Simulating the Survey of Professional Forecasters *

Anne Lundgaard Hansen[a], John J. Horton[b], Sophia Kazinnik[c], Daniela Puzzello[d], Ali Zarifhonarvar[d]

[a]*Federal Reserve Bank of Richmond*
[b]*MIT Sloan School of Management*
[c]*Stanford HAI*
[d]*Indiana University Bloomington*

First Draft: January 2024. This Draft: September 2024

Abstract

We simulate economic predictions of professional forecasters using a set of large language models (LLMs). First, we construct a unique hand-gathered dataset of individual characteristics of the participants in the Survey of Professional Forecasters (SPF). Then, we generate AI forecaster personas by feeding a set of LLMs this dataset alongside real-time macroeconomic data. Finally, we use this setup to replicate the process of generating forecasts for an extensive number of macroeconomic variables and forecasting horizons. We find that AI-generated forecasts closely align with human forecasts in both accuracy and behavioral patterns. Our simulated survey data can be used to extend the existing survey in a number of directions such as survey frequency and variables, and to identify behavioral patterns that lead to forecast disagreement.

Keywords: Large Language Models; Survey of Professional Forecasters; Behavioral Finance; Synthetic Surveys; Generative Artificial Intelligence; Simulated Economic Agents

JEL Classification: G4, G21, C8, C9

*This paper does not necessarily reflect the views of the Federal Reserve Bank of Richmond or the Federal Reserve System. We thank Thomas Stark for his time and highly valuable input. We thank the amazing Bryson Alexander and Nicole Lindsay for excellent research assistance. All remaining errors are our own.

Email addresses: anne.hansen@rich.frb.org (Anne Lundgaard Hansen), jjhorton@mit.edu (John J. Horton), kazinnik@stanford.edu (Sophia Kazinnik), dpuzzell@iu.edu (Daniela Puzzello), ali.zarif@iu.edu (Ali Zarifhonarvar)

1. Introduction

Forecasting surveys are a cornerstone in economic modeling and policy-making, serving as a key input for understanding expectations and beliefs (Rudebusch, 2002; Orphanides and Williams, 2002, 2007; Patton and Timmermann, 2010; Manzan, 2011, 2021). However, traditional forecast surveys face significant limitations.

First, surveys are costly to conduct, resulting in low-frequency data collection.[1] Second, survey data is static, reflecting participants' responses at a specific point in time, which inherently complicates establishing a time series of observations for new questions introduced into the survey. Third, researchers and policymakers are constrained by the fixed design of surveys, where existing questions cannot easily be altered, and new questions are challenging to introduce. Finally, personal characteristics of forecasters may impact macroeconomic forecasts (Benchimol et al., 2022; Huang et al., 2022). These limitations may reduce the ability of surveys to inform economic policy effectively.

A central question in forecasting is how expectations are formed. The literature increasingly questions the assumptions behind the full-information rational expectations (FIRE) model. Studies point to informational frictions (Mankiw and Reis, 2002; Maćkowiak and Wiederholt, 2015) and behavioral biases in expectation formation, such as over-extrapolating past trends (Bordalo et al., 2020) or under-extrapolating due to cognitive limitations (Gabaix, 2020). Despite extensive research, the precise mechanisms driving expectation formation remain unresolved, particularly regarding the influence of personal characteristics.

To address these challenges, we introduce a novel simulation-based framework that

[1] For instance, the Survey of Professional Forecasters (SPF) is conducted only quarterly.

replicates the Survey of Professional Forecasters by generating synthetic forecast data using Large Language Models (LLMs).[2] By creating synthetic forecasters based on hand-collected individual-level personal characteristics, we produce panel-style survey responses that mirror the SPF generation process, using identical inputs such as real-time data and survey framework. This approach allows us to overcome the constraints of traditional surveys by enabling high-frequency, adaptable, and cost-effective data collection. It also allows us to take the influence of personal characteristics into account.

Recent advances in artificial intelligence show that, in certain domains, LLMs can reliably simulate economic agents and produce outputs aligning with human reasoning (Argyle et al., 2023; Horton, 2023; Fedyk et al., 2024; Zarifhonarvar, 2024). LLMs also replicate human-like errors (Koralus and Wang-Maścianica, 2023; Hayes et al., 2024) and reflect distinct personality traits based on psychological models (Jiang et al., 2023). Moreover, their probabilistic judgments often exhibit biases similar to human reasoning patterns (Zhu and Griffiths, 2024). These findings suggest that LLMs are valuable tools for modeling both artificial and human decision-making processes.

Building on this literature, we show that LLMs can effectively simulate real-world survey responses at the individual level. Our key innovation lies in integrating individual forecaster personas with real-time macroeconomic data to replicate the SPF. This framework enables us to explore how personal characteristics influence economic expectations and forecasting behavior, providing a more detailed view of forecast dynamics than aggregate models allow.

Our contributions are threefold. First, we provide novel insights into the behav-

[2]We select the Survey of Professional Forecasters (SPF) for this study because it is the longest-running publicly available forecasting project, providing a rich dataset of professional forecasts on key macroeconomic indicators (Croushore, 1993) and making it an ideal benchmark for our exercise.

ioral aspects of forecasting by leveraging individual forecaster characteristics such as education, job experience, and institutional affiliation. We examine how these traits influence forecasters' willingness to deviate from consensus and their overall forecast accuracy. Second, we are among the first to apply generative AI in macroeconomic forecasting, opening new avenues for predictive modeling. Our findings demonstrate that LLM-generated forecasts can enhance accuracy and offer deeper insights into economic trends by capturing forecasters' behavioral responses to real-time data and macroeconomic shifts. Third, we present a novel framework that integrates synthetic forecasting agents with real-time data, replicating the forecast generation process at both the individual and aggregate levels. This approach addresses key limitations in traditional surveys, and provides a robust complement to conventional survey-based methodologies.

Our work extends the emerging literature on the application of LLMs in economic forecasting. For instance, Faria-e Castro et al. (2023) show that LLMs like Google's PaLM can generate inflation forecasts that often outperform the SPF. Similarly, Bybee (2023) use historical news data to generate economic expectations via LLMs, with outcomes aligning closely to survey-based measures. However, our paper differs in scope and approach by using LLMs to simulate individual decision-making processes and applying them to macroeconomic forecasting. Through this integration of behavioral analysis and AI-driven forecasting, we offer a more comprehensive understanding of how personal traits influence forecasting behavior, and at the same time provide a novel framework for improving forecast accuracy.

Our findings reveal that AI models, particularly the GPT 4o-mini, often outperform human forecasters in short- and medium-term forecasts, with lower mean squared errors and more accurate directional predictions. However, human forecasters retain

an advantage in long-term forecasting, especially during periods of economic volatility such as the 2008 financial crisis and the COVID-19 pandemic. These results suggest that while AI can replicate human forecasting patterns in stable conditions, further refinement is needed to improve performance during times of economic uncertainty.

The remainder of the paper is organized as follows: Section 2 covers the structure of the SPF panel. Section 3 discusses how synthetic forecaster persona are constructed. Section 4 outlines our methodology for generating synthetic forecasts using LLMs. Section 5 describes the data used. Section 6 explains the simulation framework. Section 7 presents the results of our analysis, comparing AI-generated forecasts with those from the SPF and the realized macroeconomic values. Section 8 concludes.

2. The Survey of Professional Forecasters

The Survey of Professional Forecasters (SPF) is a quarterly survey of economic experts in the United States. Initiated in the fourth quarter of 1968 by the American Statistical Association and the National Bureau of Economic Research, it has been conducted by the Federal Reserve Bank of Philadelphia since 1990. Response deadlines have typically been around the middle of each quarter.[3]

The SPF panelists are considered attentive and rational agents who provide forecasts based on extensive data analysis, quantitative models, and their professional judgment. They produce projections as part of their professional responsibilities and typically have substantial experience in macroeconomic forecasting. These participants are knowledgeable about the variables they forecast and the relevant economic theories.[4] Panelists work in various settings: some at forecasting firms serving exter-

[3]For complete survey deadlines and release dates, see the Philadelphia Fed's website.

[4]Forecasters are recruited via the SPF Call for Participants. New panelists undergo a trial period

nal clients, others at banks or financial institutions providing forecasts internally and externally. The panel also includes chief economists from industry trade groups and manufacturers, as well as academics specializing in optimal forecasting methods.

Although SPF forecasters remain anonymous and their personal characteristics are largely unknown, the individual-level SPF dataset does include each forecaster's industry classification ("financial service provider" or "non-financial service provider"), as well as assigns identifiers that allow researchers to track an individual's forecasts across different survey rounds. The anonymity of individual forecasters is understandable, as it protects the forecasters' professional interests and encourages them to provide honest and unbiased forecasts without concern for potential repercussions.[5]

The composition of the SPF panel varies over time due to respondents joining or leaving, and occasionally, respondents may not answer all survey questions, resulting in an unbalanced panel. However, the use of numerical identifiers allows for tracking individual forecasts over time. This facilitates analyses of forecast behavior at the individual level and enables comparisons between forecasters—for instance, assessing relative accuracy or tendencies to be contrarian or uncertain. Since the Philadelphia Fed began overseeing the survey, the average number of panelists has been around 40 (see Figure A.10).

Forecasters employ various methods to produce their predictions. A special survey in 2009 (Stark, 2013) revealed that most use quantitative models but adjust projections to account for current economic conditions and recent trends. Nearly all supplement

during which their submissions are reviewed for conceptual errors, such as forecasting the wrong variable or using incorrect data transformations. No judgment is made on the forecasts themselves beyond checking for conceptual accuracy. More information is available on the Philadelphia Fed's website.

[5]While we are able to gather the names and the individual characteristics of these forecasters, we only report aggregate statistics in the paper.

their models with subjective beliefs about the economy. Additionally, forecasters use different methods depending on the forecast horizon; for example, the model used for predicting real GDP in the current quarter may differ significantly from that used for forecasting average GDP growth over the next five years.

The survey form has undergone changes over time; in our sample period from 1999 to 2023, there have been eight modifications. These changes have expanded the original form, but the core essence of the survey has been preserved.

3. Synthetic SPF Personas

Although the SPF panelists are anonymized in the published datasets, the Philadelphia Fed acknowledges individual forecasters on their website by listing their names.[6] Additionally, many panelists publicly disclose their participation in the SPF on social media or professional platforms. Leveraging this information, we construct a unique dataset of individual forecaster characteristics within the SPF. Specifically, we employ two approaches to create synthetic forecast personas: one based on forecaster characteristics and another based on their biographies. Table 2 presents examples of personas generated by both methods.

3.1. Generating Synthetic Personas Based on Characteristics

We develop detailed personas for each forecaster by collecting key information about their backgrounds, such as their education, job titles, and affiliated organizations. This includes specifics like alma maters, degrees earned, professional roles, and the types of institutions they represent. When available, we also incorporate the locations of their organizations, countries of origin, and relevant social media presence.

[6] A screenshot of such acknowledgments is provided in Figure A.8.

Each characteristic is summarized in Table 1.

Gathering forecaster characteristics is crucial because empirical literature demonstrates that personal backgrounds significantly influence behavior and beliefs. For instance, individuals who grew up during the Great Depression tend to be more cautious about financial risk (Malmendier and Nagel, 2011), while those who experienced hyperinflation are less likely to invest in risky assets (Fajardo and Dantas, 2018). Personal experiences shape outlooks on forecasting macroeconomic variables such as inflation (Malmendier and Nagel, 2016). Moreover, both the geography and institutional affiliations of experts can affect their forecasting approaches (Batchelor, 2007; Hong and Kacperczyk, 2010).

To construct our dataset, we begin by compiling a list of forecaster names from the acknowledgments provided by the Philadelphia Fed for each survey. Using open-source intelligence (OSINT) techniques, we gather detailed professional and demographic information about each individual through internet searches. To infer the country of origin, we consider indicators such as the location of their high school listed on professional profiles like LinkedIn.[7]

With these forecaster profiles established, we create personalized narratives for the prompts provided to the LLMs. Each forecaster is introduced as a member of the SPF panel, with their academic and professional achievements highlighted. They are then prompted to provide predictions based on the most recent economic data available, such as inflation, GDP, or employment figures. By integrating their background and real-time data into the prompt, we encourage the LLM to generate forecasts that reflect

[7]The assumption is that attending high school in a particular country significantly influences an individual's formative years, shaping their perspectives and potentially affecting their future forecasts. While the majority of participants are American, the variation in our data provides sufficient relevance for the country-of-origin variable. We thank Nicole Lindsay for this insight.

the forecaster's expertise and judgment. This process results in forecasts tailored to the individual, capturing their unique experiences and roles in the forecasting process.

