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 (anne.hansen@rich.frb.org)*
[b]*MIT Sloan School of Management (jjhorton@mit.edu)*
[c]*Stanford HAI (kazinnik@stanford.edu)*
[d]*Indiana University Bloomington (dpuzzell@iu.edu; alizarif@iu.edu)*

January 2026

We develop a framework for simulating professional expectations formation using large language models (LLMs). Combining novel hand-collected data on Survey of Professional Forecasters (SPF) participant characteristics with real-time macroeconomic data and lagged SPF median forecasts, we prompt LLMs to generate quarterly forecasts for 23 variables over 1999-2023. The resulting synthetic panel replicates key properties of the human survey, including forecast accuracy, median forecast levels, revision dynamics, and cross-sectional dispersion. Ablation exercises show that performance is largely driven by information conditioning: removing forecaster characteristics modestly worsens performance, whereas removing real-time data and especially lagged SPF medians leads to substantial error increases. Our framework offers a scalable complement to traditional surveys, enabling counterfactual analysis under alternative information sets, retrospective "as-of" forecasting, and rapid prototyping of survey designs.

Keywords: Generative Artificial Intelligence; Survey of Professional Forecasters; Expectations Formation; Simulated Economic Agents.
JEL Codes: C53, C55, E37, C45, D84.

[1] This Draft: January 2026. First Draft: April 2024. 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, George Washington University, Federal Reserve Board, Stanford Data Science, VIMM, and AMLEDS seminar participants along with participants at the ECB Transformative Power of AI Conference, the Society of Government Economists 2025 Annual Conference, Workshop on Frontiers in Measurement and Survey Methods, AI Frontiers in Finance, 2025 Imperial College London Hedge Fund Conference, Spring 2025 Midwest Macroeconomics Meeting, Google Economics Seminar, ECONDAT 2025, BSE Summer Forum Workshop on Theoretical and Experimental Macroeconomics, AI and the Future of Work at Wharton, Amazon's Forecasting, Macroeconomics and Finance (FMF) seminar, and Bank of Spain Economics of AI conference for valuable comments. We thank Erik Brynjolfsson, Jonathan Benchimol, Leland Bybee, Leland Crane, Michael Ehrmann *(discussant)*, Miguel Faria-e Castro *(discussant)*, Francesco Furno, Dragos Gorduza *(discussant)*, Yuriy Gorodnichenko, Robin Horton, Seung J. Lee, Huiyu Li, Peter Meyer *(discussant)*, Joan Paredes *(discussant)*, Alexander W. Richter, Sascha Steffen, Eric Zivot, and in particular 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

Expectations matter. Recognizing this, economists and policymakers rely on surveys as a measurement tool to capture expectations about future economic variables and conditions (e.g., Rudebusch, 2002; Orphanides and Williams, 2002, 2007; Patton and Timmermann, 2010; Manzan, 2011, 2021; Almås et al., 2024; Caplin, 2025). However, traditional surveys face inherent limitations.

Surveys are costly to administer, constraining both sampling frequency and sample size. They often lack flexibility, e.g., it is difficult to revise existing questions or add new ones, and new questions lack historical responses, limiting their value for time series analysis. Beyond that, most macroeconomic surveys solicit numeric answers with little insight into the reasoning behind them. Finally, surveys depend heavily on institutional infrastructure and funding, the stability of which has come under stress in recent years. These constraints have prompted calls for methodological innovation (Caplin, 2025). In this paper, we show that large language models (LLMs) can help mitigate some of these limitations.

We develop a framework that uses LLMs to simulate individual participants in the Survey of Professional Forecasters (SPF). Because LLMs are trained on large-scale text data, they can capture systematic patterns in how people describe, interpret, and justify decisions, which can be leveraged to approximate human reasoning in structured tasks (Suri et al., 2024; Niu et al., 2024). Recent work further shows that LLMs can reproduce behavioral patterns of economic agents in experimental contexts (e.g., Horton, 2023; Kazinnik, 2023; Zarifhonarvar, 2025; Hansen and Lee, 2025). In our setting, this allows us to construct synthetic forecasters whose behavior can be studied under different informational and institutional environments.

To build our simulation framework, we combine three inputs: (1) a novel dataset of hand-collected SPF participant characteristics (e.g., education, affiliation, experience), (2) real-time macroeconomic data available at each survey date, and (3) lagged median SPF forecasts. Using these inputs, we prompt an LLM to generate individual-level forecasts for 23 macroeconomic variables across 100 quarterly surveys from 1999Q1 to 2023Q4, creating a synthetic panel of professional forecasters.

We validate this synthetic panel against the human SPF along several dimensions by showing that human and AI forecaster panels share similar features. Their forecasts are broadly compa-

rable. Both undershoot and overshoot the same variables in the same survey quarters (e.g., CPI inflation is highly overshot in 2008:Q3 and undershot in 2020:Q2). The two panels also show similar patterns of forecast dispersion: disagreement among forecasters is greater at longer horizons than at shorter ones. Despite broad similarities, AI-generated and human forecasts differ in statistically and economically meaningful ways. This is not surprising; the two types of forecasters operate under distinct incentives and constraints. Humans face attention costs, reputation concerns, strategic updating, and heterogeneous judgment, while AI agents respond to standardized prompts and fixed information sets. Given these differences and the limited input data we provide, exact alignment at the level of point forecasts should not be expected.

We next isolate the contribution of each input and modeling choice by running a set of ablation exercises. We show that removing individual characteristics increases mean absolute errors in predicting real GDP by about 3-39%, further excluding real time data raises errors by up to 44%, and by also omitting past SPF median forecasts produces much larger increases, exceeding 70%. These patterns indicate that our AI forecasters respond systematically to the same types of input data and information that matter for human forecasters, and that their information sets and heterogeneity can be manipulated in ways that are informative but infeasible in real-world survey panels.

We interpret our results through a simple framework of human and AI expectations formation inspired by Mullainathan and Spiess (2017) and Agrawal et al. (2018). Human forecasters combine observable variables (e.g., labor market data) with difficult-to-observe inputs such as experience, judgment, and narrative context. Their forecasts reflect both cognitive constraints and informational advantages. In contrast, traditional prediction algorithms can process observable data very efficiently but cannot access these latent components. Generative AI models bridge this gap. By training on massive text corpora, LLMs internalize regularities in how humans transform information into verbal judgments and numerical forecasts, allowing them to approximate aspects of human expectations formation without direct access to individual forecasters' latent states. In our setting, we use this capability to construct LLM-based forecaster personas, which map real-time macroeconomic data, forecaster characteristics, and past SPF forecasts into survey responses. This design allows LLMs to combine the statistical advantages of algorithms with a stylized representation of human judgment, generating synthetic survey

forecasts that both resemble human panel behavior and can be used to study how expectations would change under alternative information sets or forecaster characteristics.

In addition, by incorporating individual characteristics, our framework shifts attention from a single aggregate forecast to the cross-sectional distribution of expectations across forecasters. Although many benchmark models of expectations (e.g., full information rational expectations) treat agents as homogeneous and fully rational, a large empirical literature documents substantial and persistent heterogeneity in both household and professional forecasts (e.g., Mankiw and Reis, 2002; Mankiw et al., 2003a; Coibion and Gorodnichenko, 2015; Manzan, 2011; Gabaix, 2020; Andre et al., 2022; Huang et al., 2022; Benchimol et al., 2022; Stark, 2013, among many others).

This heterogeneity is often attributed to differences in information sets, interpretation of information, attention, behavioral biases, forecasting methods, and subjective models of the economy. We capture some of these features by explicitly mapping observed forecaster characteristics into distinct LLM-based personas, and by studying how these personas translate a common real-time information set into different survey responses. And, because our AI forecasters provide narrative justifications for every prediction, we can make the "black box" of expectations formation more transparent. We show that the simulated forecasters exhibit substantial heterogeneity in their reasoning, drawing on policy considerations, mechanical updating rules, recent events, and theoretical models. As human professional forecasters, they combine model-based projections with judgment, with judgment playing a more prominent role during periods of heightened economic uncertainty.

In this respect, the Survey of Professional Forecasters is not merely a collection of macroeconomic variables and point predictions, but a record of how individual experts with heterogeneous backgrounds, heuristics, and biases respond to a standardized flow of information and survey instructions. Each SPF round can be viewed as a repeated decision problem in which forecasters observe the same question and process macroeconomic conditions through their individual expectation-formation rules. Our goal is to explore the potential of large language models to approximate these expectation-formation rules, not just the resulting time series of forecasts. LLM-based forecasters are well suited to this task: they can read and interpret the same prompts and real-time data as SPF participants, condition on forecaster-level attributes,

and generate synthetic survey responses that mimic the joint distribution of forecasts and characteristics observed in the human panel.

Our approach 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. A growing set of studies (e.g., Kazinnik, 2023; Cook and Kazinnik, 2024; Fedyk et al., 2024; Zarifhonarvar, 2025; Wu et al., 2025, and others) similarly explores how assigning specific roles to LLMs affects their decisions and behavior in economic environments. We extend this line of research in a different direction: rather than studying how role prompts shift behavior in stylized tasks, we construct an institutionally grounded synthetic panel that mirrors the SPF's repeated, multi-horizon forecasting environment.

This work also relates to the emerging literature on the application of LLMs in economic forecasting. For example, Faria-e Castro et al. (2023) show that LLMs such as Google's PaLM can generate inflation forecasts that often outperform the SPF. In contrast, our primary objective is not to develop a superior forecasting tool, but to model the behavior of SPF participants.[2] Covering 23 variables over a 24-year period, our setting provides a rich environment for assessing LLMs as behavioral models of professional forecasters, rather than as forecasting algorithms.

This work also intersects with recent efforts to use machine learning to replicate or complement expert judgment in financial domains. Murray et al. (2024) use deep learning to mimic technical chartists. Cao et al. (2024) develop an AI equity analyst that processes corporate disclosures, quantitative fundamentals, and macro indicators; its price targets outperform those of most human analysts, and a combined man *plus* machine forecast performs best. This work differs from this line of literature in two ways: our domain is macroeconomic surveys rather than securities analysis, and we treat the LLM as an object of study, rather than as a forecasting tool.

An important and commonly cited concern with models like LLMs is temporal data contami-

[2] While our goal is not to develop a superior forecasting model, we note that several studies have shown that the Survey of Professional Forecasters tends to outperform standard model-based forecasts for key macroeconomic variables, such as inflation and GDP (e.g., Ang et al. (2007) and Stark (2010), among others). This underscores the value of using SPF as the basis for evaluating simulated forecaster behavior.

nation, which occurs when the model has been trained on information that was published after the forecast date, potentially biasing its predictions (e.g., Lopez-Lira et al., 2025; He et al., 2025). This issue is relevant for us because we interpret the LLM as approximating an SPF forecaster who only has access to information available at the time of the survey. If the model heavily relies on *ex post* realizations, its forecasts risk reflecting knowledge of the future rather than expectations formed in real time. While individual-level SPF forecasts are unlikely to appear in training data, the model may have access to realized macroeconomic outcomes that the SPF aims to predict.

We address this concern in three ways. First, we instruct the model to use only information available prior to each survey deadline. However, we recognize that simply instructing the model to ignore post-survey information may not be sufficient to eliminate the risk of temporal data contamination. We thus test whether the model is simply recalling specific forecast values from its training set. The model's performance deteriorates sharply under this test, indicating that the model is not merely retrieving memorized numbers when generating responses. Finally, we conduct an out-of-sample validation exercise using data from 2024, which lies strictly beyond the model's October 2023 training cutoff and is therefore unavailable to the model. We find no evidence of structural breaks in forecast accuracy or behavior around this cutoff, suggesting that temporal data contamination is not driving our results and does not affect the interpretation of our framework as a *model of SPF forecasters*.

The use of computational techniques to represent individual behavior is not a new idea in economics. At the same time, our approach differs from traditional methods such as agent-based modeling (ABM).[3] In standard ABM applications, agents follow relatively simple, explicitly specified rules. By contrast, our AI forecasters embed high-dimensional decision rules learned from large text corpora, enabling them to make nuanced, context-dependent predictions that respond to complex information sets.[4] This additional flexibility can improve the fit of simulated behavior to observed data, but it also raises concerns about consistency and robustness.

To address these concerns, we systematically validate our forecaster model. We replicate our

[3] See Axtell and Farmer (2025) for a comprehensive review of agent based modeling in economics and finance.

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

simulations across multiple LLM architectures (GPT, Llama, and DeepSeek) and under both deterministic and stochastic decoding in order to assess how model choice and randomness affect the distribution of forecasts and panel behavior, and to ensure that our findings are not driven by a particular model. We also vary the definition of the forecaster persona within a fixed prompt structure to isolate the effect of role framing on the expectations-formation rules inferred by the model.

Taken together, our results suggest that LLM-based forecasters can serve as a scalable and transparent complement to traditional expectations surveys, rather than as a replacement for existing forecasting tools. Because synthetic panels can be run continuously at negligible marginal cost, they allow high-frequency tracking of the cross-sectional distribution of beliefs, including disagreement and tail risks, which is especially valuable in periods of heightened uncertainty. By flexibly varying the information set and combining structured real-time macro data with unstructured inputs (e.g., news, policy statements, scenario descriptions), researchers and policymakers can study how modeled forecasters would revise expectations across alternative information environments. The same framework can prototype and test new survey questions, fill in historical responses under consistent information sets, and assess counterfactual scenarios. In this sense, the LLM-based panel functions as a laboratory for expectations formation and survey design that augments existing survey infrastructure.

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

2 A Framework of Human and AI Expectation Formation

We present a stylized framework that focuses on the formation of expectations. Suppose agents form expectations on the state of a variable, y, H periods into the future, denoted by y_{t+H}. We assume that the true data generating process is governed by a function f plus an error ε_{t+H}:

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

The function f 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. For simplicity, we assume that the forecasting error, ε_{t+H}, is unpredictable with zero mean.

Humans can access both observable and unobservable information.[5] However, they process information imperfectly, which introduces an error term $\Delta_{i,t}$, varying over individuals i and time t. The error term $\Delta_{i,t}$ 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 $\mathbb{E}^{H_i}$ of $f(x_t, z_t)$, the human expectations 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. The human error term is thus given by:

$$\Delta_{i,t} = h_{i,t} - f(x_t, z_t). \tag{3}$$

Traditional algorithms, by contrast, can only access x_t, but they process x_t efficiently:[6]

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

We, however, employ LLMs, which form expectations in a non-deterministic 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 $\mathbb{E}^{\text{AI}}(\cdot)$:

$$m_t^{\text{AI}} = \mathbb{E}^{\text{AI}}\Big[f\big(x_t, z_t\big) \mid x_t\Big]. \tag{5}$$

We argue that LLMs can proxy human-formed expectations more closely than traditional algorithms, hence $\mathbb{E}^{H_i}(\cdot) - \mathbb{E}^{\text{AI}}(\cdot) < \mathbb{E}^{H_i}(\cdot) - \mathbb{E}(\cdot)$. However, human and AI expectations remain

[5] Although humans clearly lack complete information about *every* factor that predicts y_{t+H}, we assume that humans can observe both x_t and z_t to fix ideas in a simple framework.

