Simulating the Survey of Professional Forecasters[1]

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: April 2024
This Draft: February 2025

We simulate economic forecasts of professional forecasters using large language models (LLMs). We construct synthetic forecaster personas using a unique hand-gathered dataset of participant characteristics from the Survey of Professional Forecasters. These personas are then provided with real-time macroeconomic data to generate simulated responses to the SPF survey. Our results show that LLM-generated predictions are similar to human forecasts, but often achieve superior accuracy, particularly at medium- and long-term horizons. We argue that this advantage arises from LLMs' ability to extract latent information encoded in past human forecasts while avoiding systematic biases and noise. Our framework offers a cost-effective, high-frequency alternative that complements traditional survey methods by leveraging both human expertise and AI precision.

Keywords: Large Language Models; Survey of Professional Forecasters; Behavioral Finance; Generative Artificial Intelligence; Simulated Economic Agents.
JEL Codes: G4, C8, C9.

[1] 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, Indiana University, Bank of Israel, George Washington University, Federal Reserve Board, Stanford Data Science, and VIMM seminar participants for valuable comments. We thank Erik Brynjolfsson, Jonathan Benchimol, Leland Bybee, Leland Crane, Robin Horton, Huiyu Li, and Tara Sinclair for valuable feedback. Last, but not least, we thank 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 (e.g., Rudebusch, 2002; Orphanides and Williams, 2002, 2007; Patton and Timmermann, 2010; Manzan, 2011, 2021; Almås et al., 2024; Caplin, 2025), yet traditional approaches to gathering these predictions face fundamental constraints.

First, surveys are costly to conduct, which leads to less frequent data collection.[2] Second, their static design 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; Kay et al., 2023), the standard practice of maintaining forecaster anonymity obscures our ability to study these individual-level effects and their role in formation of expectation. These limitations may reduce the effectiveness of surveys in informing economic models and policy.

Because large language models (LLMs) function as advanced prediction engines that can mimic multiple aspects of human cognition (e.g., Suri et al., 2024; Niu et al., 2024), they can potentially address these challenges. Indeed, recent work shows that LLMs can reliably simulate economic agents and produce outputs that align with human reasoning (e.g., Argyle et al., 2023; Horton, 2023; Fedyk et al., 2024; Kazinnik, 2023; Zarifhonarvar, 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).

Building on these strengths, we propose a novel simulation framework that replicates the Survey of Professional Forecasters (SPF) using LLMs together with detailed profiles of its participants. The SPF plays a central role in informing policy makers,[3] serves as a key benchmark

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

[3] The SPF is used as a benchmark to the Federal Reserve staff projections in the Tealbook. See, e.g., the latest published Tealbook associated with the December 2018 FOMC meeting at the Board of Governors' website.

for evaluating forecast accuracy in academic research,[4] and is the longest-running publicly available forecasting project, providing a rich dataset of forecasts on key macroeconomic indicators (Croushore, 1993). To build our simulation framework, we start by constructing forecaster personas using hand-collected individual-level data.[5] We then deploy a suite of LLM-based forecaster personas that generate predictions using the same real-time data and survey framework as the human SPF participants.

Empirically, we show that while human and AI-generated forecasts are similar, AI-generated forecasts often outperform human experts in accuracy. For example, AI-generated predictions of real GDP and unemployment rate are statistically significant improvements over human forecasts providing more accurate forecasts in 63-81% of the surveys from 1990 Q1 to 2023 Q4. We also show that including information on individual forecaster characteristics, real-time macroeconomic data, and past SPF median forecasts in the LLM prompt is key to achieving accurate AI forecasts. For example, mean absolute errors for real GDP increase by 4-15% when excluding personal characteristics, 8-37% when excluding real-time data, while omitting past median SPF forecasts leads to a much larger increase of 53–757%. Our findings indicate that, with relevant human-generated inputs, LLMs can serve as a powerful tool for generating economic surveys for use in economic research and policy-making. LLMs should thus be considered as tools that augment, rather than substitute, human forecasters.

We interpret these results in a framework of human and AI forecasting problems inspired by Kleinberg et al. (2018). Humans bring both observable (e.g., unemployment statistics) and unobservable (e.g., intuition and context) information to their judgments, but can still make inaccurate predictions due to cognitive biases and limitations—even as professional forecasters. By contrast, traditional machine learning algorithms rely solely on observable data as they cannot process the subtle signals that humans can detect. However, algorithms are typically better than humans at exploiting the information and statistical patterns in the observable data.

[4] See Diebold et al. (1997); Giordani and Soderlind (2003); Engelberg et al. (2009); Manzan (2011); Clements (2014); Rossi and Sekhposyan (2015); Clements and Galvão (2017), among many others.

[5] Although SPF panelists are anonymized in published datasets, the Philadelphia Fed acknowledges most of their participants by listing their names on its website. We use these names to build out identities based on publicly available information.

Generative AI models, such as LLMs, bridge this divide by learning from vast amounts of text, allowing them to approximate aspects of human intuition without directly accessing unobservable factors. As a result, they can approximate the benefits of human insight without directly accessing all unobservable factors. This process can be further optimized by incorporating human inputs such as forecaster characteristics and past SPF forecasts to improve their predictive accuracy. In this way, LLMs can combine the benefits of algorithms and humans to generate forecasts that are more accurate than human forecasts. In particular, LLMs can extract latent information encoded in past human forecasts while avoiding systematic biases and noise that contaminate human forecasts.

By including personal characteristics, our framework aims to provide a more realistic representation of how individual differences affect forecasting. Although a number of economic theories—such as the full-information rational expectations (FIRE) model—assume that agents are homogeneous in their expectations and process information rationally, empirical evidence shows that both household and professional forecasters differ widely in their predictions (Mankiw and Reis, 2002; Maćkowiak and Wiederholt, 2015; Bordalo et al., 2020; Gabaix, 2020; Andre et al., 2022; Huang et al., 2022). These differences are driven at least in part by informational frictions, behavioral biases, and diverse subjective models. We capture forecast heterogeneity by explicitly incorporating individual forecaster characteristics into our simulation.

This approach also complements recent research on LLMs as economic agents, which suggests that language models, and their biases, are significantly influenced by the roles they are asked to play. In particular, Bybee (2023) demonstrates that LLMs not only simulate expectations but also systematically replicate certain types of errors. In addition, 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.

Our work also contributes to 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. Alam et al. (2024) generate

real-time macroeconomic forecasts using an LLM with AI-simulated forecast.[6] Our approach is different, as we produce macroeconomic forecasts at the individual level and incorporate forecaster characteristics and structured real-time data into the simulation. This integration allows us to replicate not only the numeric predictions but also the behavioral tendencies of individual forecasters. Furthermore, we simulate over 20 economic variables across a 25-year period, allowing us to provide a more comprehensive assessment of large language models' forecasting performance.

The use of computational techniques to represent individual behavior is not a new idea in economics. Our approach differs from traditional methods like agent-based modeling (ABM).[7] Unlike traditional ABM agents, which rely on predefined rules and have limited decision-making capabilities, LLM-simulated agents can adapt and make nuanced decisions in complex social simulations.[8] While this increased complexity can enhance performance, it also introduces challenges related to consistency and robustness. To address these challenges, we carefully validate our results. We replicate our simulation using a number of LLM configurations, including a suite of GPT, Llama, and Deepseek models under both deterministic and stochastic settings, in order to assess the impact of model architecture and randomness on forecast accuracy. We also vary the definition of the forecaster persona within a fixed prompt structure to isolate the effects of framing on the model's predictions.[9] We show that LLMs struggle to recall historical data but are more accurate when forecasting based on structured inputs; nonetheless, to mitigate look-ahead bias we restrict the models to using only the information that was available at the time of forecasting. Finally, we report findings from an out-of-sample evaluation using data from 2024, i.e., after the cut-off date of the training data sample.

The rest of the paper is organized as follows. Section 2 outlines our framework for understanding AI and human forecasting. Section 3 details the structure and institutional context

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

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

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

[9] These robustness exercises are discussed in more detail in the appendix.

of the SPF. We describe how we extract information on the participants in the SPF panel and outline our methodology for generating synthetic forecasts using LLMs in Section 4. Section 5 describes the results, and Section 6 provides a discussion of these results in the context of our framework. Section 7 concludes.

2 A Framework of Human and AI Forecasting

Consider the problem of forecasting a vector of economic variables H periods ahead, denoted by y_{t+H}. We assume that the true forecasting process is governed by a function f plus an unpredictable zero-mean error ε_{t+H}:

$$y_{t+H} = f\big(x_t, z_t\big) + \varepsilon_{t+H}. \tag{1}$$

The function f that depends on two types of information available at time t: observable data x_t and unobservable factors z_t. The unobservable factors represent soft information such as private insights, tacit domain knowledge, internalized heuristics, and intuition.

Humans can access both observable and unobservable information. Although it is unrealistic to presume that humans have access to all information that predicts y_{t+H}, we shall assume that humans can observe both x_t and z_t to fix ideas in a simple framework.[10] However, they process information imperfectly, which introduces an error term $\Delta_{i,t}$, varying over individuals i and time t. $\Delta_{i,t}$ is not assumed to have zero mean and hence may incorporate systematic biases. By defining $\Delta_{i,t}$ as the difference between $f(x_t, z_t)$ and a subjective expectation ^{H_i} of $f(x_t, z_t)$, the human forecast can be described as:

$$h_{i,t} = \mathbb{E}^{H_i}\Big[f\big(x_t, z_t\big) \mid x_t, z_t\Big]. \tag{2}$$

Note that the human subjective expectation varies with individuals i, reflecting that humans do not necessarily agree on expectations, even given the same set of information.

[10] A more realistic model may introduce a third type of information which humans cannot access. Such extension could be introduced into our framework without changing our conclusions.

Algorithms, by contrast, can only access x_t, but they process x_t efficiently:[11]

$$m_t = \mathbb{E}\Big[f\big(x_t, z_t\big) \mid x_t\Big]. \tag{3}$$

This represents traditional algorithmic forecasting using machine learning techniques. We, however, employ LLMs, which form expectations in a nondeterministic manner, using patterns learned from training data, which mainly consists of human-generated texts. To capture this feature, we introduce an AI-specific expectations operator $\mathbb{E}^{\text{AI}}(\cdot)$:

$$m_t^{\text{AI}} = \mathbb{E}^{\text{AI}}\Big[f\big(x_t, z_t\big) \mid x_t\Big]. \tag{4}$$

We argue that LLMs can proxy human-formed expectations more closely than traditional algorithms, hence $\mathbb{E}^{H_i}(\cdot) - \mathbb{E}^{\text{AI}}(\cdot) < \mathbb{E}^{H_i}(\cdot) - \mathbb{E}(\cdot)$. But, human and AI expectations remain different, first and foremost because their conditional information is different ($\{x_t, z_t\}$ versus only x_t). To narrow the difference between human and AI-generated forecasts, we bridge the gap between the limited data x_t available to the algorithm and the unobservable factors considered by human forecasters. We accomplish this by expanding the information set available to the AI forecaster. First, we include past median human forecasts:

$$m_t^{\text{AI}} = \mathbb{E}^{\text{AI}}\Big[f\big(x_t, z_t\big) \mid x_t,\, \bar{h}_{t-1}\Big], \tag{5}$$

where the lagged median human forecast $\bar{h}_{t-1}$ is defined as:

$$\begin{aligned}
\bar{h}_{t-1} &= \text{median}\Big(h_{1,t-1},\, h_{2,t-1},\, \ldots,\, h_{N,t-1}\Big) \\
&= f\big(x_{t-1}, z_{t-1}\big) + \text{median}\Big(\Delta_{1,t-1},\, \Delta_{2,t-1},\, \ldots,\, \Delta_{N,t-1}\Big) \\
&= f\big(x_{t-1}, z_{t-1}\big) + \bar{\Delta}_{t-1}.
\end{aligned} \tag{6}$$

Since $f(x_{t-1}, z_{t-1})$ is common to all forecasters, the median essentially reflects the true past signal plus a median bias, $\bar{\Delta}_{t-1}$.

[11] While algorithms can be fed all relevant observable variables in theory, they are likely only given a subset of x_t in practice. For example, in our application, x_t consists of real-time macroeconomic data and past median SPF forecasts. This technicality could be introduced into our framework without changing our conclusions. We abstract from this extension to maintain simplicity of our framework.

Second, note that the human bias term is partly determined by personal characteristics of forecasters, such as experience, education, age, and area of expertise. Let these characteristics be represented by a vector $w_{i,t}$.[12] Specifically, we decompose the bias as:

$$\Delta_{i,t} = \gamma\big(w_{i,t}\big) + e_{i,t}, \tag{7}$$

where $e_{i,t}$ represents the residual bias not captured by $w_{i,t}$. We therefore also include these characteristics in the AI forecasting problem. This specification introduces an individual i subscript on the AI forecast, similar to the human forecast.

Altogether, the AI forecasting problem can then be expressed as:

$$m_{i,t}^{\text{AI}} = \mathbb{E}^{\text{AI}}\Big[f\big(x_t, z_t\big) \mid x_t,\, f\big(x_{t-1}, z_{t-1}\big) + \bar{\Delta}_{t-1},\, w_{i,t}\Big] \tag{8}$$

and involves an approximation error given by:

$$\Delta_{i,t}^{\text{AI}} = m_{i,t}^{\text{AI}} - f(x_t, z_t). \tag{9}$$

The AI forecaster accesses measured data along with the past median human forecast and a set of personal characteristics. The lagged human median forecasts - although imperfect and contaminated by biases and noise - contain information about z_t because this information is available to humans. The AI model leverages these forecasts as proxies for the unobserved factors. The inclusion of personal characteristics helps the model learn the systematic patterns that vary across forecasters. This formulation emphasizes that our framework augments traditional forecasting—not by substituting human judgment with algorithms, but by integrating human-derived insights into the forecasting process.