3.2. Generating Synthetic Personas Based on Biography

An alternative approach involves crafting synthetic personas using biographical narratives. We use OpenAI's Assistant API to interact with a specialized AI assistant designed for this task. Specifically, we employ the following system prompt:

> *You are designed to gather information on the background and biography of a professional forecaster who participates in the Survey of Professional Forecasters for the Federal Reserve Bank of Philadelphia. Your task is to create a persona, with a maximum of 200 words, that accurately represents their most important characteristics, forecasting expertise, background, and overall persona. Do not reference them by name, but mentioning associated organizations is acceptable. Each text should start with "You are ...". Use simple words and easy-to-read text.*

The assistant processes the biographical information obtained from each forecaster's online profile and generates a synthesized persona as output. This approach captures the essence of the forecaster's expertise and background without revealing their identity, allowing the LLM to produce forecasts that embody the forecaster's professional persona.

4. Method

4.1. General Mechanism

For each of the SPF forecast personas, we replicate the process of filling out the SPF survey as closely as possible for each quarter in the sample.

Considering N forecasters and T surveys, i.e., quarters, we set out to produce a nowcast and 1-4 quarters ahead forecasts for each forecast variable X_t, i.e.:

$$\left\{[\mathbb{E}(X_{t+s}|\mathbb{F}_{t,i})]_{s=0}^{s=4}\right\}_{\forall i \in N,\, t \in T}$$

where $\mathbb{F}_{t,i}$ is the information set available to forecaster i at the time of the deadline for the SPF survey in quarter t. This information set includes real-time macroeconomic data, the aggregate SPF forecast in the previous survey, the individual forecaster's previous forecast, and the persona-specific knowledge and expereicnes. The nowcast appears as the case where $s = 0$.

For example, consider a fictional forecaster, Alexandra Bryson, who is member of the panel betweeen 2017 Q1 and 2023 Q4. We first generate a set of forecasts for different horizons for each of the variables for Alexandra for her first forecasting date (2017 Q1). For this first forecasting date, Alexandra operated with available real-time data and the data on the aggregate forecasts made by the panel in 2016 Q4. On our next iteration in 2017 Q2, we incorporate this data in Alexandra's information set, and iterate forward until Alexandra's tenure on the panel concludes.

4.2. Large Language Models

Recent research demonstrates that language models can effectively mimic human economic behavior and thought patterns (Horton, 2023; Brand et al., 2023; Argyle et al., 2023; Jiang et al., 2023; Aher et al., 2023). Trained on vast corpora of text, these models learn the statistical structures of language, enabling them to generate contextually appropriate and coherent responses. Essentially, LLMs can be viewed as computational proxies for human respondents, able to produce outputs that resemble human reasoning and decision-making.

The concept of synthetic agent modeling, introduced by Horton (2023) and Argyle et al. (2023), merges theoretical and empirical research to simulate human behavior using LLMs. Traditional economic theory often becomes unwieldy when incorporating realistic attributes, while empirical research may be constrained by data limitations. Synthetic agent modeling bridges these approaches by positioning the model prompt as the central behavioral driver. This methodology allows for the testing of hypotheses and theories about human behavior.

Horton (2023) suggest that LLMs emerge as implicit computational representations of humans—termed *homo silicus*. Analogous to the economic concept of *homo economicus*, LLMs can be endowed with information, preferences, and biases, facilitating the simulation of behavior in diverse scenarios. LLMs thus serve as preliminary simulation environments, enabling researchers to uncover novel social science insights that can be validated in real-world settings.

Similarly, Argyle et al. (2023) demonstrate that a single language model, when properly conditioned, can produce outputs biased both toward and against specific groups and perspectives, closely mirroring human response patterns across fine-grained demographic dimensions. This indicates that language models encapsulate a multitude of biases, reflecting the diversity of human thought. Additionally, Brand et al. (2023) employ an LLM to generate realistic consumer behavior survey responses, finding that the distributions of responses to changes in product attributes and choice probabilities align with observed consumer behavior, including patterns like downward-sloping demand curves and state dependence.

Expanding on the role of LLMs in economics, Immorlica et al. (2024) argue that these models should be conceptualized as economic agents. AI systems can function similar to human consultants, possessing their own information sets, actions, and

preferences, albeit with limited agency. This perspective positions AI not merely as a tool but as an active participant in economic interactions.

Some social scientists even advocate for using LLMs as full substitutes for human participants in psychological studies (Dillion et al., 2023; Grossmann et al., 2023). Grossmann et al. (2023) contend that advanced AI models, trained on extensive datasets, capture a wide array of human experiences and viewpoints. Not burdened by human limitations such as attention spans or cognitive biases, language models can provide broader parameter responses. This approach is particularly beneficial in high-risk contexts, offering a simulated testing environment prior to real-world implementation.

In this paper, our focus is on time series forecasting. Given that language models excel at modeling sequential dependencies in text, we posit that their capabilities can generalize to capturing sequential dependencies in time series data. Combined with their ability to simulate human reasoning, LLMs provide an ideal framework for recreating the SPF and exploring the formation of economic expectations.

4.3. Temporal Restriction

Do language models generate or recall? Language models are trained on large datasets collected during a specific period, allowing them to learn patterns and relationships within the data. However, when generating responses, they do not simply recall exact pieces of information. Instead, they create new outputs based on these learned patterns, which gives the impression of generating content, even though the underlying knowledge comes from previously seen data (Gurnee and Tegmark, 2023). There are, however, nuances.

Restricting large language models (LLMs) to operate within a specific temporal scope presents a unique set of challenges in the context of forecasting. LLMs are

trained on large datasets that include information from various time periods, meaning they can often reference or generate insights from data outside the intended forecasting window. This becomes problematic when attempting to simulate real-time forecasts, as historical or future data that would not have been available to human forecasters at a given point in time might influence the model's predictions.

To address this issue, the LLMs in this study were provided with real-time datasets that reflect the economic data available to human forecasters at the time of their predictions. The LLMs were specifically instructed through prompts to base their forecasts only on the information available up to the specific quarter being forecasted (e.g., t_0 or t_1), and not to rely on data beyond that point.

Despite this effort, enforcing strict temporal adherence is challenging. Language models are naturally trained to predict the next word or sentence based on patterns they have learned from vast amounts of text. When tasked with forecasting, LLMs may unintentionally draw on patterns or knowledge that imply future data trends, even when explicitly told to disregard information beyond a specific time point.

This phenomenon can lead to what is known as "temporal leakage," where predictions are informed by data that should be outside the forecaster's knowledge at the time. The introduction of temporal leakage can result in several potential impacts on forecast accuracy.

First, the effect of temporal leakage is often less pronounced in short-term forecasts, as the model can primarily rely on current or very recent data to make predictions. However, if the model references future trends or patterns it has learned from historical data, even in a subtle way, this could give it an unintended advantage over human forecasters. In some cases, this might explain why AI models outperformed human forecasters in shorter forecast horizons, as shown in the results.

The impact of temporal leakage can be more detrimental in longer-term forecasts. While human forecasters rely heavily on judgment and expert knowledge to navigate the uncertainty of distant future predictions, LLMs might fall back on probabilistic trends that they have learned from the overall training data. This could lead to overconfidence or reliance on patterns that do not hold in real-time economic environments, ultimately reducing forecast accuracy. In our study, this was reflected in the slightly inferior performance of AI models for long-term forecasts, as compared to human forecasters who were able to better incorporate uncertainty and contextual knowledge.

To deal with this issue, we apply several mitigation strategies to limit temporal leakage and improve the realism of simulated real-time forecasts. First, the models were explicitly instructed to focus on the data available up until the quarter being forecasted, with prompts designed to emphasize the importance of ignoring future trends. This approach aimed to "condition" the model to act as a human forecaster with limited knowledge of future events.

Second, real-time data was segmented to match the exact information that would have been available to human forecasters at the time of their predictions. This ensured that LLMs were not inadvertently provided with future data, helping to reduce the risk of temporal leakage.

5. Time-Series Data

Professional forecasters rely on a large number of economic indicators when forming expectations about the future state of the economy. However, the SPF respondents only have access to data released at the time of the SPF submission deadline, which also excludes ex post data revisions.

To simulate the SPF, we aim at providing the LLMs access to the same data environment as that of human forecasters. Specifically, we feed the LLMs with past SPF forecasts and real-time macroeconomic indicator data. This section describes how we collect and process these data.

5.1. Data Collection and Processing

We use data from the Federal Reserve Bank of Philadelphia from two datasets: real-time macroeconomic data and median SPF forecast data. The final merged dataset contains quarterly real-time observations where each row represents a date and each column (vintage) represents the date when the data point was recorded. The variables and their definitions are summarized in Table 3. The median SPF forecast dataset provides historical SPF forecasts of various variables in levels at the quarterly frequency at forecasting horizons of 1-4 quarters (denoted by "3"–"6" in the dataset) and the quarterly nowcast (denoted by "2"). The dataset also provides the past observed value of the data (denoted by "1"), which is provided to the forecasters as part of the SPF submission form.

We process these data to generate a row of data for each survey quarter with (1) the latest observed values of the SPF variables, (2) the latest real-time macroeconomic data available at the time of the SPF submission deadline, (3) and the median SPF forecasts. One of the core tasks in this process is transforming percent change variables and calculating the percent change between quarterly levels. We compute the percent change using the following formula, where X represents the quarterly means of the variables:

$$\left(\frac{X_t}{X_{t-1}}\right)^4 - 1 \times 100$$

6. Survey Simulation

Building off all of the inputs we've described up to this point, we use a set of LLMs to run a simulated survey grounded in a set of hypothetical scenarios presented in the form of a prompt. The prompt serves both as a backdrop against which synthetic participants respond, and the driver for generating responses.

6.1. Prompt

We automate the process of generating forecast queries and interacting with an AI assistant to gather responses. We start by feeding in information about forecasters and economic variables. This way, we create personalized forecast queries that ask participants to predict economic variables such as GDP, inflation, and unemployment for future quarters. As described in previous sections, the queries are customized based on the forecasters' past predictions and real-time economic indicators, making them highly tailored to each individual:

```
You are a participant on a panel of Survey of Professional Forecasters. Your name is [name], you
graduated from [alma mater] with a [education] around [graduation year].

Today, you work as [title] at [affiliation]. It's [afilliation types] organization.

Your organization is based in [company_location].
You are originally from [country_origin]. [social media status].
```

Note: This is a simplified version of the prompt that is set to create a synthetic characteristics-based persona. The terms highlighted in blue are inputs.

```
We are in [date_q]. You are about to fill out the forecast form for [date_q]. Using only the
information available to you as of [date_q], please provide your best numeric forecasts for the
following variables: [variable_data].

Do this for the following quarters: t (current quarter), t+1, t+2, t+3, and t+4, as well as annual
forecasts for this and next year (annual averages). You have the most recent real-time data on key
macroeconomics variables available to you as of today: [real_time_data].

The forecasts made by the SPF panel during the previous quarter were as follows (for t-1, t, t+1,
t+2, t+3, t+4; where t is previous quarter:[variable_forecasts_text].

Do not incorporate any data that was not available to you beyond the current date in your
forecasts. Do consider all relevant information on the broad economic conditions and current
Federal Reserve actions (up to, but not beyond [release_date]).

Use available information, and your professional judgement and experience.
Your forecast is anonymous. Provide the forecasts as a sequence of numerical values only.
Please only provide your forecasts in the format: (t, t+1, t+2, t+3, t+4, this year's average,
next year's average).
```

Note: This is a simplified version of the prompt that provides each synthetics forecasters with real-time data and past predictions, and solicits predictions.

As described in the prompt above, synthetic forecasters are provided with a set of real-time economic data and past forecasts to help them generate their predictions. These variables reflect the current state of the economy, and the forecasters are also given past forecasts from previous quarters to help contextualize their predictions. Additionally, the forecasters receive personalized background information, such as their education, job title, and the type of organization they work for, which is used to simulate a professional forecasting environment.

Based on this data, the synthetic forecasters are asked to produce numeric forecasts

for a set of variables for several upcoming quarters, specifically for periods like the current quarter t, $t+1$, $t+2$, and up to $t+4$. They are also asked to provide annual forecasts for both the current and the following year. The forecasts must be numerical and presented in a specific format. Additionally, the synthetic forecasters are asked to provide a brief explanation or reasoning behind their predictions, limited to 1-2 sentences.[8]

The overall goal is for the forecasters to use the available information and their professional judgment to predict how key economic variables will evolve in the near future, without considering any data beyond the current point in time.

In this scenario, the individual is depicted as a member of a panel of Professional Survey Forecasters (SPF), actively participating in a forecasting exercise on a specified date. The narrative introduces the forecaster by detailing their academic and professional background, including their *alma mater*, highest degree earned, year of graduation, current job title, and the nature of the work of their employer. We test the hypothesis that inclusion of these characteristics in the prompt could influence the weighting of different data sources or models based on presumed expertise.

6.2. Assistants

We use the OpenAI Assistant tool that has instructions and can leverage models, tools, and files to respond to user queries. Broadly, Assistants are AI agents built using the Assistants API, designed to perform tasks and respond to user queries within applications. They can be assigned various OpenAI models, as well as model temperature, and other parameters, and can generate responses based on specific in-

[8]These instructions are assigned to them via system prompt.

structions.[9] Assistants manage persistent conversations through Threads, which store message histories and ensure smooth interactions. They can create and reference files, while also using a structured process called "Runs" to perform tasks step by step, ensuring accurate results.

