Simulating the Survey of Professional Forecasters*

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

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

First Draft: March 2024
This Draft: November 2024

We simulate economic predictions of professional forecasters using a set of large language models (LLMs). We create AI forecaster personas by combining a unique hand-gathered dataset of Survey of Professional Forecasters (SPF) participant characteristics with real-time macroeconomic data. This setup replicates the forecasting process across multiple variables and horizons while controlling for look-ahead bias. AI-generated forecasts match human forecasts in accuracy, and provide new insights into the factors driving forecast disagreement. Our method not only augments existing macroeconomic surveys by enabling higher frequency and broader variable coverage but also facilitates a more detailed analysis of the factors driving forecast disagreement among professionals.

Keywords: Large Language Models; Survey of Professional Forecasters; Behavioral Finance; Synthetic Surveys; Generative Artificial Intelligence; Simulated Economic Agents.
JEL Codes: G4, 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 Stanford Digital Economy Lab and Indiana University seminar participants for valuable comments. We thank Erik Brynjolfsson, Jonathan Benchimol, Leland Bybee, Robin Horton, Huiyu Li, and Tara Sinclair for valuable feedback. Last, but not least, we thank the amazing Bryson Alexander and Nicole Lindsay for excellent research assistance. All remaining errors are our own.

1. Introduction

Economists and policymakers rely on survey-based forecasts to gauge expectations about future economic conditions (Rudebusch, 2002; Orphanides and Williams, 2002, 2007; Patton and Timmermann, 2010; Manzan, 2011, 2021), yet traditional approaches to gathering these predictions face fundamental constraints.

First, surveys are costly to conduct, leading to infrequent sampling intervals.[1] Second, the fixed and static design of surveys constrains researchers and policymakers; existing questions cannot be easily altered, new questions are challenging to introduce, and establishing time series for any new questions is complicated by the lack of historical data. Third, while personal characteristics of forecasters can significantly influence macroeconomic predictions (e.g., Benchimol et al., 2022; Huang et al., 2022), the standard practice of maintaining forecaster anonymity obscures our ability to study these individual-level effects and their role in expectation formation. These limitations may reduce the effectiveness of surveys in informing economic policy.

To address these challenges, we propose a novel simulation framework that replicates the Survey of Professional Forecasters (SPF) using Large Language Models (LLMs) together with detailed profiles of professional forecasters.[2] We use hand-collected individual-level data to create synthetic forecasters who generate predictions using the same real-time data and survey structure as the SPF. This approach addresses key limitations of traditional surveys: it enables more frequent data collection, allows for flexible survey design, reduces costs, and accounts for forecasters' individual characteristics.

Recent advances in artificial intelligence show that LLMs can reliably simulate economic agents and produce outputs that align with human reasoning in a number of domains (see, for example, Argyle et al., 2023; Horton, 2023; Fedyk et al., 2024; Kazinnik, 2023; Zarifhonarvar,

---

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

[2] We select the Survey of Professional Forecasters 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). Further, the SPF plays a key role for both policy-makers and academics, serving as a tool for shaping monetary policy and as a benchmark for assessing forecast accuracy. These features make it an ideal benchmark for our analysis.

2024). These models not only replicate human cognitive patterns but also exhibit human-like traits—including predictable errors (Koralus and Wang-Maścianica, 2023; Hayes et al., 2024), distinct personalities (Jiang et al., 2023), and systematic biases in probabilistic judgment (Zhu and Griffiths, 2024). These findings suggest that LLMs are valuable tools for modeling decision-making processes.

Building on this literature, we show that LLMs can effectively simulate human survey responses, even at the individual level. To start with, we show that LLM-generated median forecasts align closely with those from human professional forecasters. Moreover, the distributions of individual forecasts are similar for the AI and human survey. Where there is disagreement between AI and human forecasters, we show that LLM-generated forecasts are often more precise than human predictions, especially at long (up to four quarters ahead) forecasting horizons.

Existing economic theories, including the full-information rational expectations (FIRE) model, assume that agents are homogeneous in terms of their expectations and rationality in processing information. However, existing research reveals significant heterogeneity in forecasts, driven by informational frictions and behavioral biases (Mankiw and Reis, 2002; Maćkowiak and Wiederholt, 2015; Bordalo et al., 2020; Gabaix, 2020). Coibion and Gorodnichenko (2015) link the analysis of beliefs to observable data, providing widespread evidence of information rigidities, while Coibion et al. (2018) further argue that micro-level survey data often reveal deviations from rational expectations, showing that agents do not always process information fully or efficiently, as assumed by the FIRE model, indicating the importance of incorporating real-time survey-based expectations into economic analyses.

Our framework addresses this gap by explicitly modeling how individual forecaster characteristics influence predictions. We show empirically that the inclusion of individual personas leads to more nuanced forecasts. We also explore which characteristics tend to impact forecast accuracy and the direction of the impacts. These findings complement recent research on LLMs as agents in economic settings suggesting that these models, and their biases, are significantly influenced by the roles they are asked to play. A number of existing studies (e.g., Cook and Kazinnik, 2024; Fedyk et al., 2024; Zarifhonarvar, 2024) explore how assigning specific roles to LLMs affects their decision-making processes and outputs in economic simulations. These findings indicate that the framing of a task and the context provided can lead to variations in the models' responses, reflecting the characteristics and potential biases associated with the assigned roles.

In addition, 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. Alam et al. (2024) generate real-time macroeconomic forecasts using an LLM with AI-simulated forecast.[3] Our approach is different, as we produce macroeconomic forecasts at the individual level and incorporate forecaster characteristics and structured data into the simulation. This integration allows us to replicate not only the numeric predictions but also the behavioral tendencies of individual forecasters. Unlike prior models that treat forecasters as homogeneous agents, our framework captures heterogeneity by using detailed individual-level data, which has not been previously explored in this context.

Using computational techniques to represent individual behavior is not a new idea in economics.[4] Our approach differs from traditional methods like agent-based modeling (ABM). While traditional ABM agents operate based on predefined rules and have limited capacity for complex decision-making, LLM-simulated agents can adapt and make nuanced decisions in complex social simulations.[5] At the same time, while this increased complexity can improve performance, it may also pose a challenge in terms of consistency and robustness. We therefore subject our results to thorough validation exercises and robustness checks. We test different LLM configurations, including GPT-3.5, GPT-4, and GPT-4o-mini, under both deterministic and stochastic settings to assess the impact of model architecture and randomness on forecast

---

[3] The authors built an interactive dashboard displaying these forecasts in real time: `https://aiinflationexpectations.org/`

[4] See Axtell and Farmer (2022) for comprehensive review of agent-based modeling in economics and finance.

[5] Chopra et al. (2024) discuss the performance of heuristic (ABM) agents versus LLM agents in predicting disease waves and unemployment.

accuracy. We experiment with prompt variations, varying the way in which we define our forecaster personas to evaluate framing effects.[6] Finally, we address potential temporal leakage and look-ahead bias by ensuring that the models only use information available at the time of forecasting.

Overall, our contributions are threefold. First, 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. Second, we are among the first to apply generative AI to macroeconomic forecasting, opening up new possibilities in predictive modeling. Our findings demonstrate that LLM-generated forecasts can enhance accuracy and offer deeper insights into economic trends by capturing forecasters' responses to real-time data and macroeconomic shifts. Third, we provide novel insights into the behavioral aspects of forecasting by taking into account individual forecaster characteristics such as education, job experience, and institutional affiliation. We show how these traits influence forecast accuracy.

Organization Section 2 details the structure and institutional context of the SPF. Section 3 describes how we extract information on the participants in the SPF panel, and characterize their profiles. Section 4 outlines our methodology for generating synthetic forecasts using LLMs, incorporating forecaster traits and real-time data. Section 5 shows that AI-generated forecasts generally match those of the human SPF. The section also discusses whether the AI outperforms the human panel. Section 6 uses the AI-generated forecasts to explore the role of real-time data and forecaster characteristics for shaping expectations. Section 7 concludes.

2. The Survey of Professional Forecasters

The Survey of Professional Forecasters (SPF) is a quarterly survey of U.S. economic experts, initially launched in 1968 by the American Statistical Association and the National Bureau of

---

[6] These robustness exercises are discussed in the appendix.

Economic Research and conducted by the Federal Reserve Bank of Philadelphia since 1990.[7]

As of 2024, the SPF collects point forecasts for 23 economic variables at nine horizons: the current quarter (nowcast), one to four quarters ahead, the current year, and one to three years ahead. The survey includes real-time data for the previous quarter and asks respondents to assign probabilities to potential changes in real GDP, the GDP price index, the unemployment rate, and CPI and core CPI inflation rates, along with long-term CPI and PCE inflation projections. In this paper, we focus on the 23 point forecasts for the current quarter and one to four quarters ahead, as summarized in Table 1. Response deadlines typically fall mid-quarter.[8] Although the survey form has evolved over time, with eight updates from 1999 to 2023, its core structure remains intact.

2.1. SPF Panel

SPF panelists are considered a benchmark for attentive and rational agents, drawing on extensive data analysis, quantitative models, and professional judgment to produce forecasts as part of their roles. Most panelists have substantial experience in macroeconomic forecasting and are well-versed in the variables and economic theories they predict.[9] Panelists work in diverse environments: some at forecasting firms, others at banks or financial institutions, while the panel also includes chief economists from industry trade groups and manufacturers, as well as academics specializing in forecasting methods.

Although SPF forecasters remain anonymous, the individual-level dataset includes each forecaster's industry classification ("financial" or "non-financial" service provider) and identifiers that allow researchers to track forecasts across rounds. This anonymity protects forecasters' professional interests and encourages honest, unbiased forecasts without concern for repercussions.[10]

---

[7] Complete information and documentation are available through the Federal Reserve Bank of Philadelphia (link).

[8] See the full schedule of deadlines and release dates here.

[9] New forecasters are recruited via the SPF Call for Participants and undergo a trial period, during which submissions are checked for conceptual accuracy. For more details, see the Philadelphia Fed's website.

[10] While we have access to the forecasters' names, they are not disclosed in this paper to preserve their anonymity.

The SPF panel composition changes over time as respondents join or leave, and some may skip certain questions, resulting in an unbalanced panel. However, numerical identifiers enable tracking individual forecasts, allowing for analyses of forecast behavior and comparisons between forecasters, such as accuracy or contrarian tendencies. Since the Philadelphia Fed took over the survey, the average panel size has been fluctuating around forty participants (see Figure 3.).

Forecasters use various methods to produce predictions. A 2009 survey (Stark, 2013) showed that most rely on quantitative models but adjust for current conditions and recent trends, often supplementing their models with subjective beliefs. Methods also vary by forecast horizon; for instance, the model for predicting current-quarter GDP may differ substantially from that used for forecasting average GDP growth over five years.

2.2. Who Are the Forecasters in the SPF Panel?

Although SPF panelists are anonymized in published datasets, the Philadelphia Fed acknowledges most of their participants by listing their names on its website.[11] Many panelists also share their participation on social media or professional platforms. Using this information, we construct a unique dataset of individual forecaster characteristics.

We create detailed personas for each forecaster by gathering key background information, including education, job titles, affiliated organizations, alma maters, degrees, and professional roles. When available, we also include organization locations, countries of origin, and social media presence. Table 2 summarizes each characteristic.

Empirical research shows that personal backgrounds strongly influence behavior and beliefs. For example, individuals shaped by the Great Depression are often more cautious with financial risk (Malmendier and Nagel, 2011), while those who experienced hyperinflation may avoid risky assets (Fajardo and Dantas, 2018). Personal experiences affect forecasts of macroeconomic variables like inflation (Malmendier and Nagel, 2016), and both geography and institutional affiliation can influence forecasting approaches (Batchelor, 2007; Hong and Kacperczyk, 2010).

To build our panel dataset, we start by compiling forecaster names from Philadelphia Fed

[11] A screenshot of such acknowledgments is provided in Figure 4.

acknowledgments. We then gather detailed professional and demographic information through online searches. For country of origin, we consider indicators like high school location on profiles such as LinkedIn.[12] With these profiles, we create personalized narratives for the LLM prompts, introducing each forecaster as an SPF panelist and highlighting their academic and professional achievements. They are then prompted to forecast based on recent economic data (e.g., inflation, GDP, employment). We provide the full prompt in Figure 8.

Figure 6 provides an overview of the SPF panel over the past 20 years, showing a diverse group with varied educational backgrounds, roles, and affiliations. Most panelists hold Ph.D.s in Economics or Finance, while the share of Master's degree holders has remained steady and Bachelor's and MBA/MPA degrees are less common. Chief Economists and Economists dominate the panel, but there is an increasing presence of consultants and analysts. Affiliations span consulting firms, universities, and asset management, with consulting and academia remaining prominent, while commercial banking has declined in favor of investment banking and asset management.

Gender diversity has seen slight improvement, though men remain the majority. Most panelists are from the USA, though European and Asian representation has grown. Public engagement is limited, but more forecasters are now using platforms like Twitter and participating in interviews. Panelist tenure varies widely, with some contributing for over 25 years and others for shorter periods, adding both established and fresh perspectives.

Figure 7 shows sector distribution over time, with the panel divided into Financial, Non-Financial, and Unknown categories.[13] From the late 1990s to 2024, financial sector representation has notably increased.

---

[12] We assume that attending high school in a particular country shapes formative perspectives, influencing forecasts. Most participants are American, but our data includes enough variation to make the country-of-origin variable relevant. Special thanks to Nicole Lindsay for this idea.

[13] The Philadelphia Fed classifies forecasters as 1" for financial, 2" for non-financial, and "3" for unknown, updating in real-time with any affiliation changes.

3. Simulating the SPF with LLMs

We use a set of LLMs to run a simulated survey. For our main analysis, we use GPT-4o mini model.[14] Our choice of GPT-4o mini follows from a series of preliminary experiments comparing model performance and consistency. Using CPI inflation as a test case, we evaluated three model architectures (GPT-3.5, GPT-4, and GPT-4o) under different temperature settings—zero temperature for deterministic outputs, default temperature (1.0) for balanced generation, and high temperature (>1.5) for increased variability. Given the architectural similarities between GPT-4o and GPT-4o mini, we selected GPT-4o mini as our primary model.[15] Our implementation uses an AI agent built with the OpenAI Assistants API to handle tasks like query responses, and additional data processing.[16]

3.1. Prompt

We automate the generation of forecast queries and gather responses from an AI assistant. Each query incorporates forecaster details—such as education, job title, and organization type—to create personalized predictions of economic variables for future quarters. Queries are customized based on each forecaster's past predictions and real-time economic data, as described in Section 3.2.

The prompt, shown in Figure 8, assigns the model the role of an SPF panel member, forecasting on a specified date. AI forecasters provide numeric forecasts for the current quarter (t) and the next four quarters ($t+1$ to $t+4$), formatted as instructed. They also include a brief 1-

[14] We conduct a robustness check using GPT-3.5 and GPT-4. The results are similar to our main analysis and are described in the appendix. This choice provides a balanced representation across model generations: GPT-3.5 as a baseline older-generation model, GPT-4 as a frontier model, and GPT-4o mini as a latest model implementation.

[15] Temperature in LLMs refers to a parameter that controls the randomness of the model's outputs. Lower temperatures (approaching 0) make the model more deterministic, while higher temperatures (>1) increase output variability.

[16] The Assistants API includes tools like Code Interpreter, File Search, and Function Calling for complex, multi-tool tasks.

2 sentence explanation of their predictions.[17] The goal is for the AI 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. This restriction is specifically emphasized in the prompt.

3.2. Real-Time Data

To simulate the SPF, we provide LLMs with a data environment as close as possible to that of human forecasters, including past forecasts and real-time macroeconomic indicators.

We use data from the Federal Reserve Bank of Philadelphia from two datasets: real-time macroeconomic data and median SPF forecast data. The merged dataset contains quarterly observations, with each row representing a date and each column (vintage) the date when the data point was recorded. The variables and their definitions are summarized in Table 1. 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.

For each survey quarter, we generate a data row with (1) the latest observed SPF variables, (2) the latest real-time data at the SPF submission deadline, and (3) the median SPF forecasts. This process involves transforming percent change variables and calculating percent changes between quarterly levels.[18]

4. Can LLMs Replicate the SPF Forecasts?

This section evaluates how closely LLM-generated forecasts align with human forecasts and real-time macroeconomic data, focusing on both accuracy and variability across our sample of economic indicators and time horizons.

Median forecasts Table 3 shows the mean absolute percentage errors (MAPEs) and directional accuracy between AI-generated and human SPF forecasts across variables. Directional accuracy,

[17] These instructions are given via system prompt.

[18] Percent changes are calculated per the SPF definition: $\left(\frac{X_t}{X_{t-1}}\right)^4 - 1 \times 100$.

which measures how often the AI forecast matches the sign of the human forecast, is high, with most variables achieving near-perfect agreement. However, the ability of LLMs to replicate human predictions varies by variable and forecast horizon. For stable indicators like nominal GDP and the GDP price index, AI forecasts align closely with human forecasts, maintaining low MAPE values across horizons. In contrast, volatile indicators like corporate profits and housing starts show higher MAPEs, especially at longer horizons. Generally, MAPE increases with forecast horizons, though inflation rates are an exception.