[6] 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.

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:

$$m_t^{\text{AI}} = \mathbb{E}^{\text{AI}}\Big[f\big(x_t, z_t\big) \mid x_t,\, \bar{h}_{t-1}\Big], \tag{6}$$

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

$$\begin{aligned}\bar{h}_{t-1} &= \text{median}\Big(h_{1,t-1},\, h_{2,t-1},\, \ldots,\, h_{N,t-1}\Big) \\ &= f\big(x_{t-1}, z_{t-1}\big) + \text{median}\Big(\Delta_{1,t-1},\, \Delta_{2,t-1},\, \ldots,\, \Delta_{N,t-1}\Big) \\ &= f\big(x_{t-1}, z_{t-1}\big) + \bar{\Delta}_{t-1}. \end{aligned} \tag{7}$$

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, $\bar{\Delta}_{t-1}$.

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 $w_{i,t}$.[7] Specifically, we decompose the bias as:

$$\Delta_{i,t} = \gamma\big(w_{i,t}\big) + e_{i,t}, \tag{8}$$

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 expectation can then be expressed as:

$$m_{i,t}^{\text{AI}} = \mathbb{E}^{\text{AI}}\Big[f\big(x_t, z_t\big) \mid x_t,\, \bar{h}_{t-1},\, w_{i,t}\Big] \tag{9}$$

and involves an approximation error given by:

$$\Delta_{i,t}^{\text{AI}} = m_{i,t}^{\text{AI}} - f(x_t, z_t). \tag{10}$$

[7] 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.

The AI forecaster has access to observed macroeconomic data, the past median human forecast, and a set of assigned 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 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 and judgment into the forecasting process. In this sense, our framework fits into the broader agenda of combining flexible statistical models with economically structured questions. Whited (2023) argues that reduced-form and structural methods are complements rather than substitutes, and documents several ways to use high dimensional statistical models to inform and discipline structural analysis. In this spirit, we treat the LLM as a rich reduced-form representation of forecasters' expectation formation rules that could ultimately be embedded in structural models of the economy.

Finally, we stress that AI forecasters are 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.

3 The Survey of Professional Forecasters

The SPF is a quarterly survey of U.S. economic forecasting 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. Response deadlines typically fall mid-quarter.[8] Although the survey form has evolved over time, with eight updates from 1999 to 2023, its core structure remains intact. The SPF plays a central role in informing policy makers, serves as a key benchmark for evaluating forecast accuracy in academic research, and is the longest-running publicly available forecasting project, providing a rich dataset of forecasts

[8] Complete information and documentation are available through the Federal Reserve Bank of Philadelphia (link). See the full schedule of deadlines and release dates here.

on key macroeconomic indicators (Croushore, 1993).[9]

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 variables are summarized in Table 1. Altogether, these variables cover a wide range of indicators with different properties, including both growth rates and levels. 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. We exclude the 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. Table 3 provides descriptive statistics. 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.[10]

3.2 The SPF Panel

SPF panelists are typically considered to be a benchmark for attentive agents. They typically rely on a mix of data analysis, quantitative models, and professional judgment to make predictions as part of their jobs. Most have significant experience in macroeconomic forecasting. New forecasters are recruited via SPF's *Call for Participants* and undergo a trial period, during which submissions are checked for conceptual accuracy. 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

[9] 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. As for academic research, please 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); Croushore and Stark (2019), among many.

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

that allow researchers to track forecasts across rounds. This anonymity protects forecasters' professional interests and encourages honest, unbiased forecasts without concern for repercussions (Croushore and Stark, 2019). While we have access to the forecasters' names, they are not disclosed in this paper to preserve their anonymity. The SPF panel composition changes over time as respondents join or leave, and some may skip certain questions, resulting in an unbalanced panel. The average panel size fluctuates around forty participants (see Figure A.6 in appendix).

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.

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 (example in Figure 1). Many panelists also share their participation on social media or professional platforms.

We compile forecaster characteristics exclusively from publicly available sources (e.g., the Philadelphia Fed acknowledgements and public professional profiles). We use these characteristics only to parameterize synthetic personas and do not attribute any simulated forecasts or explanations to named individuals. In the analysis, we work with anonymized identifiers and report results in aggregate. We also avoid inferring or imputing sensitive personal attributes; when a characteristic is not publicly available, we leave it missing rather than guessing.

We start by collecting forecaster names and affiliations from the Philadelphia Fed acknowledgments. Using this information, we explore publicly available sources (e.g., personal websites, LinkedIn, etc.), and 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 pres-

ence. For country of origin, we consider indicators like high school location, if available.[11] 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).

We condition the model only on forecaster characteristics that plausibly map into systematic, interpretable differences in information sets, incentives, and forecasting styles. In practice, the most consequential fields are (i) affiliation/sector (e.g., academia, commercial bank, consulting, government), which proxies for institutional objectives and exposure to particular data sources and modeling traditions; (ii) job title/role, which captures the forecaste's function (research, markets, policy analysis, macro strategy) and therefore the likely salience of specific variables and horizons; and (iii) education and training, which proxy for methodological priors and comfort with structural versus reduced-form reasoning. These inputs are not meant to "identify" individuals or reproduce private information. Rather, they provide coarse but meaningful variation that can generate heterogeneity in levels, disagreement, and updating, mirroring the fact that SPF participants differ in institutional constraints, attention, and modeling approaches even when observing similar public macro data.

Figure 2 provides an overview of the SPF panel over the past 20 years, showing a diverse group with varied educational backgrounds, roles, and affiliations. As Figure A.7 in the Appendix shows, panelist tenure varies widely, with some contributing for over 25 years and others for shorter periods, adding both established and fresh perspectives. 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

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

interviews. Most panelists hold a 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.

Taken together, these facts highlight a long-lived but evolving panel of macroeconomic experts with persistent heterogeneity in training, roles, and institutional context. In the next section, we use these observable differences as inputs to construct synthetic forecasters that mirror the human SPF participants.

4 Simulating the SPF with LLMs

We next outline our simulation procedure. We use the information described in the previous section to construct personalized profiles that mirror human participants and generate synthetic forecaster personas. To closely match the forecasting environment humans face, we embed each AI forecaster in a prompt that includes real-time macroeconomic data and historical SPF median forecasts, and have them generate point forecasts for each quarter between 1999Q1 and 2023Q4. In addition, we extend the simulation to include an out-of-sample test covering 2024Q1 through 2024Q4, allowing us to evaluate model performance beyond its training window.

4.1 Models

We run the simulated survey using a number of large language models. Our main analysis relies on GPT-4o mini with a temperature of 1.0. We also conduct robustness checks with GPT-3.5, GPT-4, Llama-3.3-70B, and DeepSeek-V3. These models span different generations: GPT-3.5 as an older baseline, GPT-4 as a frontier model, and GPT-4o mini as a recent lightweight implementation. Including open models (i.e., Llama and DeepSeek) also enhances the transparency and reproducibility of our simulation (see, e.g., Cook et al., 2025), aligning with recent calls in the literature to prioritize open infrastructure in scientific research using LLMs. We find the results to be similar across models and available upon request.

We select GPT-4o mini as our primary model based on a series of preliminary experiments

evaluating performance and stability. Using CPI inflation as a test case, we compared GPT-3.5, GPT-4, and GPT-4o under multiple temperature settings: 0 for deterministic outputs, 1.0 as the default balanced setting, and values above 1.5 to introduce greater variability.[12] Given its strong performance and its architectural similarity to GPT-4o, GPT-4o mini offers an efficient and reliable option for large-scale simulations. Our implementation uses an AI agent built with the OpenAI Assistants API to manage tasks such as generating forecasts, handling query–response chains, and performing additional data processing.[13] Ultimately, however, this framework is model-agnostic. Our implementation uses an automated pipeline to generate quarter-specific queries and to parse responses into a standardized format consistent with the SPF.

4.2 Prompt Design

Our objective in the prompt design is to elicit forecasts from LLMs that are directly comparable to those submitted by human SPF participants. The baseline prompt, described in Prompt 1, assigns the model the role of an SPF panel member making a forecast at a specific survey date. Each AI forecaster is instructed to report numerical forecasts for the current quarter (t) and the next four quarters ($t+1$ to $t+4$) in a standardized format, and to provide a brief 1–2 sentence explanation of their predictions.[14] The prompt explicitly emphasizes that the forecaster must base their judgment only on information available at that point in time and must not use data or events from the future. In this way, we aim to simulate a professional forecaster who uses an internally coherent model of the economy and a fixed information set to form expectations about the near-term evolution of key macroeconomic variables.

We automate the generation of these forecast queries and collect responses using an AI assistant. Each query embeds a forecaster profile (education, job title, organization type, and related characteristics), which induces heterogeneity in the simulated panel while holding the information environment fixed. To mirror the SPF process as closely as possible, we also provide our AI forecasters with a data environment that closely matches that of human participants.

[12] Temperature controls the randomness of model outputs. Higher values produce more dispersed responses.

[13] The Assistants API is designed for complex, multi-step workflows involving memory, file handling, and coordinated tool use.

[14] These instructions are implemented via the system prompt.

Specifically, the prompt includes past median SPF forecasts and real-time macroeconomic data. This setup largely ensures that any differences between human and AI forecasts arise from differences in forecasting behavior rather than from differences in the underlying information set. In the next section, we evaluate how closely the behavior of AI forecasters matches that of the human SPF panel.

5 Results

This section presents our simulation results and validates them against data from human forecasters. With two sets of forecasts (human and AI) for five forecasting horizons across 23 variables, the amount of data is substantial. To keep the presentation focused, we therefore discuss results for selected horizons only, and focus mainly on four key variables in our discussion: the CPI inflation rate, 3-month Treasury bill rate, unemployment rate, and real GDP index.[15] We choose to focus on these four economic variables because 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. Forecasts are winsorized at the median ± 3 standard deviations of the realized data to exclude potential erratic LLM answers.[16]

5.1 Comparing AI and Humans

First, we evaluate how closely AI and human expectations are aligned. Figure 3 compares median expectations from AI and human forecasters 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 forecasters' expectations, particularly for unemployment and real GDP. The largest discrepancy appears in CPI inflation, where AI forecasters respond more to the business cycle. Another notable difference occurs with the four-quarter ahead 3-month Treasury bill rate during the zero lower bound period,

[15] The tables provide additional results for the other variables.

[16] Winsorization only impacts the "recall" results in Table 7.

where human forecasters predict a lift-off, while AI forecasters stick to the lower bound until the 2015 rate-hike cycle.

Table 4 compares the median expectations of AI and human forecasters for all variables over three forecast horizons: nowcast and one and four quarter(s) ahead. To relate expectations to realized values, the table reports mean absolute errors (MAEs) associated with expectations from human and AI forecasters. The table shows that AI is more accurate than humans for some variables (e.g., nominal GDP and the PCE inflation rate), but less accurate for others (e.g., real GDP). The differences are, however, mostly statistically insignificant.

This comparison is based on averages across all survey quarters. Table 5 shows similar results for the percentage of surveys in which AI forecasters are more accurate than human forecasters (values above 50% are boldfaced).

However, focusing on the best-performing AI and human forecasters provides tentative evidence that AI forecasters achieve higher accuracy (see Table 6).

Next, we consider the comparison of individual-level expectations. Since the survey is anonymized, we cannot match individual human and AI forecasters one-to-one. Instead, we focus on the cross-sectional distributions of expectations. Figure 4 shows density plots of individual expectations across forecasters for four key episodes: 1999Q1 (earliest observation), 2008Q3 (Great Financial Crisis), 2020Q2 (COVID-19 pandemic), and 2023Q1 (latest survey with four-quarter-ahead realizations). Each plot is centered around realized values. In some episodes, the distributions of human and AI expectations align almost perfectly (see, e.g., the one-quarter ahead expectations for the unemployment rate and real GDP in 2008Q3). In addition, the dispersion of real GDP expectations in 2020Q2 is unusually high for both AI and human forecasters. In other episodes, there are notable differences (e.g., one-quarter-ahead expectations for real GDP and the unemployment rate in 1999Q1). The extent to which human and AI forecast distributions align is thus highly dependent on the variable, forecast horizon, and quarter. We also do not see a simple pattern in which humans and AI are systematically more or less aligned during high-uncertainty episodes such as 2008Q3 and 2020Q2. However, cross-sectional forecast dispersion exhibits a common pattern across both panels: disagreement rises with the forecast horizon, reflecting greater uncertainty about more distant outcomes (e.g., see Table 8).

Taken together, these results are consistent with our interpretation of the LLM as an imperfect but informative model of expectations formation.

5.2 The Role of Prompt and Context Engineering

We have shown that AI and human forecasters are qualitatively similar. Our framework in Section 2 attributes these results to the AI-specific expectations operator $\mathbb{E}^{\text{AI}}(\cdot)$, which combines patterns learned from the training data with the conditioning information given in the LLM prompt: forecaster characteristics, real-time data, and past forecasts. In this section, we evaluate the roles of these inputs in shaping expectations by progressively removing them from the prompt. We run an ablation exercise comparing several "stripped-down" versions of the AI forecaster to a fully informed benchmark that has access to all three information types.

We begin by simulating expectations 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 its 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 prompt input so that the model only relies on its training data. This setup is shown in Prompt 4.

Table 7 compares the results 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 forecaster loses some ability to account for systematic, forecaster-specific biases (the component $\gamma(w_{i,t})$ in the bias term) when characteristics are omitted, but it retains much of its ability to map aggregate information into expectations because it still conditions on real-time data and past median SPF forecasts, which approximate $\bar{h}_{t-1}$. Therefore, performance

only degrades slightly when characteristics are removed.

In contrast, removing real-time data often pushes the ratios further above one, showing a clearer deterioration in how well the AI forecaster tracks realized outcomes. Eliminating *both* real-time data and past median SPF forecasts is associated with even higher MAE ratios, and we can typically reject that these ratios are equal to one with p-values below one percent. For some variables, the loss in performance is extreme. For example, the four-quarter-ahead expectations for real GDP are 73% worse compared with the baseline, whereas the nowcast is worse by a factor near six. Eliminating these pieces of information forces the AI to rely almost entirely on its pre-trained knowledge without current context or the valuable proxy for z_t. In terms of the framework, this is equivalent to conditioning only on a much more limited information set, leading to a larger approximation error and thus much higher MAEs.