By employing the Assistant, we effectively endow our prompts with agency, transforming static instructions into interactive dialogues with a simulated economic forecaster. This approach aligns with the work of Zarifhonarvar (2024), who use OpenAI's Assistants to simulate economic agents in inflation expectation surveys. In tha paper, a group of Assistants receive informational treatments to influence their prior expectations, and their responses are analyzed to understand how LLMs adjust beliefs in response to new information. The authors find that Assistants with access to relevant economic data generate more accurate and consistent inflation expectations compared to those without access.

Operationally, once we define the forecaster persona and the forecasting task via the prompts outlined in the previous subsection, the Assistant acts as the agent executing the task. The script interacts with the model through the Assistants API to gather responses, effectively simulating a conversation with a professional forecaster. The Assistant is prompted to confirm its understanding of the task, provide numeric forecasts, and explain the reasoning behind its predictions.

6.3. Models and Setup

We evaluate twelve different AI models to assess their performance in economic forecasting tasks. These models represent combinations of:

[9]The Assistants API supports tools like Code Interpreter, File Search, and Function Calling, allowing assistants to handle complex tasks by accessing multiple tools in parallel.

- LLM Architectures: GPT-3.5, GPT-4, and GPT-4o-mini
- Temperature Settings: 0 (deterministic) and 1 (stochastic)
- Persona Configurations: Biography-based and Characteristic-based personas

Each model generates forecasts for up to four quarters ahead ($t0$ through $t4$), and we compare their performance against forecasts from professional forecasters in the Survey of Professional Forecasters (SPF). Here is a brief overview of the models used:

- GPT-3.5: A highly capable language model trained on extensive internet text, serving as the predecessor to GPT-4. While proficient in understanding and generating coherent responses, it is slightly less advanced in handling nuanced tasks compared to GPT-4 (Hansen and Kazinnik, 2023).
- GPT-4: An advanced iteration with enhanced capabilities in handling ambiguity, generating coherent text, and performing complex reasoning tasks. With a larger training corpus and more sophisticated parameters, GPT-4 excels in predictive tasks.
- GPT-4o-mini: A lightweight variant of GPT-4, optimized for speed and efficiency. While it offers quicker response times due to fewer parameters, it may not achieve the same level of accuracy or detail as the full GPT-4 model.

Additionally, we use two parameters to set the model temperature - zero and one. The temperature parameter controls the randomness of the model's outputs:

- Temperature = 0: Generates deterministic and focused responses, adhering closely to known patterns and facts. This setting reduces creativity and potential errors, leading to narrowly focused predictions.

- Temperature = 1: Introduces more randomness, allowing for a wider range of possible predictions. While this can result in more creative or unconventional outputs, it may also increase variability and the likelihood of less probable outcomes.

Finally, we also have two ways to define our synthetic forecasters:

- Biography-Based Persona: The model is provided with a persona that mirrors a professional forecaster's biography and experience, including background, expertise, and decision-making style. This aims to simulate the behavior and thought processes of an individual human forecaster.

- Characteristic-Based Persona: The model is assigned a set of hand-collected characteristics and attributes, allowing it to adopt different forecasting strategies based on general traits rather than simulating a specific individual's style.

By combining the three LLM architectures with the two temperature settings and the two persona configurations, we create twelve distinct models for evaluation. For each LLM architecture, the configurations are:

- Biography-Based Persona, Temperature = 0

- Biography-Based Persona, Temperature = 1

- Characteristic-Based Persona, Temperature = 0

- Characteristic-Based Persona, Temperature = 1

These combinations allow us to conduct a detailed comparison of how different model configurations perform in economic forecasting across various time horizons. We describe the outcome of this simulation in the next chapter.

7. Results

As discussed, we start by evaluating the twelve different simulation set-ups, each representing variations in the LLM architecture (GPT 3.5, GPT 4o, and GPT 4o-mini), temperature settings (temperature = 0, 1), and persona configurations (biography-based, characteristic-based) we use in the prompt. The models generate forecasts for up to four quarters ahead (t_0 through t_4), and their performance is compared against the mean and median forecasts provided by professional forecasters from the SPF.

Table 4 summarizes the description of the models, and Table 5 outlines the evaluated forecast horizons.

7.1. How Human?

Figure **??** compares AI-generated forecasts with those produced by human forecasters. The six panels display the average forecast bias over time for AI models (shown in blue) and human forecasts from the Survey of Professional Forecasters (SPF, shown in red), covering the period from around 2000 to 2024. Each panel represents a different AI model (SPF1 to SPF6), allowing us to assess how closely the AI models emulate human forecasting behavior.

Overall, the blue and red lines move in parallel, indicating that the AI models effectively approximate the biases observed in human forecasts. This suggests that the simulations are successful in capturing key aspects of human forecasting behavior. However, deviations occur during periods of economic turmoil. For example, the AI forecasts exhibit increased volatility during the 2008 financial crisis and the COVID-19 pandemic in 2020.

In models SPF2, SPF3, and SPF4, there are sharp spikes in AI forecast bias around the time of the 2008 financial crisis, where the AI models appear to overreact compared

to human forecasters. This implies that AI-generated forecasts may be more prone to larger errors during crises, potentially due to a heavier reliance on historical patterns or difficulties adapting to sudden economic changes.[10]

In the more recent period (2020 and beyond), which includes the COVID-19 pandemic, divergences between AI and human forecasts become more pronounced. While overall trends remain somewhat aligned, the AI models sometimes overshoot or undershoot relative to the SPF forecasts. This indicates that, although AI models can mimic human forecasters under normal conditions, they may struggle in rapidly changing economic environments or unprecedented events.

In sum, the AI models do a reasonable job of emulating human forecast bias during stable periods, effectively capturing human forecasting patterns. However, they tend to exhibit more volatility during times of economic uncertainty, such as the 2008 financial crisis and the COVID-19 pandemic. This suggests that while the simulations can approximate human forecasts to a large extent, further improvements are needed to enhance their performance during unpredictable scenarios where human judgment plays a more nuanced role.

7.2. Personal Characteristics

In this section, we look into how various professional character of forecasters impact the simulated forecasts. We start with Table 19, which provides a comprehensive overview of forecasting performance across various professional roles.

The largest group, Chief Economists, shows moderate forecast accuracy with a mean squared error (MSE) of 2.48 and a slight underestimation bias. In contrast, the

[10]We observe in Figure 4 that AI models using characteristic-based personas (solid blue lines) tend to align more closely with SPF forecasts, while those using biography-based personas (dashed blue lines) exhibit greater volatility. This suggests that characteristic-based AI forecasts demonstrate more stable behavior compared to biography-based models.

Consultant/Analyst role demonstrates the highest accuracy, with the lowest MSE of 1.87, though it also exhibits a relatively high underestimation bias. At the other end of the spectrum, Portfolio Managers display the most significant overestimation bias and the highest MSE (4.20), indicating the least accurate forecasts. Interestingly, the Other group has the highest MSE, suggesting overall lower forecast accuracy. The Senior Economist group shows the lowest negative bias, while the Unknown category has the lowest variability in its forecasts, indicated by the smallest standard deviation. These results suggest that domain-specific expertise and the nature of the role heavily influence forecasting tendencies and performance.

We also look at how affiliation impacts forecast accuracy. Table 18 summarizes the results. The largest groups, Banking and Financial Services, Consulting and Research, and Affiliation Unknown, dominate the count and exhibit relatively similar performance. Banking and Financial Services, with the lowest bias (-0.07) and a relatively low mMSE of 2.47, suggest that this group produces relatively accurate forecasts with minimal bias. On the other hand, Consulting and Research displays a slightly higher bias (-0.19) and MSE (2.71).

In addition, we look into education (Table 20) and gender (Table 21). Table 20 suggests that higher educational attainment, particularly among those with MBAs and Ph.D.s, is associated with better forecast accuracy, though bias varies across groups. Overall, MBAs and Ph.D.s show a tendency toward underestimation, while Bachelor's degree holders show a slight overestimation.

Table 21 compares the forecast performance between female, male, and unknown gender groups. The male group, by far the largest with 4,792 participants, shows relatively accurate forecasts with a MSE of 2.48 and a mild underestimation bias (-0.13).

Female forecasters, with a count of 542, exhibit a higher MSE (2.92) and a stronger underestimation bias (-0.34), indicating that while their forecasts tend to be more conservative, they also have less accuracy compared to their male counterparts. The higher standard deviation (1.65) for synthetic female forecasters also suggests more variability in their predictions.

7.3. Forecast Accuracy

We begin the comparison between AI and human forecasts is based on their Mean Squared Errors (MSE), directional accuracy, and forecast bias.[11] This section presents the findings across various forecasting horizons, persona configurations, and model architectures.

Table 14 provides a deeper comparison of the AI models by focusing on the percentage of wins for biography-based versus characteristic-based personas. This comparison is critical as it shows how the persona configuration affects forecast accuracy. The biography-based models, which are designed to simulate human forecasters by embedding expertise and experience into the model's output, exhibited more consistent performance in short- and medium-term forecasts. On the other hand, the characteristic-based models performed better in long-term forecasts, especially at the $t3$ and $t4$ horizons.

Notably, the GPT-4-based models, particularly the GPT 4o-mini variant, showed strong performance across multiple accuracy metrics. As shown in Table 12, these models generally performed better in terms of RMSE and MAPE across various forecast horizons.[12] This suggests that the GPT-4o-mini variant provides a more reliable

[11] Please see Online Appendix for definitions of metrics used.

[12] The GPT 4o-mini model was particularly effective at minimizing RMSE, a key measure of the magnitude of forecast errors.

prediction by reducing large outliers in the forecast data. The GPT 4-based models also outperformed human forecasters in terms of MAPE, which assesses the average absolute percentage errors in the predictions.

In addition to RMSE and MAPE, directional accuracy was another critical metric we used to evaluate the performance of the AI models. Directional accuracy measures the ability of a model to correctly predict the direction of change in the forecasted variable.[13] In this context, the GPT 4-based models also showed robust performance, especially in the $t0$ and $t1$ horizons. The characteristic-based GPT 4o-mini model, in particular, demonstrated superior directional accuracy, correctly predicting the direction of inflation more frequently than its human counterparts.

Finally, we look at forecast bias as an accuracy metric. Forecast bias measures the systematic overestimation or underestimation of a variable. A low forecast bias indicates that a model's predictions are centered around the true values, without consistently erring on the side of optimism or pessimism. The results showed that biography-based models had lower forecast bias compared to characteristic-based models, especially at the shorter forecasting horizons ($t0$ and $t1$). The systematic nature of these models made them more reliable in providing unbiased forecasts in the short term. However, as the forecasting horizon increased ($t3$ and $t4$), the characteristic-based models began to demonstrate lower forecast bias, indicating their strength in providing more balanced predictions over the long run.

The overall comparison between AI models and professional forecasters reveals that AI-driven models, particularly those based on GPT 4o-mini, generally perform better across most key metrics. The superior performance of AI models is especially

[13]This is particularly important in economic forecasting, where understanding whether a variable (such as inflation) will increase or decrease can be as valuable as the precise magnitude of that change.

apparent in the short- and medium-term forecasts, where AI models consistently had lower MSE, higher directional accuracy, and lower forecast bias. However, human forecasters still maintained a slight edge in long-term predictions, particularly in terms of directional accuracy. This suggests that while AI models are highly effective at short-term forecasting, there is still room for improvement in forecasting over extended time periods.

7.4. Winners and Losers

7.4.1. Human vs. Machine

Table 15 provides a clear breakdown of "winners" and "losers" when comparing AI and human forecasters across different forecasting tasks, using mean squared error (MSE) as the measure of accuracy. Each row represents a competition between AI and human forecasters at a specific time horizon (from t_0 to t_4) against corresponding Consumer Price Index (CPI) forecasts (CPI2 to CPI6).

Looking at the "winners" column, for example, in the t_0 vs CPI2 comparison, human forecasters win 630 times, while the AI wins 548 times. Though humans come out slightly ahead in this case, the number of ties (22) indicates close competition. As we move forward in time (t_1 vs CPI3, t_2 vs CPI4, and so on), AI begins to dominate more consistently. By the t_4 vs CPI6 comparison, AI wins 685 times, far outpacing human forecasters, who win only 509 times.

Overall, the table suggests that while human forecasters may have an edge in some short-term predictions (like t_0), AI becomes the stronger "winner" in the medium- to long-term forecast horizons. This implies that AI is better equipped for making more accurate predictions as the forecasting period extends, positioning AI as the consistent "winner" in long-range forecasting tasks.

One hypothesis for why AI tends to outperform human forecasters as the forecasting horizon increases is the inherent differences in how humans and AI process and use data. In short-term forecasts, such as t_0, human forecasters may have an edge because they can incorporate immediate context and subjective judgment. This ability to interpret qualitative nuances or last-minute changes in economic conditions could explain why human forecasters win more frequently in the t_0 vs CPI2 comparison.