Finally, we stress that AI forecasters are therefore not given an advantage over humans in our framework. On the contrary, humans have access to more information given by z_t than AI forecasters, including the potential use of sophisticated forecasting models that help them exploit patterns in the data.

[12] While most personal characteristics tend to remain constant over time, they can be time-varying, e.g., a forecaster obtaining an advanced degree or changing employment. We therefore include a subscript t in the characteristic vector.

3 The Survey of Professional Forecasters

The Survey of Professional Forecasters is a quarterly survey of U.S. economic experts, initially launched in 1968 by the American Statistical Association and the National Bureau of Economic Research and conducted by the Federal Reserve Bank of Philadelphia since 1990.[13] Response deadlines typically fall mid-quarter.[14] Although the survey form has evolved over time, with eight updates from 1999 to 2023, its core structure remains intact.

3.1 Forecast Variables

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 point forecasts are divided into three sections: Section 1: U.S. business indicators, Section 2: Real GDP and its components, and Section 3: CPI and PCE inflation, and they are summarized in Table 1. Altogether, these variables cover a wide range of indicators with different properties, including both growth rates and levels. The survey also 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 point forecasts for the current quarter and one to four quarters ahead, covering the period from 1999 to 2023.[15]

Table 3 summarizes the descriptive statistics of the forecast variables in Section 1-3. The data is from the real-time macroeconomic data set from the Federal Reserve Bank of Philadelphia, which we format to match the data transformations in the SPF.[16] The heterogeneity across variables is vast: absolute mean values vary between -0.21 (e.g., core PCE inflation rate) and 138,530 (e.g., non-farm payroll); some variables are restricted to the positive domain (e.g.,

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

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

[15] We exclude years prior to 1999 because forecasters' names were not publicly available before that time. Further, for our out of sample exercise, we focus on 2024.

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

housing starts), while others can take negative values (e.g., real net exports); some exhibit high variance (e.g., inflation rates), others less (e.g., industrial production); some are thin tailed (e.g., real GDP), while others are leptokurtic (e.g., core PCE inflation); autocorrelation coefficients ranges from practically zero (CPI inflation) to 0.85 (GDP price). Forecasting variables with wildly different properties may pose a challenge for both human and AI forecasters.

3.2 The SPF Panel

SPF panelists are typically 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.[17] Panelists work in diverse environments: some at forecasting firms, others at banks or financial institutions; the panel also includes chief economists from industry trade groups and manufacturers, as well as academics specializing in forecasting methods.

Although SPF panelists 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.[18] The SPF panel composition changes over time as respondents join or leave, and some may skip certain questions, resulting in an unbalanced panel. Since the Philadelphia Fed took over the survey in 1990, the average panel size has been fluctuating around forty participants (see Figure 2).

Forecasters use various methods to produce predictions. Stark (2013) shows 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.

---

[17] 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.

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

3.3 Who Are the Forecasters on 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.[19] Many panelists also share their participation on social media or professional platforms.

We start by compiling forecaster names from Philadelphia Fed acknowledgments. Using these names and publicly available 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 mater, degrees, and professional roles. When available, we also include organization locations, countries of origin, and social media presence. For country of origin, we consider indicators like high school location on profiles such as LinkedIn.[20] We describe these characteristics in Table 2.

Empirical research shows that personal backgrounds strongly influence people's behavior and beliefs. For example, individuals who experienced 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). Salle et al. (2023) and Malmendier and Nagel (2016) also show that memories and personal experiences of inflation shape inflation expectations. Geography and institutional affiliation could also impact forecasting approaches (Batchelor, 2007; Hong and Kacperczyk, 2010).

As Figure 3 shows, panelist tenure varies widely, with some contributing for over 25 years and others for shorter periods, adding both established and fresh perspectives. Figure 4 provides an overview of the SPF panel over the past 20 years, showing a diverse group with varied educational backgrounds, roles, and affiliations. Gender distribution has seen slight improvement over time, though men remain the majority. Most panelists are from the USA, with some European and Asian representation . While public engagement remains limited, more forecasters now use social media platforms like Twitter (X) and grant interviews. Most panelists hold a

[19] A screenshot of such acknowledgments is provided in Figure 1.

[20] 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.

Ph.D. in Economics or Finance; the share of Master's degree holders has remained steady, while Bachelor's and MBA/MPA degrees are less common. Chief Economists and Economists predominate, though consultants and analysts are increasingly visible. Affiliations span consulting firms, universities, and asset management; consulting and academia remain prominent, while commercial banking has declined in favor of investment banking and asset management. Figure 5 shows sector distribution over time, with the panel divided into Financial, Non-Financial, and Unknown categories. From the late 1990s to 2024, financial sector representation has notably increased.

4 Simulating the SPF with LLMs

We use the information described in the previous section to create personalized profiles that mirror human participants, and generate synthetic forecaster personas (AI forecasters) using an LLM prompt. To mimic the environment of the human SPF forecasters, we also include real-time information and past SPF median forecasts in the prompt. Using the prompt and a choice of LLM, we simulate the SPF by generating point forecasts of each of the synthetic forecasters for each quarter from 1999 Q1 to 2023 Q4.[21]

4.1 Models

We use a set of LLMs to run the simulated survey. For our main analysis, we use the GPT-4o mini with temperature 1.0.[22] 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 sim-

[21] We also consider the most recent year 2024 Q1 to 2024 Q4 in an out-of-sample exercise testing the model on a sample extending beyond its training data.

[22] We conduct a robustness check using GPT-3.5, GPT-4, Llama-3.3-70B-Instruct-Turbo, and DeepSeek-V3. The results are similar to our main analysis and are available upon request.

ilarities between GPT-4o and GPT-4o mini, we selected GPT-4o mini as our primary model.[23] Our implementation uses an AI agent built with the OpenAI Assistants API to handle tasks like query responses, and additional data processing.[24]

4.2 Prompt Design

The prompt, shown in Prompt 1, 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 1-2 sentence explanation of their predictions.[25] 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.

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. To replicate the SPF process as closely as possible, we also provide our AI forecasters with a data environment that is as close as possible to that of human forecasters. This data environment includes past median SPF forecasts and real-time macroeconomic data, both provided by the Federal Reserve Bank of Philadelphia from two datasets.

5 Results

This section presents our empirical results.[26] With two sets of forecasts (human and AI) for five forecasting horizons across more than 20 variables, the amount of results is massive. We therefore report results for selected horizons only, and we mainly focus on four key variables in

[23] Temperature is a parameter that controls the randomness of the model's outputs.

[24] At the time of this writing, the Assistants API are typically used for complex, multi-tool tasks.

[25] These instructions are given via system prompt.

[26] Forecasts are winsorized at the median ± 3 standard deviations of the realized data to exclude potential erratic LLM answers. Winsorization only impact the "recall" results in Table 8.l

our descriptions of the results: the CPI inflation rate, 3-month Treasury bill rate, unemployment rate, and real GDP index. 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.

5.1 Comparing AI and Humans

First, we evaluate how closely AI and human forecasts are aligned. Figure 6 compares AI and human median forecasts over time for the one- and four-quarter horizons for the CPI inflation rate, 3-month Treasury bill rate, unemployment rate, and real GDP index. The figure shows strong alignment between human and AI forecasts, 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.

Next, we consider the comparison of individual-level forecasts. Since the SPF is anonymized, we cannot match individual-level AI and human forecasts. Instead, we focus on the distributions of individual-level forecasts. Figure 7 shows density plots of individual forecasts for all forecasters for the following key periods: 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 is centered around realized values. In some cases, the distributions of human and AI forecasts align almost perfectly, see, e.g., the one-quarter ahead forecast of the unemployment rate and real GDP in 2008 Q3 and most of the 1999 Q1 forecasts. In addition, the dispersion of forecasts of real GDP in 2020 Q3 is unusually high for both AI and human forecasts. In other cases, there are notable differences, see, e.g., one-quarter ahead forecasts of the CPI inflation rate in 2008 Q3 and 2020 Q2 and the unemployment rate forecasts in 2020 Q2. The extent to which human and AI forecasts distributions align is thus highly dependent on the variable, forecast horizon, and quarter. We also note that there is no clear pattern on whether

humans and AI are more or less aligned during uncertain times such as 2008 Q3 and 2020 Q2.

For a more formal comparison, Table 4 compares distributional moments via two panels representing a one-quarter-ahead and a four-quarter-ahead horizon. Rather than focusing solely on central tendencies (e.g., median), we look at differences using multiple moments of the forecast distributions: the median, quartiles, mean, standard deviation, skewness, and kurtosis. For each horizon, we report differences across our sample of forecasting variables along with randomized tests of the statistical significance of their differences.[27]

The results in Table 4 reveal statistically significant distributional gaps between AI and human forecasts. Key differences in the table highlight that divergences between human and AI forecasts are not limited to simple level shifts. For instance, a negative median difference for indicators like nominal and real GDP suggests that one group's forecasts consistently lie lower than the other's, while differences in the quartiles reveal disparities in the lower and upper tails of the forecast distributions. Variations in standard deviation indicate differences in forecast dispersion. Moreover, significant differences in skewness and kurtosis point to contrasting distribution shapes, i.e., while central measures might be relatively similar for some variables in the sample, the forecasts differ markedly in terms of asymmetry and tails. These differences are most noticeable in the four-quarter-ahead forecasts, where greater variation in dispersion and shape suggests that uncertainty and disagreement grow over longer forecasting periods.

While most of the differences in distributional moments are statistically significant, some are not significant in an economic sense. For example, the median CPI inflation forecasts differ by just 30 basis points (one quarter ahead) and one basis point (four quarters ahead), which is small relative to the median of 2.5 as reported in Table 3. Other moments of the CPI inflation forecast distributions are, however, economically different, see, e.g., skewness and kurtosis.

In sum, even when human and AI forecasts appear similar, the underlying distributional characteristics may vary considerably, with divergences becoming more marked as the forecast

[27] Randomization tests assess whether the difference between human and AI forecasts is larger than what is expected to arise by chance. The method involves pooling and shuffling forecasts into random pseudo-groups, repeatedly recalculating distributional differences (e.g., mean, median) 10,000 times to generate a null distribution. If the actual observed difference is more extreme than most shuffled differences, it is deemed statistically significant.

horizon extends. This shows that humans and AI forecasters might have similar centers for some forecasts but differ greatly in whether they expect outliers, how spread out the forecasts are, or which side the forecast distribution is skewed toward. Next, we evaluate which type of forecaster—human or AI—is more accurate.

5.2 Evaluating Forecast Accuracy

In this section we compare the performances of human and AI forecasters. Our goal is to compare the size of the human bias $\Delta_{i,t}$ relative to the AI approximation error $\Delta_{i,t}^{\text{AI}}$.

Table 5 compares the mean absolute errors (MAEs) of AI-generated versus human predictions across all economic indicators over three forecast horizons—nowcast and one and four quarter(s) ahead. AI forecasts generally outperform human forecasts, particularly at forecasting variables for one-four quarters ahead. At the nowcasting horizon, the results are mixed and mostly statistically insignificant. For example, the AI-generated four-quarter ahead forecasts of the 3-month Treasury bill rate achieve an MAE of 1.15, which is a statistically significant improvement over the human MAE of 1.21. The corresponding nowcast MAE is higher for AI compared with human forecasts, but the difference is not statistically significant.

This comparison is based on averages across all survey quarters. Table 6 shows the percentage of surveys where AI forecasts are more accurate than human forecasts with boldfaced values above 50%. The results show that AI forecasts exhibit superior accuracy more frequently than human forecasts across nearly all variables and horizons. For example, the four-quarter ahead forecast of the 3-month Treasury bill rate is more accurate for AI than for humans in 60% of the surveys, which is statistically significant from 50%.

Finally, we compare forecast accuracy among the best performing human and AI forecasters. Specifically, for each variable and forecasting horizon, we identify the AI and human forecasters that achieve lowest MAEs on average across all survey quarters. We report the MAEs for these *best forecasters* in Table 7. These results represent an upper limit for the accuracy attainable by human and AI forecasters, respectively. Our results suggest that even the best human forecasters are outperformed by AI both in forecasting and nowcasting across most variables.

5.3 The Role of Prompt Inputs

We have shown that while AI and human forecasts are qualitatively similar, they differ quantitatively. Moreover, AI forecasts tend to outperform human forecasters. Our framework in Section 2 attributes these result to the AI-specific expectations operator $\mathbb{E}^{\text{AI}}(\cdot)$, which uses patterns learned from the training data to form predictions, along with the conditioning information given in the LLM prompt, i.e., forecaster characteristics, real-time data, and past forecasts. In this section, we evaluate the roles of these inputs for forming expectations. Specifically, we compare several variations of "stripped-down" AI forecasters to the baseline that is fully informed, i.e., has access to all three information types.

We begin by simulating forecasts for a "generic" forecaster that omits forecaster characteristics. Prompt 2 illustrates this setup, in which the model relies on real-time data, past median SPF forecasts, and pre-trained knowledge. Next, we construct another scenario with the same generic forecaster but without access to real-time data, shown in Prompt 3. In this case, the model can only draw on past median SPF forecasts along with its pre-trained knowledge, thereby mimicking a professional forecaster who lacks timely data releases. Finally, we consider a setup without any inputs so that the model only relies on its training data. This setup is shown in Prompt 4.

Table 8 compares the accuracy of these variants to the baseline results using all information, see Prompt 1. The table reports MAEs as ratios of baseline MAEs, where values of one indicate performance on par with the fully informed benchmark and values above one indicate worse performance than the baseline. For the "generic" variant (i.e., no forecaster-specific traits but with real-time data and past median SPF forecasts), MAE ratios remain near one for most variables and horizons, suggesting only a slight loss in accuracy relative to the baseline. In addition, the MAE ratios are usually not statistically different from one. In the context of our framework, the AI forecasters lose some ability to account for systematic biases (the component $\gamma(w_{i,t})$ in the bias term) when forecaster characteristics are omitted, but retains much of the forecasting ability due to information coming from real-time data and past median SPF forecasts approximating $\bar{h}_{t-1}$. Therefore, forecasting performance only degrades slightly when omitting characteristics.