We examine four key variables in detail: CPI inflation rate, 3-month Treasury bill rate, unemployment rate, and real GDP index.[19] Figures 1 and 2 compares AI and human median forecasts over time for one- and four-quarter horizons, showing strong alignment, particularly for unemployment and real GDP. The largest discrepancy appears in CPI inflation, where AI forecasts respond more to the business cycle at short horizons. Another notable difference occurs with the four-quarter ahead 3-month Treasury bill rate during the zero-lower bound period, where human forecasts predict a lift-off, while AI forecasts stay at the lower bound until the 2015 rate hike cycle. This divergence may stem from the LLM's reliance on recent data, while human forecasters likely incorporate term structure models (that do not describe extended periods of rates near the zero-lower bound well) or a preference for *normal* rates.

Individual forecasts Figure 9 shows density plots of individual forecasts for all forecasters on four key dates: 1999 Q1 (earliest observation), 2008 Q3 (Global Financial Crisis), 2020 Q2 (COVID-19 pandemic), and 2023 Q1 (latest survey with four-quarter ahead realizations). Each plot centers around realized values, enabling direct comparison between AI and human SPF forecast distributions. In 1999 Q1, AI forecasts closely align with the full distribution of human forecasts, indicating strong alignment in stable periods. During the 2008 and 2020 crisis periods, both AI and human forecasts showed wider distributions, reflecting greater disagreement. Notably, for inflation and interest rates, forecaster disagreement was larger during the Global

[19] We choose to focus on these four economic variables because they are both relevant for policy and diverse in their economic dynamics. They cover key aspects of the economy - price stability, interest rates, labor market conditions, and overall economic output. This provides a comprehensive test of the language model's ability to simulate human forecasting behavior across different dimensions of economic activity.

Financial Crisis than during COVID-19, while the opposite was true for unemployment and real GDP. In 2020 Q3, dispersion is high for both AI and human forecasts.

The models' ability to mirror the dispersion of the human forecasters is remarkable. The result suggests that, like the human SPF panel, the AI forecasters interpret signals from the real-time data in more diverse ways in distressed times, underscoring the model's responsiveness to both economic conditions and forecaster characteristics.

4.1. Are LLMs More Accurate Than Human Forecasters?

The AI and human surveys are not always perfectly aligned. The question is then, which one is more accurate in predicting the actual values of the macroeconomic variables? This section discusses this question by comparing forecast accuracy across the two surveys.

Table 4 shows MAPE values comparing forecasts with the actual realized values against the AI and the human surveys, respectively. Panel (a) shows the results averaged across all survey dates, whereas Panel (b) focuses on survey dates during NBER recessionary periods. For each forecasting horizon, the table boldfaces the lowest MAPE value between the AI and human surveys.

Overall, the results indicate that AI-generated forecasts tend to outperform human forecasts, particularly over medium- to long-term horizons. At the nowcasting horizon, however, neither approach is consistently more accurate across all variables. AI-generated forecasts tend to be more accurate for economic indicators – such as the unemployment rate, housing starts, and the GDP price index – and inflation rates, while human forecasts perform better for components of real GDP. This pattern generally holds even when focusing on forecasts made during NBER-defined recessions. During recessions, human forecasters also show higher accuracy for long-term projections of real GDP components. As expected, forecast accuracy declines during recessionary periods for both human and AI forecasts.

Figure 10 shows the time series of forecast errors for AI-generated and human-generated forecasts across the four key variables, comparing predictions one and four quarters ahead. Each plot displays the error percentage, with negative values representing underestimation and positive values representing overestimation. The forecasting errors for AI and human surveys appear quite similar, with both series heavily influenced by a few instances of extreme errors. For example, both surveys show forecast errors of 15-20% for CPI inflation during the Great Financial Crisis. For the unemployment rate, large forecast errors appear in both surveys around the onset of the COVID pandemic and during the recovery period. The most significant differences in accuracy between the AI and human forecasts are seen in the 3-month Treasury bill rate, especially during the 2005-2006 rate-hiking cycle. During zero-lower bound periods, AI forecasts are more accurate at the four-quarter horizon, as human forecasters incorrectly predict a departure from the lower bound.

4.2. Are LLMs Forecasting or Recalling?

Limiting large language models to use only information from a specific time period is a challenging task. LLMs generate output based on learned patterns, which gives the impression of generating content, even though the underlying knowledge comes from previously seen data (Gurnee and Tegmark, 2023). This could be 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, leading to what is known as "temporal leakage."

Temporal leakage can negatively affect forecast accuracy in several ways. 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. 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 are explicitly instructed to focus on the data available up until the survey date, 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 are segmented to match the exact information that would have been available to human forecasters at the time of their predictions. This ensures that LLMs are not inadvertently provided with future data, helping to reduce the risk of temporal leakage.

To test if the AI-generated forecasts are contaminated with "temporal leakage," we explicitly prompt the model to recall the value of our macroeconomic variables for each survey quarter. We then compare these recall values with the nowcast predictions generated by our AI survey. For simplification, we did not provide the model with forecaster characteristics for this exercise.[20]

Table 6 compares the MAPE of recalled values and nowcasts. The table shows that the recall errors are high, and much higher than the nowcast errors, indicating that the model struggles with accurately recalling historical data. This serves as evidence that the model is not relying on recalling values from its training data. Instead, it is likely referring to the provided real-time data, which is not impacted by "temporal leakage", when generating forecasts.

Why is the model better at forecasting than recalling data? When recalling past values, the model only relies on internalized data representations and its memory of past values. However, because LLMs aren't specifically designed to store and retrieve exact historical data, their recall ability is likely limited, especially for variables that are not heavily discussed in the training data. For instance, the variable measuring housing starts is likely to be less prevalent in the training data than inflation variables, leading to higher recall errors. In contrast, our forecasting prompt has the benefit of using more structured external data, which allows the model to rely on more resources to make more accurate predictions.

[20] The prompt is shown in Figure 11.

5. What Shapes Professional Forecasts?

The LLM-generated survey allows us to explore the role of the individual pieces of information in shaping individual forecasts. We start by investigating the forecast accuracy when omitting real-time data and individual characteristics, respectively. Then, we regress forecast accuracy on each set of characteristics to understand heteroskedasticity in accuracy across different personas. Finally, we study the reasoning behind forecast decisions provided directly by the LLM.

5.1. The Role of Personal Characteristics and Real-Time Data

In this section, we analyze the importance of providing real-time data and forecaster characteristics for forecasting performance. We start simulating forecasts for a "generic forecaster," obtained by a prompt that does not provide the LLM with forecaster characteristics. The prompt is shown in Figure 11. This prompt forces the model to make predictions using only real-time data and past SPF predictions. Next, we create a simulation for the same generic forecaster, but with no access to real-time data. The prompt for this simulation is provided in Figure 12. This prompt aims to simulate a professional forecaster's environment without the advantage of real-time data, requiring the model to make predictions using information only on past forecasts, and the general knowledge that the model is endowed with as part of its pre-training.

We compare these simulations against our baseline AI forecasts in Table 7. The table reports the mean absolute error (MAE) of the generic forecaster (respectively with and without access to real-time data) as a fraction of the MAEs from the baseline results. Values above one indicate that the generic forecaster is outperformed by the baseline AI seeded with forecaster characteristics and real-time data.

Most of the relative MAEs exceed one, and the average relative MAEs across all variables are above one across all horizons both with and without real-time data. This is evidence that information related to forecaster characteristics and real-time data improve forecast accuracy. What is more, the average relative MAEs are remarkably higher for the generic forecast *without* real-time data compared with those reported for the generic forecaster *with* access to real-time data. Hence, non-surprisingly, real-time data play a key role in the formation of forecasts.

Considering the results for each variable separately, three observations stand out. First, the generic forecaster without access to real-time data performs substantially worse for some variables, especially for the nowcast. The most prominent examples are non-farm payroll and industrial production with relative MAEs for the nowcast equal to 28.93 and 3.16. Second, there are some cases where the absence of real-time data does not severely impact accuracy; for instance, real change in private inventory and real net exports have relative MAE values below one at the nowcast horizon. These exceptions may be due to lagged effects in economic indicators, where historical data already encapsulates sufficient predictive information, reducing the reliance on real-time updates. Finally, the results for real PCE exhibit a surprising pattern: the MAEs of the generic forecaster *with* access to real-time data are 63-86% higher than the baseline MAEs, but for the generic forecaster *without* real-time data, the forecasts are *more* accurate than the baseline by a large factor (relative MAEs are 0.06–0.17). Overall, these results suggest that incorporating real-time data not only does not enhance predictive accuracy for real PCE, but actually results in worse forecasts. This suggests that PCE may be influenced by stable consumer behavior patterns that real-time data fails to capture effectively.

5.2. Which Personal Characteristics Impact Forecast Accuracy?

Rich and Tracy (2024) use data from the ECB's Survey of Professional Forecasters (ECB-SPF) to show that forecasters are not interchangeable. They find consistent differences in predictive accuracy of forecasters, both within and across target variables such as GDP growth, inflation, and unemployment. These differences are closely tied to the difficulty of the forecasting environment, with some forecasters excelling in tranquil periods and others in volatile ones. Contrary to expectations models featuring information rigidities, the study demonstrates that forecasters' behaviors are neither random nor interchangeable, emphasizing the role of environmental factors in shaping predictive performance.

To assess the impact of each forecaster characteristics on forecast accuracy in our study, we regress absolute forecast errors on characteristic dummies. With eight types of characteristics (gender, graduation year, origin, public appearance, education level, education field, affiliation, and job title), each containing between two and six outcomes (excluding the cases of "unknown" as baseline to avoid multicollinearity), we have a total of 34 dummy variable regressors. The sparse-group LASSO regression model from Babii et al. (2022) allows us to identify which variables are important for explaining forecast accuracy, under penalty of a high-dimensional regression. We implement the model as a pooled panel regression.[21]

The results are shown for all variables in Figure 13. Panel (a) reports results for the nowcast errors, while panel (b) focuses on the four-quarter ahead forecast errors. Darker colors indicate more weight on a characteristic dummy, with blue (red) indicating that the variable is associated with lower (higher) errors. Since panel (b) highlights more dark-colored fields than panel (a), it is tempting to conclude that forecaster characteristics impact errors more at long forecasting horizons. Similarly, at both horizons, the GDP price index, the AAA corporate bond yield, the 10-year Treasury bond yield, and the core PCE inflation rate involve darker-shaded areas across many characteristic dummy variables, naturally leading to the conclusion that characteristics particularly matter for the forecasting accuracy for these variables. However, for each variable, some characteristics lower the forecast accuracy, while others are associated with higher accuracy, and these effects may cancel out. This exercise thus provides different information than that in Table 7, comparing the accuracy of the generic forecaster to the forecaster seeded with forecasting characteristics.

The results show that the importance of each characteristic for forecasting performance is highly variable-dependent. However, we can infer a few general trends from the results. First, gender does not seem to play a role for forecasting performance. The coefficients on male and female indicators are near zero for most variables, and where they are different from zero, the coefficient on male and female are similar in both direction and magnitude. Second, age, as proxied by graduation year, is inversely related to forecast accuracy: more recent graduation is generally associated with higher forecast errors, likely reflecting a relative lack of experience. Third, forecasters with Ph.D. degrees do not outperform other forecasters. One explanation

[21] Ideally, we would estimate a fixed effects panel model with time fixed effects to account for time variation in forecast accuracy. But, the Babii et al. (2022) model does not allow the specification of time fixed effects. Since the characteristics for each forecaster are constant across time, and only the composition of the forecasting panel is time-varying, it is unlikely that time-variation in the characteristic dummy variables is similar to the time-variation in forecast errors. The omission of time fixed effects is therefore unlikely to impact the results.

could be that forecasters with higher education may use more complex models, which may not always outperform simpler models.[22] There is no clear distinction between forecasters with master's, bachelor's, and MBA/MPA degrees. Fourth, forecasters with educational backgrounds in economics and finance are generally less precise than other forecasters. Here, it is relevant to note that these forecasters make up a vast majority of the forecasting panel in the SPF survey, see Figure 6(f). Related to this observation, forecasters with job titles "Chief Economists" or "Economists" are also less accurate than other forecasters. Finally, consultancies and academic institutions generate less accurate forecasts than other forecasts, as measured both from the job title and affiliation variables. This observation is consistent with the observation that the Ph.D. dummy variable is associated with higher forecast errors.

Altogether, these observations suggest that more accurate forecasters rely on experience gathered through long professional careers, rather than sophisticated models learned through long and recent trainings in economics and finance. We note that these results are obtained using the AI forecasting personas. The applicability of these conclusion to human forecasters hinges on how well the AI individual forecasts matches human individual forecasts, something that is not directly testable due to lack of individual identifiers in the human survey. The comparison of densities of individual forecasts in Figure 9 suggests, however, that our results are likely to generalize to human forecasters.

5.3. Forecast Reasoning

In addition to numeric predictions from our synthetic AI forecasters, we also gather explanations about how each prediction was made using open-ended questions. These explanations can help shed light on how forecasters form predictions.

We analyze these explanations using two different methods. First, we used LDA (Latent Dirichlet Allocation), an unsupervised machine learning method that finds latent topics based

---

[22] There are indeed many examples that a simple random walk can outperform more sophisticated models in out-of-sample forecasting, see, for instance, Meese and Rogoff (1983) for exchange rates and Diebold and Li (2006) for Treasury bond yields.

on word co-occurrence.[23] We configured LDA to categorize the responses into 9 categories and then assigned topic labels based on the most common words in each category. Second, we used GPT to analyze the responses. We provided GPT with a sample of responses and asked it to identify distinct topics for categorization. Then, using the OpenAI API, we processed each response individually to assign it one of these topic labels.[24]

As shown in Table 8, both methods identified Monetary Policy as the dominant topic, accounting for 43.4% and 51.2% of documents in LDA and GPT analyses, respectively. Consumer Demand was the second most common topic for both methods (32.3% LDA, 27.1% GPT). However, there were notable differences: Economic Recovery was prominent in LDA (21.7%) but absent in GPT's categorization, while Labor Market was barely present in LDA (0.0%) but significant in GPT's analysis (14.3%). Other topics like Fiscal Policy, Housing Market, International Conditions, and Supply Chain appeared with varying frequencies between the two methods.

6. Conclusions

Our research shows that language models can effectively generate synthetic forecasts for the Survey of Professional Forecasters by incorporating individual forecaster traits and real-time macroeconomic data. By replicating the forecasting process across various economic variables, we show that LLMs not only mimic human forecasters but also uncover patterns of disagreement influenced by behavioral and personal traits. We also show that LLMs can outperform human forecasters, particularly at long forecasting horizons. Our approach addresses traditional survey limitations, such as high costs, infrequent data, and bias, while offering better insights into forecasting behavior.

Beyond replicating existing survey methodologies, our results suggest that LLMs could serve as powerful tools for enhancing macroeconomic nowcasting and forecasting frameworks. For example, by incorporating high-frequency data sources like news headlines or social media trends, LLMs have the potential to generate dynamic, real-time nowcasts that can adapt quickly

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

[24] This method efficiently uses GPT's pattern recognition abilities for systematic response labeling.

to new information. Moreover, combining LLMs with human expertise could further boost the accuracy and reliability of long-term economic forecasts.

As these models improve and evolve, they can enhance how policymakers, researchers, and practitioners analyze macroeconomic trends, leading to better decision-making.

References

Jahangir Alam, Huiyu Li, and Tatevik Sekhposyan. Inflation and Macroeconomic Expectations Through the Lens of Artificial Intelligence. *SSRN Working Paper*, 2024.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 31(3), 2023.

Robert L Axtell and J Doyne Farmer. Agent-based modeling in economics and finance: Past, present, and future. *Journal of Economic Literature*, pages 1–101, 2022.

Andrii Babii, Eric Ghysels, and Jonas Striaukas. Machine Learning Time Series Regressions With an Application to Nowcasting. *Journal of Business & Economic Statistics*, 40:1094–1106, 2022.

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

Jonathan Benchimol, Makram El-Shagi, and Yossi Saadon. Do Expert Experience and Characteristics Affect Inflation Forecasts? *Journal of Economic Behavior & Organization*, 201: 205–226, 2022.

Pedro Bordalo, Nicola Gennaioli, Yueran Ma, and Andrei Shleifer. Overreaction in Macroeconomic Expectations. *American Economic Review*, 110(9):2748–2782, 2020.

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

Ayush Chopra, Shashank Kumar, Nurullah Giray-Kuru, Ramesh Raskar, and Arnau Quera-Bofarull. On the Limits of Agency in Agent-based Models. *arXiv preprint arXiv:2409.10568*, 2024.

Olivier Coibion and Yuriy Gorodnichenko. Information Rigidity and the Expectations Formation Process: A Simple Framework and New Facts. *American Economic Review*, 105(8): 2644–2678, 2015.

Olivier Coibion, Yuriy Gorodnichenko, and Rupal Kamdar. The formation of expectations, inflation, and the phillips curve. *Journal of Economic Literature*, 56(4):1447–1491, 2018.

Thomas R. Cook and Sophia Kazinnik. Social bias in financial applications of large language models. *SSRN Working Paper*, 2024.

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

Francis X. Diebold and Canlin Li. Forecasting the Term Structure of Government Bond Yields. *Journal of Econometrics*, 130:337–364, 2006.