7.4.2. Across Machine Methods

Table 14 shows this forecasting accuracy across methods. The upper chart illustrates the MSE for each synthetic forecaster method across forecast horizons (t_0 to t_4). Across all horizons, synthetic forecaster method MSE values rise as the forecast horizon increases, indicating that forecasts become less accurate over time. At the shortest horizon (t_0), the MSE values are generally lower, but as the horizon extends to t_4, the error increases across all synthetic forecaster methods, suggesting the challenge of long-term forecasting.

Within each horizon, there's notable variation in the performance of the synthetic forecaster method. Some models consistently produce lower MSEs, indicating more accurate predictions. However, no clear pattern emerges between the temperature settings or the forecasting methods (bio-based vs. characteristics-based) when comparing individual MSE values. The differences suggest that various factors, such as the AI model's setup or the methodology (bio vs. characteristics), may influence the accuracy differently across different horizons.

The lower chart depicts the bias in the forecasts for each model across the same horizons. Bias measures whether the forecasts tend to consistently overestimate (positive bias) or underestimate (negative bias) the actual outcomes. The vertical axis represents the level of bias, while the forecast horizon remains on the horizontal axis.

At the shorter horizons (e.g., t_0), there is a mix of positive and negative bias across the models, with some forecasters consistently over-forecasting and others under-forecasting. As the horizon extends to t_4, this mix of bias persists, indicating that some AI models are more prone to systematic forecasting errors than others. However, the overall distribution of bias does not seem to change drastically with increasing forecast horizons.

Interestingly, the temperature settings (0 or 1) and forecasting methods (bio-based vs. characteristics-based) do not seem to consistently impact the bias in a significant or predictable way. Both methods and temperature settings show a mix of over- and under-forecasting tendencies, though individual forecasters exhibit different tendencies.

The two charts highlight the increasing difficulty of making accurate forecasts as the forecast horizon grows. MSE values climb for all AI-generated models, reflecting a decline in accuracy over time. Bias remains mixed across all horizons, with no clear trend based on the temperature settings or the forecasting methods used. This suggests that while some forecasters perform better than others, the bio-based and characteristics-based methods, as well as the temperature settings, may have complex, context-dependent effects on the accuracy and bias of the forecasts.

8. Explaining Forecast Decisions

We collect not only numeric predictions from our synthetic forecasters, but also explanations as to why each prediction was made in a certain way.

We use two approaches to categorize these explanations. First, we use an LLM, as shown in Table 16. We ask ChatGPT to review 1,000 responses and group them based

on keywords it identifies.[14] The resulting categories are broad and intuitive, such as *Consumer Demand*, *Monetary Policy*, *Supply Chain & Logistics*, and *Wage Growth & Labor Market*, reflecting common considerations behind forecast decisions. We see, for example, that Monetary Policy stands out as a frequent topic, while categories like *Supply Chain & Logistics* and *Economic Recovery* are also common.

The second approach, summarized in Table 17, uses LDA, a machine learning method that finds latent topics based on word co-occurrence.[15] LDA produces more granular topics, such as *Short-term Economic Conditions*, *Supply Chain Disruptions*, *Consumer Behavior*, and *Economic Indicators*. For instance, *Economic Indicators* dominate in models like SPF3 and SPF4, showing the synthetic forecasters focus on quantitative measures of economic performance. LDA also captures niche topics like *Consumer Behavior* and *Labor Market Conditions*.

9. Discussion

We develop a framework using large language models to generate synthetic predictions for the Survey of Professional Forecasters (SPF). By feeding LLMs a dataset of individual forecaster traits and real-time macroeconomic data, we replicate the forecast generation process across various economic variables. Our results show that LLMs can effectively mimic human forecasters, revealing patterns of disagreement influenced by behavioral and personal traits. This method addresses traditional survey limitations—such as cost, frequency, and bias—while offering deeper insights into economic forecasting behaviors.

Our study highlights AI's potential for reliable economic forecasts, particularly

[14]This method is quick and leverages ChatGPT's ability to detect patterns in text, categorizing responses based on key words or themes.

[15]LDA uncovers hidden structures within the text, leading to more specific, data-driven topics.

excelling at short-term predictions, though there remains room for improvement over longer horizons. We also find that persona configurations impact AI forecast performance, with different configurations performing better at various horizons. Lastly, GPT-4 models demonstrate strong accuracy, especially regarding prediction errors.

In sum, AI models, particularly when using personas, outperform human forecasters in short-term predictions and remain competitive in medium-term forecasts. Going forward, hybrid approaches that combine AI with human judgment could enhance long-term forecasting accuracy and reliability.

References

Aher, G. V., Arriaga, R. I., and Kalai, A. T. (2023). Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. In *International Conference on Machine Learning*, pages 337–371. PMLR.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., and Wingate, D. (2023). Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 31(3).

Batchelor, R. (2007). Bias in Macroeconomic Forecasts. *International Journal of Forecasting*, 23(2):189–203.

Benchimol, J., El-Shagi, M., and Saadon, Y. (2022). Do Expert Experience and Characteristics Affect Inflation Forecasts? *Journal of Economic Behavior & Organization*, 201:205–226.

Bordalo, P., Gennaioli, N., Ma, Y., and Shleifer, A. (2020). Overreaction in Macroeconomic Expectations. *American Economic Review*, 110(9):2748–2782.

Brand, J., Israeli, A., and Ngwe, D. (2023). Using GPT for Market Research. *SSRN Working Paper*.

Bybee, J. L. (2023). The Ghost in the Machine: Generating Beliefs with Large Language Models. *SSRN Working Paper*.

Croushore, D. D. (1993). Introducing: The Survey Of Professional Forecasters. *Business Review-Federal Reserve Bank Of Philadelphia*, 6:3.

Dillion, D., Tandon, N., Gu, Y., and Gray, K. (2023). Can AI Language Models Replace Human Participants? *Trends in Cognitive Sciences*.

Fajardo, J. and Dantas, M. (2018). Understanding The Impact Of Severe Hyperinflation Experience On Current Household Investment Behavior. *Journal Of Behavioral And Experimental Finance*, 17:60–67.

Faria-e Castro, M., Louis, F. S., and Leibovici, F. (2023). Artificial Intelligence and Inflation Forecasts. Technical report.

Fedyk, A., Kakhbod, A., Li, P., and Malmendier, U. (2024). ChatGPT and Perception Biases in Investments: An Experimental Study. *Available at SSRN 4787249*.

Gabaix, X. (2020). A Behavioral New Keynesian Model. *American Economic Review*, 110(8):2271–2327.

Grossmann, I., Feinberg, M., Parker, D. C., Christakis, N. A., Tetlock, P. E., and Cunningham, W. A. (2023). AI and the Transformation of Social Science Research. *Science*, 380(6650).

Gurnee, W. and Tegmark, M. (2023). Language Models Represent Space and Time. *arXiv preprint arXiv:2310.02207*.

Hansen, A. L. and Kazinnik, S. (2023). Can ChatGPT Decipher Fedspeak? *SSRN Working Paper*.

Hayes, W. M., Yax, N., and Palminteri, S. (2024). Relative Value Biases in Large Language Models. *arXiv preprint arXiv:2401.14530*.

Hong, H. and Kacperczyk, M. (2010). Competition and Bias. *The Quarterly Journal of Economics*, 125(4):1683–1725.

Horton, J. J. (2023). Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus? Technical report, National Bureau of Economic Research.

Huang, R., Pilbeam, K., and Pouliot, W. (2022). Are Macroeconomic Forecasters Optimists or Pessimists? A Reassessment of Survey Based Forecasts. *Journal of Economic Behavior and Organization*, 197.

Immorlica, N., Lucier, B., and Slivkins, A. (2024). Generative AI As Economic Agents. *ArXiv Preprint ArXiv:2406.00477*.

Jiang, H., Zhang, X., Cao, X., Kabbara, J., and Roy, D. (2023). PersonaLLM: Investigating the Ability of GPT-3.5 to Express Personality Traits and Gender Differences. *arXiv preprint arXiv:2305.02547*.

Koralus, P. and Wang-Maścianica, V. (2023). Humans In Humans Out: On GPT Converging Toward Common Sense in Both Success and Failure. *arXiv preprint arXiv:2303.17276*.

Maćkowiak, B. and Wiederholt, M. (2015). Business Cycle Dynamics Under Rational Inattention. *The Review of Economic Studies*, 82(4):1502–1532.

Malmendier, U. and Nagel, S. (2011). Depression Babies: Do Macroeconomic Experiences Affect Risk Taking? *The Quarterly Journal Of Economics*, 126(1):373–416.

Malmendier, U. and Nagel, S. (2016). Learning from Inflation Experiences. *The Quarterly Journal of Economics*, 131(1):53–87.

Mankiw, N. G. and Reis, R. (2002). Sticky information vs. sticky prices: a proposal

to replace the new keynesian phillips curve. *The Quarterly Journal of Economics*, 117(4):1295–1328.

Manzan, S. (2011). Differential Interpretation in the Survey of Professional Forecasters. *Journal of Money, Credit and Banking*, 43(5).

Manzan, S. (2021). Are Professional Forecasters Bayesian? *Journal of Economic Dynamics and Control*, 123(104045).

Orphanides, A. and Williams, J. C. (2002). Robust Monetary Policy Rules with Unknown Natural Rates. *Brookings Papers on Economic Activity*.

Orphanides, A. and Williams, J. C. (2007). Robust Monetary Policy with Imperfect Knowledge. *Journal of Monetary Economics*, 54(5).

Patton, A. J. and Timmermann, A. (2010). Why Do Forecasters Disagree? Lessons from the Term Structure of Cross-Sectional Dispersion. *Journal of Monetary Economics*, 57.

Rudebusch, G. D. (2002). Assessing Nominal Income Rules for Monetary Policy with Model and Data Uncertainty. *Economic Journal*, 83(2):402–432.

Stark, T. (2013). SPF Panelists' Forecasting Methods: A Note on the Aggregate Results of a November 2009 Special Survey. *Federal Reserve Bank of Philadelphia*.

Zarifhonarvar, A. (2024). Experimental Evidence on Large Language Models. *Available at SSRN 4825076*.

Zhu, J.-Q. and Griffiths, T. L. (2024). Incoherent Probability Judgments in Large Language Models. *arXiv preprint arXiv:2401.16646*.

Figures

Figure 1: Comparison of the forecasting accuracy between AI (blue line) and SPF (red line) forecasters over time, using a 1-year moving average. (1-year moving average)

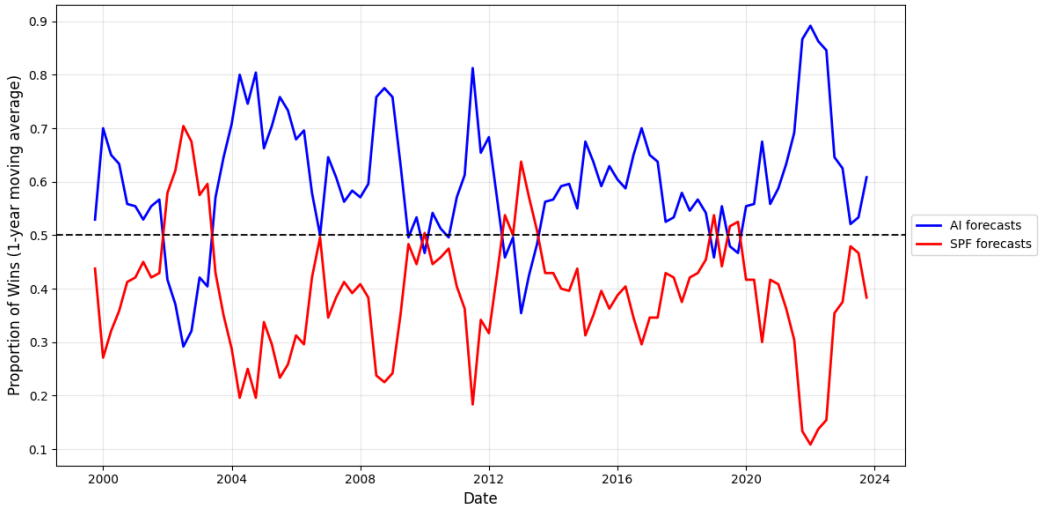


Figure 2: The number of quarters in the full survey where each forecaster, either AI (represented in blue) or SPF (represented in red), has a lower mean squared error (MSE).