In contrast, removing real-time data often pushes the ratios further above one, showing a clearer deterioration in forecast quality. Eliminating *both* real-time data and past median SPF forecasts is associated with even higher MAEs ratios, and we can typically reject that these ratios are equal to one with less than one percent p-values. For some variables, the loss in performance is extreme. For example, the four-quarter ahead forecast of real GDP is 53% worse compared with the baseline, whereas the nowcast is worse by a factor greater than seven. Eliminating these pieces of information means the AI must rely on its pre-trained knowledge without current context or the valuable proxy for z_t. In the framework, this is equivalent to conditioning only on an incomplete information set, leading to a larger approximation error and thus much higher MAEs.

Overall, the performance of AI forecasters becomes progressively worse as more inputs are stripped away from thee prompt. Particularly past median SPF forecasts strongly contribute to predictive accuracy. Hence, these results confirm that providing forecaster characteristics, real-time data, and past SPF forecasts meaningfully improves the quality of AI-generated economic predictions.

5.4 Are LLMs Forecasting or Recalling?

In this section, we address and rule out an alternative explanation for why AI forecasters might outperform humans: the possibility that our AI forecasters inadvertently use future information to inform their forecasts.

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 repeating the patterns 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."[28]

[28] The term "temporal leakage" is often used to describe all forms of time-related data contamination, including lookahead bias.

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 further check for contamination by temporal leakage, we prompt the model to *recall*—rather than forecast—the actual values of our macroeconomic variables for each survey quarter, see Prompt 5 for details. The final column in Table 8 shows the MAEs from this recall exercise as ratios of the baseline nowcast MAEs. The accuracy drops considerably, with MAEs ratios often surging into double-digit numbers.[29] For example, the LLM recalls the values real GDP with a MAE that is 38 times larger than the baseline nowcasting MAE.

Finally, we test forecast accuracy in a genuine out-of-sample exercise on 2024 Q1-Q4 data, which excludes the training data sample and therefore removes concerns of temporal leakage by construction. Tables 9 and 10 report MAEs for respectively median and best forecasts for the nowcasts and one- and two-quarter-ahead forecasts.[30] The results are mixed: the median human forecasts often outperform AI, but the AI survey has generally lower forecasting errors when considering only the best AI and human forecasters. We note that these conclusions are based on just a few quarters of observations, and it is therefore difficult to establish statistical significance due to lack of power.

[29] We note that these results are achieved after removing outliers given by forecasts that are more than three standard deviations away from the median. Without removing outliers, the MAE ratios are often in the magnitude of thousands.

[30] At the time of writing, we cannot evaluate the performance of four-quarter forecasts due to lack of realized data values.

6 Discussion

Human and AI forecasters differ largely in how they process information. Humans incorporate unobservable insights z_t directly into their judgment, enabling them to adapt to unforeseen shocks or structural changes. This reflects a dual process of cognition: a rapid, intuitive System 1 that quickly identifies patterns and a slower, deliberative System 2 that carefully integrates broader context and qualitative signals (Kahneman, 2003). However, even with this nuanced approach, human forecasts are susceptible to individual biases $\Delta_{i,t}$. By contrast, AI models primarily process the observable data x_t. Their operation is akin to a System 1 process—fast, automatic, and heavily reliant on historical data patterns.

In our framework, the relative accuracy of human versus AI forecasts depends on the magnitude of human bias $\Delta_{i,t}$ relative to the AI approximation error, $\Delta^{\text{AI}}_{i,t}$. Our results indicate that $\Delta^{\text{AI}}_{i,t} < \Delta_{i,t}$, but only when AI models are enriched with lagged human forecasts $\bar{h}_{t-1}$ and forecaster characteristics $w_{i,t}$. Our empirical results thus reinforce the framework's central claim: accurate forecasting relies on integrating both observable data x_t and proxies for the unobservable factors z_t. We show that including lagged human forecasts (a proxy for z_t) and forecaster-specific characteristics helps the AI better approximate the function $f(x_t, z_t)$, thereby reducing forecast errors. The sensitivity of the MAE ratios to the removal of inputs makes it clear that integration of real-time data and credible proxies for unobservable human insights is key. This integrated approach effectively narrows the gap between the algorithm's conditional expectation (which is limited to x_t) and the full, nuanced human forecast (which benefits from z_t).

The evidence suggests that a hybrid system where AI models are enriched with lagged human forecasts $\bar{h}_{t-1}$ and forecaster characteristics $w_{i,t}$ can effectively combine the strengths of both approaches. Such an integrated framework holds promise for improving macroeconomic forecasting by mitigating the data-bound limitations of AI while compensating for the biases inherent in human judgment. This approach leverages the rapid, pattern-recognition capabilities of System 1 alongside the reflective, context-sensitive insights of System 2.

It is a remarkable result that we can beat human forecasters by prompting an LLM with rel-

evant inputs. After all, by using professional human forecasters rather than, e.g., households, we set the bar for clearing human performance high. In addition to experience and relevant education, these forecasters typically have access to sophisticated forecasting models, which are not provided to the AI model. However, even if LLMs are not specialized in time series forecasting per se, they have been trained on a massive corpora that includes past economic data, discussions of macroeconomic dynamics, news articles about recessions and expansions, policy announcements, and more. By prompting the LLM with relevant, up-to-date data we effectively leverage these learned patterns, enabling the LLM to combine fresh data with its broad background knowledge about how certain variables tend to move together or respond to shocks. Given this, we argue that LLMs can support policymakers, researchers, and practitioners in analyzing macroeconomic trends, contributing to more informed decision-making.

7 Conclusion

In this paper we simulate the forecasting process of the Survey of Professional Forecasters using LLMs integrated with real-time data, historical forecasts, and hand-collected forecaster-specific characteristics. Our empirical results reveal that AI and human forecasts are qualitatively similar, but AI-generated forecasts often achieve superior accuracy compared to human forecasts. We show evidence that the improved accuracy of AI forecasts relative to humans stems from the AI model's ability to extract latent information from past human forecasts.

Our approach addresses traditional survey limitations, such as high costs, infrequent data, and bias, while offering better insights into forecasting behavior. The results of our study suggest that LLMs could serve as powerful tools for enhancing macroeconomic nowcasting and forecasting frameworks and as a laboratory for pre-testing and refining hypotheses related to experimental treatments. 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 to new information. Our approach can also be applied to examine the impact of different FOMC communication strategies on the formation of expectations, as well as their influence on disagreement and uncertainty. Combining LLMs with human expertise could further boost the accuracy and reliability of long-term economic forecasts. Building

on Schoenegger et al. (2024), who demonstrate that ensembles of LLMs can rival the performance of human crowds, a promising approach is that of "hybrid wisdom"—where human domain knowledge and generative AI work together.

References

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

Ingvild Almås, Orazio Attanasio, and Pamela Jervis. Presidential Address: Economics and Measurement: New Measures to Model Decision Making. *Econometrica*, 92:947–978, 2024.

Peter Andre, Carlo Pizzinelli, Christopher Roth, and Johannes Wohlfart. Subjective Models of the Macroeconomy: Evidence From Experts and Representative Samples. *Review of Economic Studies*, 89:2958–2991, 2022.

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.

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.

Andrew Caplin. Data Engineering for Cognitive Economics. *Journal of Economic Literature*, forthcoming, 2025.

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.

Michael P. Clements. Forecast Uncertainty - Ex Ante and Ex Post: U.S. Inflation and Output Growth. *Journal of Business & Economic Statistics*, 32(2), 2014.

Michael P. Clements and Ana Beatriz Galvão. Model and Survey Estimates of the Term Structure of Us Macroeconomic Uncertainty. *International Journal of Forecasting*, 33(3): 591–604, 2017.

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, Anthony S. Tay, and Kenneth F. Wallis. Evaluating Density Forecasts of Inflation: The Survey of Professional Forecasters. *NBER Working Paper Seriees*, 6228, 1997.

Joseph Engelberg, Charles F. Manski, and Jared Williams. Comparing the Point Predictions and Subjective Probability Distributions of Professional Forecasters. *Journal of Business & Economic Statistics*, 27(1), 2009.

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.

Paolo Giordani and Paul Soderlind. Inflation Forecast Uncertainty. *European Economic Review*, 47(6):1037–1059, December 2003.

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

Daniel Kahneman. Maps of Bounded Rationality: A Perspective on Intuitive Judgement and Choice. *American Economic Review*, 93(5):1449–1475, 2003.

Benjamin Kay, Jane Ryngaert, Aeimit Lakdawala, and Michael Futch. Partisan Bias in Professional Macroeconomic Forecasts. Technical report, SSRN Working Paper, 2023.

Sophia Kazinnik. Bank Run, Interrupted: Modeling Deposit Withdrawals with Generative AI. *SSRN Working Paper*, 2023.

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, and Jens Ludwig. Human Decisions and Machine Predictions. *Quarterly Journal of Economics*, 133(1):237–293, 2018.

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.

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.

Qian Niu, Junyu Liu, Ziqian Bi, Pohsun Feng, Benji Peng, Keyu Chen, Ming Li, Lawrence KQ Yan, Yichao Zhang, Caitlyn Heqi Yin, Cheng Fei, Tianyang Wang, Yunze Wang, Silin Chen, and Ming Liu. Large Language Models and Cognitive Science: A Comprehensive Review of Similarities, Differences, and Challenges. *arXiv preprint arXiv:2409.02387*, 2024.

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.

Barbara Rossi and Tatevik Sekhposyan. Macroeconomic Uncertainty Indices Based on Nowcast and Forecast Error Distributions. *American Economic Review*, 105(5), 2015.

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

Isabelle Salle, Yuriy Gorodnichenko, and Olivier Coibion. Lifetime Memories of Inflation: Evidence from Surveys and the Lab. *NBER Working Paper Series*, 31996, 2023.

Philipp Schoenegger, Indre Tuminauskaite, Peter S. Park, and Philip E. Tetlock. Wisdom of the Silicon Crowd: LLM Ensemble Prediction Capabilities Match Human Crowd Accuracy. 2024.

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

Gaurav Suri, Lily R Slater, Ali Ziaee, and Morgan Nguyen. Do Large Language Models Show Decision Heuristics Similar to Humans? A Case Study Using GPT-3.5. *Journal of Experimental Psychology: General*, 2024.

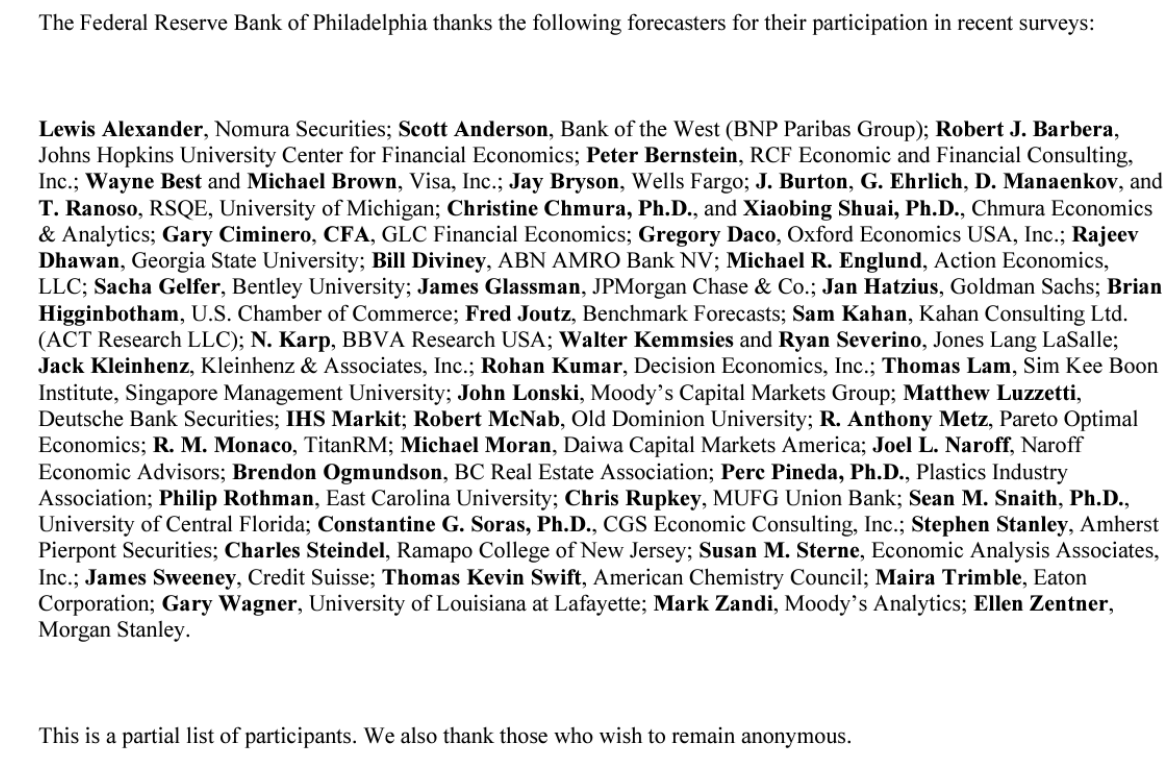
Ali Zarifhonarvar. Experimental Evidence on Large Language Models. *Available at SSRN 4825076*, 2024.

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

Figures

Figure 1: 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 2: 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.