Overall, the performance of the AI forecaster deteriorates progressively as more inputs are removed from the prompt, highlighting that context engineering plays a key role in determining the performance of LLMs. In particular, past median SPF forecasts strongly contribute to the AI's ability to reproduce human expectations and to track realizations. The critical role of lagged human forecasts reflects a fundamental feature of our approach. On the one hand, conditioning on lagged human forecasts enables the LLM to access indirect signals about factors that are difficult to quantify (professional judgment, institutional knowledge, narrative context). On the other hand, this dependence implies that the current LLM framework is best suited as a complement to traditional surveys, valuable for exploring counterfactual scenarios and belief distributions, rather than as a replacement.

5.3 Are LLMs Forecasting or Recalling?

In this section, we address a critical concern: whether our results reflect temporal data contamination; that is, whether the LLM inadvertently uses information from after the forecast date that was included in its training data. Limiting large language models to use only information from a specific time period is a challenging task. LLMs generate output based on learned patterns, which gives the impression of generating content, even though the underlying knowledge comes from repeating the patterns from previously seen data (Gurnee and Tegmark, 2023). This could be problematic in our context, 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."[17]

To deal with this issue, we apply several mitigation strategies to limit temporal leakage and improve the realism of a real-time simulation. 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 aims 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 Figure 5 for details). The final column in Table 7 shows the MAEs from this recall exercise as ratios of the baseline nowcast MAEs. The accuracy drops considerably, with MAE ratios often surging into double digits. For example, the LLM recalls the values of real GDP with a MAE that is 20 times larger than the baseline nowcasting MAE (see Table 7).

Finally, we evaluate forecast accuracy in an out-of-sample exercise on 2024 Q1:Q4 data, which excludes the training data sample and therefore eliminates any concerns about temporal leakage by construction. Figure 5 shows forecast errors for one-quarter ahead forecasts around the training cut-off date for the four key variables. The figure indicates that there is no clear structural break in the forecast error, or in how the error of AI-generated forecasts differ from the human forecast error, around the cut-off date, suggesting that the performance of the model is not driven by temporal leakage.[18] We provide additional tests in Appendix B that show no significant difference between AI and human forecasters out of sample.

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

[18] We observe larger deviations for the 3-month bill rate and real GDP in the final quarter of 2024 (which is not around the cutoff), which may be due to the model's lack of access to real-time major developments—such as election-related news, updates on monetary policy, and other current events.

5.4 Cross-Sectional Heterogeneity

A central feature of professional forecast panels is persistent heterogeneity: individual forecasters systematically deviate from consensus in ways that reflect their backgrounds, institutional affiliations, information processing strategies, or informational rigidities (Mankiw et al., 2003b; Patton and Timmermann, 2010; Coibion and Gorodnichenko, 2015; Manzan, 2011). We assess whether our LLM-based forecaster personas replicate this pattern by estimating forecaster-specific mean deviations (individual fixed effects), defined as the average deviation of forecaster i's forecast from the quarterly mean.

For each variable and quarter, we calculate forecaster i's deviation from the cross-sectional mean:

$$\text{deviation}_{it} = \text{forecast}_{it} - \frac{1}{N_t} \sum_{j=1}^{N_t} \text{forecast}_{jt} \tag{11}$$

where N_t is the number of forecasters in quarter t. We then compute each forecaster's fixed effect as the time-series average of these deviations, i.e., $\text{FE}_i = \frac{1}{T_i} \sum_{t=1}^{T_i} \text{deviation}_{it}$.

Figure 6 shows the distribution of fixed effects across the four main variables. We observe substantial cross-sectional variation: for real GDP, fixed effects range from -160 to $+187$ billion dollars (2012 dollars), with a standard deviation of 28.4. For CPI inflation, the standard deviation is 0.27 percentage points. This heterogeneity is consistent with the dispersion documented in studies of human SPF forecasters (Patton and Timmermann, 2010).

We further examine whether disagreement varies systematically with forecast horizon. Table 8 reports cross-sectional standard deviations at short (h+1) and long (h+4) horizons. As in the SPF, disagreement in our AI panel is greater at longer horizons for real variables. For unemployment, the ratio of long-horizon to short-horizon dispersion is 1.53 in the LLM panel versus 2.29 in the SPF (both $p < 0.001$). For real GDP, the corresponding ratios are 1.40 and 2.49. However, neither panel exhibits this pattern for CPI inflation: the SPF actually shows *less* dispersion at longer horizons ($\sigma_{h+4}/\sigma_{h+1} = 0.72$, $p < 0.001$), while the LLM panel shows no significant difference. This asymmetry, increasing disagreement with horizon for real variables but not for inflation, likely reflects anchored long-run inflation expectations, and the LLM forecasters replicate this qualitative feature of human forecast disagreement.

5.5 Forecast Revision Dynamics

Human forecasters update their beliefs gradually in response to new information and past forecast errors, exhibiting sticky expectations and convergence toward consensus (Coibion and Gorodnichenko, 2012; Bordalo et al., 2020). We assess whether AI forecasters replicate these patterns by analyzing forecast revisions across consecutive quarters for both panels.

For each forecaster-quarter pair, we define the forecast revision as:

$$\text{revision}_{it} = \text{forecast}_{it} - \text{forecast}_{i,t-1} \tag{12}$$

We test two hypotheses central to the expectations formation literature: (1) forecasters correct past errors, and (2) forecasters converge toward the consensus view. Table 9 reports these regressions separately for human SPF participants (Panel A) and AI forecasters (Panel B).

Both human and AI forecasters exhibit significant error correction, but with different magnitudes. Human forecasters display near-complete error correction: coefficients range from 0.88 (CPI) to 0.98 (RGDP), with R^2 values between 0.64 and 0.94. A coefficient of one implies that forecasters fully revise their predictions in response to past errors; e.g., if a forecaster underestimated GDP by \$100 billion, they revise upward by approximately \$100 billion in the following quarter.

AI forecasters also correct past errors, but only partially. Error correction coefficients range from 0.44 (CPI) to 0.65 (UNEMP), approximately 40–65% of the human response. This suggests that AI forecasters place less weight on realized forecast errors when updating their beliefs, potentially reflecting stronger priors from their training data or less sensitivity to recent information.

As for the consensus tracking behavior, human forecasters show statistically significant but economically modest convergence toward the SPF median: coefficients range from -0.30 (TBILL) to -0.87 (UNEMP), but R^2 values are uniformly low (0.004–0.14). Distance from consensus explains little of the variation in human forecast revisions, suggesting that professional forecasters maintain independent views rather than systematically anchoring on the panel median.

AI forecasters exhibit considerably stronger consensus tracking. For RGDP, the coefficient is -1.02 with $R^2 = 0.78$: forecasters who deviate from the SPF median by \$X billion revise their

forecast by approximately $X billion toward consensus. TBILL shows even stronger anchoring (-1.14, $R^2 = 0.72$). CPI displays moderate consensus tracking (-0.69, $R^2 = 0.55$), while unemployment shows the weakest convergence (-0.51, $R^2 = 0.09$).

These results reveal a fundamental difference in how human and AI forecasters form expectations. Human forecasters behave as independent information processors: they aggressively correct past errors (learning from mistakes) but place comparatively little weight on the consensus view (maintaining idiosyncratic beliefs). AI forecasters exhibit the opposite pattern: they update more gradually in response to past errors but strongly anchor on the professional consensus, particularly for output and interest rate forecasts.

This difference has important implications for interpreting our framework. The strong consensus anchoring by AI forecasters validates our ablation results in Section 5.2: removing past SPF forecasts significantly increases forecast errors precisely because AI forecasters treat the consensus as a highly informative summary of latent information used by human forecasters. At the same time, this anchoring behavior also means that AI forecasts may inherit biases present in the SPF consensus, and may be slower to deviate from consensus during periods of structural change when independent judgment becomes especially valuable. The variable-specific patterns (stronger anchoring for RGDP and TBILL than for CPI and unemployment) suggest that AI forecasters weight the consensus signal differently depending on the variable, potentially reflecting differences in how these indicators are discussed in training data. We view this feature not as a limitation, but as an economically meaningful property of the simulation: by explicitly controlling access to consensus information and historical context, the framework allows researchers to study when consensus anchoring improves forecast accuracy and when it may hinder adaptation.

5.6 Counterfactual Applications

A key advantage of LLM-based simulation over traditional surveys is the ability to conduct counterfactual experiments. We demonstrate this capability by examining a policy relevant question: how would macroeconomic forecasts have differed if policymakers had access to LLM-based forecasting during major economic crises?

We compare LLM consensus forecasts (median across our simulated panel) to actual SPF

consensus forecasts across four major crisis episodes: the 2008–09 financial crisis, the 2020 COVID-19 shock, and the 2021–22 inflation surge. For each episode, we compute forecast errors by comparing both LLM and SPF predictions to subsequently realized values. This exercise addresses a practical counterfactual: would LLM-based forecasts have provided policymakers with more accurate signals during periods of extreme uncertainty?

Table 10 presents the results. During the COVID-19 lockdown quarter (2020:Q2) professional SPF forecasters predicted unemployment of 3.8%, anchoring close to the pre-pandemic rate. By contrast, the LLM panel predicted unemployment of 14.5%. The realized unemployment rate reached 13.0%, implying forecast errors of -9.2 percentage points for the SPF versus $+1.5$ percentage points for the LLM, a six-fold improvement in accuracy. This difference has significant policy implications: in March 2020, policymakers debating the scale of fiscal stimulus and monetary accommodation would have received dramatically different signals from an LLM-based forecast.

The LLM's advantage persists in other crisis quarters. During the financial crisis trough (2009:Q1), the LLM predicted CPI deflation of -2.5%, nearly matching the realized -2.4%, while SPF forecasters predicted -9.2%, overshooting the actual deflation by 6.8 percentage points. For unemployment forecasts during crises, the LLM achieves a mean absolute error of 1.12 percentage points compared to 1.91 for the SPF. During normal periods, both forecasting approaches perform similarly, with mean absolute errors of approximately 0.2 percentage points for unemployment.

Figure 9 illustrates this pattern. The left panel shows unemployment forecast errors across crisis quarters, highlighting the COVID outlier where the LLM's forecast was remarkably accurate while professionals severely underestimated the shock. The right panel shows that for CPI inflation, results are more mixed: the SPF outperforms during the 2022 inflation shock, where real-time access to current data releases provided professional forecasters an advantage.

Why might LLM forecasts outperform during crises? We propose two mechanisms. First, LLMs draw on extensive historical information about past crises, including pandemics, financial collapses, and their macroeconomic consequences. Professional forecasters in real-time face genuine uncertainty and may anchor to recent experience. Second, LLMs process information differently than human forecasters: they do not exhibit the same cognitive biases (such as

anchoring to pre-crisis levels) that may have led SPF forecasters to underestimate the COVID shock's magnitude.

These results suggest a specific use case for LLM-based forecasting in policy applications: as a complement to traditional surveys for detecting tail risks during periods of extreme uncertainty. While LLMs cannot replace real-time data access for nowcasting, they may provide valuable signals about the potential magnitude of unprecedented shocks precisely when such information is most valuable to policymakers.

5.7 Reasoning Patterns

A distinctive feature of LLM-based forecasting is the ability to elicit reasoning alongside numerical predictions. After each forecast, we prompt the model to provide a brief rationale, generating explanations that clarify the considerations underlying the resulting forecasts.

Since 2008, the European Central Bank has conducted special surveys at roughly five-year intervals to better understand the processes and methodologies used by professional forecasters. In this section, we use evidence from these surveys to assess the extent to which the simulated reasoning data align with documented forecasting practices among human forecasters. Specifically, we classify each reasoning response into four categories using keyword frequency rules: *Mechanical* (references to trends, historical averages, and mean-reversion), *Policy* (references to Federal Reserve actions, fiscal or monetary policy, and institutional responses), *Model* (references to economic theory such as the Phillips curve, Okun's law, or structural models), and *Event* (references to specific shocks, crises, or salient events such as the pandemic or financial crisis).[19]

Following the ECB methodology (European Central Bank, 2024), we aggregate these categories into two broader types: *model-based* reasoning (combining Model and Mechanical, as both involve mathematical representations of economic relationships) and *judgment-based* reasoning (combining Policy and Event, reflecting experience-driven interpretation of current conditions). The ECB defines model-based forecasting as relying on "any mathematical representation of relationships between economic variables" and judgment-based forecasting as relying on "experience and intuition" (European Central Bank, 2019).

[19] Full classification details and keyword lists are provided in Appendix C.

Table 11 reports the prevalence of each reasoning type across normal and crisis periods, showing both the detailed categories and the ECB-style aggregation. During normal periods, model-based reasoning dominates (99.6%), with judgment-based reasoning at 63.3%. However, during crisis periods, LLM forecasters shift markedly toward judgment-based reasoning. The most dramatic shift occurs in event-driven reasoning, which rises from 19.7% in normal periods to 82.4% during the financial crisis and 77.0% during COVID-19. Mechanical reasoning drops substantially (from 98.6% to 72.1–84.7%), while the inflation shock period shows a distinct pattern: policy-based reasoning surges to 87.3% as forecasters emphasize Federal Reserve tightening.

The ECB-style aggregation reveals the broader pattern. During crises, judgment-based reasoning rises sharply: from 63.3% in normal periods to 87.6% during the financial crisis, 79.9% during COVID-19, and 87.7% during the inflation shock. Model-based reasoning declines most dramatically during COVID-19 (to 72.5%), likely reflecting the unprecedented nature of the pandemic shock that rendered historical patterns less informative. All differences between crisis and normal periods are statistically significant at conventional levels ($p < 0.001$).

These patterns are qualitatively consistent with evidence from special surveys of human professional forecasters. The ECB's Survey of Professional Forecasters finds that "the role of judgment in point forecasts has increased" during "periods of economic volatility" such as the global financial crisis and pandemic (European Central Bank, 2024). Our results suggest that LLM forecasters, like their human counterparts, adapt their reasoning approach to economic conditions, relying more heavily on judgment and event-specific narratives during periods of heightened uncertainty. While more work is needed to fully characterize and validate LLM reasoning, this analysis suggests that our framework could be extended to simulate not only forecasts but also the reasoning processes that professional forecasters employ.

6 Conclusion

In this paper, we introduce a new framework for modeling expert expectations' formation using large language models. Rather than treating the LLM primarily as a forecasting tool, we use it as a model of forecasters: an expectations operator that maps real-time data, lagged forecasts, and forecaster characteristics into survey responses.

Using this framework, we construct a synthetic panel of forecasters designed to mirror the institutional environment of the Survey of Professional Forecasters. We show that the resulting AI-generated panel reproduces salient features of the human survey, including median forecast levels. Importantly, the synthetic panel does not collapse to a trivial replication of the SPF: while aligned in many dimensions, AI and human forecasts differ in systematic and interpretable ways that reflect differences in information availability, information processing, and updating behavior.