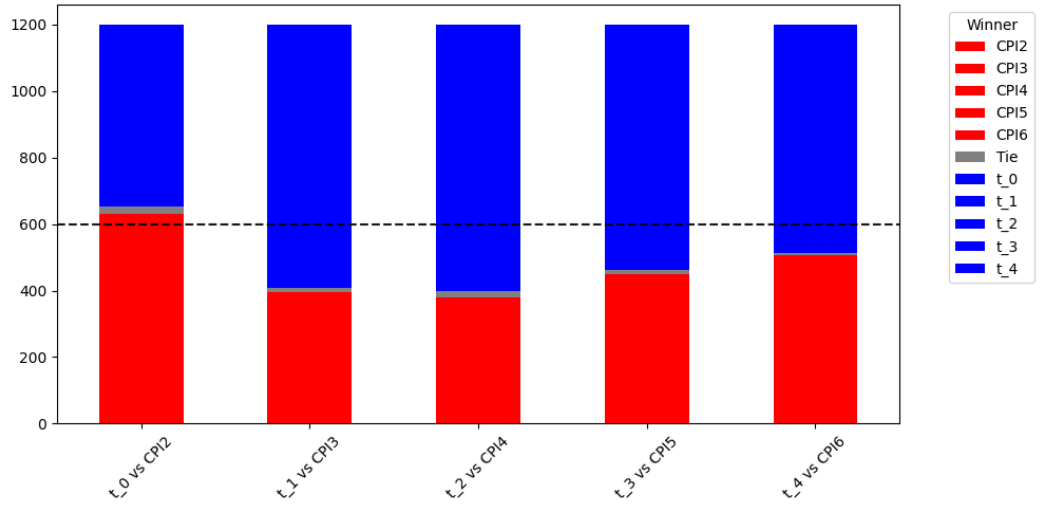


Figure 3: The chart shows how forecast accuracy and bias vary across different time horizons for each model.

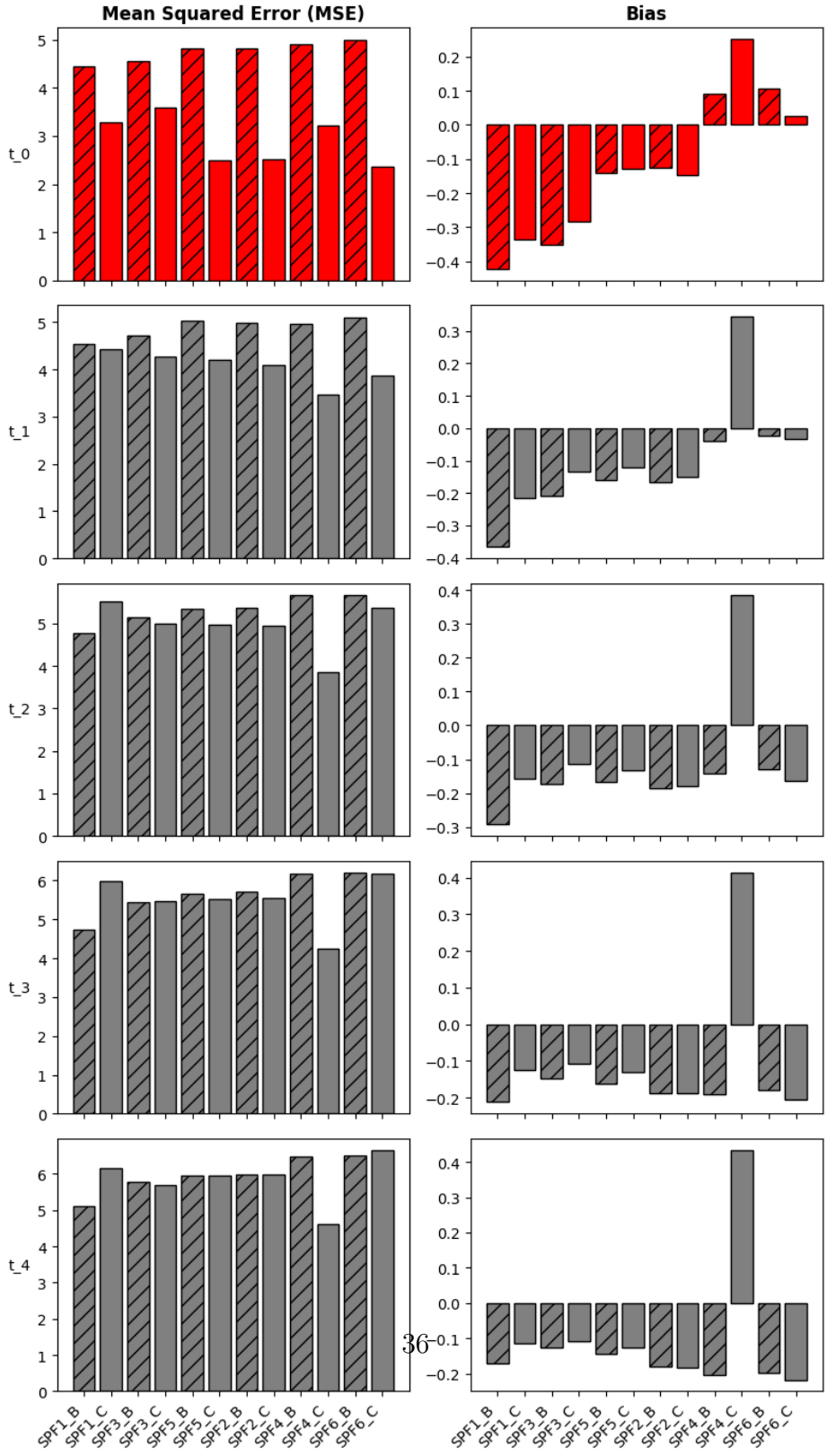


Figure 4: Comparison of MSE Trends (GPT4, Temperature=1, Biography-Based vs Characteristic-Based Personas)