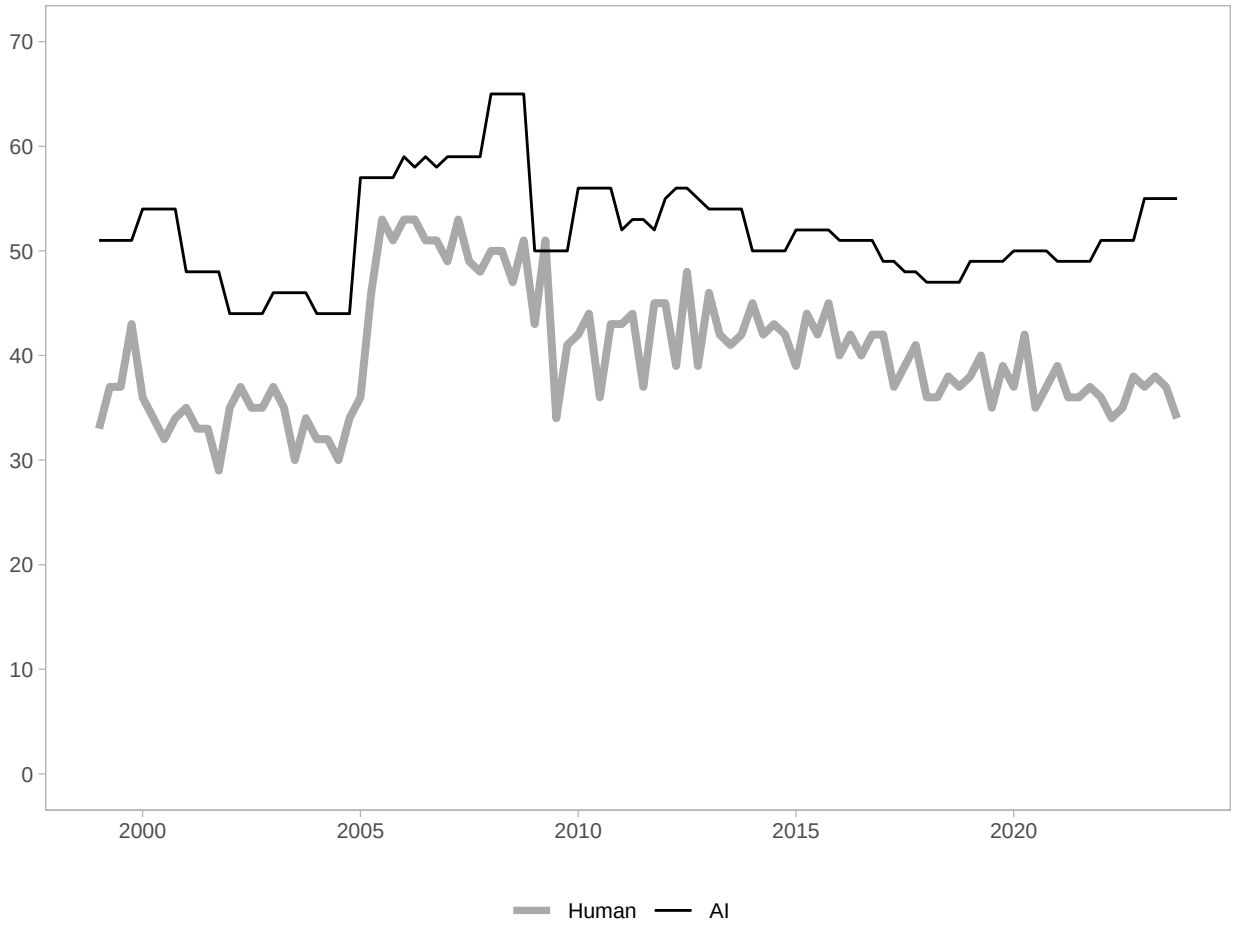


Figure 3: Duration of participation of each forecaster individual forecaster in the SPF panel over time

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.

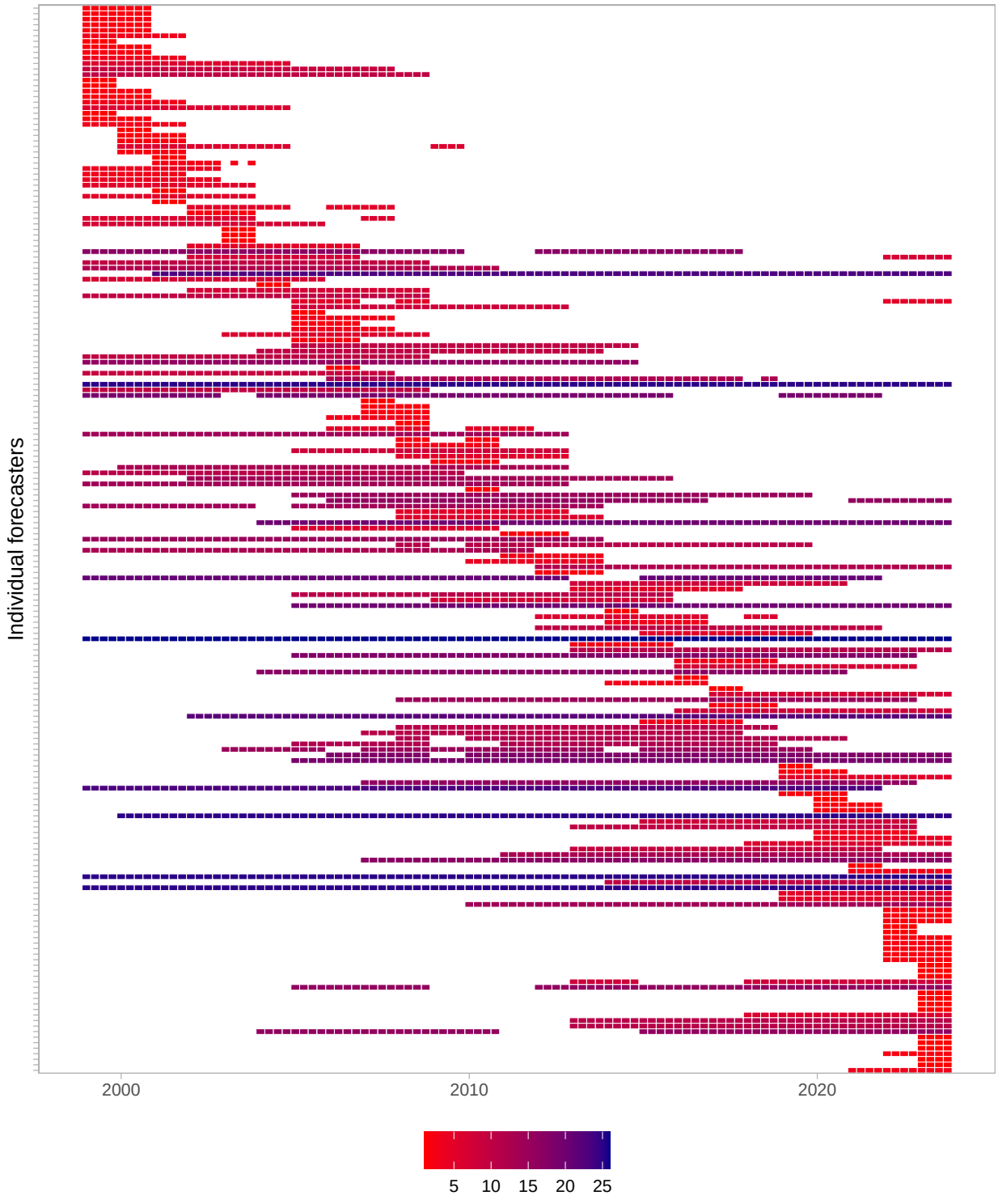
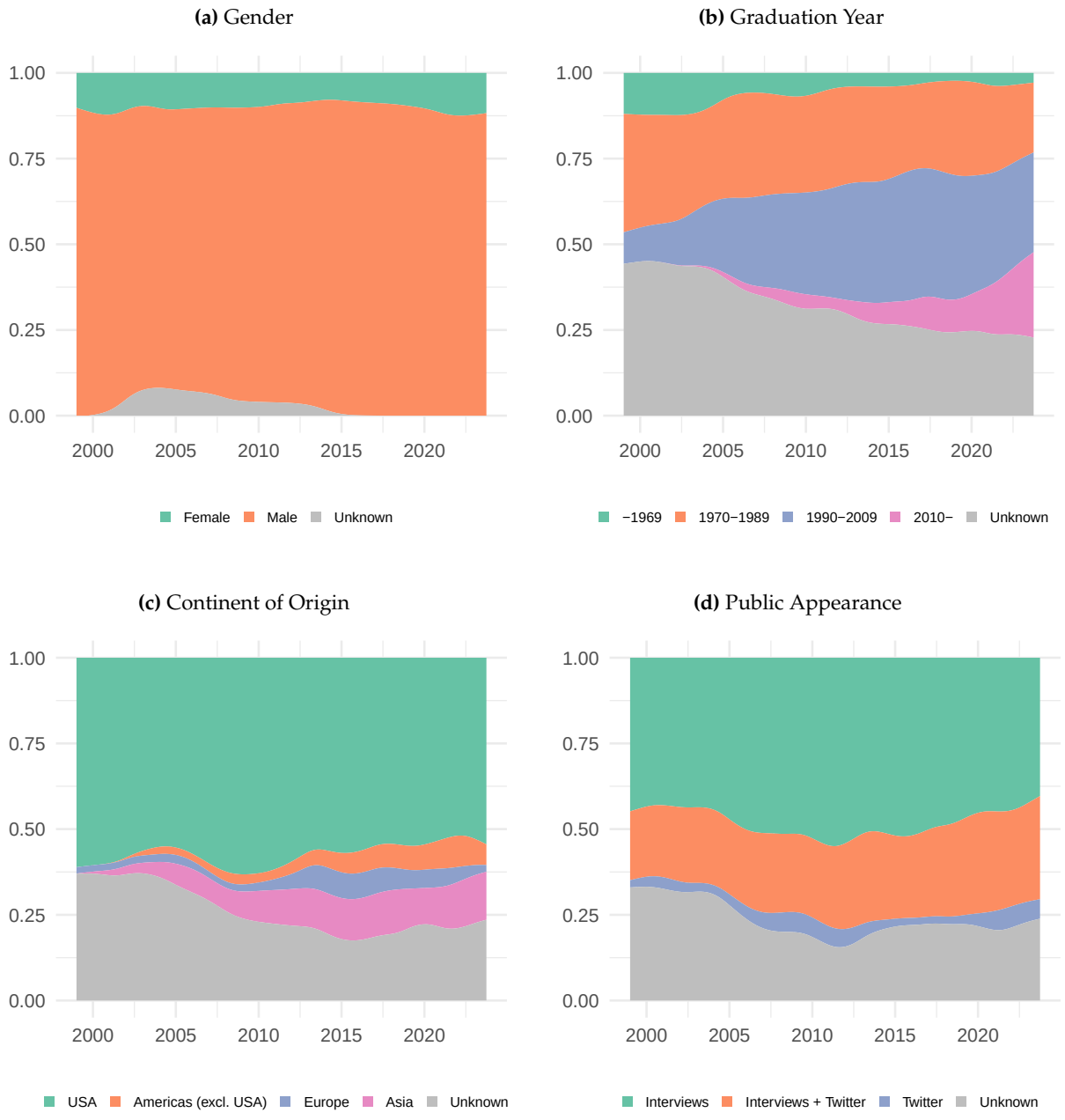


Figure 4: Characteristics of the SPF panel over time

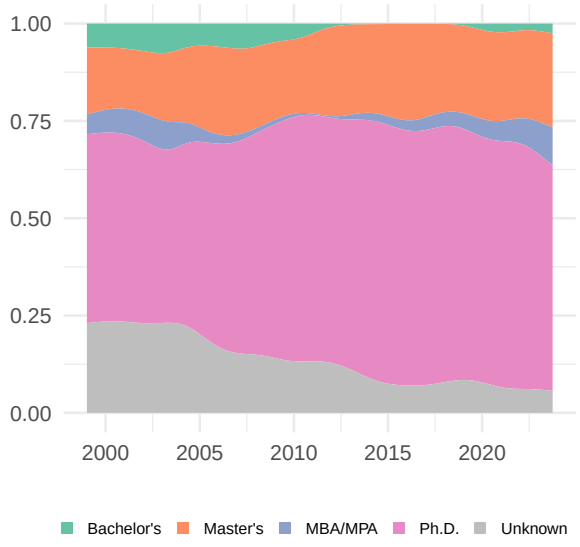
The figures show the characteristics of the AI SPF panel over time: (a) gender, (b) graduation year, (c) continent of origin, (d) public appearance; on next page: (e) education level, (f) education field, (g) affiliation type, and (h) job title.