Our ablation exercises clarify why this occurs. Performance improves notably when the LLM is augmented with lagged human forecasts, real-time data, and forecaster characteristics. Removing these inputs systematically worsens performance, with the largest deterioration occurring when past SPF medians are excluded. This finding highlights the role of lagged human forecasts as informative summaries of latent factors—such as professional judgment, narrative context, and institutional knowledge—that are difficult to observe directly. Conditioning on these signals allows the model to approximate aspects of human expectations formation. We therefore view the framework not as a replacement for human surveys, but as a complementary measurement tool that enables controlled analysis of how expectations respond to alternative information environments.

A key advantage of LLM-based panels is their flexibility. Traditional surveys cannot retroactively modify panel composition, control information availability, or experiment with alternative question designs. LLM-based panels enable all of these counterfactual analyses with appropriate context engineering. This, in turn, allows researchers to prototype new survey designs at minimal cost and simulate counterfactual forecast distributions for past periods under custom information sets. More broadly, our results point to a substantive role for generative AI in macroeconomics: not as a forecasting oracle, but as a laboratory for studying expectations formation. By combining institutional realism with flexible context engineering, the framework allows researchers to analyze expectations in a disciplined and transparent way. We see this approach as a step toward integrating modern machine learning tools into the measurement of expectations in a manner that complements economic theory and existing survey-based methods.

Integration of human expertise with AI-based forecaster panels, analogous to a "man + ma-

chine" approach in other domains, promises richer measurements of expectations and deeper understanding of belief formation in macroeconomic environments. The human SPF provides ground truth for calibration and validation; the LLM panel provides a flexible experimental platform for testing hypotheses that are challenging to address with traditional surveys. This complementarity is the core contribution of our framework.

References

Ajay Agrawal, Joshua Gans, and Avi Goldfarb. Prediction, judgment, and complexity: a theory of decision-making and artificial intelligence. In *The economics of artificial intelligence: An agenda*, pages 89–110. University of Chicago Press, 2018.

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.

Andrew Ang, Geert Bekaert, and Min Wei. Do macro variables, asset markets, or surveys forecast inflation better? *Journal of Monetary Economics*, 54(4):1163–1212, 2007.

Robert L Axtell and J Doyne Farmer. Agent-based modeling in economics and finance: Past, present, and future. *Journal of Economic Literature*, 63(1):197–287, 2025.

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.

Sean Cao, Wei Jiang, Junbo Wang, and Baozhong Yang. From man vs. machine to man+ machine: The art and AI of stock analyses. *Journal of Financial Economics*, 160:103910, 2024.

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.

Olivier Coibion and Yuriy Gorodnichenko. What can survey forecasts tell us about information rigidities? *Journal of Political Economy*, 120(1):116–159, 2012.

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

Thomas R. Cook and Sophia Kazinnik. Social Bias in Financial Applications of Large Language Models. *SSRN Working Paper*, 2024.

Thomas R Cook, Anne Lundgaard Hansen, Sophia Kazinnik, and Peter McAdam. Under Pressure: Strategic Signaling in Bank Earnings Calls. *Available at SSRN 5382397*, 2025.

Dean Croushore and Tom Stark. Fifty Years of the Survey of Professional Forecasters. *Economic Insights*, 4(4):1–11, 2019.

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.

European Central Bank. The ecb survey of professional forecasters: Forecast processes and methodologies: Results of the 2018 special survey. Ecb survey report, European Central Bank, February 2019. Special Survey.

European Central Bank. The ecb survey of professional forecasters: Forecast processes and methodologies: results of the 2023 special survey. Technical Report QB-BR-24-005-EN-N, European Central Bank, Frankfurt am Main, Germany, April 2024. Survey conducted on the occasion of the 25th anniversary of the ECB Survey of Professional Forecasters (SPF).

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.

Anne Lundgaard Hansen and Seung Jung Lee. Financial Stability Implications of Generative AI: Taming the Animal Spirits. *arXiv preprint arXiv:2510.01451*, 2025.

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

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.

Shih-Tang Huang, C James Hueng, and Jyh-Lin Shiao. Macroeconomic attention and forecaster heterogeneity. *Journal of Economic Behavior & Organization*, 201:75–98, 2022.

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

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

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 versus sticky prices: A proposal to replace the new keynesian phillips curve. *Quarterly Journal of Economics*, 117(4):1295–1328, 2002.

Gregory Mankiw, Ricardo Reis, and Justin Wolfers. Disagreement about Inflation Expectations. *NBER Macroeconomics Annual*, 18:209–248, 2003a.

Gregory Mankiw, Ricardo Reis, and Justin Wolfers. Disagreement about inflation expectations. *NBER Macroeconomics Annual*, 18:209–248, 2003b.

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.

Sendhil Mullainathan and Jann Spiess. Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106, 2017.

Scott Murray, Yusen Xia, and Houping Xiao. Charting by machines. *Journal of Financial Economics*, 153:103791, 2024.

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.

Tom Stark. Realistic evaluation of real-time forecasts in the survey of professional forecasters. *Federal Reserve Bank of Philadelphia Working Paper Series*, 2010.

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.

Toni M Whited. Integrating structural and reduced-form methods in empirical finance. *Journal of Financial Econometrics*, 21(3):597–615, 2023.

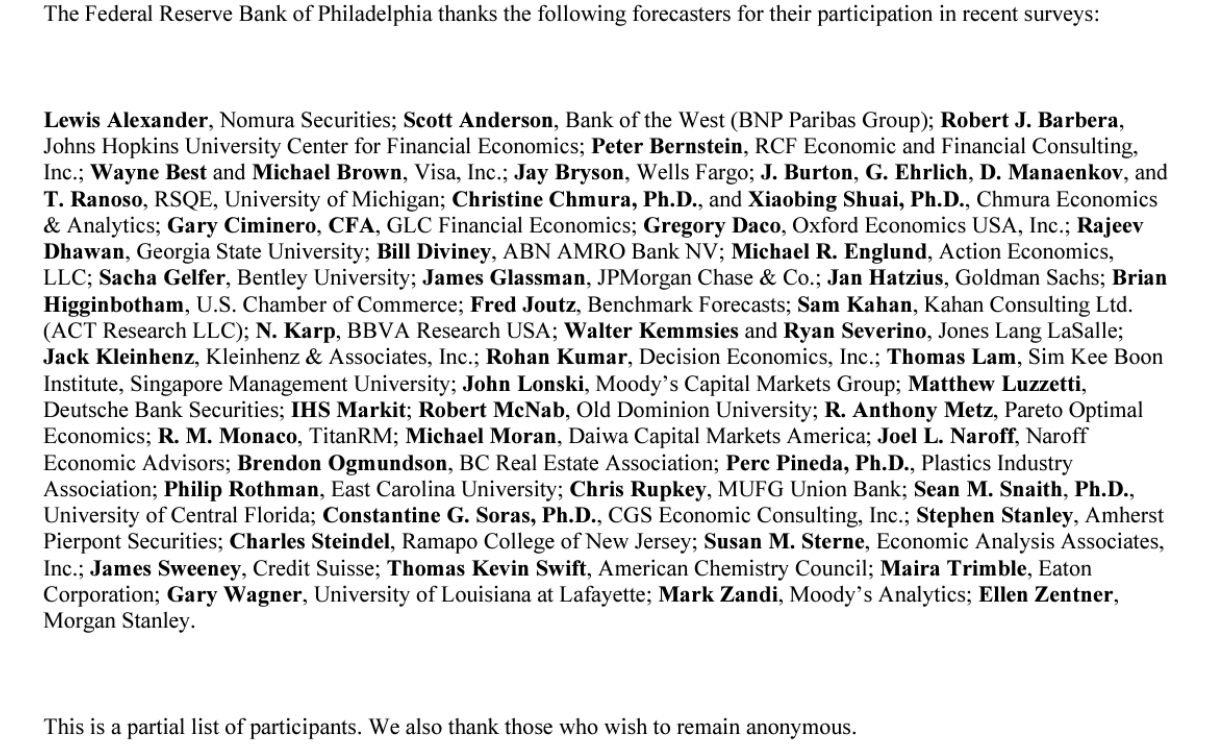
Jing Cynthia Wu, Jin Xi, and Shihan Xie. LLM survey framework: Coverage, reasoning, dynamics, identification. SSRN working paper, 2025.

Ali Zarifhonarvar. Generating inflation expectations with large language models. *Journal of Monetary Economics*, page 103859, 2025.

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.



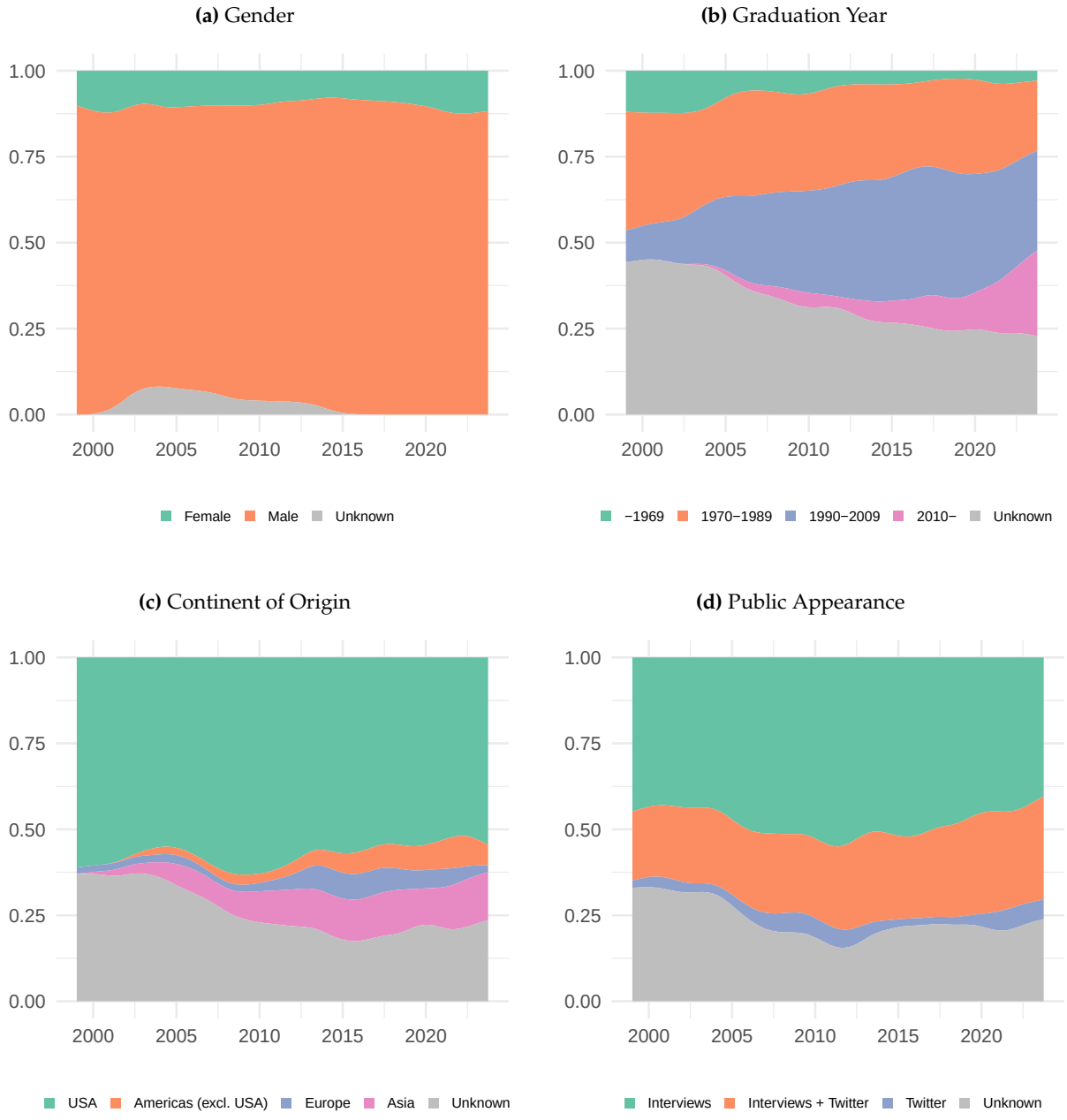
The Federal Reserve Bank of Philadelphia thanks the following forecasters for their participation in recent surveys:

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: Characteristics of the SPF panel over time

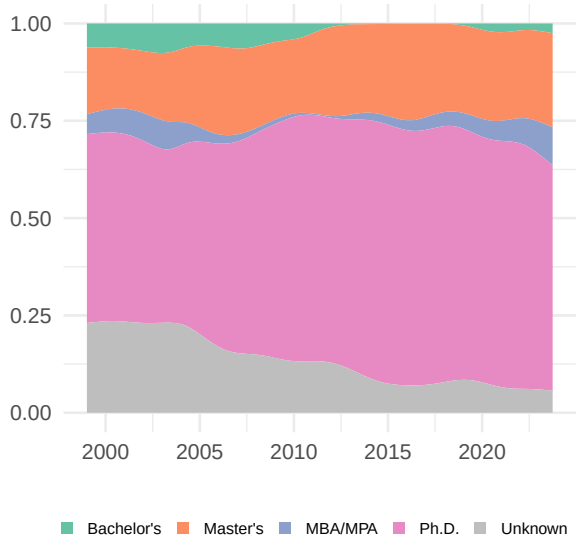
The figures show the characteristics of the AI SPF panel over time: (a) gender, (b) graduation year, (c) continent of origin, (d) public appearance; on next page: (e) education level, (f) education field, (g) affiliation type, and (h) job title.