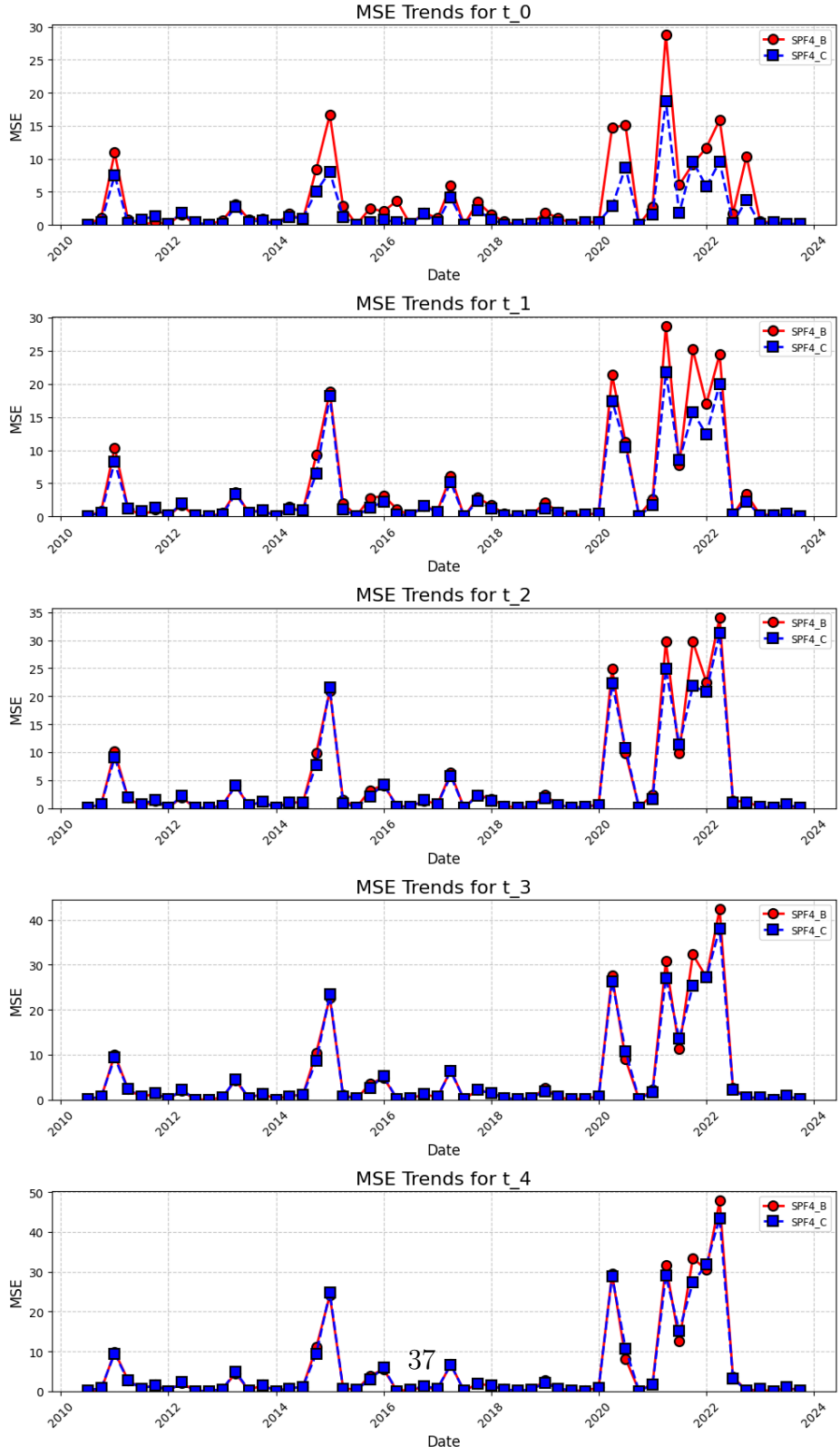


Figure 5: Median SPF vs. Median AI (Characteristics) vs. Realized CPI

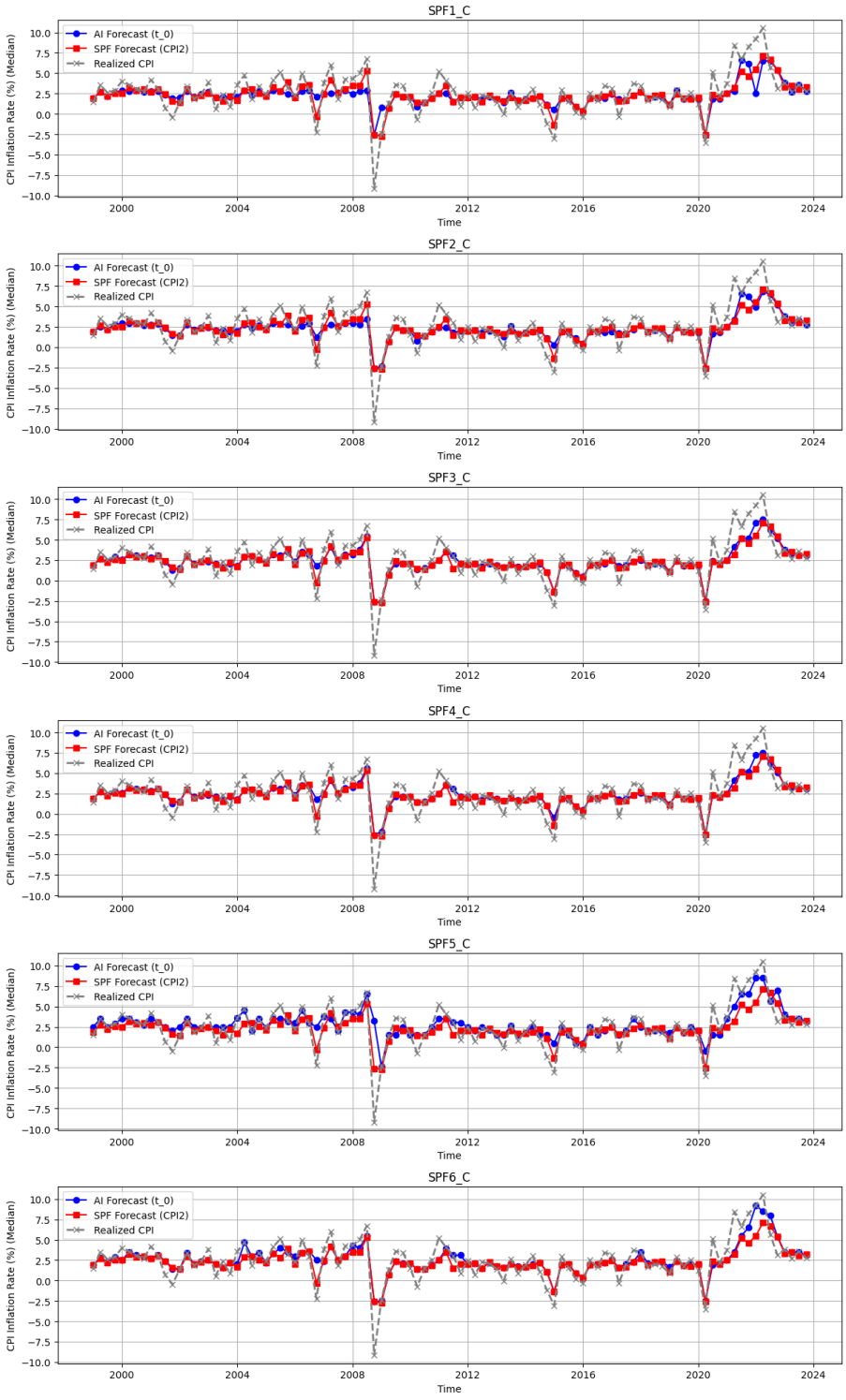


Figure 6: Mean SPF vs. Mean AI (Characteristics) vs. Realized CPI

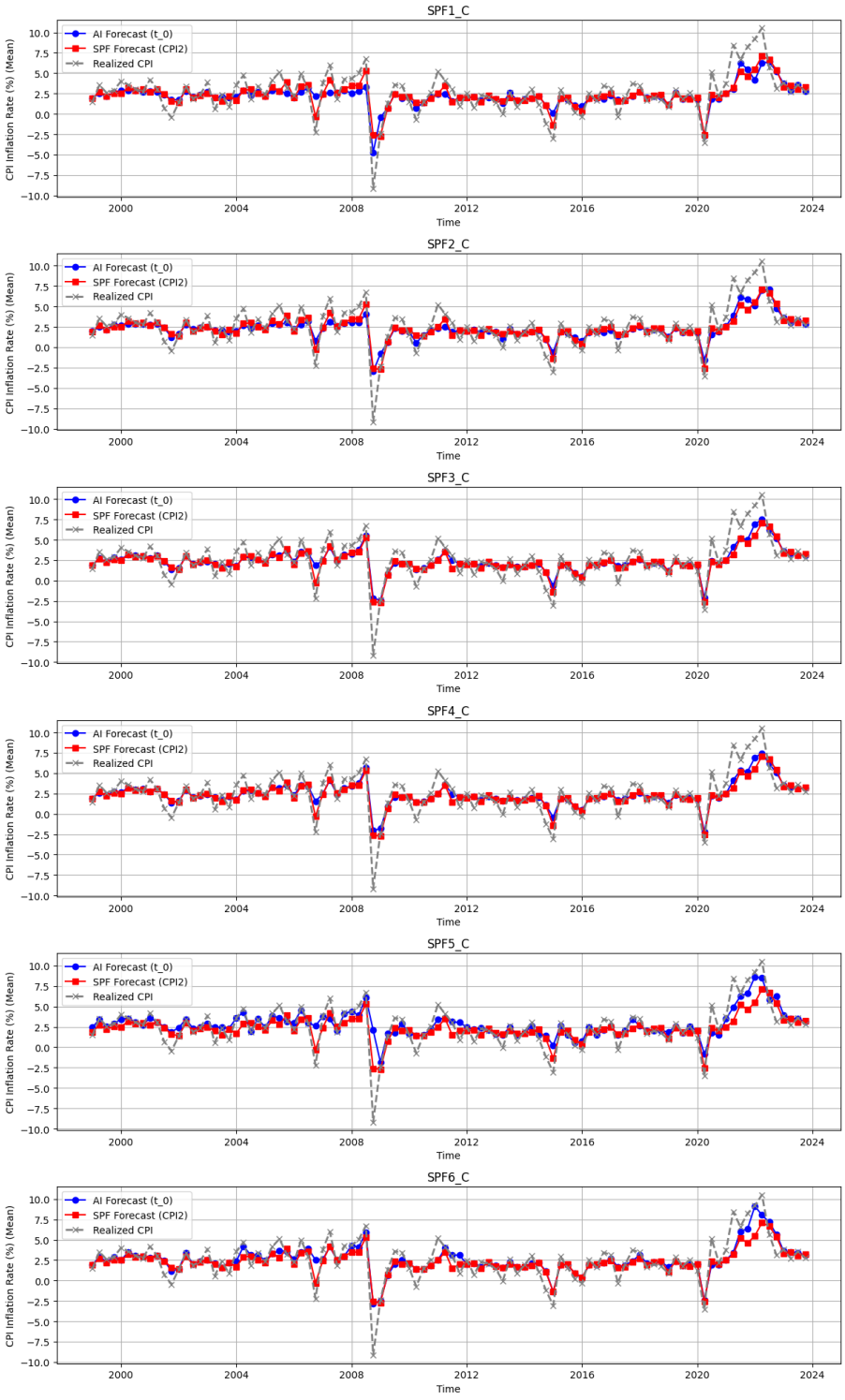
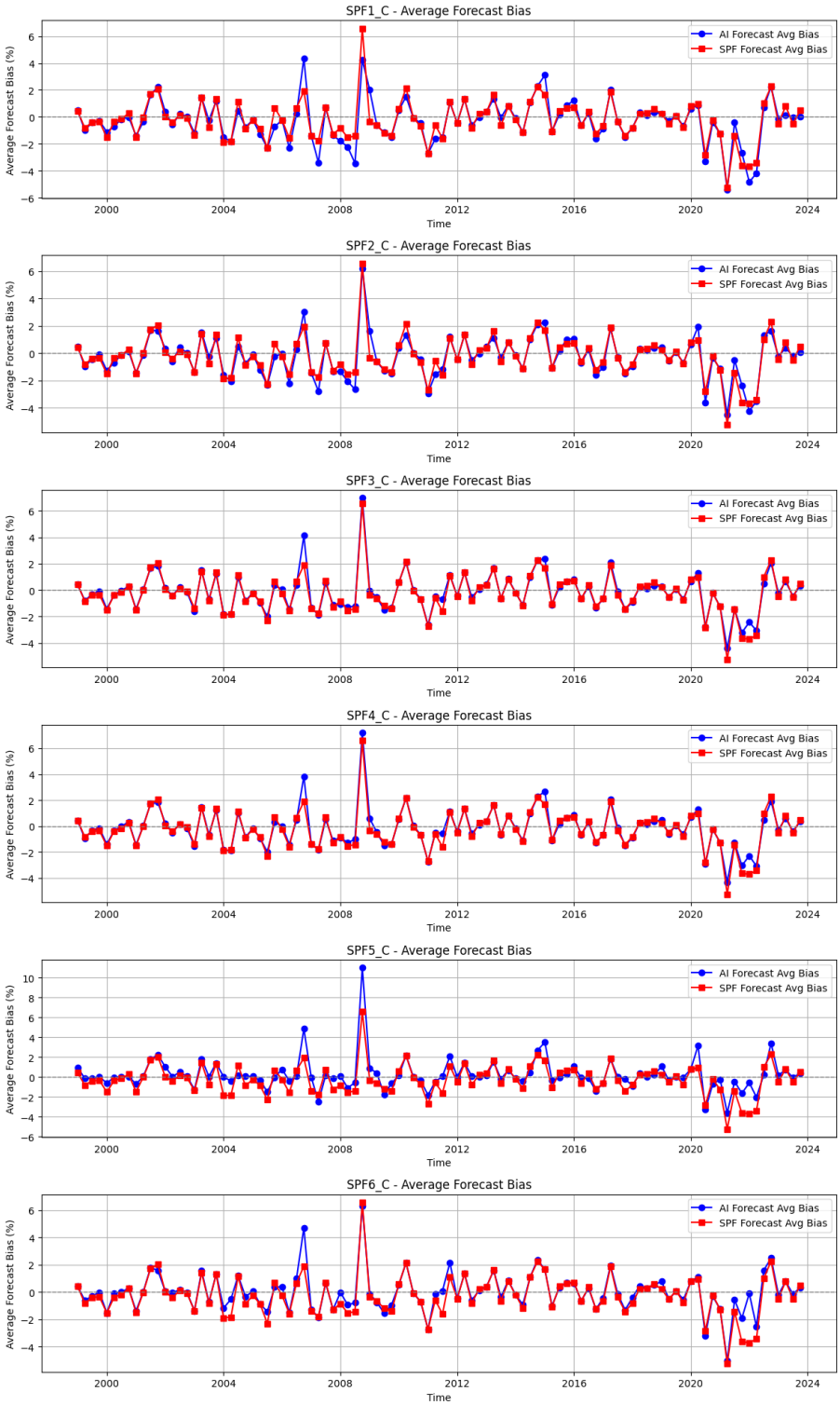


Figure 7: Average Bias per AI Model for Nowcast (t_0) vs SPF



Tables

Table 1: Forecaster Characteristics

Characteristic	Description
Name	Full name of the forecaster as listed in the SPF acknowledgement panel.
Gender (inferred)	Gender inferred based on publicly available information (e.g., name, social media).
Affiliation	The institution or company the forecaster is associated with (e.g., university, bank).
Affiliation type	Type of organization the forecaster works for, such as academic, financial, forecasting center, government, etc.
Position in given year	The forecaster's job title or role when the forecast was made (e.g., professor, economist).
Highest degree earned	The highest degree the forecaster has earned (e.g., Bachelor, Master, PhD).
Degree area/field	The field of study for the forecaster's degree (e.g., Economics, Finance).
Graduation year	The year the forecaster completed their highest degree, used to estimate experience and age.
Alma mater	The institution where the forecaster earned their highest degree.
Company location	The geographic location of the forecaster's affiliated institution or company.
Gives interviews	Whether the forecaster gives public interviews or engages with media (yes/no).
Has Twitter	Whether the forecaster has a public Twitter profile (yes/no).

Note: Gender is inferred from publicly available sources and may not be fully accurate. This table does not reflect the full breadth of forecaster characteristics.

Table 2: Example of Characteristic-Based and Bio-Based Personas

Characteristic-Based Persona	**Bio-Based Persona**
You are a participant on a panel of Survey of Professional Forecasters. Your name is [redacted], you graduated from University of New Hampshire with a M.A. in Economics around 2014. Today, you work as a Senior Quantitative Analyst at Columbia Threadneedle Investments, an Asset Management organization based in Boston, Massachusetts. You are originally from the USA.	You are a Senior Quantitative Analyst at Columbia Threadneedle with a Master's degree in Economics. Your expertise lies in macroeconomic forecasting, and you actively participate in the Survey of Professional Forecasters for the Federal Reserve Bank of Philadelphia. Known for your meticulous approach to data analysis, you excel in interpreting complex economic information and translating it into actionable forecasts. Your experience at Columbia Threadneedle has honed your forecasting skills, making you a trusted source of economic predictions in the financial industry. Your dedication to staying informed about the latest economic developments sets you apart as a reliable and knowledgeable forecaster.

Note: The name in the characteristic-based prompt is removed here for privacy reasons. However, we feed the models all names that were featured on the acknowledgement SPF lists.

Table 3: Forecasted SPF Variables

Variable	Definition
	Section 1. U.S. Business Indicators
NGDP	Nominal Gross Domestic Product (GDP), levels, seasonally adjusted, annual rate, in billions USD.
PGDP	Gross Domestic Product (GDP) price index, chain weighted, seasonally adjusted, 2010 base year.
CPROF	Level of nominal corporate profits after tax excluding inventory valuation adjustment (IVA) and capital consumption adjustment (CCAdj). Seasonally adjusted, annual rate, in billions USD.
UNEMP	Civilian unemployment rate, levels, seasonally adjusted, percentage points.
EMP	Nonfarm payroll employment, seasonally adjusted, in thousands of jobs.
INDPROD	Industrial production index, seasonally adjusted, index, 2010 base year.
HOUSING	Housing starts, seasonally adjusted, annual rate, in millions USD.
TBILL	T-Bill rate, 3-month, levels, percentage points.
BOND	AAA corporate bond yield, levels, percentage points. Both quarterly and annual forecasts are for averages of daily levels.
	Section 2. Real GDP and Its Components
TBOND	Treasury bond rate, 10-year, levels, percentage points.
RGDP	Real Gross Domestic Product, seasonally adjusted, annual rate, 2010 base year.
RCONSUM	Real personal consumption expenditures, seasonally adjusted, annual rate, 2010 base year.
RNRESIN	Real nonresidential fixed investment, seasonally adjusted, annual rate, 2010 base year.
RRESINV	Real residential fixed investment, seasonally adjusted, annual rate, 2010 base year.
RFEDGOV	Real federal government consumption and gross investment, seasonally adjusted, annual rate, 2010 base year.
RSLGOV	Real state and local government consumption and gross investment, seasonally adjusted, annual rate, 2010 base year.
RCBI	Real change in private inventories, levels, chain-weighted real change in private inventories. Seasonally adjusted, annual rate, 2010 base year.
REXPORT	Real net exports of goods and services, levels, chain-weighted real net exports. Seasonally adjusted, annual rate, 2010 base year.
	Section 3. CPI and PCE Inflation
CPI	Headline consumer price index (CPI), seasonally adjusted, annual rate, percentage points. Quarterly and annual forecasts for changes in quarterly and fourth-quarter levels.
CORECPI	Core CPI Inflation Rate - Core CPI inflation rate, seasonally adjusted, annual rate, percentage points.
PCE	PCE inflation rate, headline chain-weighted, seasonally adjusted, annual rate, percentage points.
COREPCE	Core PCE inflation rate, core chain-weighted PCE inflation rate, seasonally adjusted, annual rate, percentage points.

Note: This table lists all of the variables and definitions that are forecasted by the SPF panel.

Table 4: Description of the Large Language Models (LLMs)

Model	**Temperature = 0**		**Temperature = 1**	
	Biography Based	**Characteristic Based**	**Biography Based**	**Characteristic Based**
`GPT-3.5`	SPF1 B	SPF1 C	SPF2 B	SPF2 C
`GPT-4`	SPF3 B	SPF3 C	SPF4 B	SPF4 C
`GPT-4o-mini`	SPF5 B	SPF5 C	SPF6 B	SPF6 C

Note: There are 12 models, each with a different temperature setting, based on three types of GPT models, with two way of persona definition.

