV, K. LEE, Z. ZHOU, X. WANG, B. SAETA, M. DIAZ, O. FIRAT, M. CATASTA, J. WEI, K. MEIER-HELLSTERN, D. ECK, J. DEAN, S. PETROV, AND N. FIEDEL (2022): "PaLM: Scaling Language Modeling with Pathways," .

FAUST, J. AND J. H. WRIGHT (2013): "Forecasting inflation," in *Handbook of economic forecasting*, Elsevier, vol. 2, 2–56.

HORTON, J. J. (2023): "Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?" Working Paper 31122, National Bureau of Economic Research.

PHILADELPHIA FED (2021): *Survey of Professional Forecasters Documentation*, Federal Reserve Bank of Philadelphia.

SALEWSKI, L., S. ALANIZ, I. RIO-TORTO, E. SCHULZ, AND Z. AKATA (2023): "In-Context Impersonation Reveals Large Language Models' Strengths and Biases," .

Appendix

This appendix studies the sensitivity of our findings to alternative specifications of the analysis.

A Alternative Temperature Values

This section contrasts the performance of PaLM's conditional forecasts relative to the SPF when the LLM is asked to provide forecasts under alternative temperature levels — a parameter that controls the degree of randomness in the model's response. This parameter ranges from zero, when there is no randomness and answers are deterministic, to a maximal value of one — the model's default value is 0.70, which is what we use for our baseline results.

A.1 Zero Temperature

Table 5: MSE comparison for different time periods: SPF vs. PaLM

Period	**SPF, Mean**	**SPF, Median**	**PaLM, Mean**	**PaLM, Median**
2019	0.49	0.50	0.52	0.52
2020	3.94	3.81	4.83	4.83
2021	14.97	15.44	15.77	15.77
2022	2.85	3.05	0.68	0.68
2023	0.02	0.02	2.99	2.99
All	5.70	5.84	5.80	5.80

Table 6: MSE comparison over different horizons: SPF vs. PaLM

Period	**SPF, Mean**	**SPF, Median**	**PaLM, Mean**	**PaLM, Median**
t	0.29	0.30	0.41	0.41
$t+1$	1.84	1.88	1.13	1.13
$t+2$	4.94	5.07	4.62	4.62
$t+3$	10.00	10.27	9.98	9.98
$t+4$	13.79	14.06	15.46	15.46

A.2 Maximum Temperature

Table 7: MSE comparison for different time periods: SPF vs. PaLM

Period	**SPF, Mean**	**SPF, Median**	**PaLM, Mean**	**PaLM, Median**
2019	0.49	0.50	0.52	0.55
2020	3.94	3.81	2.76	3.08
2021	14.97	15.44	8.04	8.46
2022	2.85	3.05	0.59	0.60
2023	0.02	0.02	7.00	7.46
all	5.70	5.84	3.22	3.43

Table 8: MSE comparison over different horizons: SPF vs. PaLM

Period	**SPF, Mean**	**SPF, Median**	**PaLM, Mean**	**PaLM, Median**
t	0.29	0.30	0.92	0.76
$t+1$	1.84	1.88	0.94	0.90
$t+2$	4.94	5.07	2.23	2.44
$t+3$	10.00	10.27	5.06	5.35
$t+4$	13.79	14.06	8.21	9.13

B Non-Revised CPI Data

This section contrasts the performance of PaLM's conditional forecasts relative to the SPF when comparing forecasts relative to real-time data releases instead of revised data, as in our baseline.

Table 9: MSE comparison for different time periods: SPF vs. PaLM

Period	**SPF, Mean**	**SPF, Median**	**PaLM, Mean**	**PaLM, Median**
2019	0.45	0.45	0.55	0.59
2020	3.90	3.77	2.95	3.42
2021	15.22	15.70	7.41	6.75
2022	2.93	3.12	0.66	0.74
2023	0.01	0.01	3.11	3.11
all	5.76	5.90	3.07	3.05

Table 10: MSE comparison over different horizons: SPF vs. PaLM

Period	**SPF, Mean**	**SPF, Median**	**PaLM, Mean**	**PaLM, Median**
t	0.31	0.31	0.42	0.33
$t+1$	1.88	1.92	0.80	0.87
$t+2$	4.99	5.11	2.44	2.64
$t+3$	10.10	10.36	5.03	4.91
$t+4$	13.91	14.18	7.97	7.75