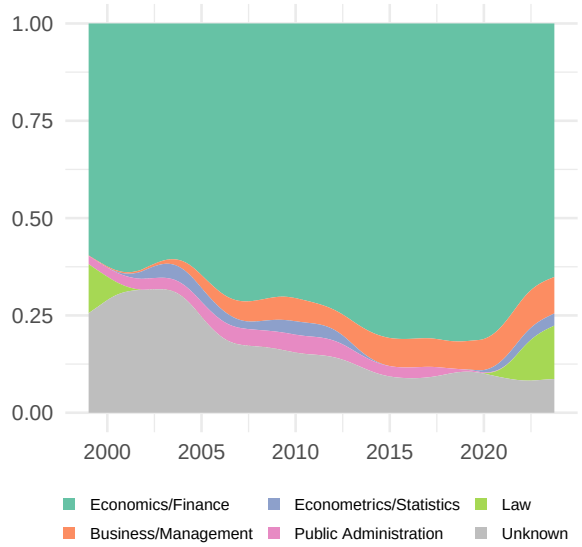
(Figure continues on next page)

Figure 2: Characteristics of the AI SPF panel over time *(continued)*

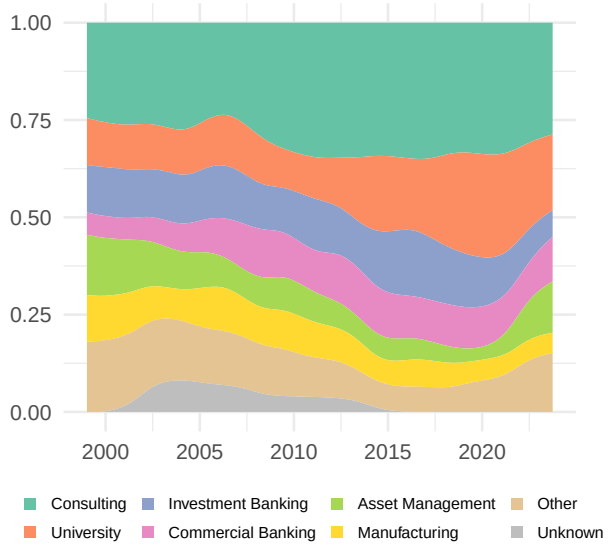
(e) Education level



(f) Education field



(g) Affiliation type



(h) Job title

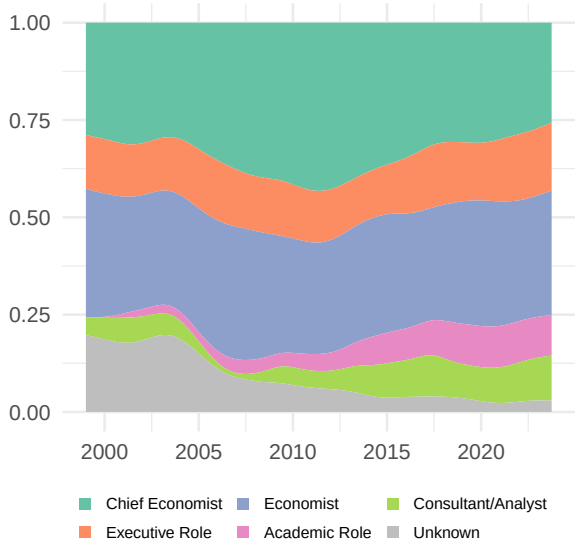
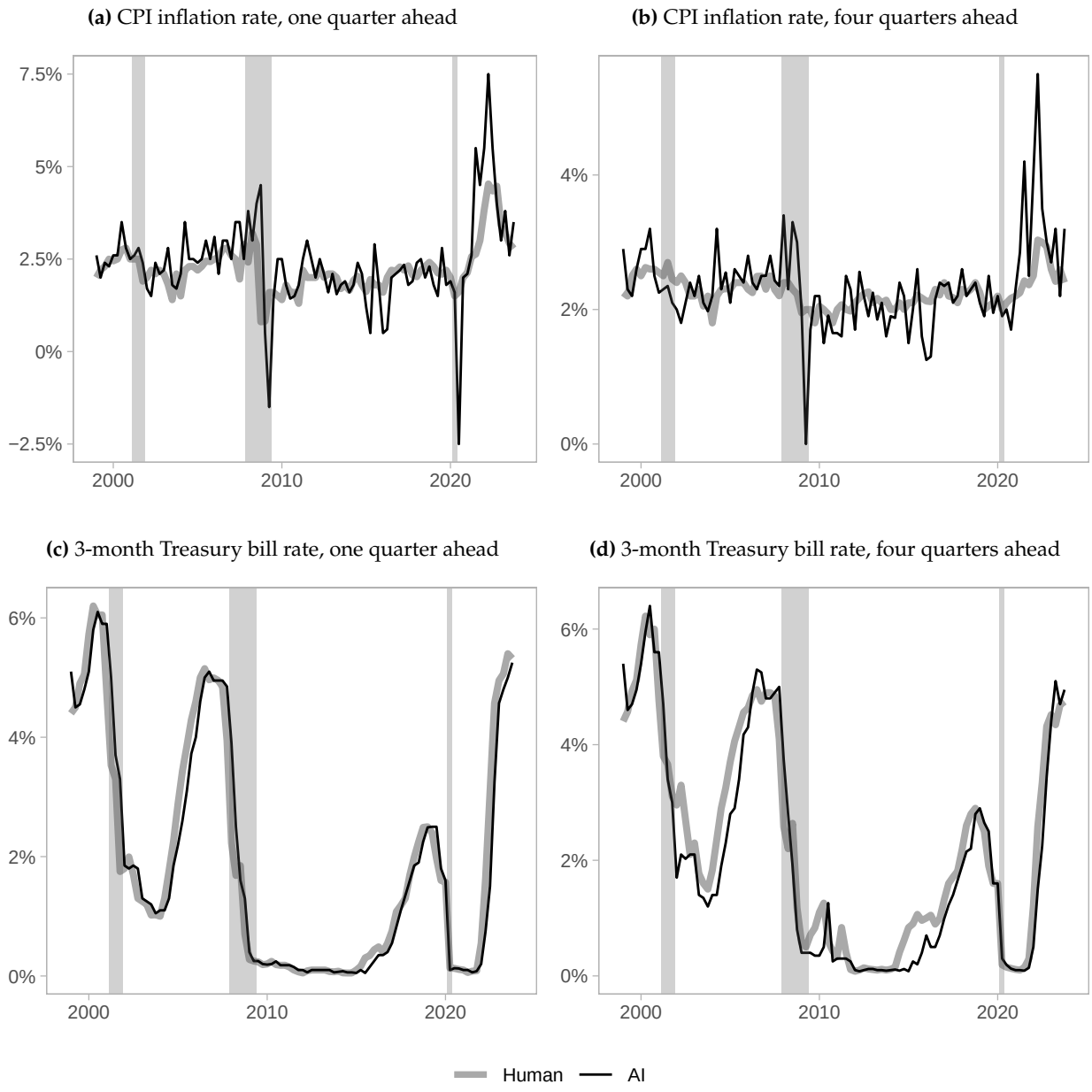


Figure 3: SPF and AI-generated median forecasts

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.

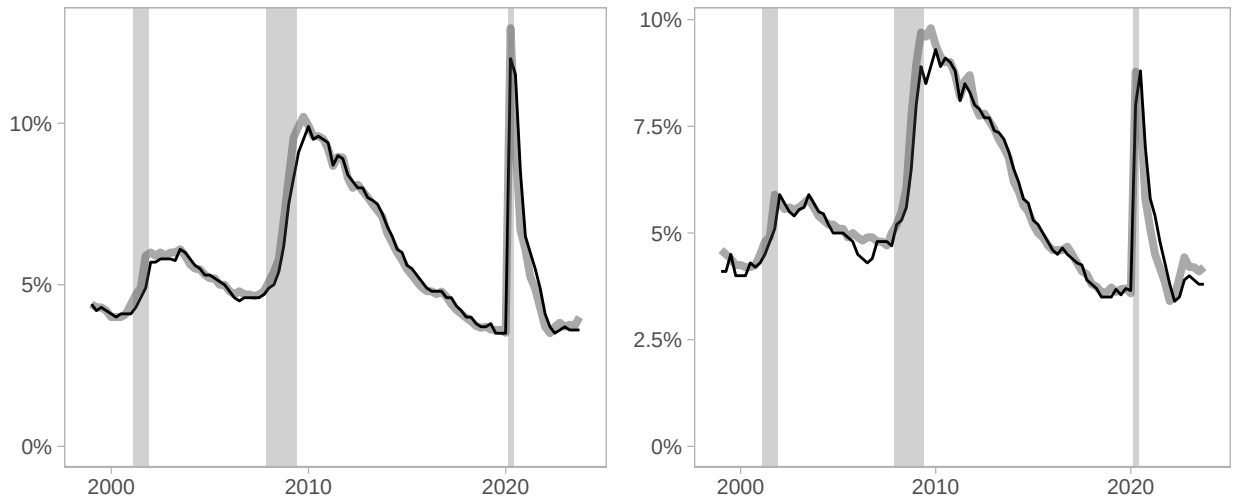


(Figure continues on next page)

Figure 3: 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

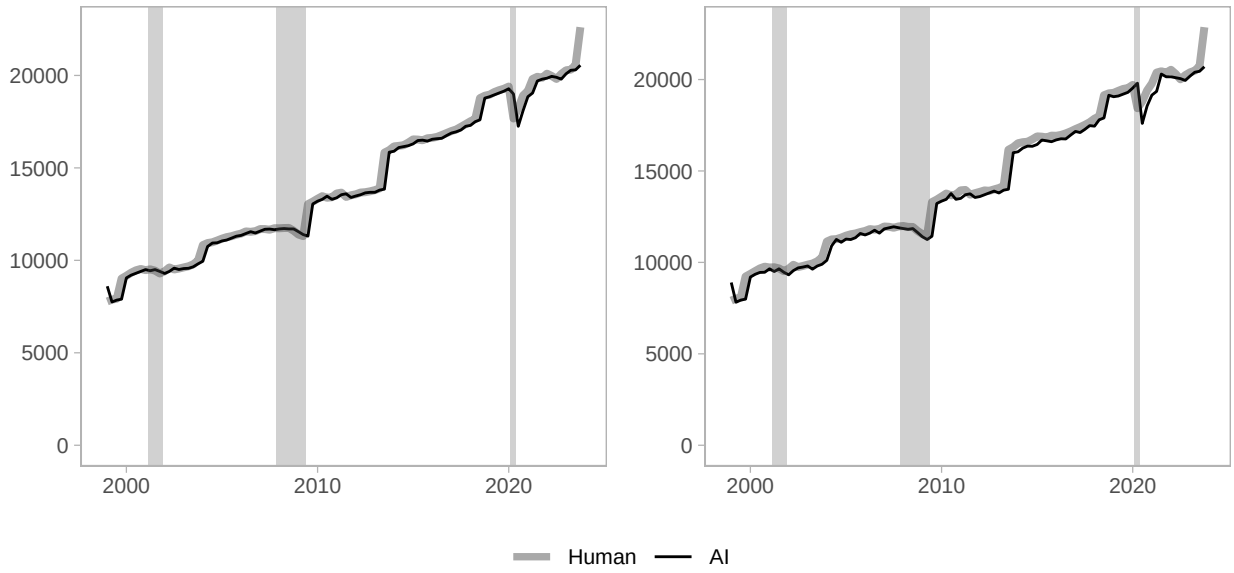
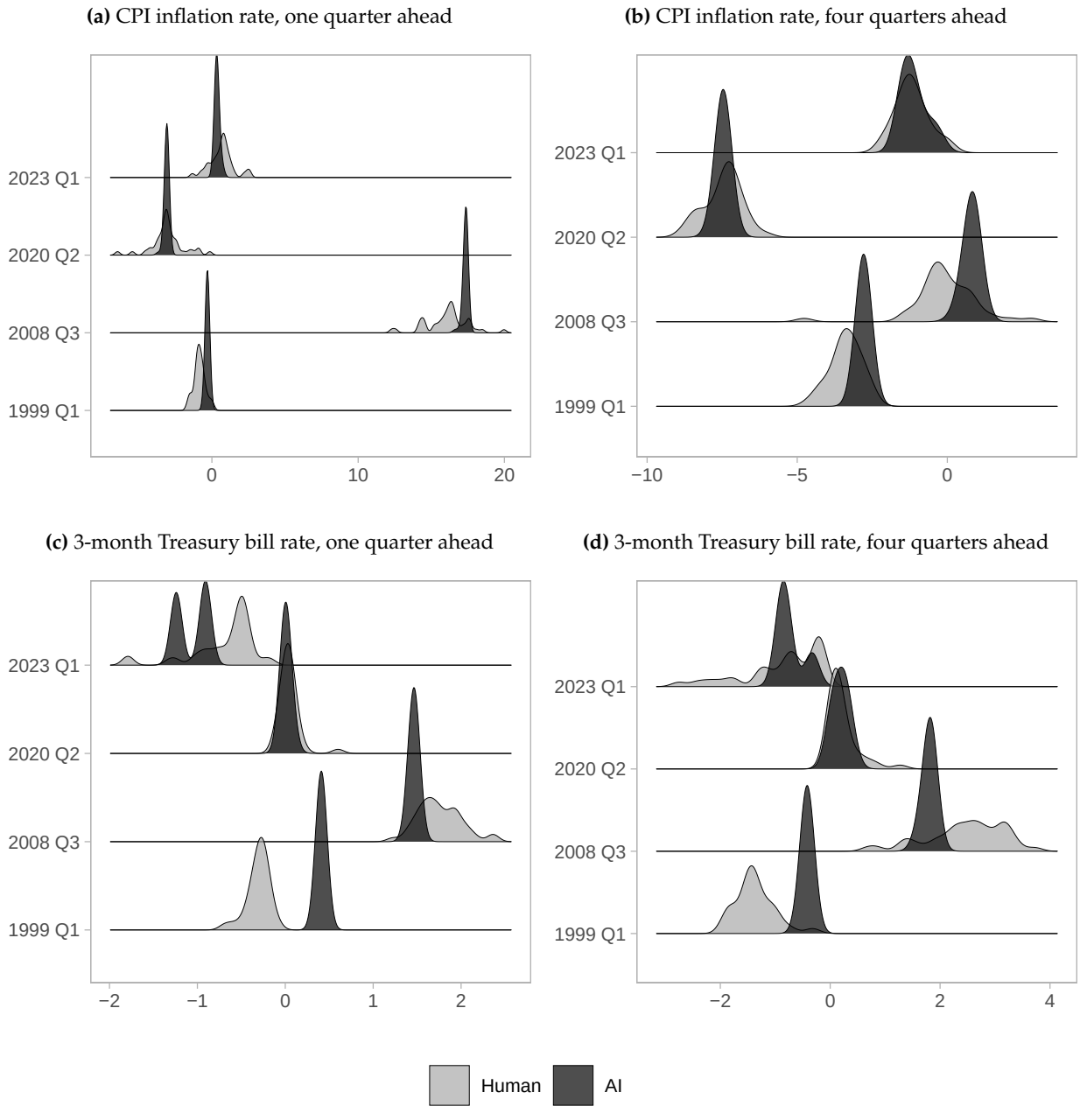


Figure 4: Densities of human and AI-generated SPF individual forecast errors

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 4: Densities of human and AI-generated SPF individual forecast errors *(continued)*