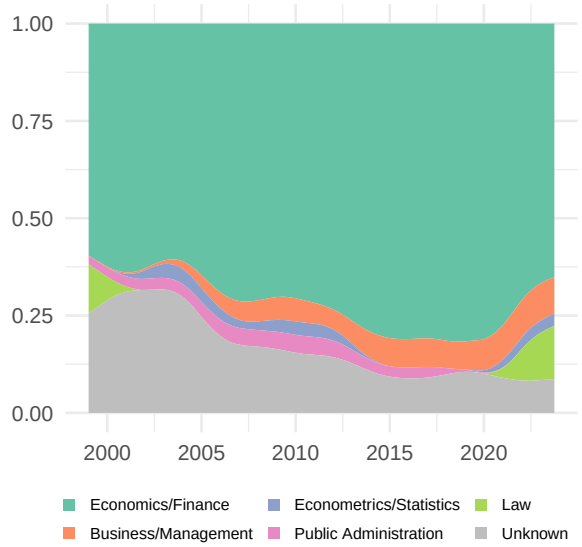
(Figure continues on next page)

Figure 4: Characteristics of the AI SPF panel over time *(continued)*

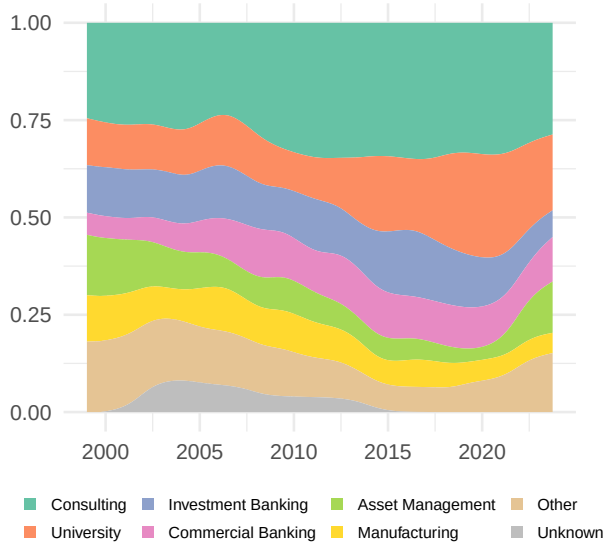
(e) Education level



(f) Education field



(g) Affiliation type



(h) Job title

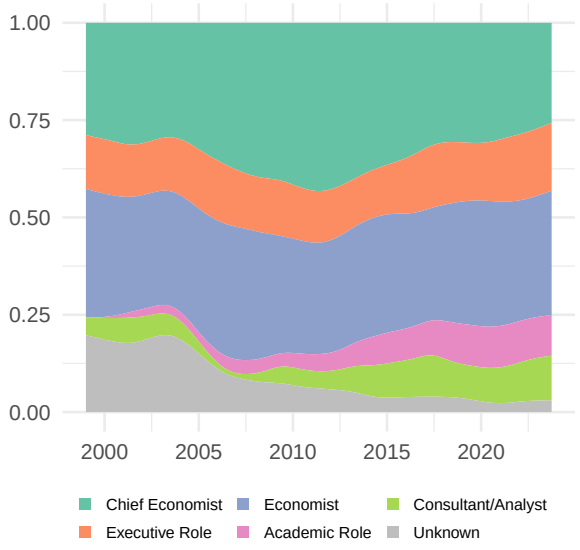


Figure 5: Distribution of sectors across the human SPF panel over time

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

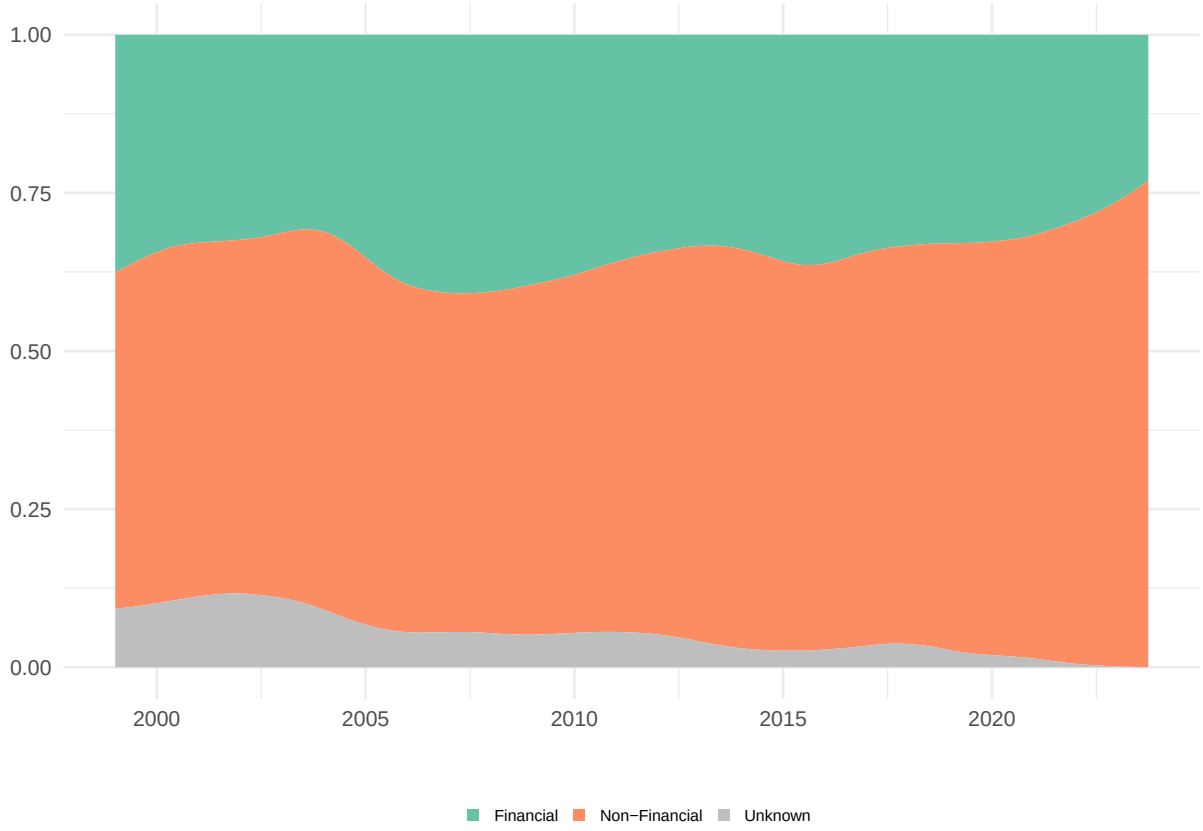
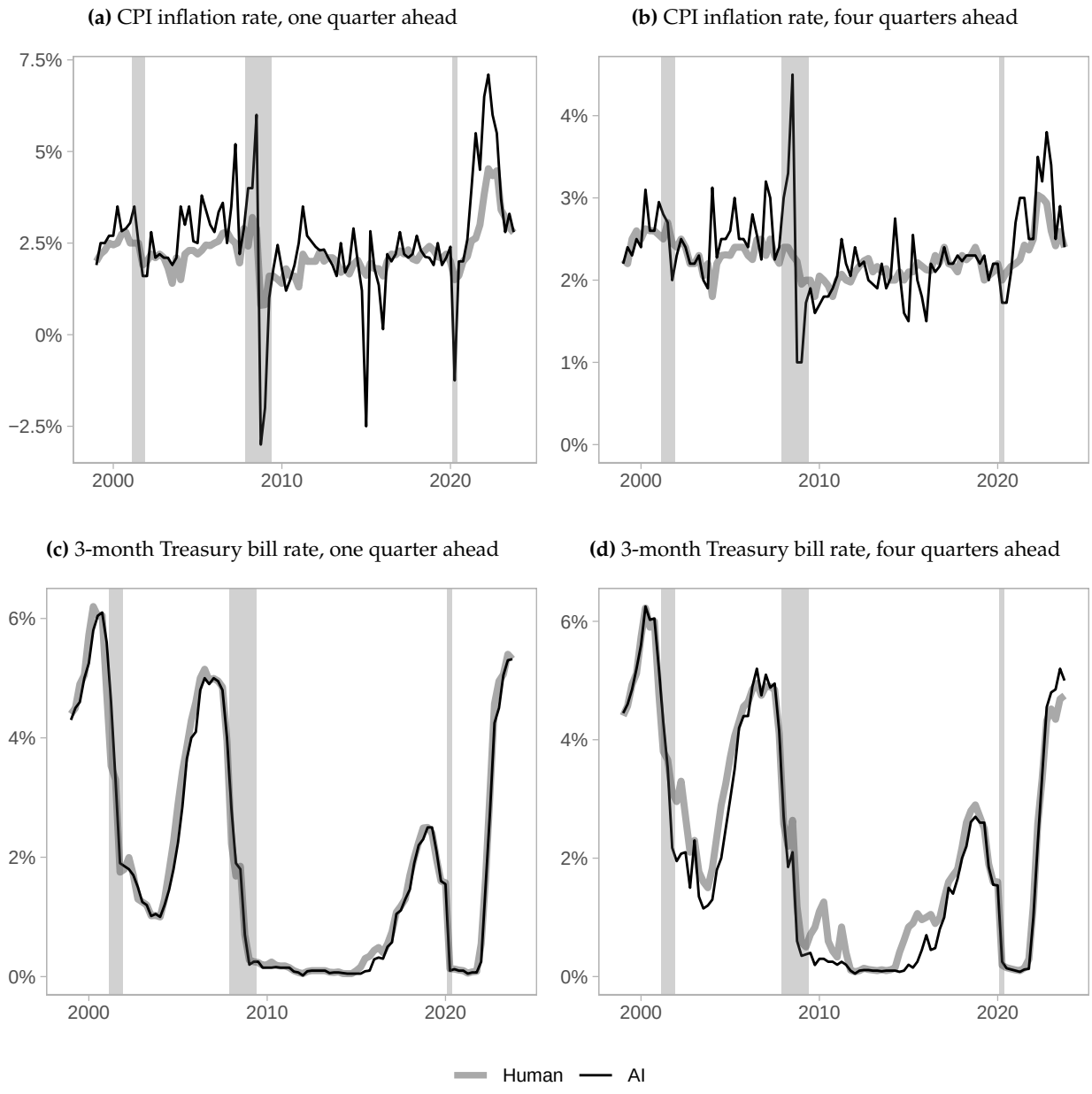


Figure 6: SPF and AI-generated median forecasts

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

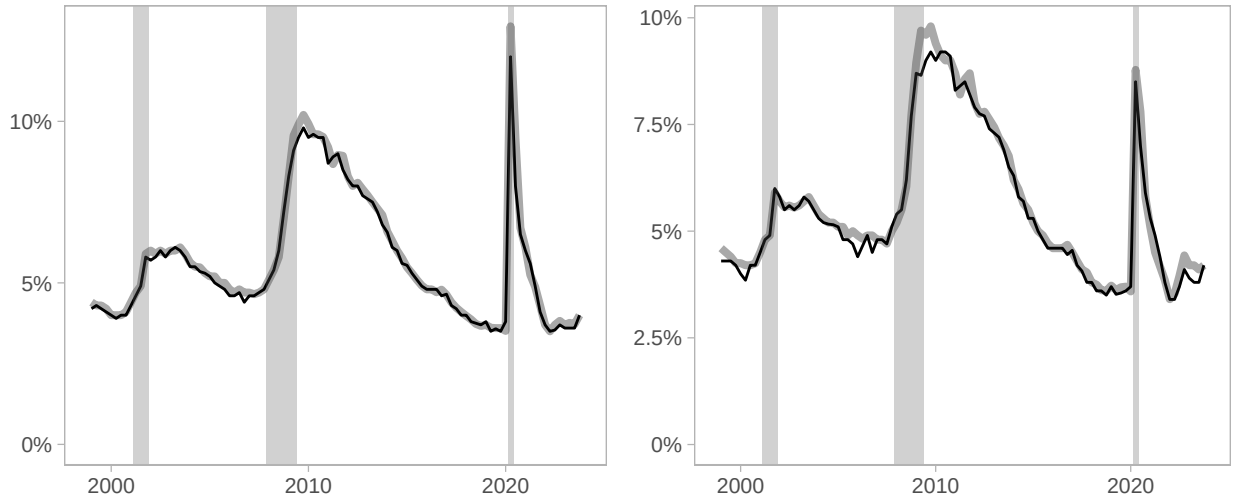


(Figure continues on next page)

Figure 6: SPF and AI-generated median forecasts *(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