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

Miguel Faria-e Castro, FRB St Louis, and Fernando Leibovici. Artificial Intelligence and Inflation Forecasts. Technical report, 2023.

Anastassia Fedyk, Ali Kakhbod, Peiyao Li, and Ulrike Malmendier. ChatGPT and Perception Biases in Investments: An Experimental Study. *Available at SSRN 4787249*, 2024.

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

Wes Gurnee and Max Tegmark. Language Models Represent Space and Time. *arXiv preprint arXiv:2310.02207*, 2023.

Anne Lundgaard Hansen and Sophia Kazinnik. Can ChatGPT Decipher Fedspeak? *SSRN Working Paper*, 2023.

William M Hayes, Nicolas Yax, and Stefano Palminteri. Relative Value Biases in Large Language Models. *arXiv preprint arXiv:2401.14530*, 2024.

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

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

Rong Huang, Keith Pilbeam, and William Pouliot. Are Macroeconomic Forecasters Optimists or Pessimists? A Reassessment of Survey Based Forecasts. *Journal of Economic Behavior and Organization*, 197, 2022.

Hang Jiang, Xiajie Zhang, Xubo Cao, Jad Kabbara, and Deb Roy. PersonaLLM: Investigating the Ability of GPT-3.5 to Express Personality Traits and Gender Differences. *arXiv preprint arXiv:2305.02547*, 2023.

Sophia Kazinnik. Bank run, interrupted: Modeling deposit withdrawals with generative ai. *Interrupted: Modeling Deposit Withdrawals with Generative AI (October 30, 2023)*, 2023.

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

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

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

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

N Gregory Mankiw and Ricardo Reis. Sticky information vs. sticky prices: a proposal to replace the new keynesian phillips curve. *The Quarterly Journal of Economics*, 117(4): 1295–1328, 2002.

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

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

Richard A. Meese and Kenneth Rogoff. Empirical Exchange Rate Models of the Seventies: Do They Fit Out of Sample? *Journal of International Economics*, 14:3–24, 1983.

Athanasios Orphanides and John C. Williams. Robust Monetary Policy Rules with Unknown Natural Rates. *Brookings Papers on Economic Activity*, 2002.

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

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

Robert W Rich and Joseph S Tracy. All Forecasters Are Not the Same: Time-varying Predictive Ability Across Forecast Environments. *American Enterprise Institute Working Paper*, 2024.

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

Tom Stark. SPF Panelists' Forecasting Methods: A Note on the Aggregate Results of a November 2009 Special Survey. *Federal Reserve Bank of Philadelphia*, 2013.

Ali Zarifhonarvar. Experimental Evidence on Large Language Models. *Available at SSRN 4825076*, 2024.

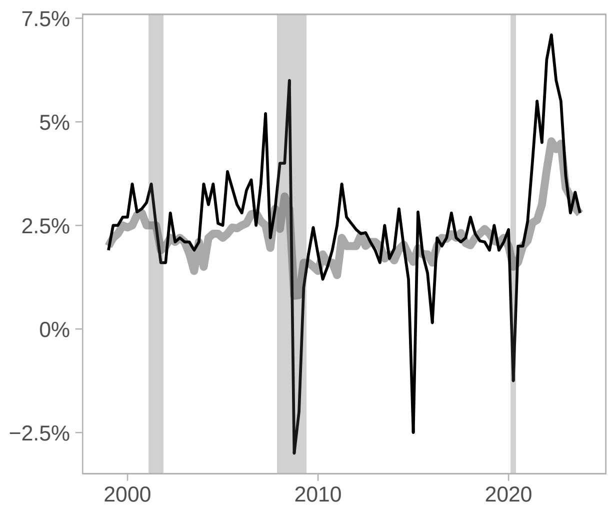
Jian-Qiao Zhu and Thomas L Griffiths. Incoherent Probability Judgments in Large Language Models. *arXiv preprint arXiv:2401.16646*, 2024.

Figures

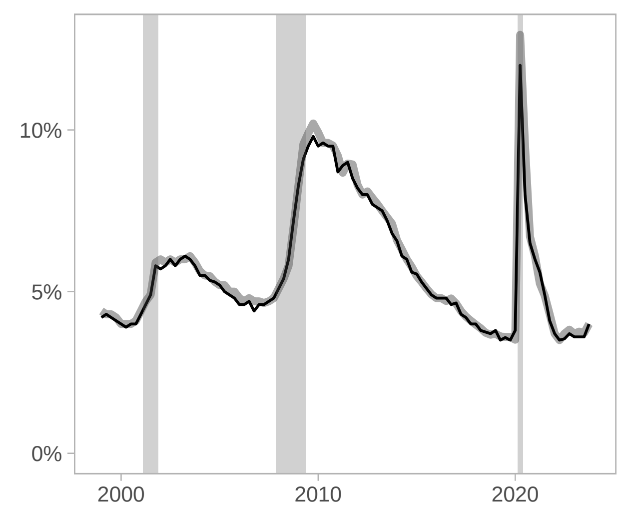
Figure 1: SPF and AI-generated short-term median forecasts.

The figure shows differences in median forecasts of the SPF and AI-generated SPF for (a) CPI inflation rate one quarter ahead, (b) Unemployment rate, one quarter ahead, (c) 3-month Treasury bill rate one quarter ahead, (d) Real GDP index, one quarter ahead. Shaded areas show NBER recession quarters.

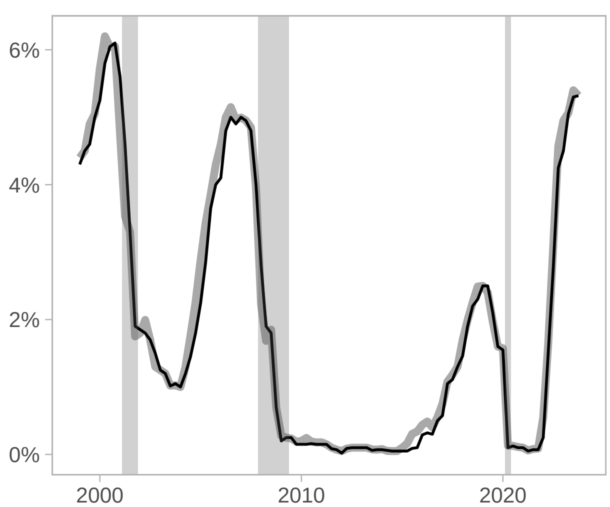
(a) CPI inflation rate, one quarter ahead



(b) Unemployment rate, one quarter ahead



(c) 3-month Treasury bill rate, one quarter ahead



(d) Real GDP index, one quarter ahead

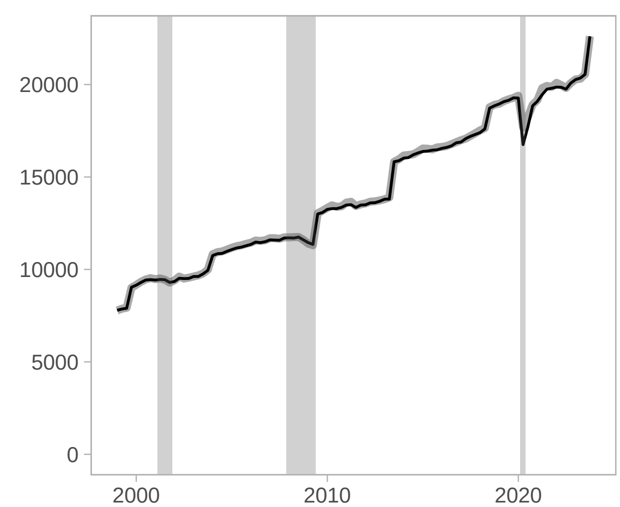
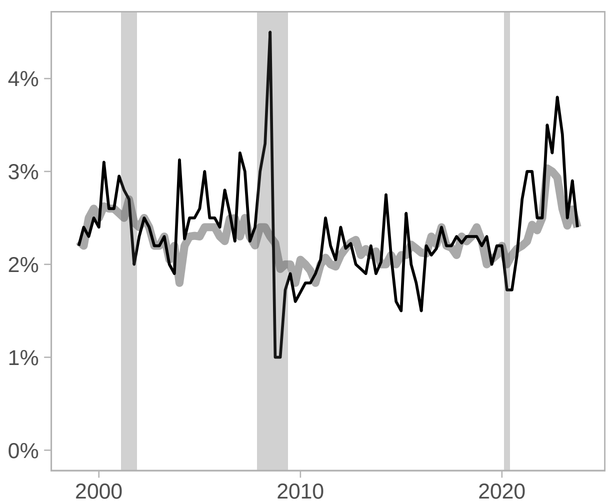


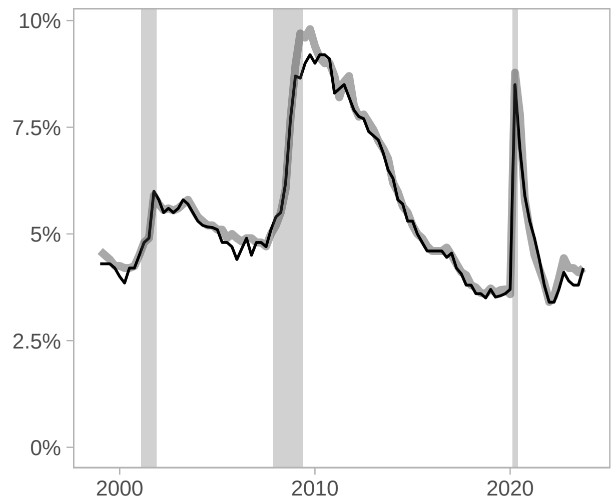
Figure 2: SPF and AI-generated medium-term median forecasts.

The figure shows differences in median forecasts of the SPF and AI-generated SPF for (a) CPI inflation rate four quarters ahead, (b) Unemployment rate, four quarters ahead, (c) 3-month Treasury bill rate four quarters ahead, (d) Real GDP index, four quarters ahead. Shaded areas show NBER recession quarters.

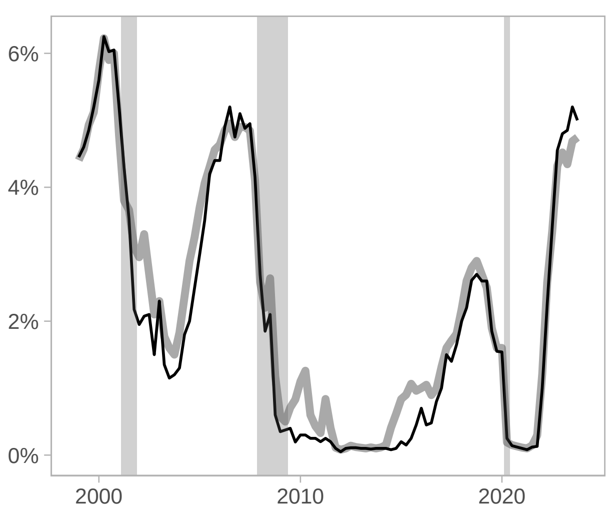
(a) CPI inflation rate, four quarters ahead



(b) Unemployment rate, four quarters ahead



(c) 3-month Treasury bill rate, one quarter ahead



(d) Real GDP index, one quarter ahead

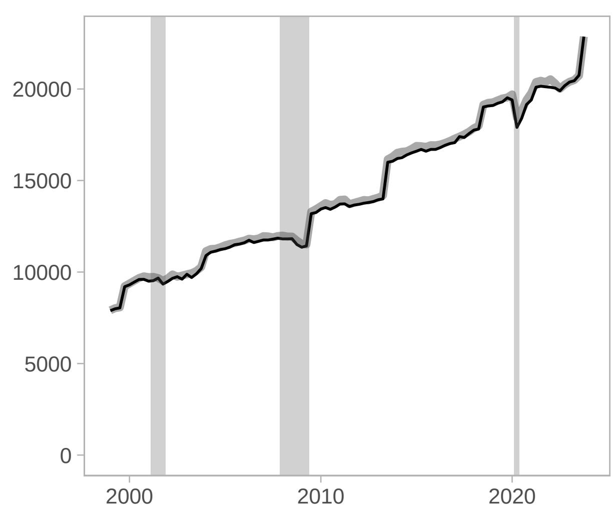


Figure 3: Number of forecasters in the SPF panel over time

The figure shows the number of human and AI forecasters in the SPF panel over time. The number of human forecasters is averaged across variables and forecasting horizons.

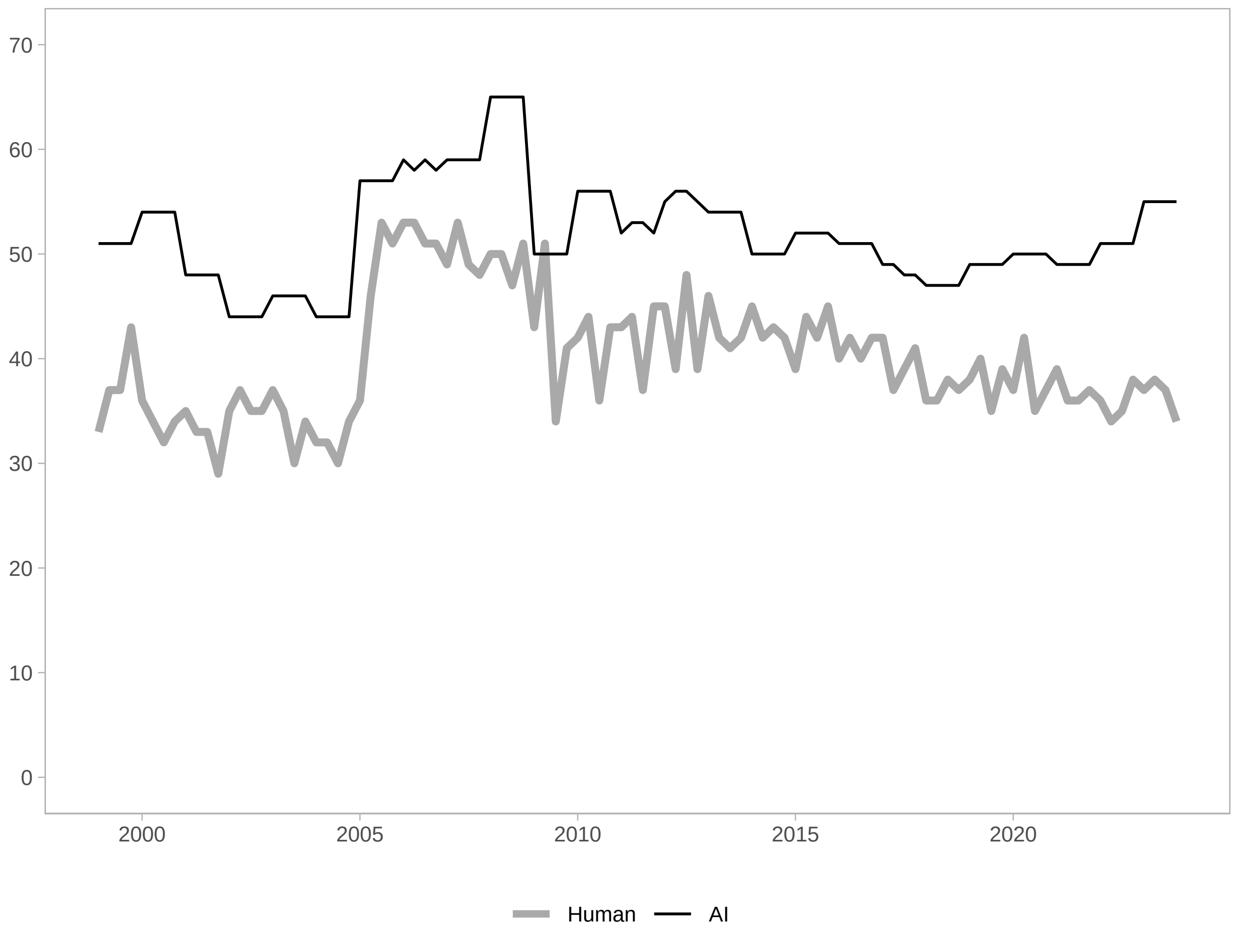


Figure 4: Example of forecaster acknowledgments

The figure shows an example of acknowledgments to forecasters who contributed to recent Federal Reserve Bank of Philadelphia 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 5: Duration of participation of each forecaster individual forecaster in the SPF panel over time

The figure tracks the duration of individual forecasters' participation in the SPF panel, ranging from a few years to over 25 years. It shows a mix of long-term participants and those with much shorter tenures.