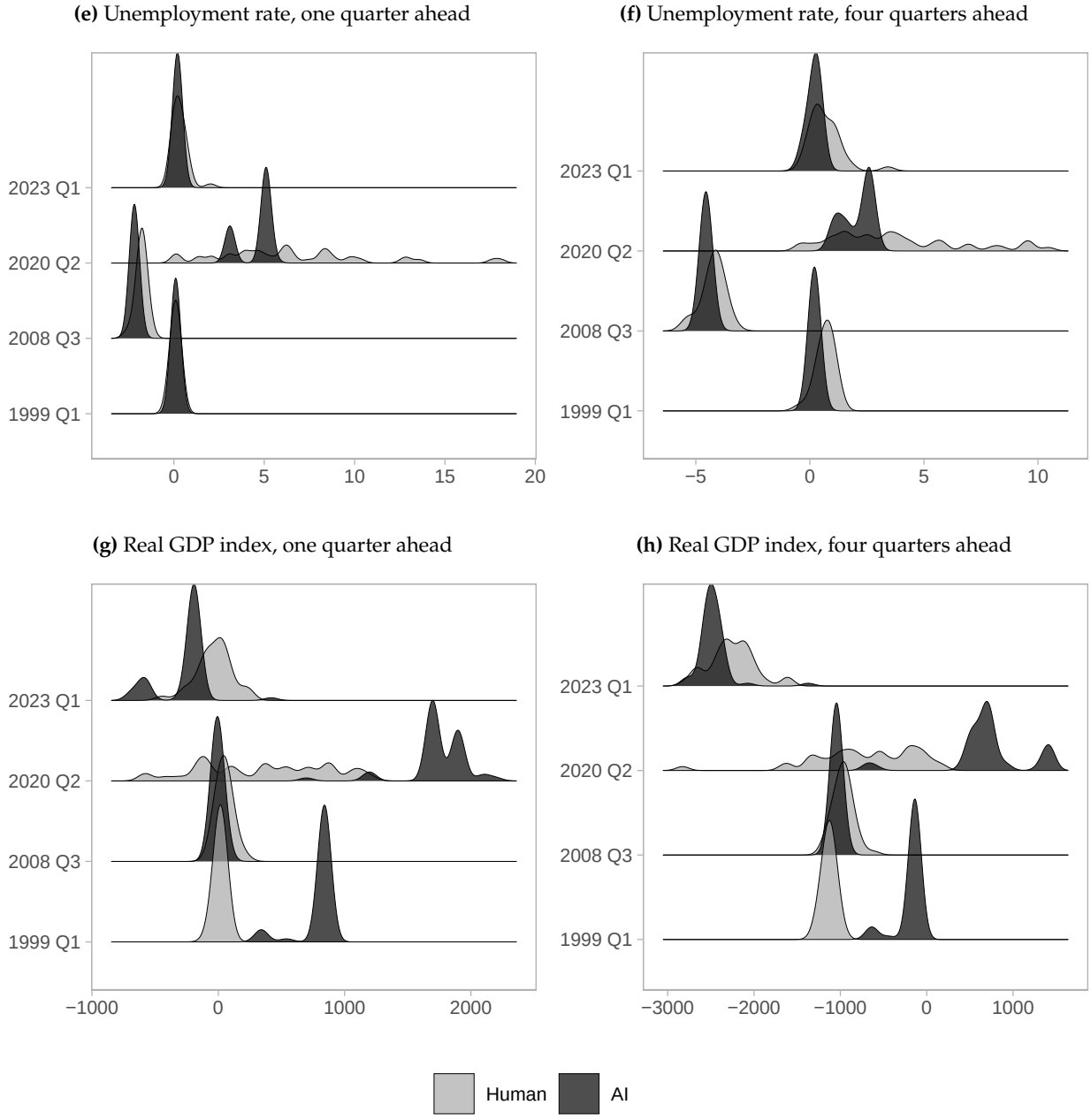


Figure 5: SPF and AI-generated median forecast errors around the LLM training data cut-off date

The figure shows the errors of one-quarter human and AI-generated SPF median forecasts for the (a) CPI inflation rate, (b) 3-month Treasury bill rate, (c) unemployment rate, and (d) real GDP index. The dashed line indicates October 2023, the training data cut-off date of the LLM (GPT-4o mini).

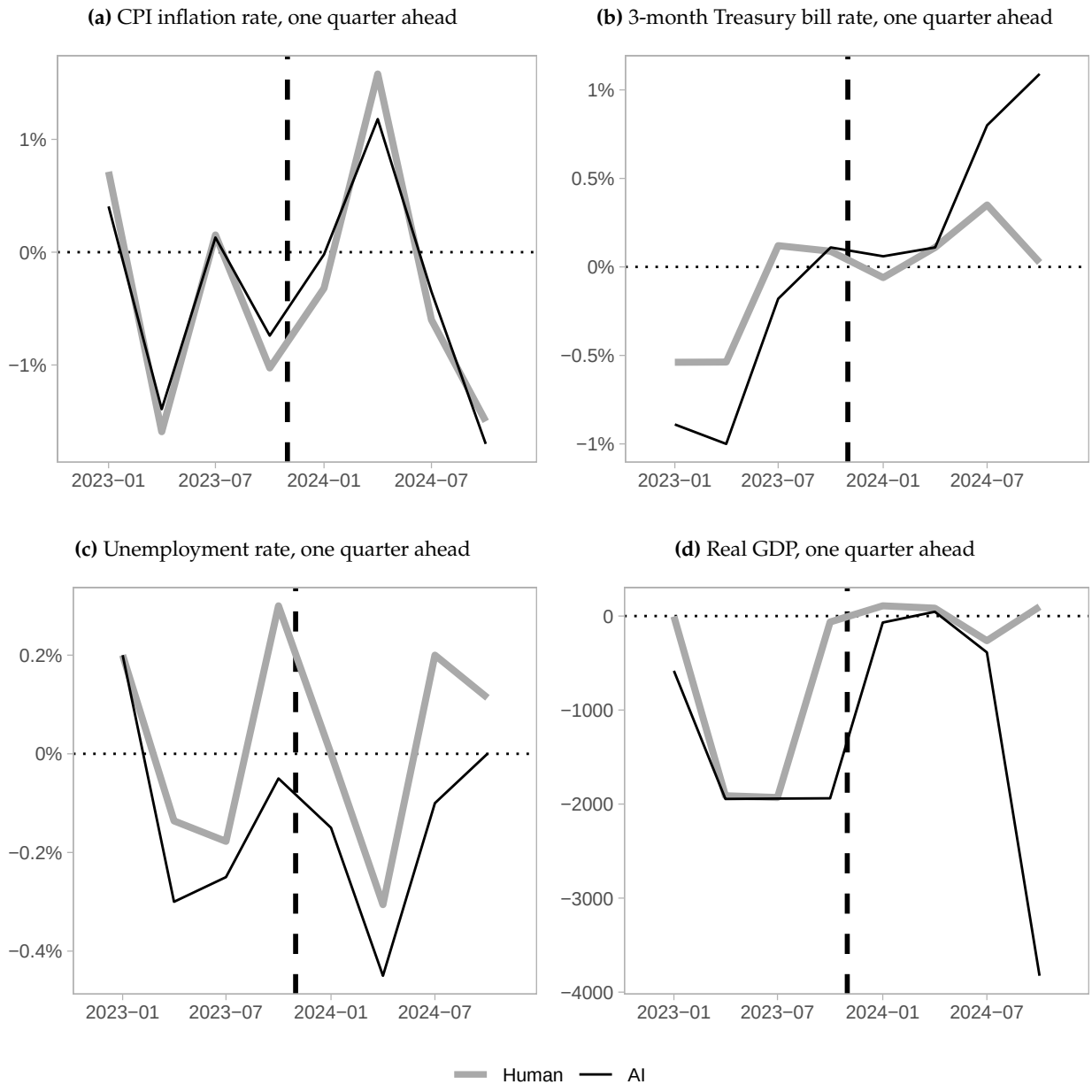
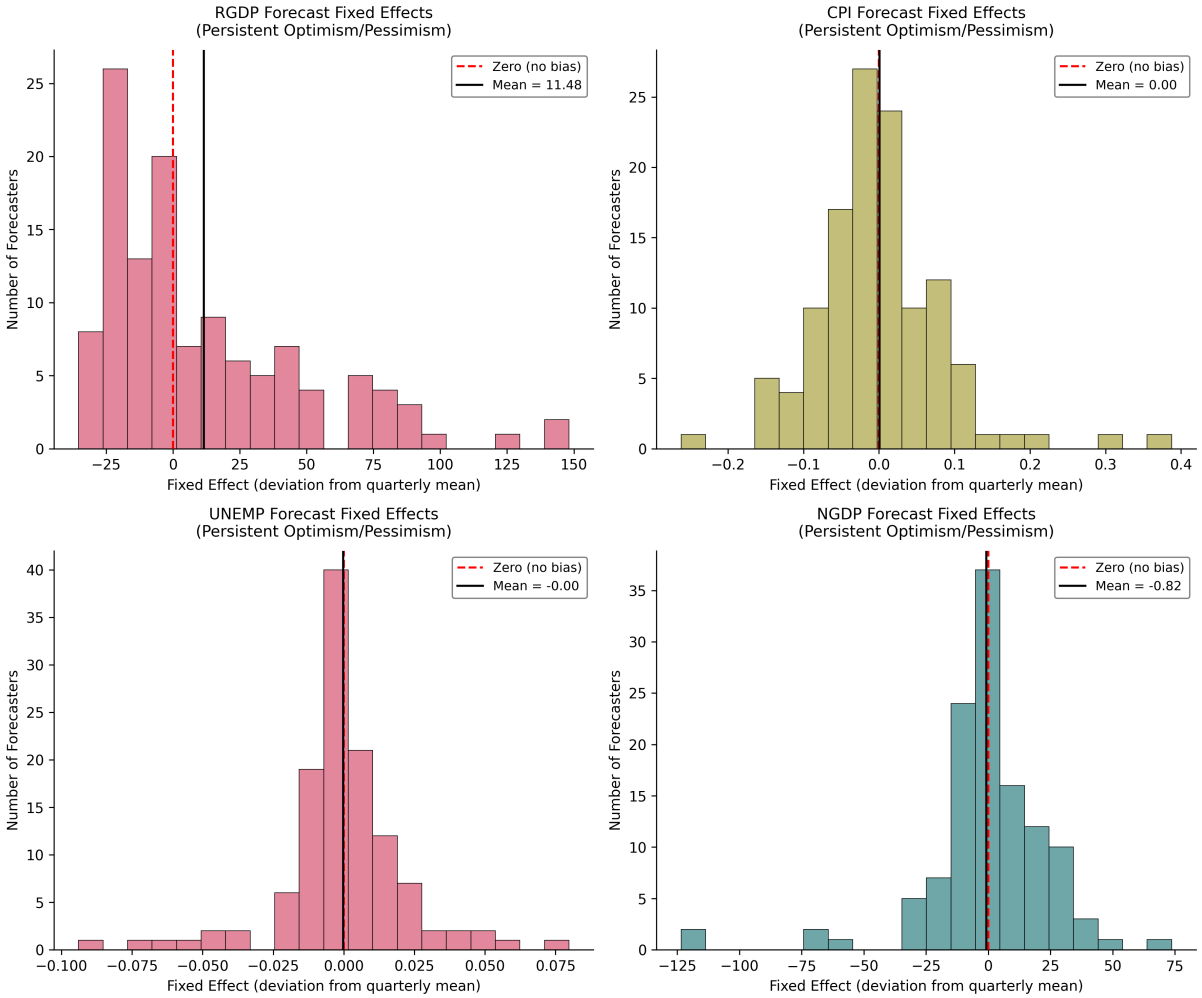
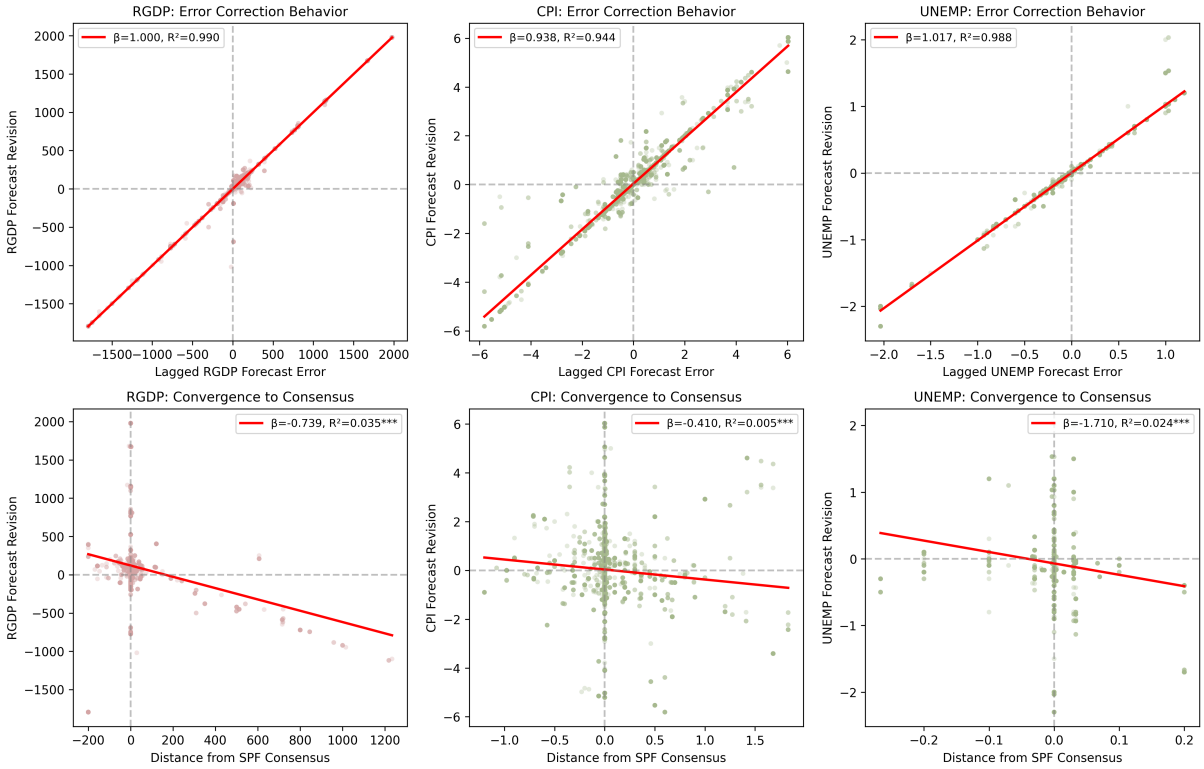


Figure 6: Distribution of Forecaster Fixed Effects



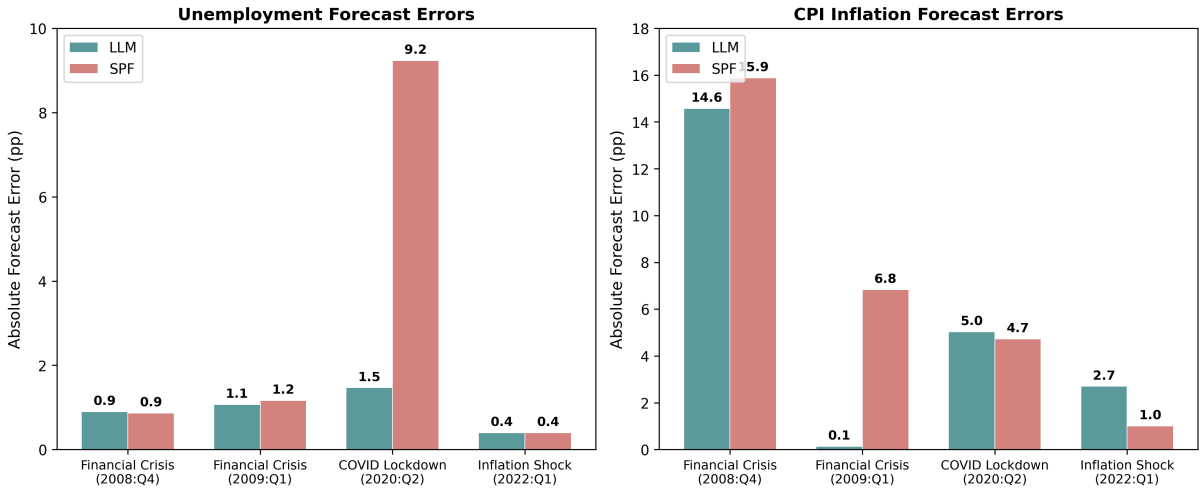
Notes: The figure shows the distribution of forecaster-specific fixed effects for four macroeconomic variables. Fixed effects are calculated as the average deviation of each forecaster's prediction from the quarterly cross-sectional mean. Each panel displays a histogram of fixed effects across 194 forecasters, with vertical lines indicating zero (red dashed) and the sample mean (black solid). RGDP and NGDP are measured in billions of 2012 dollars; CPI and UNEMP are measured in percentage points. Forecasters are restricted to those with at least 10 observations across the sample period (1999–2023).