Table 5: Time Horizon for each Forecast

	Variable	**Description**
AI Forecaster	`t_0`	Forecast for the current quarter
	`t_1`	Forecast for the next quarter
	`t_2`	Forecast for two quarters ahead
	`t_3`	Forecast for three quarters ahead
	`t_4`	Forecast for four quarters ahead
Professional Forecaster	`CPI2`	Forecast for the current quarter
	`CPI3`	Forecast for the next quarter
	`CPI4`	Forecast for two quarters ahead
	`CPI5`	Forecast for three quarters ahead
	`CPI6`	Forecast for four quarters ahead

Table 6: Summary Statistics of the Survey of Professional Forecasters and Realized CPI

Statistic	CPI2	CPI3	CPI4	CPI5	CPI6	Realized CPI
Mean	2.34	2.25	2.23	2.25	2.27	2.57
Std	1.52	0.63	0.40	0.30	0.23	2.64
Min	-2.72	0.80	1.29	1.60	1.80	-9.17
Median	2.29	2.20	2.20	2.21	2.25	2.63
Max	7.10	4.53	3.73	3.40	3.03	10.53

Note: The table summarizes CPI forecasts for different horizons (CPI2-CPI6) and the realized CPI. Forecasts are generated by the panel of professional forecasters.

Table 7: Summary Statistics of All SPF Models of AI Forecasts

Statistic	t_0	t_1	t_2	t_3	t_4
Mean	2.44	2.45	2.43	2.43	2.44
Std	1.67	1.16	0.87	0.72	0.65
Min	-17.90	-15.12	-14.26	-13.95	-14.38
Median	2.30	2.30	2.30	2.30	2.30
Max	16.20	12.80	12.40	12.20	12.50

Note: The table summarizes forecast statistics for different time horizons (t_0 to t_4) across 12 models. These models vary in temperature settings and are based on three types of GPT models, with two persona definition methods.

Table 8: Summary statistics for GPT 3.5 - Bio Based

	SPF1_B					SPF2_B				
Statistic	t_0	t_1	t_2	t_3	t_4	t_0	t_1	t_2	t_3	t_4
Mean	2.16	2.22	2.28	2.32	2.37	2.21	2.36	2.39	2.41	2.44
Std	1.97	1.57	1.25	1.02	0.90	1.86	1.40	1.12	0.99	0.95
Min	-9.17	-9.17	-9.17	-9.17	-9.17	-9.20	-15.12	-14.26	-13.95	-14.38
Median	2.10	2.20	2.30	2.30	2.30	2.20	2.30	2.30	2.30	2.30
Max	11.20	10.80	10.40	9.90	9.50	16.20	12.70	12.20	12.20	12.50

Note: The table summarizes forecast statistics for different time horizons (t_0 to t_4) across two sets of models (SPF1_B and SPF2_B), varying by temperature settings and using GPT 3.5 with bio-based personas.

Table 9: Summary statistics for GPT 3.5 - Characteristic Based

	SPF1_C					SPF2_C				
Statistic	t_0	t_1	t_2	t_3	t_4	t_0	t_1	t_2	t_3	t_4
Mean	2.23	2.35	2.41	2.43	2.46	2.28	2.43	2.44	2.45	2.46
Std	1.54	1.04	0.83	0.75	0.72	1.70	1.22	0.99	0.88	0.83
Min	-9.17	-9.17	-9.17	-9.17	-9.17	-17.90	-9.20	-7.00	-5.00	-3.65
Median	2.15	2.30	2.30	2.30	2.30	2.20	2.30	2.30	2.30	2.30
Max	9.20	8.50	7.20	7.50	7.80	12.70	12.80	12.40	11.90	12.20

Note: The table summarizes forecast statistics for different time horizons (t_0 to t_4) across two sets of models (SPF1_C and SPF2_C), varying by temperature settings and using GPT 3.5 with characteristic-based personas.

Table 10: Summary statistics for GPT 4 - Bio Based

	SPF3_B					SPF4_B				
Statistic	t_0	t_1	t_2	t_3	t_4	t_0	t_1	t_2	t_3	t_4
Mean	2.43	2.41	2.40	2.41	2.43	2.44	2.40	2.39	2.38	2.39
Std	1.29	0.86	0.66	0.53	0.45	1.28	0.88	0.67	0.53	0.45
Min	-2.72	0.50	1.00	1.40	1.30	-2.72	0.20	-0.20	0.60	0.50
Median	2.10	2.20	2.30	2.30	2.40	2.10	2.20	2.25	2.30	2.30
Max	7.50	6.80	6.10	5.40	4.90	7.80	6.90	6.10	5.50	5.20

Note: The table summarizes forecast statistics for different time horizons (t_0 to t_4) across two sets of models (SPF3_B and SPF4_B), varying by temperature settings and using GPT 4 with bio-based personas.

Table 11: Summary statistics for GPT 4 - Characteristic Based

	SPF3_C					SPF4_C				
Statistic	t_0	t_1	t_2	t_3	t_4	t_0	t_1	t_2	t_3	t_4
Mean	2.43	2.42	2.41	2.41	2.41	2.43	2.43	2.40	2.39	2.39
Std	1.44	0.94	0.72	0.58	0.50	1.49	0.95	0.70	0.55	0.47
Min	-3.50	0.50	0.80	1.10	1.30	-3.70	0.25	0.80	1.20	1.30
Median	2.30	2.20	2.30	2.30	2.30	2.20	2.20	2.20	2.30	2.30
Max	7.80	6.50	5.80	5.20	4.90	7.80	6.90	5.90	5.50	5.20

Note: The table summarizes forecast statistics for different time horizons (t_0 to t_4) across two sets of models (SPF3_C and SPF4_C), varying by temperature settings and using GPT 4 with characteristic-based personas.

Table 12: Summary statistics for GPT 4o-mini - Bio Based

	SPF5_B					SPF6_B				
Statistic	t_0	t_1	t_2	t_3	t_4	t_0	t_1	t_2	t_3	t_4
Mean	2.66	2.53	2.43	2.38	2.36	2.68	2.55	2.44	2.39	2.37
Std	1.88	1.13	0.67	0.45	0.34	1.82	1.13	0.69	0.47	0.37
Min	-9.20	-5.00	0.50	1.00	1.50	-10.00	-5.50	-2.00	1.00	1.00
Median	2.40	2.40	2.40	2.30	2.40	2.40	2.40	2.30	2.30	2.30
Max	9.20	7.50	5.80	4.50	4.20	10.50	8.50	7.50	6.00	5.00

Note: The table summarizes forecast statistics for different time horizons (t_0 to t_4) across two sets of models (SPF5_B and SPF6_B), varying by temperature settings and using GPT 4o-mini with bio-based personas.

Table 13: Summary statistics for GPT 4o-mini - Characteristic Based

	SPF5_C					SPF6_C				
Statistic	t_0	t_1	t_2	t_3	t_4	t_0	t_1	t_2	t_3	t_4
Mean	2.87	2.98	3.03	3.07	3.10	2.60	2.54	2.41	2.36	2.35
Std	1.60	1.38	1.16	1.03	0.95	1.80	1.13	0.61	0.41	0.32
Min	-5.00	-3.00	-1.80	-1.20	-0.50	-9.20	-6.00	0.50	1.00	1.00
Median	2.50	2.70	2.80	2.80	2.90	2.40	2.40	2.30	2.30	2.30
Max	9.20	8.50	7.80	7.50	7.30	9.50	8.50	6.00	5.50	5.20

Note: The table summarizes forecast statistics for different time horizons (t_0 to t_4) across two sets of models (SPF5_C and SPF6_C), varying by temperature settings and using GPT 4o-mini with characteristic-based personas.

Table 14: AI Winners by Accuracy

	t_0	t_1	t_2	t_3	t_4
MAE	SPF5_C	SPF3_B	SPF3_B	SPF3_C	SPF3_C
RMSE	SPF6_C	SPF4_B	SPF4_B	SPF4_B	SPF4_B
MAPE (%)	SPF5_C	SPF3_C	SPF3_C	SPF3_C	SPF3_C
Directional Accuracy (%)	SPF3_C	SPF6_B	SPF1_C	SPF5_C	SPF2_C
Forecast Bias	SPF1_B	SPF1_B	SPF1_B	SPF1_B	SPF1_B

Note: The table summarizes the AI models that performed best according to different accuracy metrics across time horizons (t_0 to t_4), with MAE, RMSE, MAPE, directional accuracy, and forecast bias considered.

Table 15: Winners (AI vs Original SPF. Average of Each Measure)

	t_0	t_1	t_2	t_3	t_4
MAE	OSPF	OSPF	OSPF	OSPF	OSPF
RMSE	OSPF	OSPF	OSPF	OSPF	OSPF
MAPE (%)	AI	OSPF	OSPF	OSPF	OSPF
Directional Accuracy (%)	OSPF	OSPF	AI	OSPF	AI
Forecast Bias	AI	AI	OSPF	OSPF	OSPF

Note: The table compares AI-generated forecasts with Original SPF forecasts across different accuracy metrics and time horizons (t_0 to t_4). Measures include MAE, RMSE, MAPE, directional accuracy, and forecast bias.

Table 16: Topics Categorized by ChatGPT

	S1B	S1C	S2B	S2C	S3B	S3C	S4B	S4C	S5B	S5C	S6B	S6C
Consumer Demand	1,533	1,621	1,214	1,364	357	351	475	470	68	357	196	146
Monetary Policy	1,233	815	1,372	1,059	2,636	2,695	2,605	2,575	2,462	1,894	2,415	2,164
Inflation Expectations	720	1,014	738	1,058	1	15	18	29	0	12	2	6
Other/Unspecified	712	555	553	613	5	3	8	15	0	1	1	5
Supply Chain	585	862	584	615	841	849	669	656	1,638	1,550	1,434	1,526
Energy, Commodity	342	212	375	188	18	8	52	39	2	0	16	4
Global Factors	154	146	57	39	12	9	12	11	0	16	6	1
Economic Recovery	119	65	343	273	1,546	1,348	1,577	1,487	1,346	1,536	1,439	1,395
Labor Market	87	68	221	173	95	119	93	127	0	2	5	11
Fiscal Policy	23	40	49	37	4	0	6	7	0	0	2	2
Sector-Specific Inflation	8	0	9	3	0	1	0	0	0	0	0	0

Note: The table categorizes topics identified by ChatGPT across twelve sets of models (SPF1B to SPF6C). The values represent the frequency of mentions of each topic in the respective model sets.

Table 17: LDA Based Topics

	S1B	S1C	S2B	S2C	S3B	S3C	S4B	S4C	S5B	S5C	S6B	S6C
Economic Conditions	622	585	656	750	18	25	23	30	1	2	16	38
Supply Chain	724	907	604	723	63	27	51	46	147	213	111	127
Consumer Behavior	17	6	42	45	116	84	202	144	432	1704	551	456
Economic Indicators	47	51	327	316	2,443	2,991	2,400	2,640	69	31	181	277
Demand Side	162	67	332	192	805	391	706	528	2,347	2,247	1,921	1,645
Labor Market	1,765	1,659	1,357	1,266	203	94	157	100	48	44	161	126
Economic Outlook	875	1,418	618	836	3	2	9	10	76	8	22	51
Monetary Policy	97	33	182	119	1,454	1,403	1,375	1,311	1,972	664	1,949	1,901
Economic Recovery	230	51	274	174	397	349	515	508	393	454	484	495
Trends & Indicators	977	621	1,123	1,001	13	32	77	99	31	1	120	144

Note: The table categorizes topics identified through LDA across twelve sets of models (S1B to S6C). The values represent the frequency of mentions of each topic in the respective model sets.

Table 18: Simulated Forecast Statistics by Affiliation

Affiliation	**Count**	**Mean**	**Median**	**S. D.**	**Bias**	**MSE**
Academic Institutions	704	2.435	2.1	1.533	-0.161	2.574
Banking and Financial Services	1,108	2.365	2.1	1.481	-0.074	2.473
Consulting and Research	1,512	2.432	2.1	1.527	-0.192	2.711
Government and Non-Profit	132	2.671	2.3	1.677	-0.087	3.047
Other Industries	808	2.434	2.2	1.45	-0.159	2.075
Affiliation Unknown	1,152	2.467	2.3	1.406		

Note: The table summarizes simulated forecast statistics across different affiliations. The statistics include the count of forecasts, mean, median, standard deviation, bias, and mean squared error (MSE) for each subgroup.

Table 19: Simulated Forecast Statistics by Title

Title	**Count**	**Mean**	**Median**	**S. D.**	**Bias**	**MSE**
Academic Role	258	2.424	2.1	1.709	-0.216	2.58
Chief Economist	1,798	2.381	2.1	1.501	-0.142	2.485
Consultant/Analyst Role	274	2.467	2.125	1.497	-0.256	1.873
Executive Role	908	2.498	2.3	1.54	-0.172	2.713
Other	126	2.709	2.3	1.647	-0.184	2.999
Other Economist	234	2.504	2.3	1.422	-0.1	2.277
Portfolio Manager	34	1.815	1.9	1.453	0.434	4.201
Senior Economist	356	2.408	2.1	1.536	-0.069	2.768
Role Unknown	1,428	2.434	2.3	1.36	-0.111	2.101

Note: The table summarizes simulated forecast statistics across different job titles. The statistics include the count of forecasts, mean, median, standard deviation, bias, and mean squared error (MSE) for each title type.