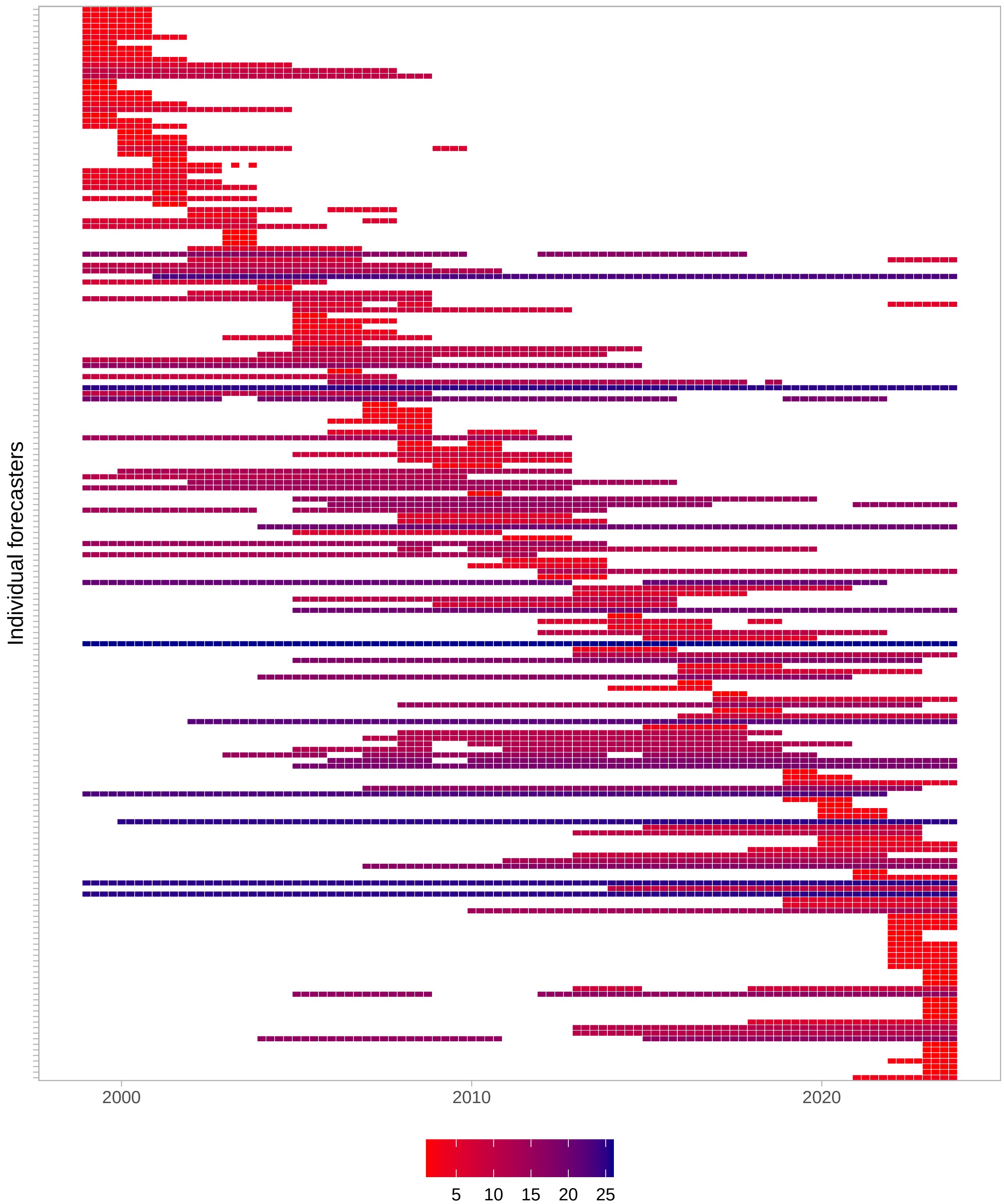
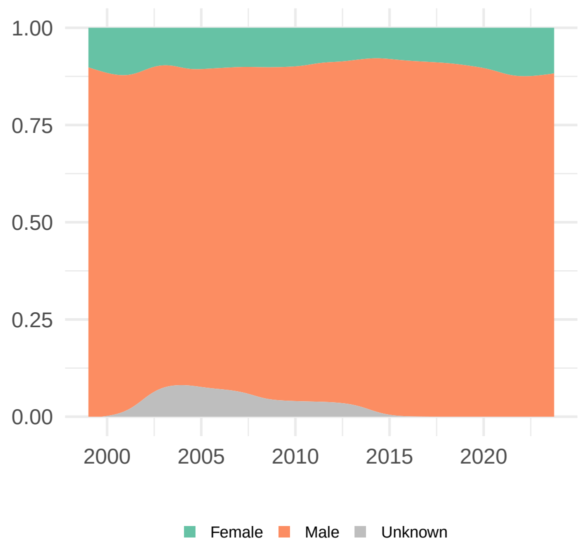


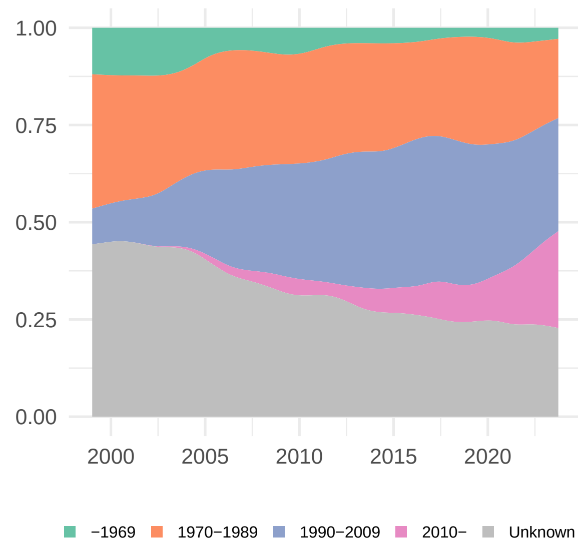
Figure 6: Characteristics of the SPF panel over time

The figures show the characteristics of the AI SPF panel over time: (a) gender, (b) graduation year, (c) continent of origin, (d) public appearance; on next page: (e) education level, (f) education field, (g) affiliation type, and (h) job title.

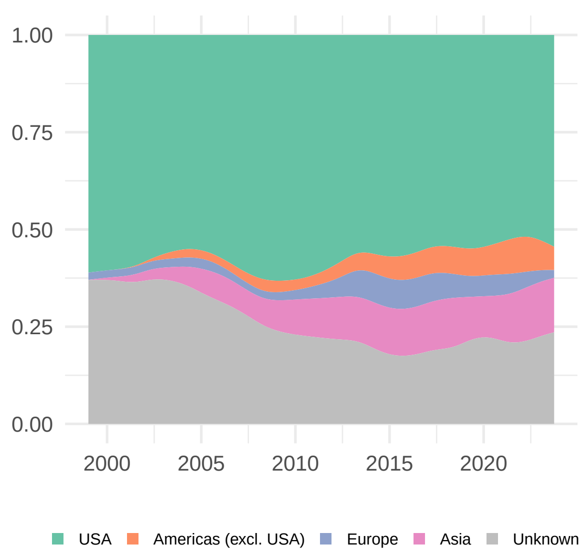
(a) Gender



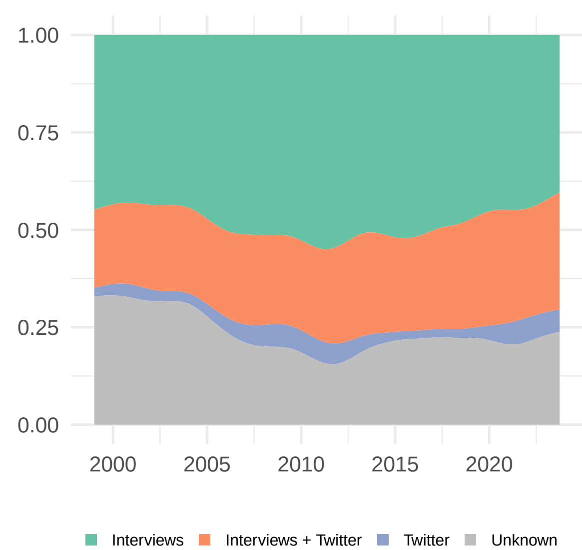
(b) Graduation Year



(c) Continent of Origin



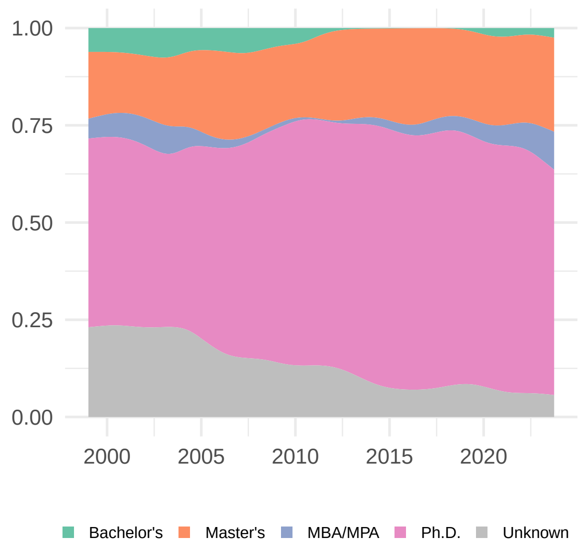
(d) Public Appearance



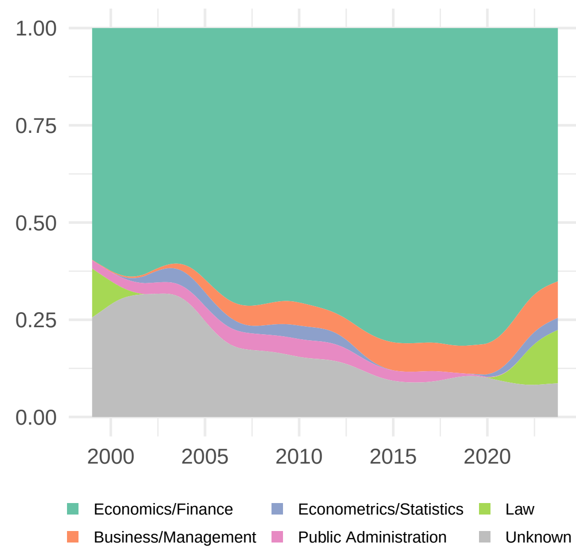
(Figure continues on next page)

Figure 6: Characteristics of the AI SPF panel over time *(continued)*

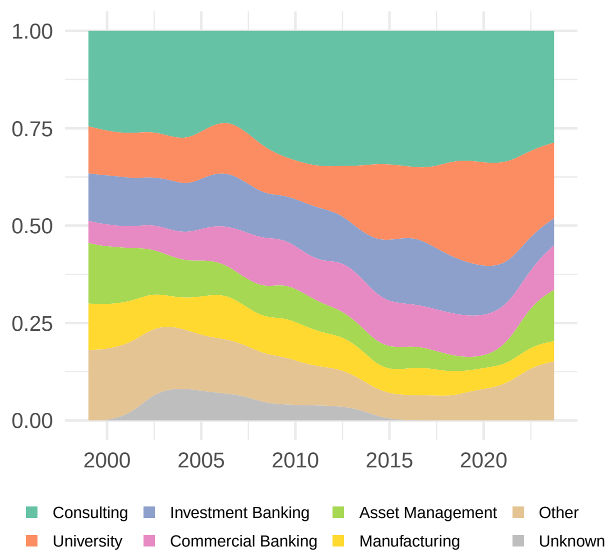
(e) Education level



(f) Education field



(g) Affiliation type



(h) Job title

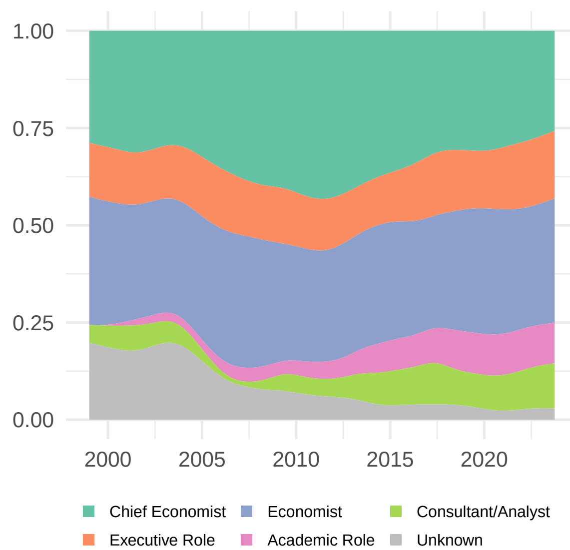


Figure 7: Distribution of sectors across the human SPF panel over time

The figures show the characteristics of the human SPF panel categorized into financial and non-financial sectors.

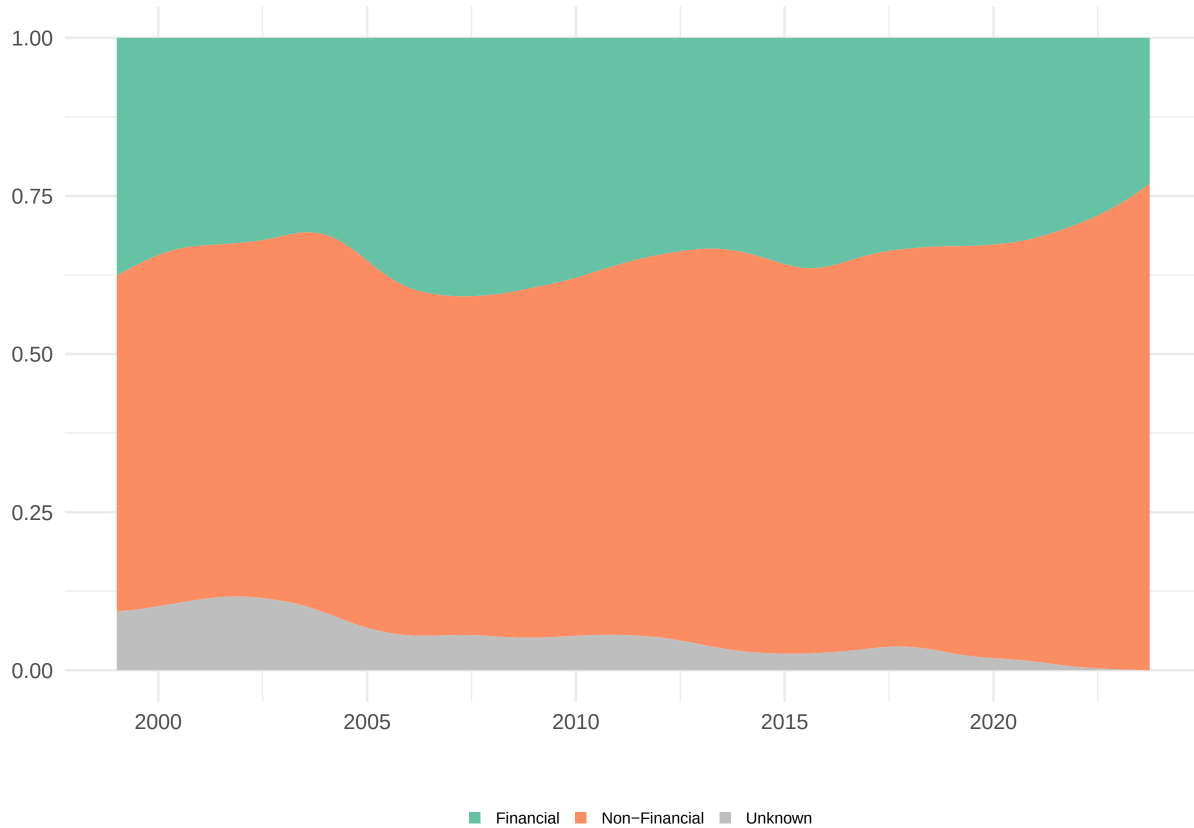


Figure 8: Prompt

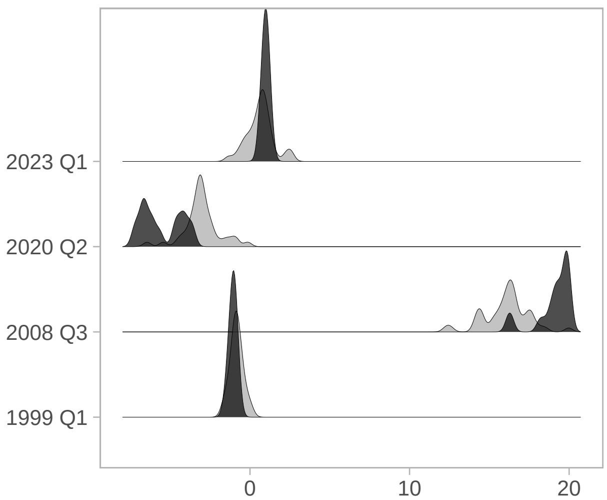
The figure shows the prompt used to generate AI forecasts given the variables highlighted in brackets and blue text consisting of real-time data, forecaster characteristics, previous forecasts, survey dates, and forecasting variable names.

```
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].

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).
```

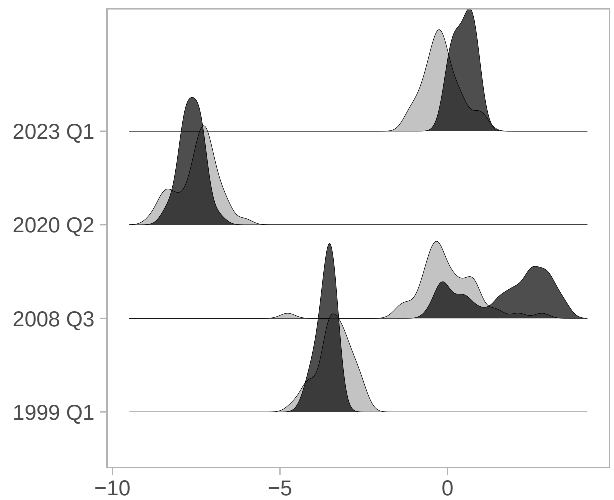
Figure 9: Densities of human and AI-generated SPF individual forecasts

The figure shows densities centered around realized variables of human and AI-generated SPF forecasts across individual forecasters for four different quarters. The densities are shown for the following variables and forecasting horizons: (a) CPI inflation rate one quarter ahead, (b) CPI inflation rate four quarters ahead, (c) 3-month Treasury bill rate one quarter ahead, (d) 3-month Treasury bill rate four quarters ahead, (e) unemployment rate one quarter ahead, (f) unemployment rate four quarters ahead, (g) real GDP index one quarter ahead, and (h) real GDP index four quarters ahead.