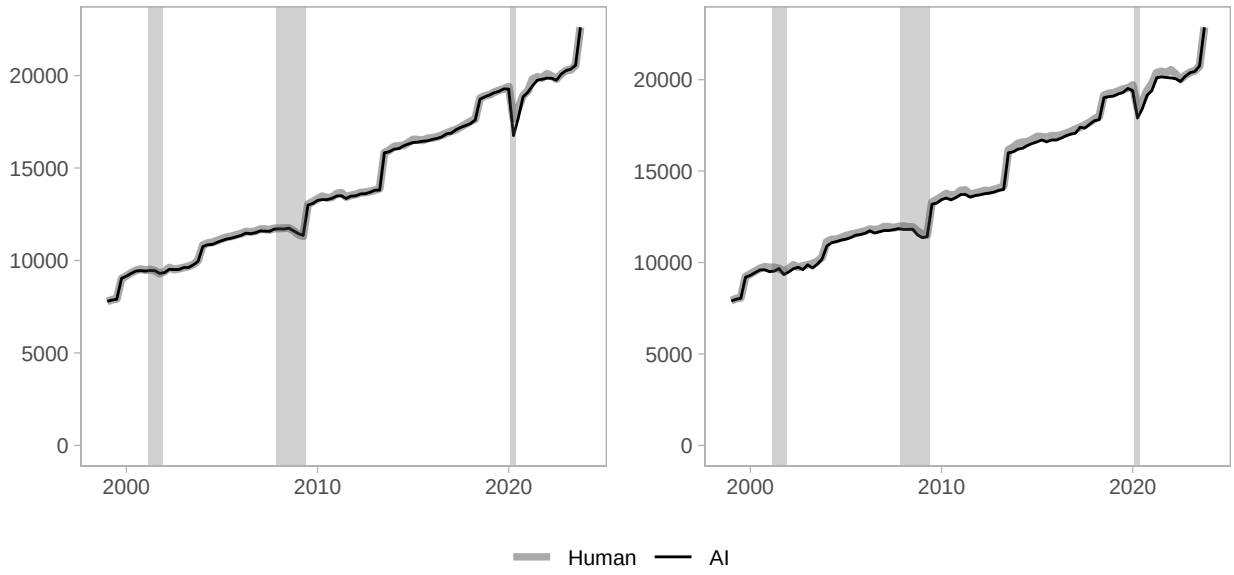
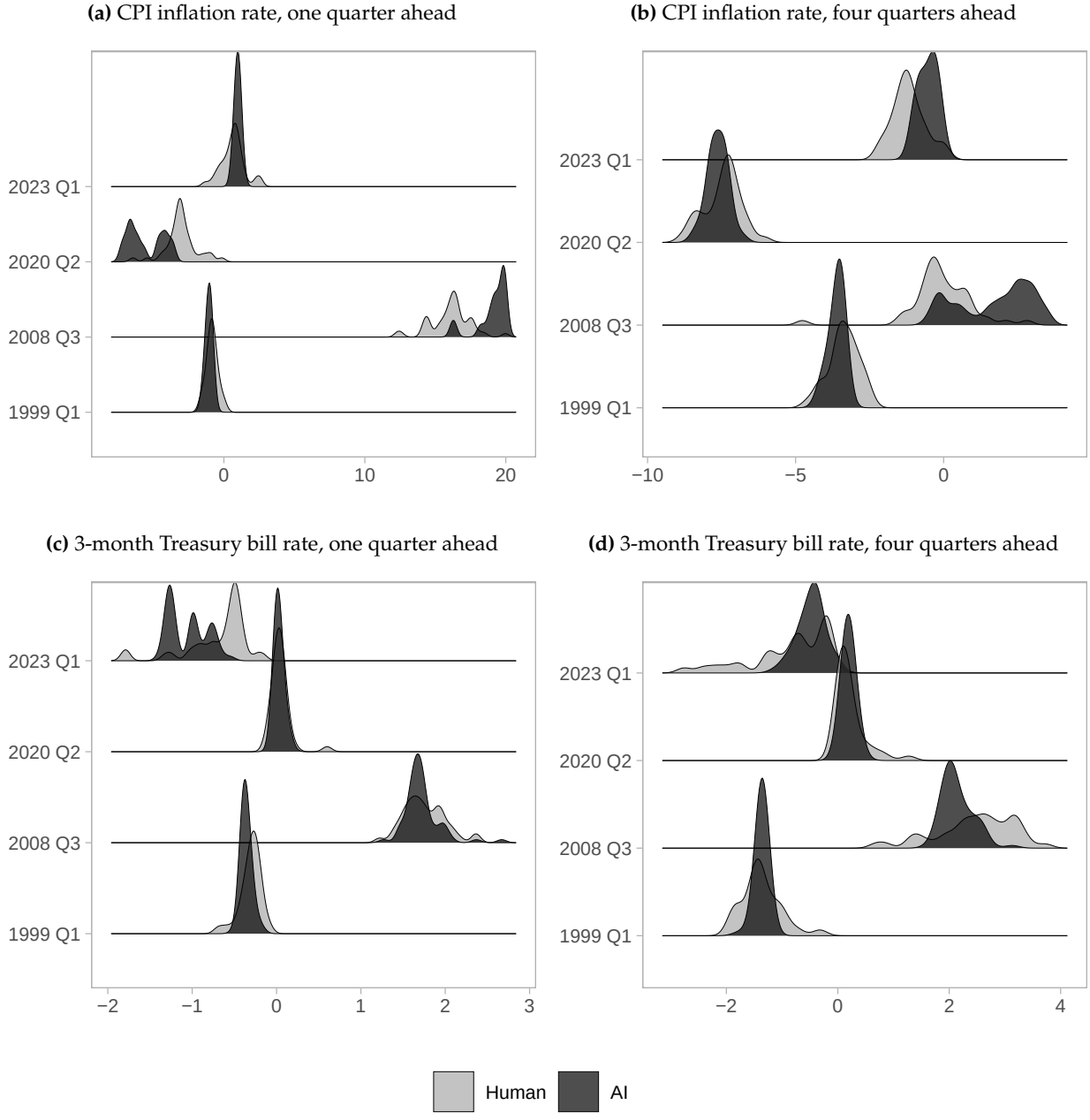


Figure 7: Densities of human and AI-generated SPF individual forecast errors

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.

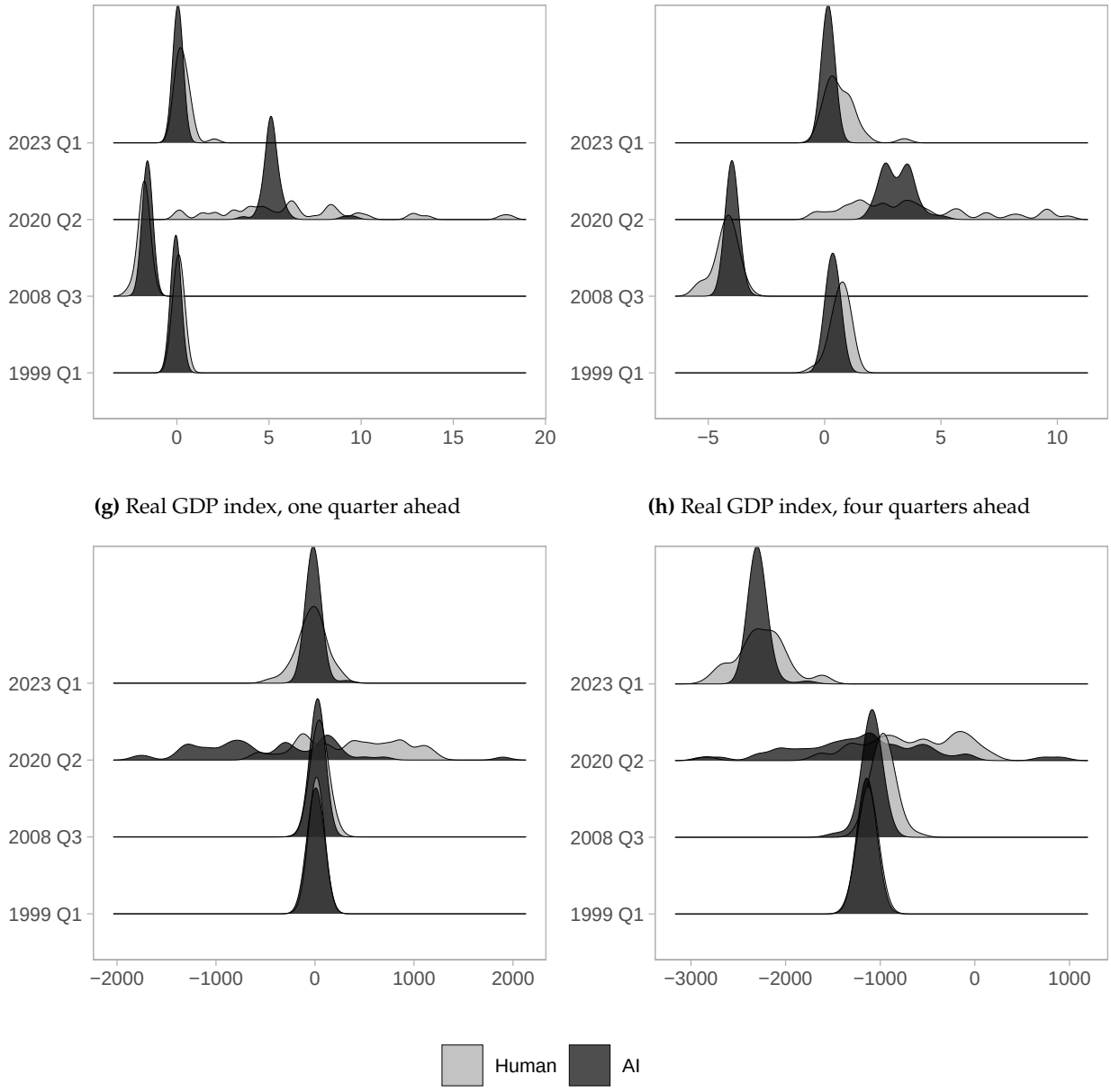


(Figure continues on next page)

Figure 7: Densities of human and AI-generated SPF individual forecast errors *(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

Tables

Table 1: Forecasting variables

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

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

The table shows the number of observations (N), mean, standard deviation (SD), skewness, kurtosis, median, minimum, maximum, and first-order autocorrelation coefficient (ACF(1)) of the data for each of the forecast variables.

	N	Mean	SD	Skewness	Kurtosis	Median	Min	Max	ACF(1)
Section 1: US Business Indicators									
Nominal GDP (x1000)	5188	16.00	4.75	0.59	−0.35	15.01	8.68	28.63	0.84
GDP Price	5188	92.09	13.37	0.42	−0.36	91.40	70.60	123.30	0.85
Corporate Profits (x1000)	5188	1.37	0.57	0.53	−0.30	1.27	0.49	2.90	0.82
Unemployment Rate	5188	5.76	2.04	1.40	2.30	5.00	3.40	14.70	0.65
Non-Farm Payroll (x1000)	5188	138.53	8.07	0.75	−0.63	136.21	127.68	158.43	0.82
Industrial Production	5188	110.47	13.87	1.32	0.88	106.00	92.50	147.30	0.80
Housing Starts	5188	1.30	0.44	−0.14	−1.01	1.25	0.51	2.11	0.79
Treasury Bill Rate (3M)	5188	1.88	2.00	0.76	−0.91	1.15	0.02	6.29	0.76
AAA Corp Bond Yield	5188	4.84	1.34	0.24	−0.54	4.93	2.14	7.78	0.81
Treasury Bond Rate (10Y)	5188	3.36	1.37	0.16	−0.80	3.39	0.62	6.66	0.78
Section 2: Real GDP and Its Components									
Real GDP (x1000)	5188	14.02	3.79	0.31	−1.02	13.33	7.68	22.92	0.84
Real PCE (x1000)	5188	9.78	2.63	0.28	−0.95	9.42	5.25	15.73	0.84
Real Non-Res Fixed Inv (x1000)	5188	1.82	0.65	0.72	−0.93	1.44	0.99	3.39	0.84
Real Res Fixed Inv	5188	493.43	123.21	0.14	−1.20	504.30	293.30	769.50	0.79
Real Federal C&GI (x1000)	5188	0.96	0.28	−0.09	−1.17	1.04	0.45	1.51	0.84
Real State/Local C&GI (x1000)	5188	1.52	0.37	0.22	−1.00	1.46	0.85	2.42	0.83
Real Change in Private Inv	5188	24.70	69.31	−1.13	3.72	34.50	−287.00	193.20	0.35
Real Net Exports	5188	−609.49	285.54	−1.44	1.27	−532.20	−1544.70	−250.00	0.81
Section 3: CPI and PCE Inflation									
CPI Inflation Rate	5188	2.51	3.12	−1.12	7.39	2.46	−13.39	10.73	0.01
Core CPI Inflation Rate	5188	2.34	1.50	2.16	7.92	2.19	−1.09	10.21	0.31
PCE Inflation Rate	5188	0.66	7.61	−4.05	16.27	2.14	−40.91	6.91	0.02
Core PCE Inflation Rate	5188	−0.21	2.92	−9.40	87.38	0.03	−28.31	1.72	−0.02

Table 4: Differences in distributions between individual-level human and AI-generated SPF forecasts

The table shows the differences in distributional moments between individual-level human and AI-generated SPF forecasts for the horizons of (a) one quarter ahead and (b) four quarters ahead. The differences are tested for statistical significance using randomized tests based on 10,000 random draws. Significance levels: * p-value≤0.10, ** p-value≤0.05, *** p-value≤0.01.

(a) One-quarter-ahead horizon

	Median	P25	P75	Mean	SD	Skewness	Kurtosis
Section 1: US Business Indicators							
Nominal GDP	-186.13***	−28.56	−292.71***	−14.91	120.65***	0.10***	0.14***
GDP Price	-0.12***	−0.32***	−0.20***	−0.09***	0.21***	−0.08***	−0.31***
Corporate Profits	83.90**	220.54***	18.55***	72.21***	−0.51	0.13***	0.21***
Unemployment Rate	-0.10***	−0.20***	−0.10***	−0.12***	−0.15***	−0.66***	−5.92***
Non-Farm Payroll	-232.00***	−271.72***	−609.50***	−245.82***	851.64***	−2.43***	36.59***
Industrial Production	-1.54***	−0.60***	−1.62***	−1.37***	−0.14***	0.13***	0.37***
Housing Starts	0.00	−0.02***	0.03***	2.43***	57.56***	26.97***	768.73***
Treasury Bill Rate (3M)	0.10	−0.04***	0.24***	0.01*	0.03***	−0.06***	−0.12***
AAA Corp Bond Yield	-0.21***	−0.10***	−0.28***	−0.18***	−0.00	0.16***	0.05***
Treasury Bond Rate (10Y)	0.10***	−0.15***	−0.08***	−0.03***	−0.00	−0.03***	−0.05***
Section 2: Real GDP and Its Components							
Real GDP	-110.20***	−34.79***	−138.01***	−86.09***	14.63**	0.05***	0.07***
Real PCE	-1328.60***	−2731.61***	−539.35***	−2397.67***	2163.47***	−0.68***	−0.01
Real Non-Res Fixed Inv	-24.68***	−14.84***	−6.43	5.83**	11.38***	0.01	−0.02
Real Res Fixed Inv	-4.49***	−2.36***	−0.69***	−0.65	−0.40**	0.04***	−0.08***
Real Federal C&GI	-0.85*	−4.12***	−5.00***	3.38***	3.99***	0.02***	0.02***
Real State/Local C&GI	-0.62***	−2.00***	−4.09***	1.72	5.68***	0.04***	0.07***
Real Change in Private Inv	-5.70***	−13.28***	5.00***	−4.47***	14.12***	0.51***	−15.00***
Real Net Exports	5.00***	−4.03***	2.83***	−2.51**	3.57***	0.08***	−0.43***
Section 3: CPI and PCE Inflation							
CPI Inflation Rate	0.30***	0.30***	0.50***	0.36***	0.57***	−1.00***	0.63***
Core CPI Inflation Rate	0.01***	0.00	0.10***	0.18***	0.27***	0.42***	−1.77***
PCE Inflation Rate	0.10***	0.11***	0.30***	0.23***	0.41***	0.03	−3.94***
Core PCE Inflation Rate	0.00	−0.05*	−0.07***	0.11***	0.20***	0.57***	−0.90***

(Table continues on next page)