Figure 8: Forecast Revision Dynamics



Notes: Panel A shows the relationship between forecast revisions (current forecast minus lagged forecast) and lagged forecast errors (lagged forecast minus realized value) for three variables. Each point represents a forecaster-quarter observation. The red lines show linear regression fits. Positive slopes indicate error correction: forecasters adjust future forecasts in response to past mistakes. Panel B shows the relationship between forecast revisions and the forecaster's distance from SPF consensus (lagged forecast minus lagged SPF median). Negative slopes indicate convergence toward consensus. Both panels restrict to consecutive forecast pairs (one quarter apart). Outliers beyond the 1st and 99th percentiles are excluded for visualization clarity.

Figure 9: Forecaster Quality and Consensus Spreads Across Crisis Episodes



Notes: The figure compares absolute forecast errors between LLM consensus (teal) and actual SPF consensus (coral) across four major crisis episodes. Left panel shows unemployment forecast errors in percentage points. Right panel shows CPI inflation forecast errors. The most striking result appears in the COVID lockdown quarter (2020:Q2): LLM forecasters predicted 14.5% unemployment (error: 1.5 pp) while professional SPF forecasters predicted only 3.8% (error: 9.2 pp); actual unemployment reached 13.0%. For CPI during the financial crisis trough (2009:Q1), the LLM achieved near-perfect accuracy (error: 0.1 pp) compared to the SPF's substantial overshoot (error: 6.8 pp). These results suggest that LLM-based forecasts may provide valuable signals about the magnitude of unprecedented shocks, particularly for labor market variables.

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	AAA corporate bond yield, levels, percentage points.
BOND BAA	BAA corporate bond yield, levels, percentage points.
TBOND	Treasury bond rate, 10-year, levels, percentage points.
Section 2: Real GDP and Its Components	
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)	5308	16.03	4.72	0.55	−0.44	15.01	8.68	27.62	0.90
GDP Price	5308	92.22	13.34	0.41	−0.37	91.40	70.60	122.85	0.90
Corporate Profits (x1000)	5308	1.38	0.57	0.52	−0.31	1.27	0.49	2.90	0.87
Unemployment Rate	5308	5.75	2.04	1.40	2.32	5.00	3.40	14.70	0.71
Non-Farm Payroll (x1000)	5308	138.61	8.04	0.72	−0.71	136.21	127.68	156.92	0.88
Industrial Production	5308	110.36	13.76	1.34	0.96	106.00	92.50	147.30	0.86
Housing Starts	5308	1.29	0.44	−0.13	−1.01	1.25	0.51	2.11	0.86
Treasury Bill Rate (3M)	5308	1.88	2.00	0.76	−0.91	1.15	0.02	6.29	0.83
AAA Corp Bond Yield	5308	4.83	1.34	0.25	−0.53	4.93	2.14	7.78	0.88
BAA Corp Bond Yield	5308	5.88	1.42	0.17	−0.79	6.01	3.24	8.88	0.85
Treasury Bond Rate (10Y)	5308	3.35	1.37	0.16	−0.79	3.39	0.62	6.66	0.84
Section 2: Real GDP and Its Components									
Real GDP (x1000)	5308	14.05	3.78	0.29	−1.06	13.33	7.68	22.49	0.90
Real PCE (x1000)	5308	9.81	2.62	0.27	−0.98	9.42	5.25	15.49	0.91
Real Non-Res Fixed Inv (x1000)	5308	1.82	0.65	0.68	−1.01	1.44	0.99	3.27	0.90
Real Res Fixed Inv	5308	493.98	122.31	0.10	−1.27	504.30	293.30	734.00	0.86
Real Federal C&GI (x1000)	5308	0.97	0.28	−0.11	−1.18	1.04	0.45	1.49	0.90
Real State/Local C&GI (x1000)	5308	1.52	0.37	0.19	−1.07	1.46	0.85	2.34	0.89
Real Change in Private Inv	5308	24.95	69.41	−1.12	3.69	34.50	−287.00	193.20	0.37
Real Net Exports	5308	−610.58	285.40	−1.43	1.27	−533.10	−1544.70	−250.00	0.86
Section 3: CPI and PCE Inflation									
CPI Inflation Rate	5308	2.52	3.14	−1.14	7.41	2.48	−13.39	10.73	0.00
Core CPI Inflation Rate	5308	2.33	1.49	2.21	8.36	2.19	−1.09	10.21	0.31
PCE Inflation Rate	5308	0.69	7.61	−4.05	16.31	2.14	−40.91	6.91	0.03
Core PCE Inflation Rate	5308	−0.21	2.95	−9.29	85.64	0.03	−28.31	2.00	−0.04

Table 4: 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 4 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	**44.39**	187.68	0.00***	**152.96**	178.49	0.00***	426.85	**381.28**	0.00***
GDP Price	**22.02**	22.07	0.30	**22.01**	22.12	0.41	**21.97**	22.27	0.62
Corporate Profits	**46.59**	71.52	0.24	**98.86**	101.25	0.56	**182.94**	186.90	0.70
Unemployment Rate	0.53	**0.38**	0.81	0.73	**0.59**	0.97	1.05	**0.94**	0.69
Non-Farm Payroll	984.75	**465.16**	0.06*	1487.62	**805.58**	0.21	2867.16	**1952.10**	0.27
Industrial Production	2.33	**1.64**	0.18	3.53	**2.98**	0.17	**5.90**	6.32	0.00***
Housing Starts	0.10	**0.09**	0.44	0.13	**0.12**	0.24	**0.21**	**0.21**	0.07*
Treasury Bill Rate (3M)	0.47	**0.26**	0.95	0.67	**0.44**	0.71	1.29	**1.21**	0.12
AAA Corp Bond Yield	0.41	**0.28**	0.99	0.48	**0.44**	0.63	**0.71**	0.73	0.14
BAA Corp Bond Yield	0.46	**0.30**	0.03**	0.56	**0.49**	0.02**	**0.80**	0.81	0.00***
Treasury Bond Rate (10Y)	0.47	**0.32**	0.85	0.57	**0.51**	0.70	**0.83**	0.88	0.03**
Section 2: Real GDP and Its Components									
Real GDP	173.30	**126.26**	0.00***	314.34	**209.62**	0.02**	650.31	**570.78**	0.00***
Real PCE	**61.41**	90.67	0.00***	166.62	**130.56**	0.01***	395.96	**332.31**	0.00***
Real Non-Res Fixed Inv	**20.51**	25.91	0.03**	54.78	**48.91**	0.06*	**131.69**	134.24	0.01***
Real Res Fixed Inv	**10.10**	10.49	0.78	25.86	**18.12**	0.76	59.55	**54.28**	0.76
Real Federal C&GI	12.63	**6.79**	0.84	29.70	**19.39**	0.65	57.52	**45.84**	0.50
Real State/Local C&GI	12.72	**8.10**	0.02**	31.43	**24.47**	0.15	78.72	**70.18**	0.12
Real Change in Private Inv	**14.26**	24.94	0.49	42.52	**38.14**	0.51	56.55	**49.10**	0.39
Real Net Exports	21.17	**16.55**	0.29	53.99	**43.31**	0.26	114.92	**91.89**	0.05*
Section 3: CPI and PCE Inflation									
CPI Inflation Rate	2.40	**1.99**	0.93	2.24	**2.14**	0.74	**2.04**	2.06	0.87
Core CPI Inflation Rate	0.90	**0.82**	0.59	**0.81**	0.87	0.30	1.02	**1.00**	0.37
PCE Inflation Rate	**1.93**	2.49	0.75	**2.64**	2.73	0.75	3.16	**3.14**	0.89
Core PCE Inflation Rate	2.43	**2.31**	0.80	2.37	**2.21**	0.94	2.19	**2.15**	0.66

Table 5: 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 4 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	**0.94**	0.00***	**0.70**	0.00***	**0.55**	0.32
GDP Price	**0.80**	0.00***	**0.86**	0.00***	**0.88**	0.00***
Corporate Profits	**0.64**	0.01***	**0.58**	0.13	**0.55**	0.32
Unemployment Rate	0.33	0.02**	0.37	0.10	0.41	0.17
Non-Farm Payroll	0.07	0.00***	0.10	0.00***	0.20	0.00***
Industrial Production	0.33	0.00***	0.41	0.07*	**0.60**	0.06*
Housing Starts	0.49	0.92	0.45	0.37	**0.51**	0.92
Treasury Bill Rate (3M)	0.24	0.00***	0.28	0.00***	**0.50**	1.00
AAA Corp Bond Yield	0.23	0.00***	0.46	0.49	**0.55**	0.42
BAA Corp Bond Yield	0.23	0.00***	0.39	0.11	**0.55**	0.46
Treasury Bond Rate (10Y)	0.34	0.00***	0.46	0.51	**0.57**	0.20
Section 2: Real GDP and Its Components						
Real GDP	**0.63**	0.01**	**0.55**	0.32	**0.50**	1.00
Real PCE	**0.86**	0.00***	**0.51**	0.92	0.34	0.00***
Real Non-Res Fixed Inv	**0.70**	0.00***	**0.52**	0.76	**0.56**	0.22
Real Res Fixed Inv	**0.69**	0.00***	0.39	0.04**	0.43	0.16
Real Federal C&GI	**0.76**	0.00***	0.42	0.11	0.36	0.01***
Real State/Local C&GI	**0.59**	0.09*	0.37	0.01**	0.38	0.02**
Real Change in Private Inv	**0.74**	0.00***	0.47	0.62	0.40	0.06*
Real Net Exports	0.49	0.92	0.44	0.27	0.29	0.00***
Section 3: CPI and PCE Inflation						
CPI Inflation Rate	0.45	0.38	0.48	0.80	**0.55**	0.54
Core CPI Inflation Rate	0.41	0.19	**0.62**	0.16	0.37	0.07*
PCE Inflation Rate	**0.75**	0.00***	**0.63**	0.08*	**0.57**	0.35
Core PCE Inflation Rate	**0.50**	1.00	0.44	0.46	**0.55**	0.53

Table 6: 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 4 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	**16.04**	115.74	0.00***	**49.00**	165.42	0.09*	**122.43**	304.78	0.02**
GDP Price	**4.88**	7.31	0.05**	**3.73**	6.13	0.04**	**0.65**	2.20	0.05**
Corporate Profits	**9.23**	55.20	0.02**	**19.67**	66.44	0.06*	**47.47**	125.77	0.03**
Unemployment Rate	**0.06**	**0.06**	0.65	**0.12**	0.18	0.22	**0.12**	0.44	0.10
Non-Farm Payroll	328.08	**85.00**	1.00	850.00	**45.00**	0.47	1971.00	**121.00**	0.87
Industrial Production	**0.32**	0.92	0.37	**0.58**	1.23	0.33	**1.50**	2.43	0.05**
Housing Starts	**0.03**	0.04	0.66	**0.05**	0.07	0.28	**0.03**	0.28	0.02**
Treasury Bill Rate (3M)	0.03	**0.02**	0.63	0.04	**0.03**	0.39	0.09	**0.07**	0.60
AAA Corp Bond Yield	0.21	**0.16**	0.96	**0.15**	0.21	0.79	0.24	**0.15**	0.59
BAA Corp Bond Yield	0.66	**0.27**	0.56	0.91	**0.39**	0.19	**0.25**	0.35	0.18
Treasury Bond Rate (10Y)	**0.09**	0.15	0.99	**0.12**	0.16	0.23	**0.17**	0.50	0.15
Section 2: Real GDP and Its Components									
Real GDP	**39.12**	103.89	0.04**	**56.80**	110.74	0.03**	**55.70**	272.73	0.01***
Real PCE	**5.80**	50.58	0.00***	**34.83**	53.36	0.14	93.62	**53.36**	0.35
Real Non-Res Fixed Inv	**6.39**	28.26	0.00***	**33.66**	41.69	0.57	**49.35**	100.42	0.02**
Real Res Fixed Inv	**1.10**	2.64	0.25	**7.89**	13.19	0.36	24.71	**23.87**	0.12
Real Federal C&GI	**1.60**	2.01	0.96	12.51	**6.03**	0.82	20.63	**13.27**	0.74
Real State/Local C&GI	**4.29**	5.85	0.00***	**6.35**	9.46	0.02**	**14.04**	17.78	0.05*
Real Change in Private Inv	**2.88**	21.49	0.03**	**8.03**	19.96	0.66	**17.38**	36.48	0.20
Real Net Exports	**8.54**	15.16	0.12	16.00	**10.80**	0.04**	18.48	**14.40**	0.92
Section 3: CPI and PCE Inflation									
CPI Inflation Rate	0.94	**0.87**	0.88	**0.46**	0.87	0.71	**0.32**	0.34	1.00
Core CPI Inflation Rate	0.68	**0.19**	0.87	0.70	**0.25**	0.88	**0.38**	**0.38**	0.74
PCE Inflation Rate	**0.21**	0.54	1.00	**0.35**	2.08	0.17	**0.33**	0.71	0.73
Core PCE Inflation Rate	1.21	**1.10**	0.52	1.28	**1.14**	0.34	0.70	**0.59**	0.42