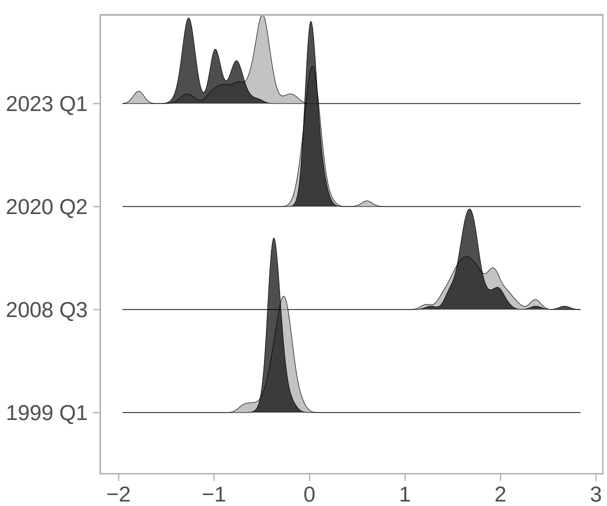
(a) CPI inflation rate, one quarter ahead



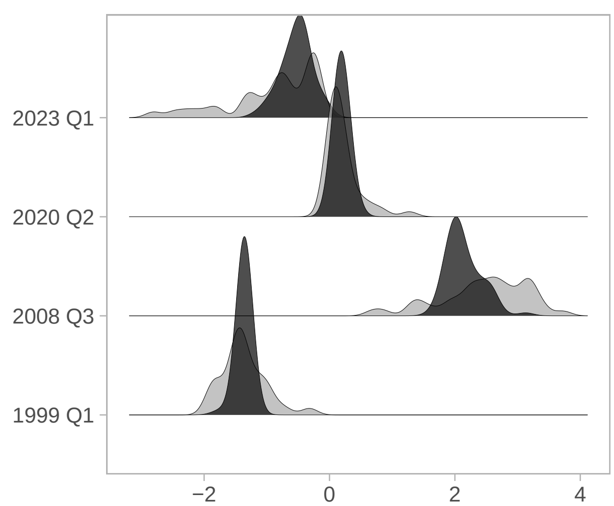
(b) CPI inflation rate, four quarters ahead

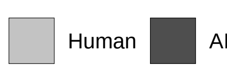


(c) 3-month Treasury bill rate, one quarter ahead



(d) 3-month Treasury bill rate, four quarters ahead





(Figure continues on next page)

Figure 9: Densities of human and AI-generated SPF individual forecasts *(continued)*

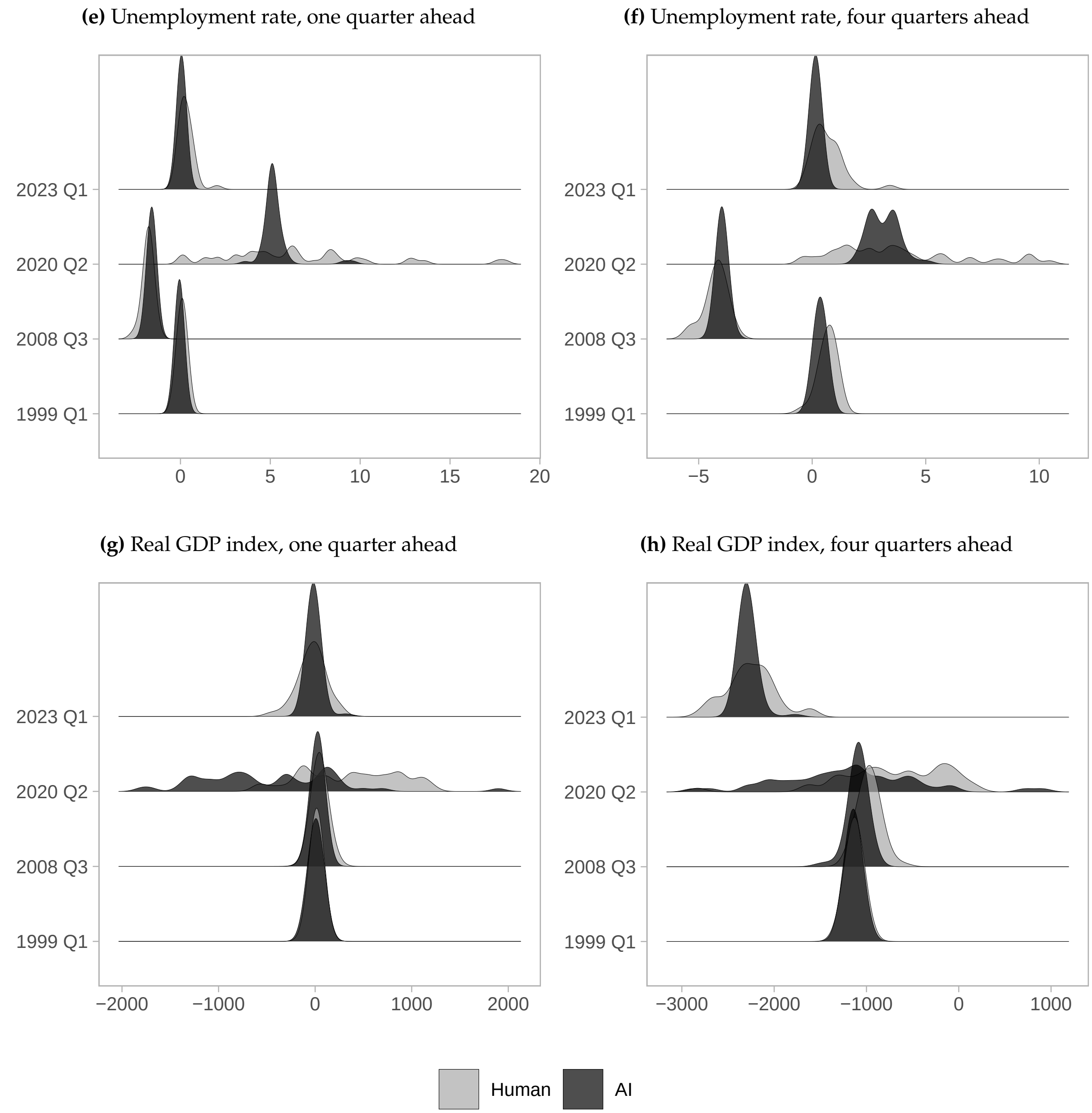
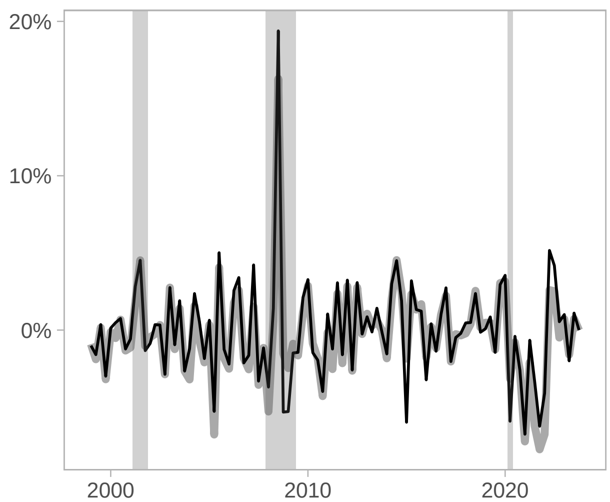


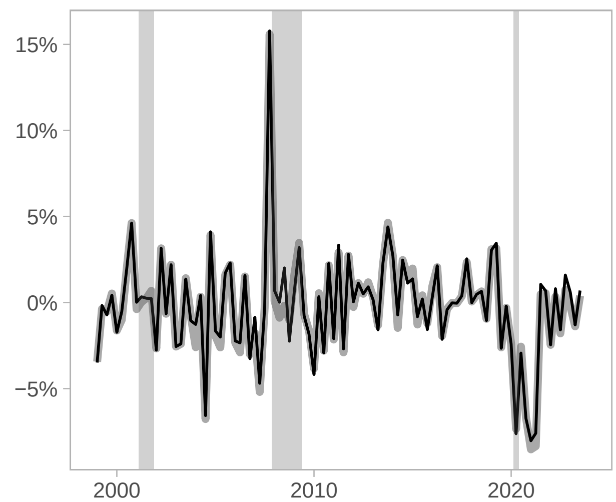
Figure 10: Median forecast errors in human and AI-generated SPF

The figure shows forecast errors in SPF and AI-generated SPF median forecasts for (a) CPI inflation rate one quarter ahead, (b) CPI inflation rate four quarters ahead, (c) 3-month Treasury bill rate one quarter ahead, (d) 3-month Treasury bill rate four quarters ahead, (e) unemployment rate one quarter ahead, (f) unemployment rate four quarters ahead, (g) real GDP index one quarter ahead, and (h) real GDP index four quarters ahead. Shaded areas show NBER recession quarters.

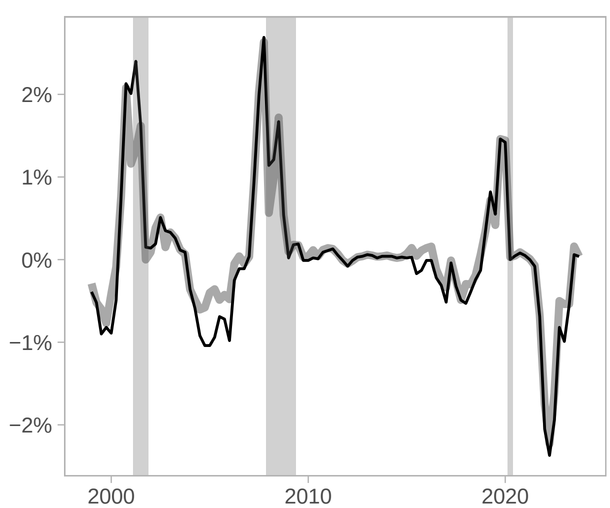
(a) CPI inflation rate, one quarter ahead



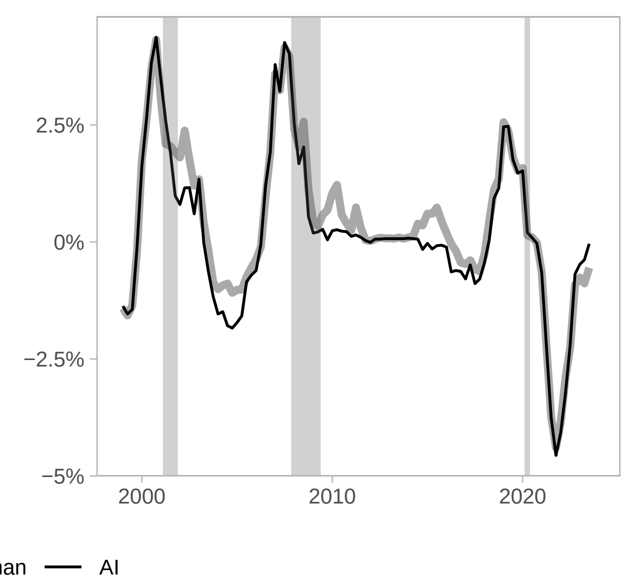
(b) CPI inflation rate, four quarters ahead



(c) 3-month Treasury bill rate, one quarter ahead

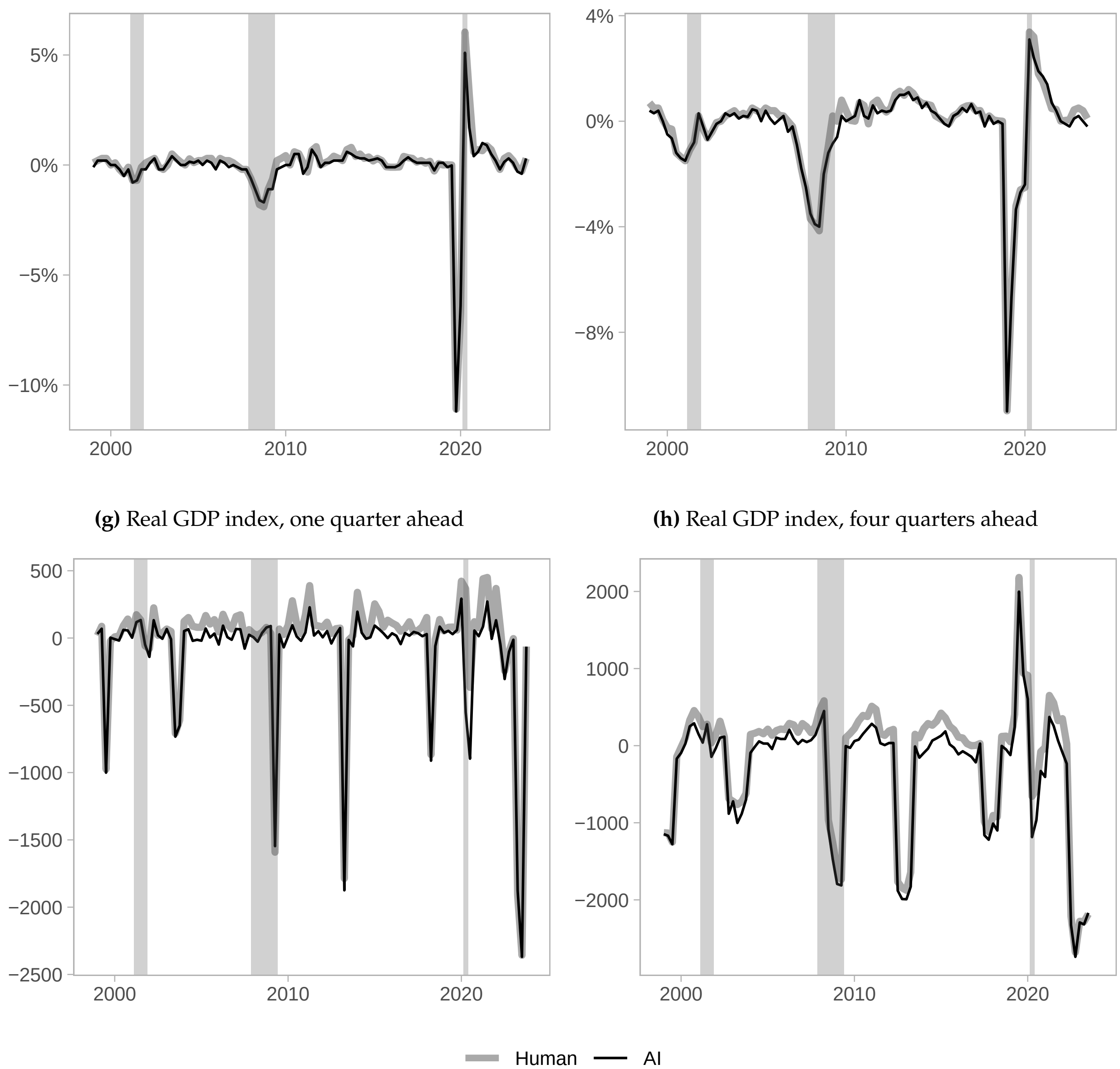


(d) 3-month Treasury bill rate, four quarters ahead



(Figure continues on next page)

Figure 10: Median forecast errors in human and AI-generated SPF *(continued)*



(e) Unemployment rate, one quarter ahead

(f) Unemployment rate, four quarters ahead

(g) Real GDP index, one quarter ahead

(h) Real GDP index, four quarters ahead

Figure 11: Prompt for generic forecaster

The figure shows the prompt used to generate AI forecasts for the generic forecaster, i.e., without inputting forecaster characteristics. The inputted variables are highlighted in brackets and blue text and consist of real-time data, previous forecasts, survey dates, and forecasting variable names.

```
You are a participant on a panel of Survey of Professional Forecasters.

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).
```