Table 4: Differences in distributions between individual-level human and AI-generated SPF forecasts *(continued)*

(b) Four-quarter-ahead horizon

	Median	P25	P75	Mean	SD	Skewness	Kurtosis
Section 1: US Business Indicators							
Nominal GDP	-348.46***	−8.42	−382.51***	−132.06***	110.42***	0.10***	0.14***
GDP Price	-0.50***	−0.59***	−0.50***	−0.41***	0.05**	−0.14***	−0.55***
Corporate Profits	46.70	124.00***	−23.12***	52.78***	−4.79***	0.11***	0.14***
Unemployment Rate	-0.11***	−0.20***	−0.00	−0.12***	−0.10***	−0.29***	−1.39***
Non-Farm Payroll	-784.00***	−809.81***	−1096.72***	−724.64***	817.69***	−2.41***	37.02***
Industrial Production	-2.43***	−1.61***	−2.61***	−2.51***	−0.55***	0.12***	0.33***
Housing Starts	-0.03***	−0.10***	0.03***	2.44***	57.88***	26.58***	738.95***
Treasury Bill Rate (3M)	-0.20***	−0.36***	0.25***	−0.14***	0.14***	0.08***	−0.10***
AAA Corp Bond Yield	-0.37***	−0.21***	−0.40***	−0.31***	−0.02***	0.22***	0.01
Treasury Bond Rate (10Y)	-0.16***	−0.30***	−0.22***	−0.21***	−0.00	0.06***	−0.13***
Section 2: Real GDP and Its Components							
Real GDP	-200.29***	−126.66***	−204.93***	−194.03***	0.37	0.06***	0.06***
Real PCE	-1429.50***	−2843.63***	−656.78***	−2512.49***	2170.74***	−0.66***	0.03***
Real Non-Res Fixed Inv	-59.70***	−29.64***	−48.31***	−26.50***	7.26***	0.04***	0.00
Real Res Fixed Inv	-14.44***	−9.05***	−2.80***	−8.73***	0.75***	0.02***	−0.18***
Real Federal C&GI	0.45	−4.22***	−11.14***	−0.78	2.94***	0.02***	0.01**
Real State/Local C&GI	-11.01***	−2.34**	−13.90***	−6.67***	3.62***	0.05***	0.08***
Real Change in Private Inv	-3.79***	−14.58***	2.02***	−6.12***	6.28***	−0.12***	−29.46***
Real Net Exports	20.17***	21.48***	15.00***	14.44***	−6.38***	0.04***	−0.40***
Section 3: CPI and PCE Inflation							
CPI Inflation Rate	0.01***	0.09***	0.10***	0.10***	0.04***	−1.75***	11.43***
Core CPI Inflation Rate	0.00	−0.02***	0.00	0.05***	0.11***	1.21***	−2.54***
PCE Inflation Rate	0.00	0.00	0.10***	0.08***	0.03***	−0.00	1.24***
Core PCE Inflation Rate	-0.06***	−0.07***	−0.10***	0.01***	0.11***	1.35***	−1.34***

Table 5: Forecast accuracy of human and AI-generated SPF forecasts

The table shows forecast accuracy reported as mean absolute errors computed as the average of absolute differences between realized values and median human and AI-generated SPF forecasts across survey quarters. Accuracy is reported for different forecast horizons, with 0, 1 and 2 denoting nowcast, one quarter and four quarters ahead, respectively. For each forecasting horizon, the table boldfaces the lowest MAE value between the AI and human surveys. The P-val column reports p-values and significance levels for randomized tests of the statistical significance of the difference between AI and human MAEs. The tests are based on 10,000 random draws. Significance levels: * p-value$\leq$0.10, ** p-value$\leq$0.05, *** p-value$\leq$0.01.

Horizon (quarters)	0			1			4		
	AI	Human	P-val	AI	Human	P-val	AI	Human	P-val
Section 1: US Business Indicators									
Nominal GDP	248.09	**187.45**	0.90	**161.71**	178.31	0.23	**340.95**	379.87	0.00***
GDP Price	**21.87**	22.16	0.00***	**21.85**	22.20	0.00***	**21.68**	22.35	0.00***
Corporate Profits	87.60	**71.78**	0.01***	**61.80**	101.39	0.02**	**165.31**	186.30	0.21
Unemployment Rate	**0.31**	0.38	0.00***	**0.52**	0.57	0.00***	**0.91**	0.94	0.00***
Non-Farm Payroll	**252.33**	465.23	0.01**	933.18	**804.54**	0.05*	2327.18	**1938.36**	0.00***
Industrial Production	**0.54**	1.64	0.00***	**2.14**	2.99	0.00***	**4.92**	6.27	0.00***
Housing Starts	**0.05**	0.09	0.05**	**0.10**	0.12	0.02**	**0.16**	0.21	0.80
Treasury Bill Rate (3M)	0.35	**0.26**	0.31	0.53	**0.43**	0.01***	**1.15**	1.21	0.00***
AAA Corp Bond Yield	**0.18**	0.28	0.07*	**0.37**	0.44	0.00***	**0.59**	0.73	0.00***
Treasury Bond Rate (10Y)	0.36	**0.32**	0.63	**0.48**	0.51	0.00***	**0.76**	0.88	0.00***
Section 2: Real GDP and Its Components									
Real GDP	**90.82**	126.20	0.00***	**169.81**	209.17	0.00***	**524.39**	568.19	0.00***
Real PCE	1393.03	**90.64**	0.00***	1454.12	**130.32**	0.00***	1710.50	**330.69**	0.00***
Real Non-Res Fixed Inv	**21.91**	25.92	0.11	**36.11**	48.83	0.00***	**116.08**	133.96	0.00***
Real Res Fixed Inv	**10.40**	10.43	0.86	**13.89**	18.10	0.39	**49.81**	54.14	0.00***
Real Federal C&GI	9.94	**6.79**	0.84	**14.32**	19.34	0.04**	46.88	**45.62**	0.00***
Real State/Local C&GI	9.66	**8.09**	0.23	**20.04**	24.39	0.00***	**68.31**	69.76	0.00***
Real Change in Private Inv	35.35	**24.91**	0.19	**19.46**	38.13	0.10*	51.62	**48.89**	0.02**
Real Net Exports	26.97	**16.79**	0.54	**24.40**	42.52	0.02**	**90.27**	91.17	0.00***
Section 3: CPI and PCE Inflation									
CPI Inflation Rate	**1.84**	1.98	0.00***	2.36	**2.14**	0.01***	**2.03**	2.06	0.04**
Core CPI Inflation Rate	**0.67**	0.82	0.02**	0.92	**0.88**	0.00***	**0.97**	1.00	0.03**
PCE Inflation Rate	**2.40**	2.49	0.54	**2.27**	2.72	0.55	**3.13**	**3.13**	0.85
Core PCE Inflation Rate	2.37	**2.30**	0.01**	2.31	**2.21**	0.15	**2.12**	2.14	0.01**

Table 6: Percentage survey quarters where AI is more accurate than the human SPF

The table shows forecast accuracy in terms of the percentage of survey quarters where the AI-generated SPF has lower forecast errors than the human SPF (Pct). Forecast errors are computed as the differences between realized values and median human and AI-generated SPF forecasts. Values are reported for different forecast horizons, with 0, 1 and 2 denoting nowcast, one quarter and four quarters ahead, respectively. The table boldfaces values above 50%. The P-val column reports p-values and significance levels for randomized tests of the percentages being equal to 50%. The tests are based on 10,000 random draws. Significance levels: * p-value≤0.10, ** p-value≤0.05, *** p-value≤0.01.

Horizon (quarters)	0		1		4	
	Pct	P-val	Pct	P-val	Pct	P-val
Section 1: US Business Indicators						
Nominal GDP	0.38	0.03**	**0.69**	0.00***	**0.72**	0.00***
GDP Price	**0.98**	0.00***	**0.98**	0.00***	**1.00**	0.00***
Corporate Profits	0.45	0.60	**0.70**	0.00***	**0.58**	0.19
Unemployment Rate	**0.81**	0.00***	**0.74**	0.01***	**0.63**	0.22
Non-Farm Payroll	**0.80**	0.00***	0.35	0.01***	0.31	0.00***
Industrial Production	**0.80**	0.00***	**0.64**	0.01***	**0.70**	0.00***
Housing Starts	**0.78**	0.00***	**0.70**	0.00***	**0.68**	0.00***
Treasury Bill Rate (3M)	**0.51**	1.00	0.47	0.97	**0.60**	0.08*
AAA Corp Bond Yield	**0.85**	0.00***	**0.73**	0.00***	**0.82**	0.00***
Treasury Bond Rate (10Y)	**0.54**	0.99	**0.64**	0.08*	**0.68**	0.00***
Section 2: Real GDP and Its Components						
Real GDP	**0.70**	0.00***	**0.75**	0.00***	**0.63**	0.01**
Real PCE	**0.73**	0.00***	0.47	0.63	0.33	0.00***
Real Non-Res Fixed Inv	**0.64**	0.02**	**0.73**	0.00***	**0.65**	0.01***
Real Res Fixed Inv	**0.57**	0.42	**0.73**	0.00***	**0.63**	0.02**
Real Federal C&GI	**0.50**	1.00	**0.80**	0.00***	0.45	0.40
Real State/Local C&GI	**0.54**	0.58	**0.74**	0.00***	**0.55**	0.34
Real Change in Private Inv	0.45	0.56	**0.80**	0.00***	**0.50**	1.00
Real Net Exports	**0.50**	1.00	**0.84**	0.00***	**0.53**	0.73
Section 3: CPI and PCE Inflation						
CPI Inflation Rate	**0.69**	0.01***	0.47	0.74	**0.55**	0.78
Core CPI Inflation Rate	**0.63**	0.15	0.44	0.50	**0.55**	0.47
PCE Inflation Rate	**0.65**	0.07*	**0.81**	0.00***	**0.61**	0.16
Core PCE Inflation Rate	**0.57**	0.61	**0.51**	0.92	**0.64**	0.05*

Table 7: Forecast accuracy of best human and AI SPF forecasters

The table shows forecast accuracy reported as mean absolute errors computed as the average of absolute differences between realized values and the best human and AI-generated SPF forecasts across survey quarters. The best forecast is computed from the forecaster that achieves lowest average error across survey quarters. These best forecasters are identified separately for each variable and horizon. Accuracy is reported for different forecast horizons, with 0, 1 and 2 denoting nowcast, one quarter and four quarters ahead, respectively. For each forecasting horizon, the table boldfaces the lowest MAE value between the AI and human surveys. The P-val column reports p-values and significance levels for randomized tests of the statistical significance of the difference between AI and human MAEs. The tests are based on 10,000 random draws. Significance levels: * p-value≤0.10, ** p-value≤0.05, *** p-value≤0.01.

Horizon (quarters)	0			1			4		
	AI	Human	P-val	AI	Human	P-val	AI	Human	P-val
Section 1: US Business Indicators									
Nominal GDP	**58.20**	76.62	0.60	**66.55**	93.06	0.94	92.93	**83.50**	0.88
GDP Price	**7.13**	7.31	0.43	**5.91**	6.13	0.26	**1.61**	2.20	0.29
Corporate Profits	**11.43**	28.74	0.39	**8.75**	32.49	0.25	39.27	**38.39**	0.42
Unemployment Rate	**0.05**	0.06	0.91	**0.07**	0.13	0.52	**0.08**	0.15	0.27
Non-Farm Payroll	356.67	**85.00**	0.41	499.25	**45.00**	0.33	**452.76**	1083.55	0.14
Industrial Production	**0.03**	0.44	0.15	**0.31**	1.23	0.20	**0.85**	2.43	0.09*
Housing Starts	**0.01**	0.04	0.52	**0.03**	0.07	0.17	**0.04**	0.06	0.72
Treasury Bill Rate (3M)	**0.02**	**0.02**	0.76	**0.03**	**0.03**	0.61	**0.07**	**0.07**	0.69
AAA Corp Bond Yield	**0.07**	0.21	0.20	**0.21**	0.22	0.93	0.16	**0.15**	0.72
Treasury Bond Rate (10Y)	**0.09**	0.10	0.79	**0.13**	0.16	0.53	**0.19**	0.24	0.05*
Section 2: Real GDP and Its Components									
Real GDP	**10.67**	85.76	0.01***	**20.53**	76.74	0.06*	**58.27**	190.17	0.03**
Real PCE	**10.10**	25.13	0.21	**43.43**	62.55	0.18	**42.68**	67.67	0.05*
Real Non-Res Fixed Inv	**4.79**	61.81	0.34	**6.85**	40.18	0.05**	31.75	**17.27**	0.05*
Real Res Fixed Inv	**1.70**	4.91	0.95	**1.90**	4.56	0.64	**9.50**	11.49	0.90
Real Federal C&GI	**1.42**	1.49	0.97	**2.68**	5.99	0.79	**7.55**	9.00	0.48
Real State/Local C&GI	**2.43**	5.02	0.03**	**2.40**	11.32	0.16	**5.85**	14.75	0.24
Real Change in Private Inv	**4.22**	10.24	0.34	**6.11**	24.24	0.06*	**22.40**	26.84	0.27
Real Net Exports	**4.50**	79.05	0.33	**7.99**	10.80	0.16	**13.42**	23.09	0.11
Section 3: CPI and PCE Inflation									
CPI Inflation Rate	**0.58**	0.87	0.87	0.67	**0.66**	0.55	**0.30**	0.34	0.79
Core CPI Inflation Rate	0.66	**0.19**	0.41	0.64	**0.25**	0.79	**0.14**	0.75	0.35
PCE Inflation Rate	**0.36**	0.89	0.54	**0.21**	0.53	0.35	**0.26**	0.71	0.51
Core PCE Inflation Rate	**1.00**	1.10	0.71	**1.08**	1.24	0.51	**0.23**	0.59	0.20

Table 8: Relative accuracy of AI forecasts with fewer prompt inputs

The table shows the accuracy of AI forecasts generated by prompts that provide less information relative to the baseline AI forecasts. Forecast accuracy is computed as mean absolute errors, i.e., the average of absolute differences between realized values and the median human and AI-generated SPF forecasts across survey quarters. The "Generic" forecast uses a prompt that does not provide forecaster characteristics, but do provide real-time data and past median SPF forecasts. The "Generic, w/o real-time data" excludes both forecaster characteristics and real-time data from the prompt. "Generic, w/o real-time data, w/o past SPF data" excludes all data from the prompt including forecaster characteristics, real-time data, and past median SPF forecasts. Finally "Recall" prompts the model to recall values from the training data rather than forecasting values. Values exceeding one (boldfaced) indicate less accuracy compared with the baseline AI forecaster seeded with forecaster characteristics, real-time data, and past median SPF forecasts. Randomized tests evaluate whether values are significantly different from one. The tests are based on 10,000 random draws. Significance levels: * p-value≤0.10, ** p-value≤0.05, *** p-value≤0.01.