Table 20: Simulated Forecast Statistics by Education Group

Education	**Count**	**Mean**	**Median**	**S. D.**	**Bias**	**MSE**
Bachelor's Degree	52	2.007	2.1	1.453	0.053	2.122
MBA	114	2.679	2.55	1.476	-0.327	1.955
Master's Degree	654	2.42	2.1	1.479	-0.137	2.428
Other	268	2.596	2.3	1.649	-0.357	2.97
Ph.D.	2,558	2.406	2.1	1.549	-0.133	2.565
Education Unknown	1,770	2.446	2.3	1.363	-0.108	2.35

Note: The table summarizes simulated forecast statistics across different education groups. The statistics include the count of forecasts, mean, median, standard deviation, bias, and mean squared error (MSE) for each education level.

Table 21: Simulated Forecast Statistics by Gender Group

Gender	**Count**	**Mean**	**Median**	**Standard Deviation**	**Bias**	**MSE**
Female	542	2.713	2.5	1.651	-0.336	2.917
Male	4,792	2.408	2.1	1.467	-0.127	2.485
Unknown	82	1.963	2.1	1.107	-0.018	1.731

Note: The table summarizes simulated forecast statistics across different gender groups. The statistics include the count of forecasts, mean, median, standard deviation, bias, and mean squared error (MSE) for each group.

Appendix A. Survey of Professional Forecasters Statistics

Figure A.8: An example of forecasters acknowledgement list from last quarter of 2020, published by the Federal Reserve Bank of Philadelphia thanking forecasters for their participation in recent surveys.

The Federal Reserve Bank of Philadelphia thanks the following forecasters for their participation in recent surveys:

Lewis Alexander, Nomura Securities; **Scott Anderson**, Bank of the West (BNP Paribas Group); **Robert J. Barbera**, Johns Hopkins University Center for Financial Economics; **Peter Bernstein**, RCF Economic and Financial Consulting, Inc.; **Wayne Best** and **Michael Brown**, Visa, Inc.; **Jay Bryson**, Wells Fargo; **J. Burton**, **G. Ehrlich**, **D. Manaenkov**, and **T. Ranoso**, RSQE, University of Michigan; **Christine Chmura, Ph.D.**, and **Xiaobing Shuai, Ph.D.**, Chmura Economics & Analytics; **Gary Ciminero**, **CFA**, GLC Financial Economics; **Gregory Daco**, Oxford Economics USA, Inc.; **Rajeev Dhawan**, Georgia State University; **Bill Diviney**, ABN AMRO Bank NV; **Michael R. Englund**, Action Economics, LLC; **Sacha Gelfer**, Bentley University; **James Glassman**, JPMorgan Chase & Co.; **Jan Hatzius**, Goldman Sachs; **Brian Higginbotham**, U.S. Chamber of Commerce; **Fred Joutz**, Benchmark Forecasts; **Sam Kahan**, Kahan Consulting Ltd. (ACT Research LLC); **N. Karp**, BBVA Research USA; **Walter Kemmsies** and **Ryan Severino**, Jones Lang LaSalle; **Jack Kleinhenz**, Kleinhenz & Associates, Inc.; **Rohan Kumar**, Decision Economics, Inc.; **Thomas Lam**, Sim Kee Boon Institute, Singapore Management University; **John Lonski**, Moody's Capital Markets Group; **Matthew Luzzetti**, Deutsche Bank Securities; **IHS Markit**; **Robert McNab**, Old Dominion University; **R. Anthony Metz**, Pareto Optimal Economics; **R. M. Monaco**, TitanRM; **Michael Moran**, Daiwa Capital Markets America; **Joel L. Naroff**, Naroff Economic Advisors; **Brendon Ogmundson**, BC Real Estate Association; **Perc Pineda, Ph.D.**, Plastics Industry Association; **Philip Rothman**, East Carolina University; **Chris Rupkey**, MUFG Union Bank; **Sean M. Snaith**, **Ph.D.**, University of Central Florida; **Constantine G. Soras, Ph.D.**, CGS Economic Consulting, Inc.; **Stephen Stanley**, Amherst Pierpont Securities; **Charles Steindel**, Ramapo College of New Jersey; **Susan M. Sterne**, Economic Analysis Associates, Inc.; **James Sweeney**, Credit Suisse; **Thomas Kevin Swift**, American Chemistry Council; **Maira Trimble**, Eaton Corporation; **Gary Wagner**, University of Louisiana at Lafayette; **Mark Zandi**, Moody's Analytics; **Ellen Zentner**, Morgan Stanley.

This is a partial list of participants. We also thank those who wish to remain anonymous.

Figure A.9: Annual distribution of gender representation within the SPF from 1999 to 2023, showcasing the frequency of female and male participants. The graph illustrates the frequency of male (blue) and female (pink) participants over time.

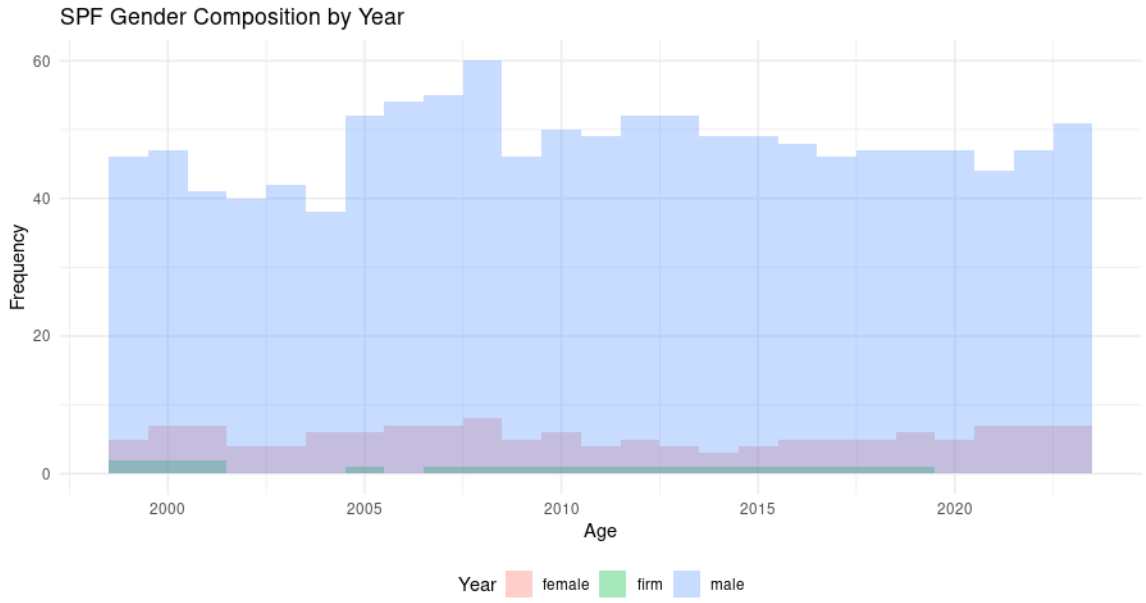


Figure A.10: The count of unique affiliations providing forecasts from 1999 to 2023. The graph displays two lines: one for both anonymous and non-anonymous forecasts (red solid line) and another for non-anonymous forecasts only (blue dashed line).

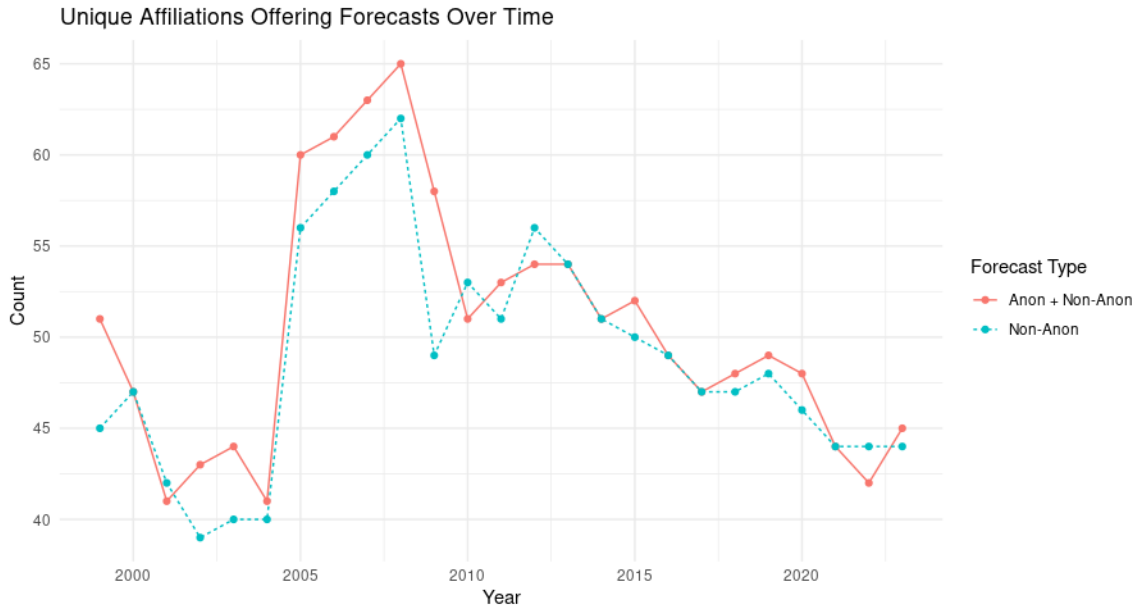
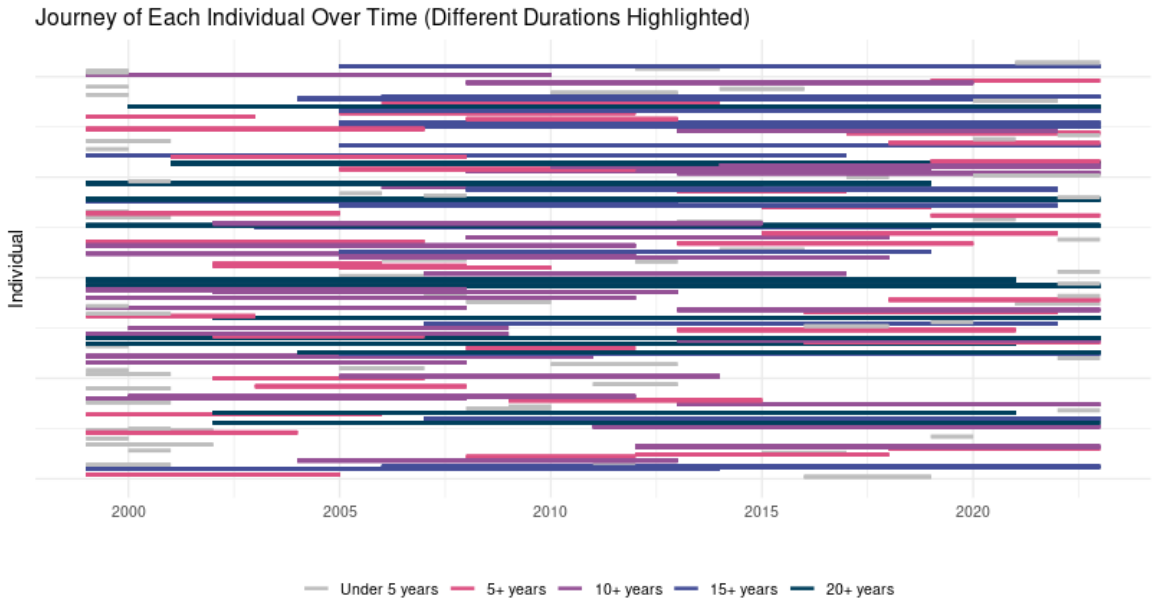


Figure A.11: The chart visualizes how long each individual (199 in total) has participated in the SPF, categorized by different durations to show long-term and short-term involvement.



Appendix B. Measures of Accuracy

Mean Absolute Error (MAE)

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- Where y_i is the actual value; $\hat{y}_i$ is the forecasted value; and n is the number of observations.

Root Mean Squared Error (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Mean Absolute Percentage Error (MAPE)

$$\text{MAPE } (\%) = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

Directional Accuracy

The Directional Accuracy measures how often the forecasted values predict the direction (up or down) of the actual values correctly.

$$\text{Directional Accuracy } (\%) = \frac{100}{n} \sum_{i=1}^{n} \mathbb{1} \left(\frac{y_i - y_{i-1}}{|y_i - y_{i-1}|} = \frac{\hat{y}_i - \hat{y}_{i-1}}{|\hat{y}_i - \hat{y}_{i-1}|} \right)$$

where $\mathbb{1}(\cdot)$ is the indicator function that equals 1 if the forecasted and actual values move in the same direction, and 0 otherwise.

Bias

$$\text{Forecast Bias} = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)$$