Figure 12: Prompt for generic forecaster and no real-time data

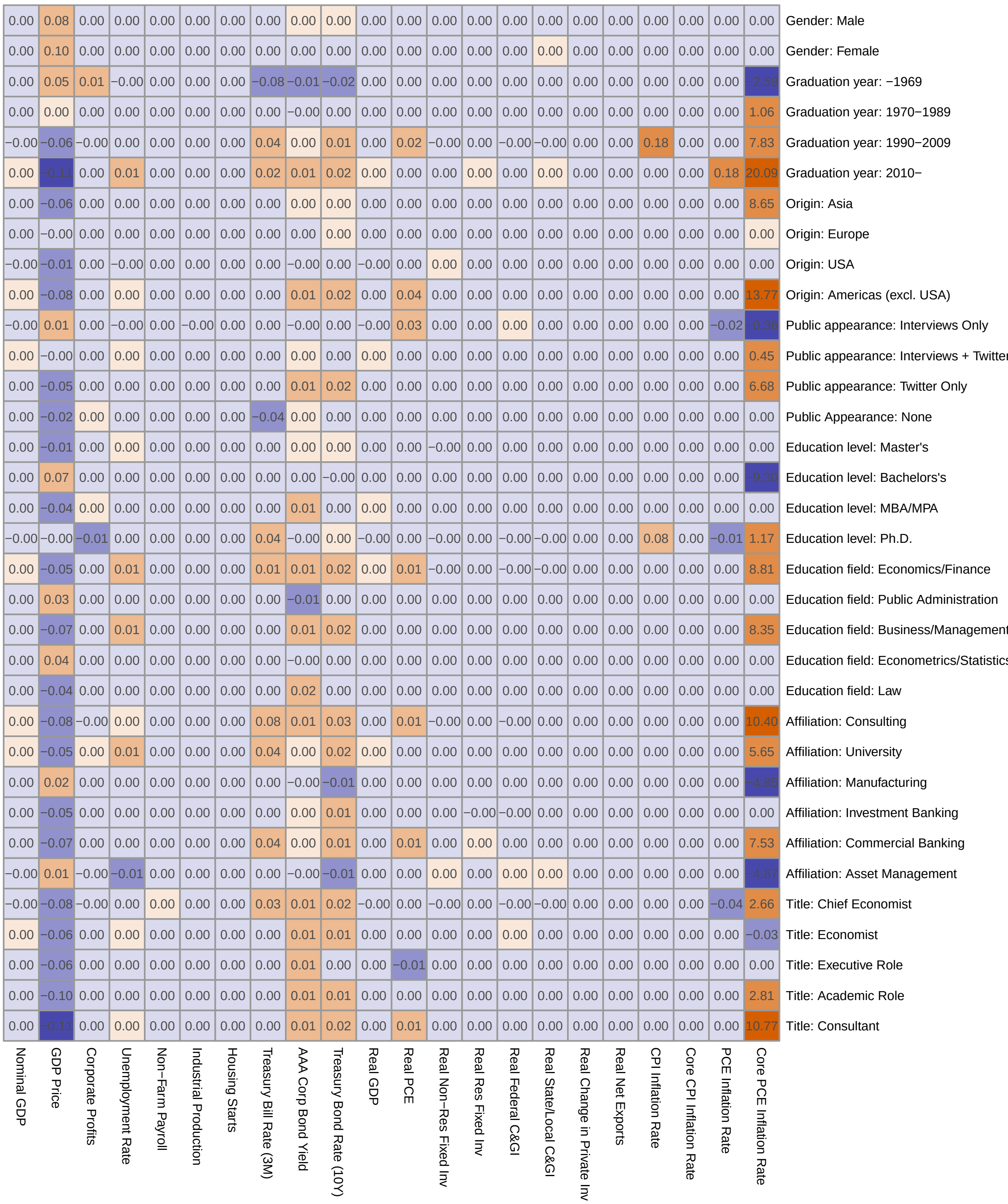
The figure shows the prompt used to generate AI forecasts for the generic forecaster, i.e., without inputting forecaster characteristics. The inputted variables are highlighted in brackets and blue text and consist of previous forecasts, survey dates, and forecasting variable names.

```
You are a participant on a panel of Survey of Professional Forecasters.

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).

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).
```

Figure 13: Coefficients from regressing individual forecast errors on characteristics

The figure shows coefficients from regressing individual AI-generated forecasts on characteristic dummy variables. All coefficients should be interpreted relative to the baseline where the characteristic is unknown. The coefficients are estimated using the sparse-group LASSO approach from Babii et al. (2022).

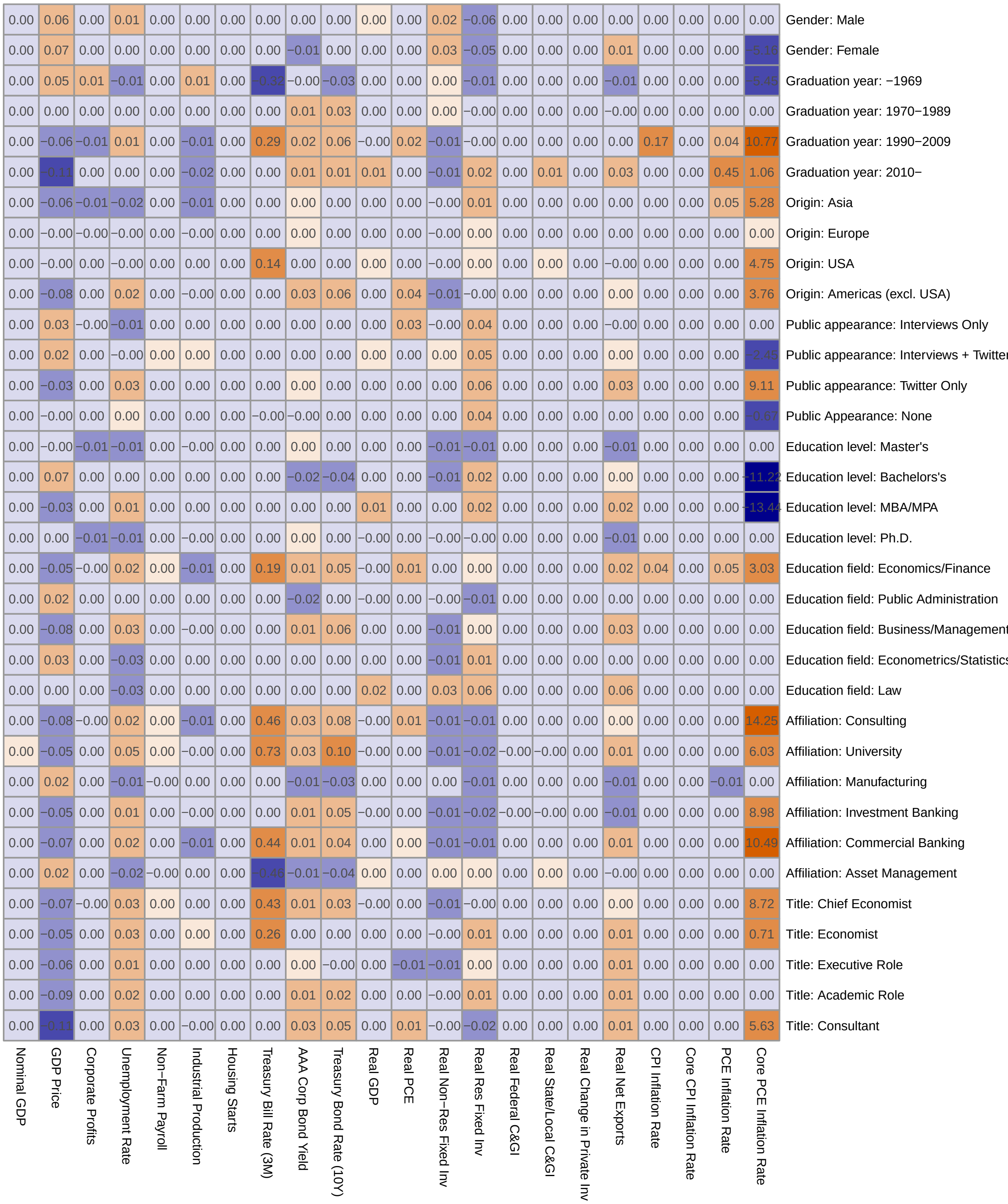
(a) Nowcast



(Figure continues on next page)

Figure 13: Coefficients from regressing individual forecast errors on characteristics *(continued)*

(b) Four-quarters ahead forecast



Tables

Table 1: Forecasting variables

The table lists the variables for which the SPF panel provides point forecasts and their definitions.

Variable	Definition
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.
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.
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.

Table 2: Definitions of forecaster characteristics

The table shows definitions of forecaster characteristics, including inferred gender, affiliation, education, position, and public media engagement. The table is based on publicly available sources and may not fully capture the breadth of 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).

Table 3: Differences between SPF and AI-generated SPF median forecasts

The table shows mean absolute percentage error and mean directional accuracy between human and AI-generated median forecasts across forecast variables and horizons.

	Mean absolute percentage error				Mean directional accuracy			
Horizon (quarter)	0	1	2	4	0	1	2	4
Nominal GDP	0.61	0.55	0.62	0.83	1.00	1.00	1.00	1.00
GDP Price	0.25	0.30	0.39	0.56	1.00	1.00	1.00	1.00
Corporate Profits	5.99	6.07	5.92	5.69	1.00	1.00	1.00	1.00
Unemployment Rate	1.82	2.14	2.27	2.78	1.00	1.00	1.00	1.00
Non-Farm Payroll	0.18	0.27	0.39	0.60	1.00	1.00	1.00	1.00
Industrial Production	1.17	1.28	1.49	2.06	1.00	1.00	1.00	1.00
Housing Starts	5.05	5.49	6.12	7.78	1.00	1.00	1.00	1.00
Treasury Bill Rate (3M)	10.65	11.76	13.57	22.00	1.00	1.00	1.00	1.00
AAA Corp Bond Yield	3.13	3.41	3.62	4.65	1.00	1.00	1.00	1.00
Treasury Bond Rate (10Y)	4.02	3.65	4.12	7.90	1.00	1.00	1.00	1.00
Real GDP	0.39	0.57	0.73	1.16	1.00	1.00	1.00	1.00
Real PCE	16.04	16.20	16.38	16.81	0.98	0.99	0.99	0.99
Real Non-Res Fixed Inv	1.00	1.35	1.88	2.72	1.00	1.00	1.00	1.00
Real Res Fixed Inv	1.37	1.64	1.96	2.88	1.00	1.00	1.00	1.00
Real Federal C&GI	0.75	0.71	0.76	0.85	1.00	1.00	1.00	1.00
Real State/Local C&GI	0.39	0.43	0.50	0.70	1.00	1.00	1.00	1.00
Real Change in Private Inv	120.86	88.70	178.97	57.17	0.83	0.88	0.87	0.87
Real Net Exports	4.49	3.90	4.08	4.45	1.00	1.00	1.00	1.00
CPI Inflation Rate	39.20	38.01	23.00	12.35	0.97	0.96	0.98	1.00
Core CPI Inflation Rate	18.86	16.84	13.90	11.29	0.99	0.99	1.00	1.00
PCE Inflation Rate	27.68	33.63	22.18	13.65	1.00	0.94	0.99	1.00
Core PCE Inflation Rate	17.78	16.61	14.13	11.21	0.99	1.00	1.00	1.00

Table 4: Forecast accuracy (mean absolute percentage errors) of SPF and AI-generated SPF forecasts

The table shows forecast accuracy (mean absolute percentage errors) of SPF and AI-generated SPF forecasts for various economic indicators across different forecast horizons. For each forecasting horizon, the table boldfaces the lowest MAPE value between the AI and human surveys.

Horizon (quarters)	0		1		2		4	
	AI	Human	AI	Human	AI	Human	AI	Human
Nominal GDP	1.49	**1.16**	**1.01**	1.15	**1.30**	1.51	**2.04**	2.31
GDP Price	**25.71**	26.03	**25.56**	25.94	**25.37**	25.86	**25.45**	26.15
Corporate Profits	6.91	**6.65**	**5.08**	8.50	**7.91**	10.69	**11.81**	13.41
Unemployment Rate	**4.40**	5.68	**7.40**	8.40	**9.76**	10.32	**13.53**	14.14
Non-Farm Payroll	**0.18**	0.35	0.63	**0.58**	0.95	**0.83**	1.55	**1.36**
Industrial Production	**0.48**	1.53	**1.98**	2.81	**3.14**	3.92	**4.59**	5.91
Housing Starts	**4.38**	7.53	**8.55**	10.42	**11.08**	12.53	**14.49**	19.63
Treasury Bill Rate (3M)	54.11	**53.99**	**98.38**	106.60	**150.79**	175.08	**360.18**	492.28
AAA Corp Bond Yield	**4.16**	6.31	**8.47**	10.09	**10.60**	12.60	**14.82**	18.47
Treasury Bond Rate (10Y)	13.76	**12.34**	**19.36**	20.48	**24.97**	26.94	**35.02**	40.02
Real GDP	**0.60**	0.86	**1.18**	1.46	**2.06**	2.38	**3.48**	3.84
Real PCE	15.92	**0.89**	16.67	**1.32**	17.37	**2.03**	18.77	**3.23**
Real Non-Res Fixed Inv	**1.27**	1.50	**2.06**	2.80	**3.56**	4.41	**6.23**	7.38
Real Res Fixed Inv	**2.03**	2.06	**2.65**	3.54	**5.25**	6.27	**9.65**	10.61
Real Federal C&GI	1.04	**0.72**	**1.52**	2.01	**2.81**	2.86	4.69	**4.53**
Real State/Local C&GI	0.63	**0.54**	**1.30**	1.58	**2.35**	2.50	**4.14**	4.20
Real Change in Private Inv	139.27	**116.79**	**105.69**	196.60	290.52	**204.84**	**217.02**	248.83
Real Net Exports	4.79	**2.87**	**3.77**	6.56	**8.31**	9.11	**13.03**	**13.03**
CPI Inflation Rate	184.45	**157.25**	225.19	**185.41**	196.22	**182.09**	**181.38**	185.22
Core CPI Inflation Rate	**40.26**	55.95	**55.93**	57.55	**55.38**	59.46	**55.16**	65.14
PCE Inflation Rate	**79.55**	83.61	**80.10**	100.36	**97.86**	112.57	**113.01**	131.74
Core PCE Inflation Rate	**4793.90**	6077.94	**4543.93**	5800.41	**4515.85**	5803.62	**4378.30**	5895.91

Table 5: Forecast accuracy (mean absolute percentage errors) of SPF and AI-generated SPF forecasts in NBER recessions

The table shows forecast accuracy (mean absolute percentage errors) of SPF and AI-generated SPF forecasts in NBER recessions for various economic indicators across different forecast horizons.

Horizon (quarters)	0		1		2		4	
	AI	Human	AI	Human	AI	Human	AI	Human
Nominal GDP	2.53	**1.42**	1.38	**1.25**	2.58	**1.86**	3.81	**3.54**
GDP Price	**38.24**	38.70	**38.37**	38.90	**38.54**	39.18	**38.66**	39.61
Corporate Profits	4.73	**4.64**	**5.23**	7.62	**11.77**	14.94	13.78	**13.05**
Unemployment Rate	**7.79**	12.20	**19.47**	20.52	23.33	**22.44**	25.71	**25.49**
Non-Farm Payroll	**0.47**	1.34	1.56	**1.49**	2.07	**1.92**	**2.91**	3.42
Industrial Production	**1.27**	2.49	**4.07**	4.59	**6.23**	6.62	**9.22**	11.39
Housing Starts	**6.58**	13.00	**16.07**	19.72	**23.59**	28.83	**30.43**	44.68
Treasury Bill Rate (3M)	95.31	**76.28**	234.85	**233.43**	**396.74**	403.69	**732.16**	933.52
AAA Corp Bond Yield	**3.92**	6.69	**5.49**	7.37	**6.31**	7.07	**7.04**	8.97
Treasury Bond Rate (10Y)	14.81	**12.03**	**15.42**	16.51	**16.22**	16.52	**17.43**	21.24
Real GDP	**1.19**	1.25	2.05	**2.02**	4.13	**3.97**	6.47	**6.10**
Real PCE	20.20	**1.28**	21.47	**2.15**	22.82	**3.35**	24.78	**5.49**
Real Non-Res Fixed Inv	**1.95**	2.81	**3.57**	4.76	**5.45**	6.90	**9.79**	11.75
Real Res Fixed Inv	**3.48**	5.17	**3.97**	4.28	**7.26**	8.09	12.44	**11.90**
Real Federal C&GI	1.57	**0.72**	**2.35**	2.90	5.28	**5.02**	10.32	**10.01**
Real State/Local C&GI	0.54	**0.40**	**2.57**	2.68	**4.34**	4.41	7.79	**7.47**
Real Change in Private Inv	135.96	**50.01**	**60.84**	82.87	747.24	**133.91**	163.44	**160.21**
Real Net Exports	6.91	**3.30**	**3.97**	6.86	12.86	**11.68**	17.23	**16.58**
CPI Inflation Rate	118.21	**112.56**	151.22	**98.70**	94.76	**53.66**	**70.56**	89.82
Core CPI Inflation Rate	**44.08**	174.36	**160.12**	164.38	**126.37**	150.38	**115.36**	154.31
PCE Inflation Rate	128.28	**125.90**	**108.71**	126.78	**150.32**	161.83	**165.07**	171.56
Core PCE Inflation Rate	1603.80	**1490.51**	3218.46	**3213.86**	**3049.08**	3796.07	**5845.02**	7103.10

Table 6: Recall versus nowcast accuracy (mean absolute percentage errors)

The table shows accuracy (mean absolute percentage errors) of AI-generated SPF nowcasts and recall of realized values.