	Generic		Generic, w/o real-time data		Generic, w/o real-time data, w/o past SPF data		Recall
Horizon (quarters)	0	4	0	4	0	4	0
Section 1: US Business Indicators							
Nominal GDP	0.92	**1.09**	0.78	**1.12⋆**	**2.88*** **	**2.91*** **	**15.95*** **
GDP Price	**1.00**	**1.00**	**1.01**	**1.03**	0.72	0.78***	**1.10**
Corporate Profits	**1.14**	0.99	0.82	**1.13**	**2.67**	**1.96*** **	**6.19**
Unemployment Rate	**1.02**	**1.02*** **	**1.20**	**1.02**	**1.12**	**1.10*** **	**1.03**
Non-Farm Payroll	0.97	**1.04** **	**6.42**	**1.25**	**6.58*** **	**1.04*** **	**41.48**
Industrial Production	**1.59**	**1.10**	**3.14** **	**1.28**	**13.09**	**1.44**	**13.95*** **
Housing Starts	**1.04**	**1.10**	**1.66**	**1.26**	**1.91**	**1.91*** **	**19.02*** **
Treasury Bill Rate (3M)	**1.09**	**1.01**	0.74	**1.03**	**1.07*** **	**1.08*** **	0.91***
AAA Corp Bond Yield	**1.51** **	**1.11**	**1.51**	**1.25**	**4.85*** **	**1.30*** **	**3.73*** **
Treasury Bond Rate (10Y)	0.98	**1.01**	0.87	**1.16**	**1.02**	**1.14**	0.72
Section 2: Real GDP and Its Components							
Real GDP	**1.15**	**1.04**	**1.37**	**1.08**	**7.57*** **	**1.53*** **	**38.04*** **
Real PCE	**1.85**	**1.59**	0.07	0.20	0.62***	0.55***	**1.95*** **
Real Non-Res Fixed Inv	**1.41*** **	**1.03*** **	**1.18**	**1.18**	**11.98*** **	**2.63*** **	**34.15**
Real Res Fixed Inv	**1.20**	0.99	**1.00**	**1.10**	**17.43*** **	**4.56*** **	**19.51*** **
Real Federal C&GI	**1.12**	**1.00**	0.70	0.98	**66.25*** **	**14.40*** **	**75.39*** **
Real State/Local C&GI	**1.39**	0.96	0.83	**1.03**	**37.54** **	**5.93*** **	**66.03*** **
Real Change in Private Inv	0.86	**1.05**	0.71	0.95	**4.78** **	**3.49**	**2.96*** **
Real Net Exports	0.98	**1.07**	0.62	**1.00**	**8.97*** **	**3.23*** **	**23.86** **
Section 3: CPI and PCE Inflation							
CPI Inflation Rate	0.90	**1.02**	**1.09**	**1.02**	**1.09**	**1.13** **	**1.09**
Core CPI Inflation Rate	0.84	**1.07**	**1.20**	**1.08⋆**	**1.08**	**1.08**	0.99
PCE Inflation Rate	**1.11**	0.95	**1.00**	**1.13**	**1.01*** **	**1.07**	**1.02**
Core PCE Inflation Rate	**1.05**	**1.04**	**1.02**	**1.08**	**1.07**	**1.25**	**1.06**
Average	**1.14**	**1.06**	**1.31**	**1.06**	**8.88**	**2.52**	**16.82**

Table 9: Out-of-sample forecast accuracy (mean absolute percentage errors) of SPF and AI-generated SPF forecasts

Forecast accuracy (mean absolute errors) of SPF and AI-generated SPF forecasts

The table shows forecast accuracy for the out-of-sample period, 2024 Q1 - 2024 Q4. Forecast accuracy is reported as mean absolute errors (MAEs) of SPF and AI-generated SPF forecasts for various economic indicators across different forecast horizons. For each forecasting horizon, the table boldfaces the lowest MAE value between the AI and human surveys. The P-val column reports p-values and significance levels for randomized tests of the statistical significance of the difference between AI and human MAEs. The tests are based on 10,000 random draws. Significance levels: * p-value≤0.10, ** p-value≤0.05, *** p-value≤0.01.

Horizon (quarters)	0			1			2		
	AI	Human	P-val	AI	Human	P-val	AI	Human	P-val
Section 1: US Business Indicators									
Nominal GDP	**0.23**	308.70	0.12	**162.96**	213.90	0.25	281.19	**107.12**	0.50
GDP Price	**0.00**	0.70	0.12	**0.23**	0.66	0.26	**0.42**	0.64	0.50
Corporate Profits	310.17	**283.06**	0.12	430.90	**402.71**	0.25	476.35	**457.83**	0.50
Unemployment Rate	0.13	**0.07**	0.24	0.21	**0.17**	0.25	0.28	**0.20**	1.00
Non-Farm Payroll	397.00	**125.76**	0.12	343.82	**174.59**	0.25	402.04	**246.47**	0.51
Industrial Production	**0.38**	0.41	0.12	**0.60**	0.66	0.25	**0.46**	0.80	0.50
Housing Starts	**0.03**	**0.03**	1.00	**0.06**	0.07	0.25	**0.08**	0.09	0.50
Treasury Bill Rate (3M)	0.22	**0.09**	0.50	0.37	**0.17**	0.25	0.44	**0.30**	0.50
AAA Corp Bond Yield	**0.33**	**0.33**	0.51	0.26	**0.23**	0.49	0.19	**0.15**	0.50
Treasury Bond Rate (10Y)	0.27	**0.14**	0.87	0.38	**0.32**	0.25	0.25	**0.17**	0.50
Section 2: Real GDP and Its Components									
Real GDP	**0.12**	118.32	0.13	**138.20**	151.52	0.25	212.50	**170.62**	0.50
Real PCE	**0.00**	93.56	0.12	**117.93**	125.42	0.25	133.10	**118.92**	0.50
Real Non-Res Fixed Inv	**0.00**	19.79	0.12	58.63	**39.50**	0.26	108.05	**89.67**	0.50
Real Res Fixed Inv	**0.00**	2.32	0.12	**13.93**	14.12	0.25	14.35	**9.77**	0.50
Real Federal C&GI	**0.00**	5.33	0.13	6.23	**5.97**	0.25	**5.60**	5.70	0.49
Real State/Local C&GI	**0.00**	11.50	0.13	11.60	**2.11**	0.26	24.60	**14.51**	0.51
Real Change in Private Inv	**0.00**	8.63	0.75	20.23	**16.61**	0.25	22.05	**20.79**	1.00
Real Net Exports	**0.00**	11.15	0.12	46.77	**42.52**	0.26	86.87	**78.10**	0.50
Section 3: CPI and PCE Inflation									
CPI Inflation Rate	1.20	**1.00**	0.25	1.25	**0.95**	0.49	1.28	**1.18**	1.00
Core CPI Inflation Rate	0.95	**0.63**	0.25	0.69	**0.61**	0.49	0.46	**0.36**	1.00
PCE Inflation Rate	1.19	**0.85**	0.50	1.15	**0.68**	1.00	0.64	**0.63**	1.00
Core PCE Inflation Rate	1.07	**0.61**	0.49	0.78	**0.54**	0.50	**0.08**	**0.08**	1.00

Table 10: Out-of-sample forecast accuracy (mean absolute errors) of best human and AI SPF forecasters

The table shows forecast accuracy for the out-of-sample period, 2024 Q1 - 2024 Q4. Forecast accuracy is reported as mean absolute errors (MAEs) of the best SPF and AI-generated SPF forecasts for various economic indicators across different forecast horizons. The best forecast is the forecast with lowest absolute error for each variable, horizon, and quarter For each forecasting horizon, the table boldfaces the lowest MAE value between the AI and human surveys. The P-val column reports p-values and significance levels for randomized tests of the statistical significance of the difference between AI and human MAEs. The tests are based on 10,000 random draws. Significance levels: * p-value$\leq$0.10, ** p-value$\leq$0.05, *** p-value$\leq$0.01.

Horizon (quarters)	0			1			2		
	AI	Human	P-val	AI	Human	P-val	AI	Human	P-val
Section 1: US Business Indicators									
Nominal GDP	**0.23**	308.70	0.12	**27.20**	213.90	0.25	**85.45**	107.12	0.51
GDP Price	**0.00**	0.70	0.12	**0.17**	0.66	0.25	**0.24**	0.64	0.50
Corporate Profits	**276.12**	283.06	0.13	**333.60**	402.71	0.25	**456.30**	457.83	0.51
Unemployment Rate	0.08	**0.07**	0.24	0.18	**0.17**	0.25	0.25	**0.20**	1.00
Non-Farm Payroll	279.57	**125.76**	0.13	302.94	**174.59**	0.25	343.45	**246.47**	0.50
Industrial Production	**0.27**	0.41	0.12	**0.15**	0.66	0.24	**0.30**	0.80	0.49
Housing Starts	**0.03**	**0.03**	1.00	**0.06**	0.07	0.25	0.07	**0.00**	1.00
Treasury Bill Rate (3M)	**0.04**	0.09	0.50	0.18	**0.17**	0.25	0.32	**0.30**	0.50
AAA Corp Bond Yield	**0.03**	0.33	0.50	**0.21**	0.23	0.50	0.10	**0.08**	1.00
Treasury Bond Rate (10Y)	0.15	**0.14**	0.87	**0.23**	0.32	0.24	0.05	**0.00**	1.00
Section 2: Real GDP and Its Components									
Real GDP	**0.07**	118.32	0.12	**27.18**	151.52	0.25	**163.80**	170.62	0.49
Real PCE	**0.00**	93.56	0.12	**30.48**	125.42	0.25	**94.11**	118.92	0.50
Real Non-Res Fixed Inv	**0.00**	19.79	0.12	**28.95**	39.50	0.25	98.68	**89.67**	0.50
Real Res Fixed Inv	**0.00**	2.32	0.13	**12.47**	14.12	0.24	10.90	**9.77**	0.50
Real Federal C&GI	**0.00**	5.33	0.12	**3.87**	5.97	0.25	**3.84**	5.70	0.50
Real State/Local C&GI	**0.00**	11.50	0.12	5.13	**2.11**	0.25	19.48	**14.51**	0.50
Real Change in Private Inv	**0.00**	8.63	0.75	**16.52**	16.61	0.24	**10.90**	20.79	1.00
Real Net Exports	**0.00**	11.15	0.13	**38.52**	42.52	0.25	81.54	**78.10**	0.50
Section 3: CPI and PCE Inflation									
CPI Inflation Rate	**0.93**	1.00	0.25	**0.94**	0.95	0.50	**1.18**	**1.18**	1.00
Core CPI Inflation Rate	0.67	**0.63**	0.25	**0.43**	0.61	0.50	**0.36**	**0.36**	1.00
PCE Inflation Rate	0.96	**0.85**	0.51	0.80	**0.68**	1.00	**0.24**	0.63	1.00
Core PCE Inflation Rate	0.83	**0.61**	0.50	**0.53**	0.54	0.51	**0.00**	0.08	1.00

Prompts

Prompt 1: Baseline prompt

This prompt is 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).

Prompt 2: Prompt for a generic forecaster

This prompt is used to generate AI forecasts for a generic forecaster, i.e., without including forecaster characteristics. The inputvariables 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).
```

Prompt 3: Prompt for a generic forecaster and no real-time data

This prompt is used to generate AI forecasts for a generic forecaster, i.e., without including forecaster characteristics. The input 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).
```

Prompt 4: Prompt for a generic forecaster and no data

This prompt is used to generate AI forecasts for a generic forecaster, i.e., without including forecaster characteristics or any data on economic variables. The input variables are highlighted in brackets and blue text and consist of only 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).

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).
```

Prompt 5: Prompt for recall task

This prompt is used to have the model recall specific values of economic variables.

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

We are in [date_q]. What is the value for the
following variables: [variable_data] for this quarter: [date_q]

Please provide your best guess for the requested variables.
```

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 A1 presents examples of personas generated by both methods:

Table A1: 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 acknowledgment 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 different AI models and settings to assess their performance in economic forecasting tasks. These models represent combinations of:

- LLM Architectures: GPT-3.5, GPT-4, GPT-4o-mini, Llama-3.3-70B-Instruct-Turbo, and DeepSeek-V3.
- Temperature Settings: 0 (deterministic) and 1 (stochastic)
- Persona Prompts: main and expanded prompts, as discussed in Section A.1

The results support the main conclusions from the paper and are available upon request.

B 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 mitigate black-box concerns surrounding LLMs by shedding light on the underlying processes through which AI 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

on word co-occurrence.[31] 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.[32]

Table A2: Comparison of Topic Distribution between LDA and GPT Methods

The table compares the topic distributions identified using Latent Dirichlet Allocation (LDA) and GPT-based topic classification methods on the forecaster explanations for their predictions.

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	0.0
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**

As shown in Table A2, 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.

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

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