	Recall	Nowcast
Nominal GDP	2017.44	**1.49**
GDP Price	39.62	**25.71**
Corporate Profits	47.16	**6.91**
Unemployment Rate	4.70	**4.40**
Non-Farm Payroll	1097.83	**0.18**
Industrial Production	15.51	**0.48**
Housing Starts	2303510.64	**4.38**
Treasury Bill Rate (3M)	**31.33**	54.11
AAA Corp Bond Yield	14.12	**4.16**
Treasury Bond Rate (10Y)	**8.61**	13.76
Real GDP	2148.38	**0.60**
Real PCE	2157.98	**15.92**
Real Non-Res Fixed Inv	1932.33	**1.27**
Real Res Fixed Inv	5376.60	**2.03**
Real Federal C&GI	7967.66	**1.04**
Real State/Local C&GI	2772.79	**0.63**
Real Change in Private Inv	11646.09	**139.27**
Real Net Exports	2394.60	**4.79**
CPI Inflation Rate	**177.00**	184.45
Core CPI Inflation Rate	47.55	**40.26**
PCE Inflation Rate	82.42	**79.55**
Core PCE Inflation Rate	5111.42	**4793.90**

Table 7: Relative accuracy of generic forecasters with and without real-time data

The table shows the accuracy of a synthetic forecaster without characteristics (generic forecaster) with and without real-time data relative to baseline results generated by AI with forecaster characteristics and real-time data. Relative accuracy is computed as mean absolute errors divided by baseline mean absolute errors from the "AI" columns of Table 4. Values exceeding one indicate less accuracy compared with an AI forecaster seeded with forecaster characteristics and real-time data.

	Generic forecaster				Generic forecaster w/o real-time data			
Horizon (quarters)	0	1	2	4	0	1	2	4
Nominal GDP	0.92	0.92	1.03	1.09	0.78	1.14	1.15	1.11
GDP Price	1.00	1.00	1.00	1.00	1.01	1.02	1.02	1.03
Corporate Profits	1.14	1.19	1.10	0.99	0.82	1.63	1.37	1.13
Unemployment Rate	1.02	0.99	0.97	1.02	1.29	1.10	1.04	1.03
Non-Farm Payroll	0.97	1.01	1.01	1.04	28.93	8.70	5.96	3.93
Industrial Production	1.62	1.16	1.12	1.10	3.16	1.42	1.25	1.29
Housing Starts	1.04	1.00	0.99	1.10	1.65	1.22	1.08	1.28
Treasury Bill Rate (3M)	1.09	1.02	1.00	1.02	0.74	0.82	0.92	1.04
AAA Corp Bond Yield	1.51	1.16	1.14	1.10	1.51	1.18	1.18	1.24
Treasury Bond Rate (10Y)	0.98	1.03	1.02	1.01	0.87	1.06	1.07	1.16
Real GDP	1.14	1.03	1.03	1.04	1.37	1.22	1.12	1.08
Real PCE	1.86	1.75	1.69	1.63	0.06	0.08	0.11	0.17
Real Non-Res Fixed Inv	1.39	0.93	1.02	1.03	1.17	1.33	1.23	1.17
Real Res Fixed Inv	1.20	1.07	1.04	0.99	1.00	1.30	1.17	1.10
Real Federal C&GI	1.12	0.93	0.98	1.00	0.70	1.34	1.02	0.97
Real State/Local C&GI	1.37	0.85	0.93	0.96	0.85	1.21	1.07	1.03
Real Change in Private Inv	0.87	1.45	1.17	1.04	0.70	2.05	1.00	0.94
Real Net Exports	0.97	1.19	1.08	1.08	0.59	1.69	1.13	1.00
CPI Inflation Rate	0.90	1.05	1.04	1.01	1.10	0.92	0.94	1.01
Core CPI Inflation Rate	0.85	1.01	1.00	1.06	1.20	1.04	0.95	1.05
PCE Inflation Rate	1.11	0.83	1.06	0.96	1.00	0.92	1.12	1.13
Core PCE Inflation Rate	1.05	1.03	1.04	1.04	1.02	1.02	1.04	1.07
Average	1.14	1.07	1.07	1.06	2.34	1.52	1.27	1.18

Table 8: Comparison of Topic Distribution between LDA and GPT Methods

Topic	**LDA**		**GPT**	
	Count	**%**	**Count**	**%**
Monetary Policy	2,286	43.4	2,698	51.2
Consumer Demand	1,701	32.3	1,425	27.1
Economic Recovery	1,144	21.7	0	0.0
Labor Market	1	0.0	751	14.3
Fiscal Policy	0	0.0	286	5.4
Housing Market	5	0.1	68	1.3
International Conditions	131	2.5	14	0.3
Supply Chain	0	0.0	13	0.2
Commodity Prices	0	0.0	13	0.2
Total	**5,268**	**100.0**	**5,268**	**100.0**

Appendices

A. Robustness Checks

A.1. Expanded Prompt

As a robustness check, we employ an alternative approach to creating synthetic forecast personas: an expanded prompt based on forecaster biographies generated using an LLM and personal characteristics. Table 9 presents examples of personas generated by both methods:

Table 9: Fictional Example of Main and Expanded Persona Prompts

Main Persona Prompt	**Expanded Persona Prompt**
You are a participant on a panel of Survey of Professional Forecasters. Your name is Alexandra Bryson, you graduated from University of New Hampshire with a M.A. in Economics around 2024. Today, you work as a Senior Quantitative Analyst at Ethan Investments, an Asset Management organization based in Boston, Massachusetts. You are originally from the USA.	You are a Senior Quantitative Analyst at Ethan Investments 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 Ethan Investments 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 example uses fictional characteristics for privacy reasons. However, we feed the models all names that were featured on the acknowledgement SPF lists.

The first approach is used in the main analysis and described in the main body of the paper. The second approach ("Expanded Persona Prompt") is 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.

A.2. Comparison Across Models

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

- LLM Architectures: GPT-3.5, GPT-4, and GPT-4o-mini
- Temperature Settings: 0 (deterministic) and 1 (stochastic)
- Persona Prompts: main and expanded prompts, as discussed in Section A.1

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.

- GPT-4o-mini: A lightweight variant of GPT-4, optimized for speed and efficiency.

By combining the three LLM architectures with the two temperature settings and the two persona configurations, we create twelve distinct models for evaluation. 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 section.

A.3. Comparison of Outcomes

The models used here include three variants: GPT-3.5, GPT-4, and GPT-4o mini. Each variant is presented with two temperature settings: temperature 0 (producing more deterministic forecasts) and temperature 1 (introducing greater variability). These combinations are labeled as follows: 1 and 2 for GPT-3.5, 3 and 4 for GPT-4, and 5 and 6 for GPT-4o mini. In each case, odd-numbered models (1, 3, and 5) use temperature 0, while even-numbered models (2, 4, and 6) use temperature 1. Additionally, C denotes our main approach to defining forecasters, while B represents an expanded prompt used to simulate alternative forecast definitions.
Consequently, the subplots in Figure 14 display the median nowcasts for each of these 12 configurations, with `SPF_C` using our main prompt approach and `SPF_B` representing forecasts generated with the expanded prompt approach. Figure 14 shows that in many subplots, the AI forecasts track human forecasts quite closely, showing that the AI models, even with different temperature settings and prompt definitions, can approximate human judgment in forecasting CPI inflation. The variability introduced by different model configurations yields slightly different forecast paths, with higher temperatures sometimes adding volatility and expanded prompt introducing slight shifts in AI forecasting behavior. These graphs collectively suggest that while AI models can emulate human forecasting patterns, the alignment with human forecasts may vary based on model type, temperature, and prompt approach – although these variations are small.
Figure 13 shows the median forecasts for CPI inflation rate four quarters ahead, with the set-up identical to the above configuration. These one-year-ahead forecasts generally reveal less alignment between AI models and human forecasts than the one-quarter-ahead forecasts.

The higher temperature settings tend to add variability in the AI forecasts, which leads to more noticeable deviations from human predictions. The expanded prompt approach also appears to influence the AI's forecast patterns, sometimes leading to more pronounced divergences from human forecasts. Ultimately, we see that variability is amplified in longer forecast horizons.

Figure 12 displays shows the forecast errors (realized minus predicted values) for CPI inflation rate four quarters ahead, also with the set-up identical to the above configuration. We see that in many cases, the forecast errors for AI models appear to align closely with human errors over time, particularly in lower temperature settings and with the main prompt definition. This suggests that, under certain configurations, AI models are capable of making prediction errors similar to human forecasters, reflecting similar judgment patterns. However, as temperature increases, the AI model errors become more variable and can diverge more significantly from human errors, indicating that the added randomness affects the model's consistency in capturing the same errors as human forecasters. Finally, Figure 11 underscores the difficulty AI models face in maintaining alignment with human forecast errors over extended time horizons. While deterministic settings and the main prompt design (C) help stabilize errors, the combination of higher temperatures and expanded prompts leads to increased variability and divergence from human forecasts.

Figure 14: Median of human and AI-generated SPF individual forecasts

The figure shows the median forecasts for CPI inflation rate. The black line represents human forecasts, and the gray line represents AI-generated forecasts. Shaded areas indicate NBER recession periods.

(a) CPI inflation rate, one quarter ahead-SPF1_B

(b) CPI inflation rate, one quarter ahead-SPF2_B

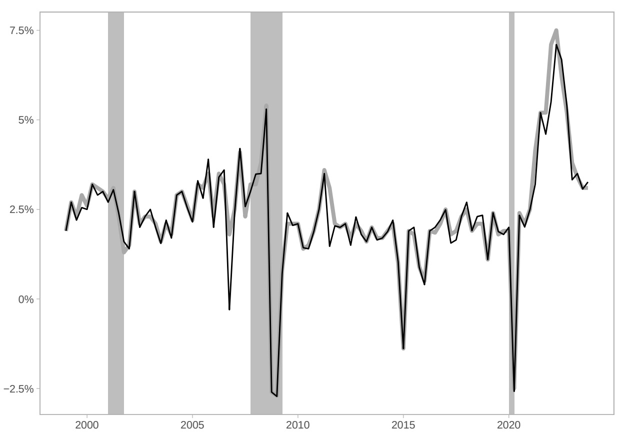
(c) CPI inflation rate, one quarter ahead-SPF1_C

(d) CPI inflation rate, one quarter ahead-SPF2_C

(e) CPI inflation rate, one quarter ahead-SPF3_B

(f) CPI inflation rate, one quarter ahead-SPF4_B

(g) CPI inflation rate, one quarter ahead-SPF3_C



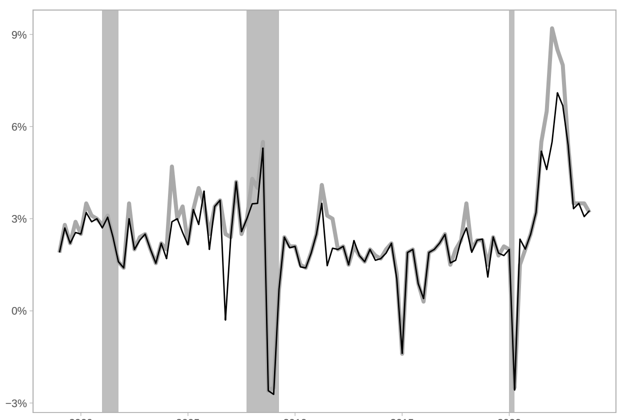
(h) CPI inflation rate, one quarter ahead-SPF4_C

Human AI

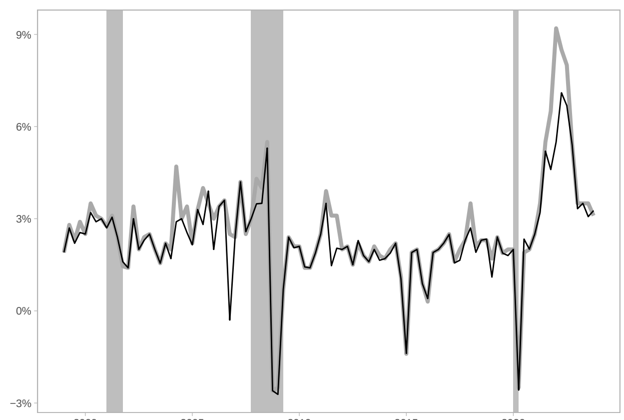
(i) CPI inflation rate, one quarter ahead-SPF5_B

(j) CPI inflation rate, one quarter ahead-SPF6_B

(k) CPI inflation rate, one quarter ahead-SPF5_C



(l) CPI inflation rate, one quarter ahead-SPF6_C

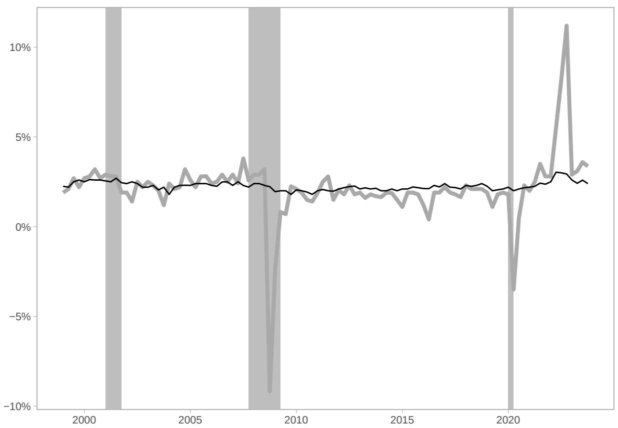


Human AI

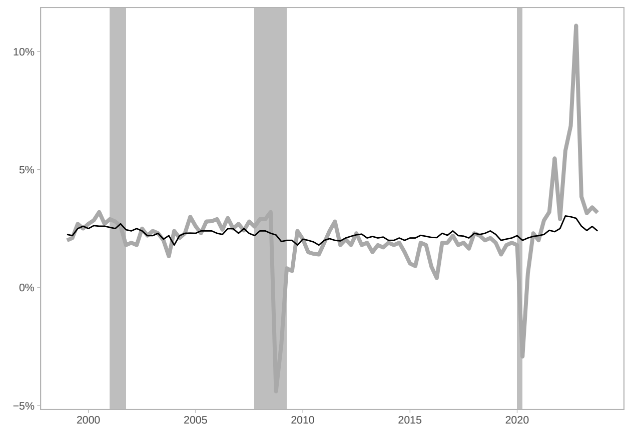
Figure 13: Median of human and AI-generated SPF individual forecasts (4-quarter ahead)

The figure shows the median forecasts for CPI inflation rate four quarters ahead. The black line represents human forecasts, and the gray line represents AI-generated forecasts. Shaded areas indicate NBER recession periods.

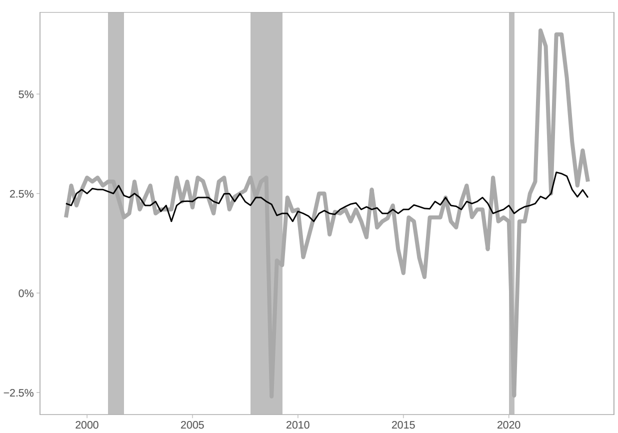
(a) CPI inflation rate, four quarters ahead-SPF1_B



(b) CPI inflation rate, four quarters ahead-SPF2_B



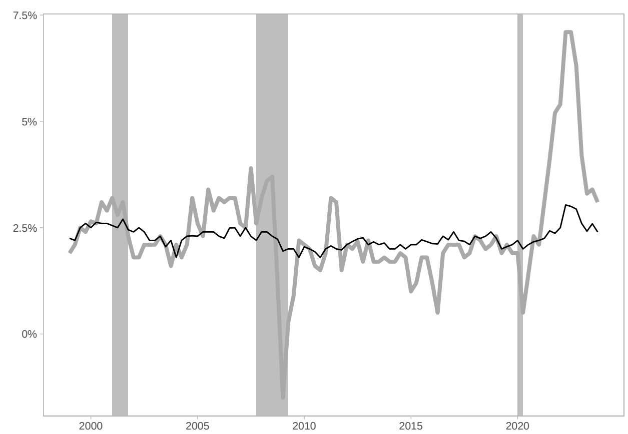
(c) CPI inflation rate, four quarters ahead-SPF1_C



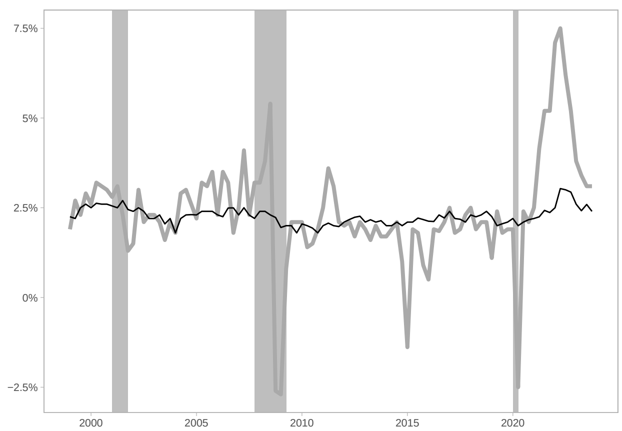
(d) CPI inflation rate, four quarters ahead-SPF2_C

(e) CPI inflation rate, four quarters ahead-SPF3_B

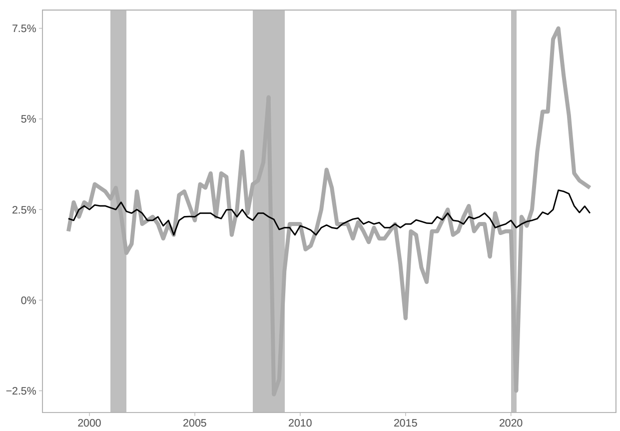
(f) CPI inflation rate, four quarters ahead-SPF4_B



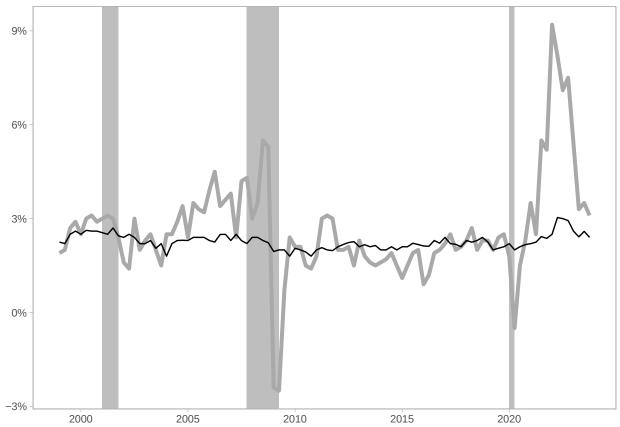
(g) CPI inflation rate, four quarters ahead-SPF3_C



(h) CPI inflation rate, four quarters ahead-SPF4_C

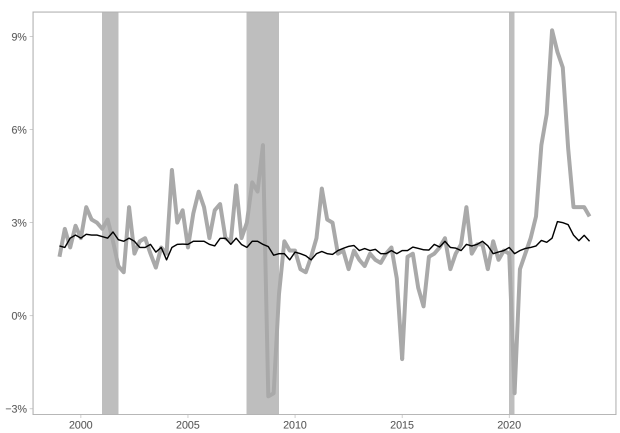


(i) CPI inflation rate, four quarters ahead-SPF5_B



(j) CPI inflation rate, four quarters ahead-SPF6_B

(k) CPI inflation rate, four quarters ahead-SPF5_C



(l) CPI inflation rate, four quarters ahead-SPF6_C

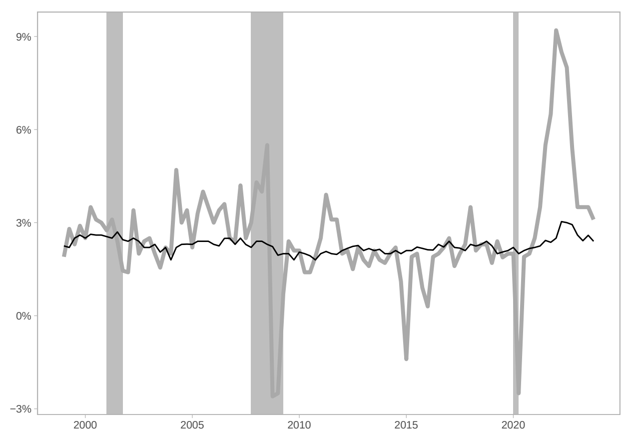
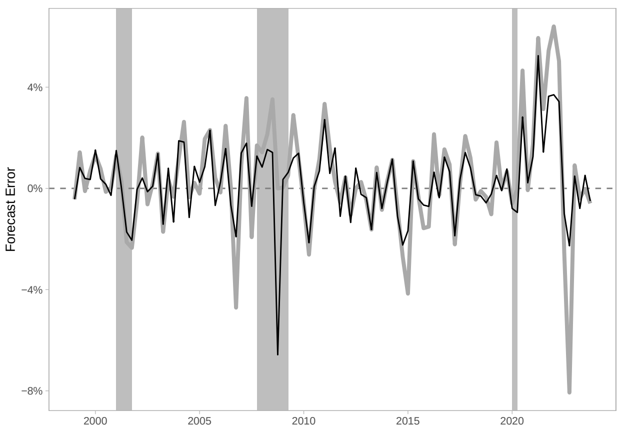


Figure 12: Forecast errors of human and AI-generated SPF individual forecasts

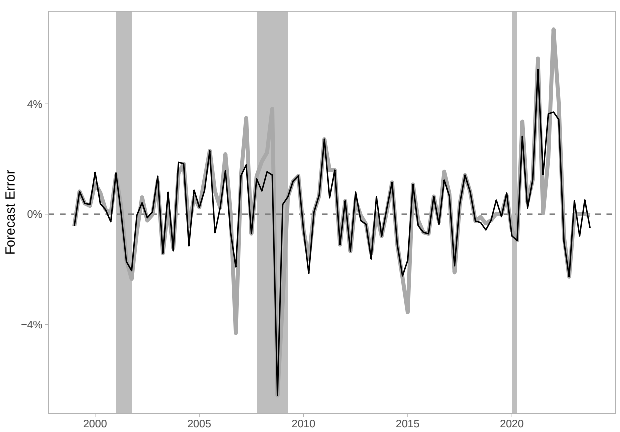
The figure shows the forecast errors (realized - predicted) for CPI inflation rate. The black line represents human forecast errors, and the gray line represents AI-generated forecast errors. Shaded areas indicate NBER recession periods.

(a) Forecast errors, one quarter ahead-SPF1_B

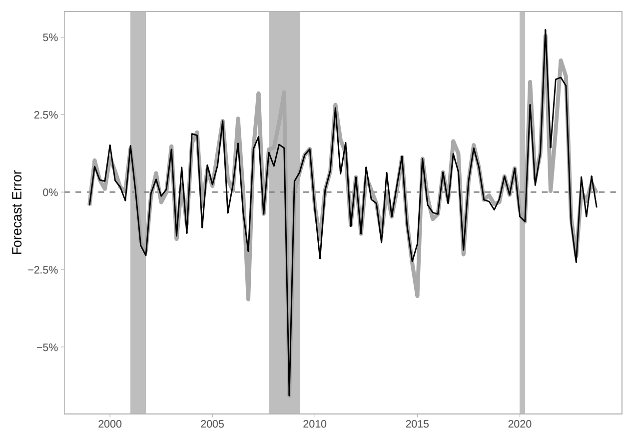


(b) Forecast errors, one quarter ahead-SPF2_B

(c) Forecast errors, one quarter ahead-SPF1_C



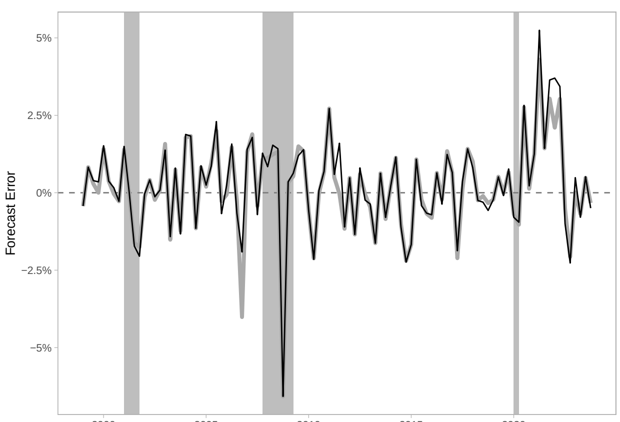
(d) Forecast errors, one quarter ahead-SPF2_C



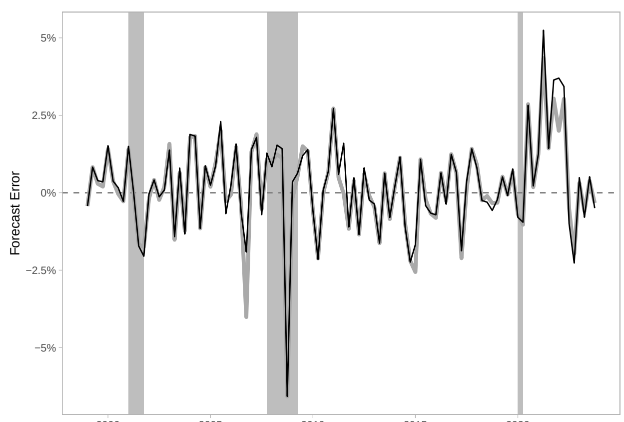
(e) Forecast errors, one quarter ahead-SPF3_B

(f) Forecast errors, one quarter ahead-SPF4_B

(g) Forecast errors, one quarter ahead-SPF3_C

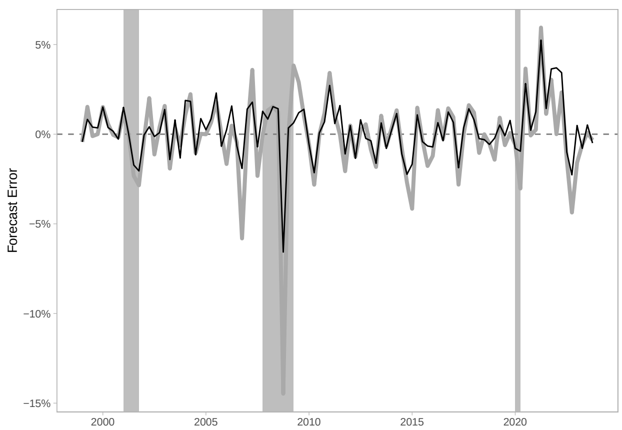


(h) Forecast errors, one quarter ahead-SPF4_C

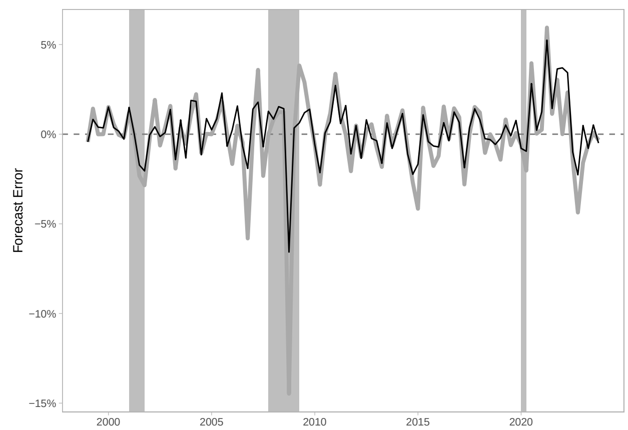


Human AI

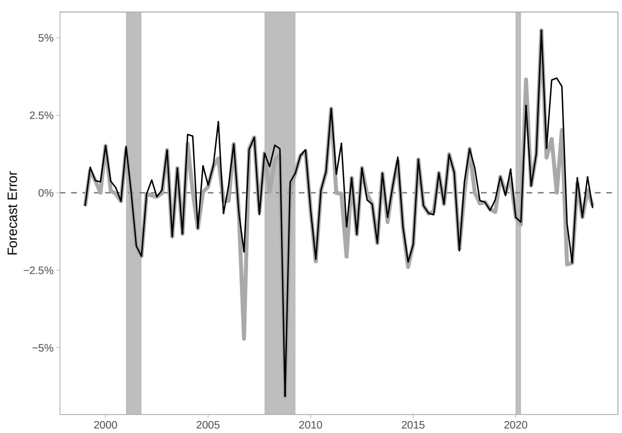
(i) Forecast errors, one quarter ahead-SPF5_B



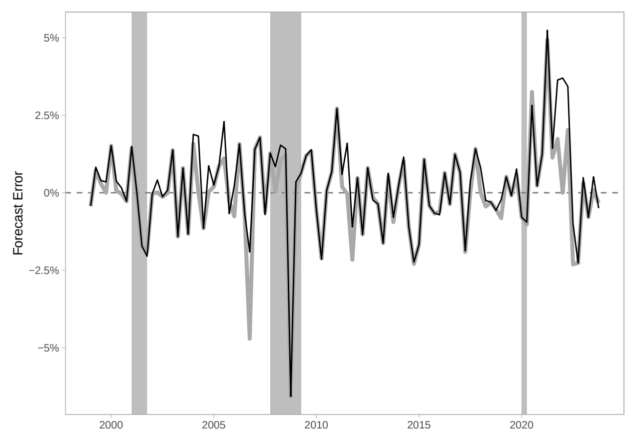
(j) Forecast errors, one quarter ahead-SPF6_B



(k) Forecast errors, one quarter ahead-SPF5_C



(l) Forecast errors, one quarter ahead-SPF6_C



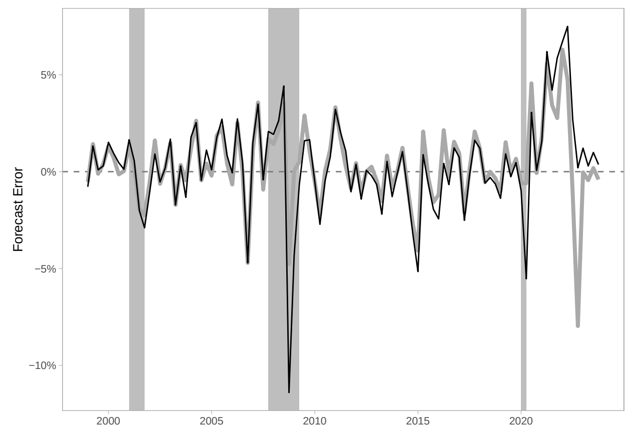
Human AI

Figure 11: Forecast errors of human and AI-generated SPF individual forecasts (4-quarter ahead)

The figure shows the forecast errors (realized - predicted) for CPI inflation rate four quarters ahead. The black line represents human forecast errors, and the gray line represents AI-generated forecast errors. Shaded areas indicate NBER recession periods.

(a) Forecast errors, four quarters ahead-SPF1_B

(b) Forecast errors, four quarters ahead-SPF2_B



(c) Forecast errors, four quarters ahead-SPF1_C

(d) Forecast errors, four quarters ahead-SPF2_C

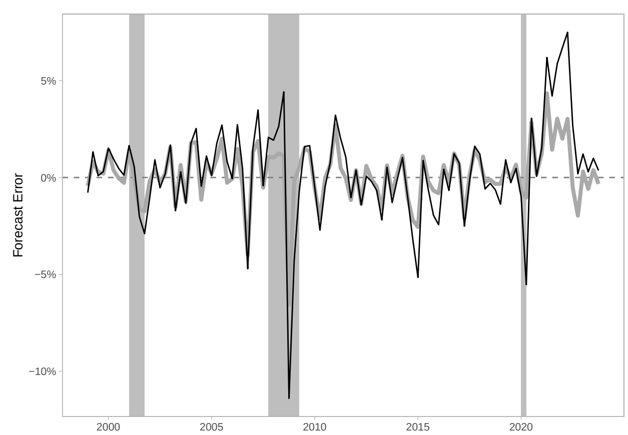
Human AI

(e) Forecast errors, four quarters ahead-SPF3_B

(f) Forecast errors, four quarters ahead-SPF4_B

(g) Forecast errors, four quarters ahead-SPF3_C

(h) Forecast errors, four quarters ahead-SPF4_C



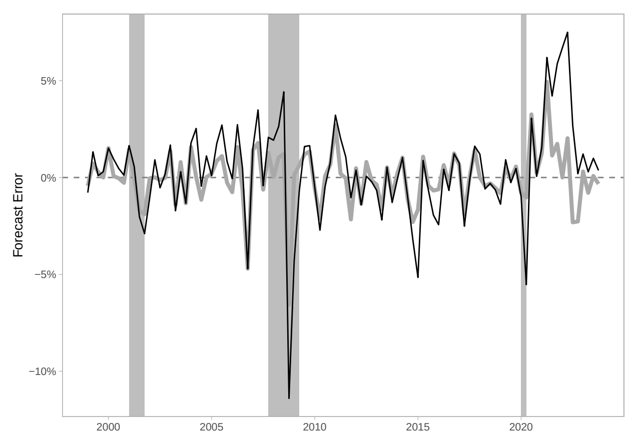
Human AI

(i) Forecast errors, four quarters ahead-SPF5_B

(j) Forecast errors, four quarters ahead-SPF6_B

(k) Forecast errors, four quarters ahead-SPF5_C

(l) Forecast errors, four quarters ahead-SPF6_C



Human AI