Table 7: 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 accuracy of baseline AI forecasts (ratio). Forecast accuracy is computed as mean absolute errors, i.e., the average of absolute differences between realized values and the median human or AI-generated SPF forecasts across survey quarters. The "Generic" forecast uses a prompt that does not provide forecaster characteristics, but does 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	**2.21**	**1.02**	**3.92**	0.95	**6.32*****	**1.16*****	**89.03*****
GDP Price	0.97	**1.01**	0.97	**1.01**	0.72***	0.77***	**1.09*****
Corporate Profits	**1.74**	**1.01**	**2.84**	**1.17**	**6.20**	**2.25*****	**11.64**
Unemployment Rate	0.98*	**1.01**	**1.04*****	**1.04**	0.66	0.90***	0.61
Non-Farm Payroll	**1.22**	**1.01**	**1.59**	**1.35*****	**1.35***	**2.57*****	**10.91*****
Industrial Production	**1.05**	**1.00**	**1.36*****	**1.22**	**3.66*****	**1.18*****	**3.13**
Housing Starts	**1.04**	**1.04****	**1.24*****	**1.05**	**1.01**	**1.60*****	**10.49*****
Treasury Bill Rate (3M)	**1.03**	**1.01**	**1.06**	**1.03**	0.72	0.94***	0.68***
AAA Corp Bond Yield	**1.04**	**1.02**	**1.03**	**1.12**	**1.97*****	0.97***	**1.65*****
BAA Corp Bond Yield	**1.04**	**1.05**	**1.06**	**1.14**	**1.88**	**1.00**	**1.51*****
Treasury Bond Rate (10Y)	**1.01**	**1.01**	**1.00**	**1.10*****	0.71	**1.02**	0.55
Section 2: Real GDP and Its Components							
Real GDP	**1.39**	**1.03**	**1.44**	**1.00**	**5.96*****	**1.73*****	**20.15*****
Real PCE	**1.21**	0.98*	**2.33**	0.96	**12.06*****	**2.09*****	**41.67*****
Real Non-Res Fixed Inv	**1.02**	**1.01**	**2.38*****	**1.12**	**22.04*****	**4.29*****	**35.33*****
Real Res Fixed Inv	0.87***	**1.00**	**2.05**	**1.06**	**12.10*****	**3.66***	**19.08*****
Real Federal C&GI	**1.37**	0.95	**2.51****	**1.04**	**20.68*****	**5.15*****	**56.83*****
Real State/Local C&GI	**2.60**	0.95	**3.46**	**1.05**	**27.36*****	**4.72*****	**48.63*****
Real Change in Private Inv	0.98	0.96	**2.59*****	0.87	**3.73*****	**1.29*****	**7.27*****
Real Net Exports	**1.00**	0.95	**2.28**	0.93	**6.62*****	**1.39*****	**30.44****
Section 3: CPI and PCE Inflation							
CPI Inflation Rate	0.90***	**1.02**	0.91***	**1.02**	0.89***	**1.02*****	0.85*
Core CPI Inflation Rate	0.94	**1.05**	0.99	**1.08**	0.93	**1.05**	0.80
PCE Inflation Rate	**1.26**	**1.00**	**1.40**	0.91	**1.34**	**1.00**	**1.27**
Core PCE Inflation Rate	0.98	**1.01**	0.97**	**1.05**	0.98	**1.14*****	**1.05**
Average	**1.21**	**1.00**	**1.76**	**1.06**	**6.08**	**1.87**	**17.16**

Table 8: Cross-Sectional Disagreement by Forecast Horizon

Mean cross-sectional standard deviation at short (h+1) and long (h+4) horizons. Ratio = $\sigma_{h+4}/\sigma_{h+1}$. Positive t-statistics indicate greater disagreement at longer horizons. Paired t-tests (two-sided). Sample: 1999–2023. $^{***}p < 0.01$.

Panel	Variable	σ_{h+1}	σ_{h+4}	Ratio	t-stat	N
Human	CPI	0.972	0.701	0.72	-6.82^{***}	100
	UNEMP	0.155	0.356	2.29	20.98^{***}	100
	RGDP	40.16	99.90	2.49	14.54^{***}	100
	TBILL	0.126	0.363	2.88	18.11^{***}	100
AI	CPI	0.266	0.253	0.95	-0.64	94
	UNEMP	0.075	0.115	1.53	4.88^{***}	72
	RGDP	122.3	171.0	1.40	9.14^{***}	91
	TBILL	0.069	0.116	1.68	6.28^{***}	73

Table 9: Forecast revision regressions: Human vs AI

The table reports OLS regressions of forecast revisions on lagged forecast errors (Error Correction) and distance from SPF consensus (Consensus Tracking). Revision$_{it}$ is the change in forecaster i's forecast from $t-1$ to t. SPF regressions use individual forecaster microdata from the Philadelphia Fed (1999–2023). Robust standard errors in parentheses. $^{***}p < 0.01$.

	Error Correction				**Consensus Tracking**			
	RGDP	CPI	UNEMP	TBILL	RGDP	CPI	UNEMP	TBILL
Panel A: Human								
Lagged Error$_{i,t-1}$	0.981^{***}	0.881^{***}	0.956^{***}	0.975^{***}				
	(0.008)	(0.016)	(0.007)	(0.008)				
Distance from Consensus$_{i,t-1}$					-0.612^{***}	-0.690^{***}	-0.868^{***}	-0.296^{***}
					(0.134)	(0.035)	(0.064)	(0.102)
Observations	3,437	3,452	3,615	3,309	3,437	3,452	3,615	3,309
R^2	0.941	0.636	0.911	0.922	0.015	0.140	0.072	0.004
Panel B: AI								
Lagged Error$_{i,t-1}$	0.610^{***}	0.443^{***}	0.647^{***}	0.594^{***}				
	(0.028)	(0.018)	(0.008)	(0.008)				
Distance from Consensus$_{i,t-1}$					-1.016^{***}	-0.693^{***}	-0.510^{***}	-1.143^{***}
					(0.004)	(0.017)	(0.049)	(0.011)
Observations	5,188	5,188	5,188	5,188	5,188	5,188	5,188	5,188
R^2	0.568	0.354	0.946	0.827	0.776	0.551	0.086	0.716

Table 10: Counterfactual: Human vs AI consensus during major economic crises

The table compares LLM consensus forecasts (median of simulated panel) to actual SPF consensus (median of professional forecasters) during major economic crises. Bold values indicate lower absolute forecast error. The COVID lockdown result is striking: LLM forecasters predicted 14.5% unemployment while professional forecasters predicted only 3.8%; actual unemployment reached 13.0%.

	Unemployment (%)			CPI Inflation (%)		
Crisis	Actual	LLM	SPF	Actual	LLM	SPF
Financial Crisis (2008:Q4)	6.9	6.0	6.0	-9.2	5.4	6.7
Financial Crisis (2009:Q1)	8.1	7.0	6.9	-2.4	-2.5	-9.2
COVID Lockdown (2020:Q2)	13.0	14.5	3.8	-3.5	1.5	1.2
Inflation Shock (2022:Q1)	3.8	4.2	4.2	9.2	6.5	8.2
Absolute Forecast Error (pp)						
Financial Crisis (2008:Q4)	—	0.9	0.9	—	**14.6**	15.9
Financial Crisis (2009:Q1)	—	**1.1**	1.2	—	**0.1**	6.8
COVID Lockdown (2020:Q2)	—	**1.5**	9.2	—	5.0	**4.7**
Inflation Shock (2022:Q1)	—	0.4	0.4	—	2.7	**1.0**

Table 11: Reasoning Categories by Economic Period

Percentage of forecaster responses containing each reasoning type. Categories are non-exclusive. *Detailed categories*: Mechanical (trends, historical patterns, mean-reversion); Policy (Fed actions, fiscal/monetary policy); Model (economic theory such as Phillips curve, Okun's law); Event (specific shocks or crises). *ECB aggregation*: Following ECB methodology, Model-Based combines Mechanical ∪ Model; Judgment-Based combines Policy ∪ Event. All differences between crisis and normal periods are statistically significant ($p < 0.001$).

		Detailed Categories				**ECB Aggregation**	
Period	**N**	Mechanical	Policy	Model	Event	Model-Based	Judgment-Based
Normal	4,626	98.6%	55.2%	86.7%	19.7%	99.6%	63.3%
Financial Crisis	346	84.7%	34.4%	37.6%	82.4%	88.4%	87.6%
COVID-19	204	72.1%	59.8%	26.5%	77.0%	72.5%	79.9%
Inflation Shock	212	85.4%	87.3%	97.2%	30.2%	98.1%	87.7%

Prompts

Prompt 1: Baseline Prompt

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 [affiliation type] 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 [release_date], please provide your best numeric forecasts for the following variables: [variable_definitions].
Do this for the following quarters: t (current quarter), t+1, t+2, t+3, t+4, and annual forecasts for this and next year (annual averages).

You have the most recent real-time data on key macroeconomics variables available to you (from [prev_quarter]): [real_time_data].

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

Do not incorporate any data that was not available to you beyond the current date.
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.

For each variable, provide the forecasts as a sequence of numerical values in the following format:
[variable_name]: t, t+1, t+2, t+3, t+4, this year's avg, next year's avg

Please only provide your forecasts in this format.

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 input variables are highlighted in brackets and blue text and consist of real-time data, previous forecasts, survey dates, and forecasting variable names.

```
You are a professional economic forecaster participating in the Survey of Professional Forecasters.
You have expertise in macroeconomic analysis and forecasting. Use your professional judgment and the
data provided to make your forecasts.

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 [release_date], please provide your best numeric forecasts for:
[variable_definitions].

For quarters: t (current), t+1, t+2, t+3, t+4, and annual averages for this year and next year.

Real-time data available (from [prev_quarter]): [real_time_data].

Previous SPF panel forecasts ([prev_quarter], for t, t+1, t+2, t+3, t+4): [variable_forecasts_text].

Do not incorporate any data beyond [release_date]. Consider all relevant economic conditions and
Federal Reserve actions up to [release_date].

Format your response as: [variable_name]: t, t+1, t+2, t+3, t+4, this year's avg, next year's avg

Provide only numerical values.
```

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 professional economic forecaster participating in the Survey of Professional Forecasters.
You have expertise in macroeconomic analysis and forecasting. Use your professional judgment and the
data provided to make your forecasts.

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 [release_date], please provide your best numeric forecasts for:
[variable_definitions].

For quarters: t (current), t+1, t+2, t+3, t+4, and annual averages for this year and next year.

Previous SPF panel forecasts ([prev_quarter]): [var_forecasts_text].

Do not incorporate any data beyond [release_date]. Consider all relevant economic conditions and
Federal Reserve actions up to [release_date].

Format your response as: [variable_name]: t, t+1, t+2, t+3, t+4, this year's avg, next year's avg

Provide only numbers.
```

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 professional economic forecaster participating in the Survey of Professional Forecasters.
You have expertise in macroeconomic analysis and forecasting. Use your professional judgment and the
data provided to make your forecasts.

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 [release_date], please provide your best numeric forecasts for:
[variable_definitions].

For quarters: t (current), t+1, t+2, t+3, t+4, and annual averages for this year and next year.

Do not incorporate any data beyond [release_date]. Consider all relevant economic conditions and
Federal Reserve actions up to [release_date].

Format your response as: [variable_name]: t, t+1, t+2, t+3, t+4, this year's avg, next year's avg.

Provide only numbers.
```

Prompt 5: Prompt for recall task

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

```
You are a professional forecaster participating in the Survey of Professional Forecasters (SPF).

Please provide your best guess for the requested variables.

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

Online Appendix for

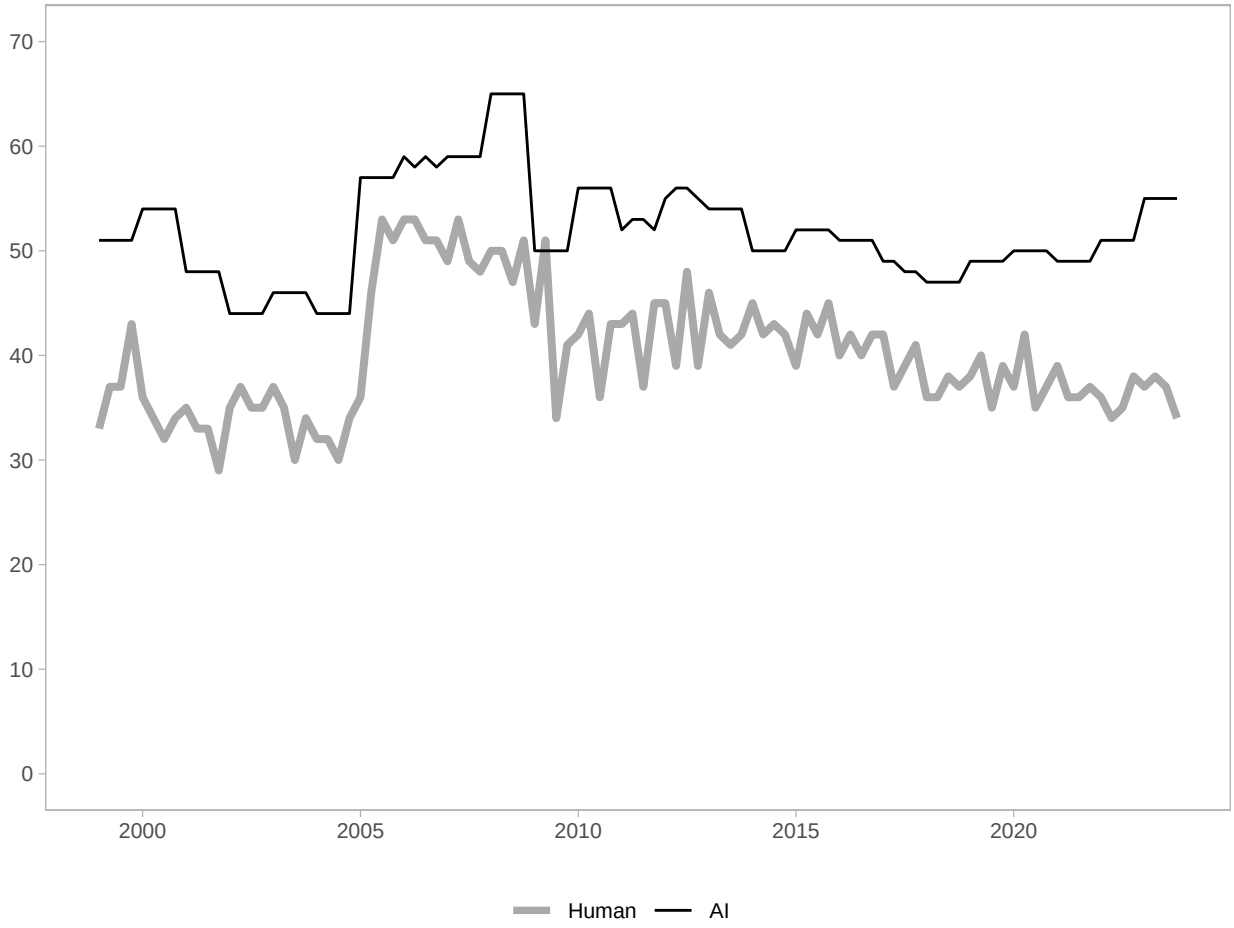
"Simulating the Survey of Professional Forecasters"

This Internet Appendix provides additional material on the data, methodology, and some robustness exercises related to the main text.

A Descriptive Figures

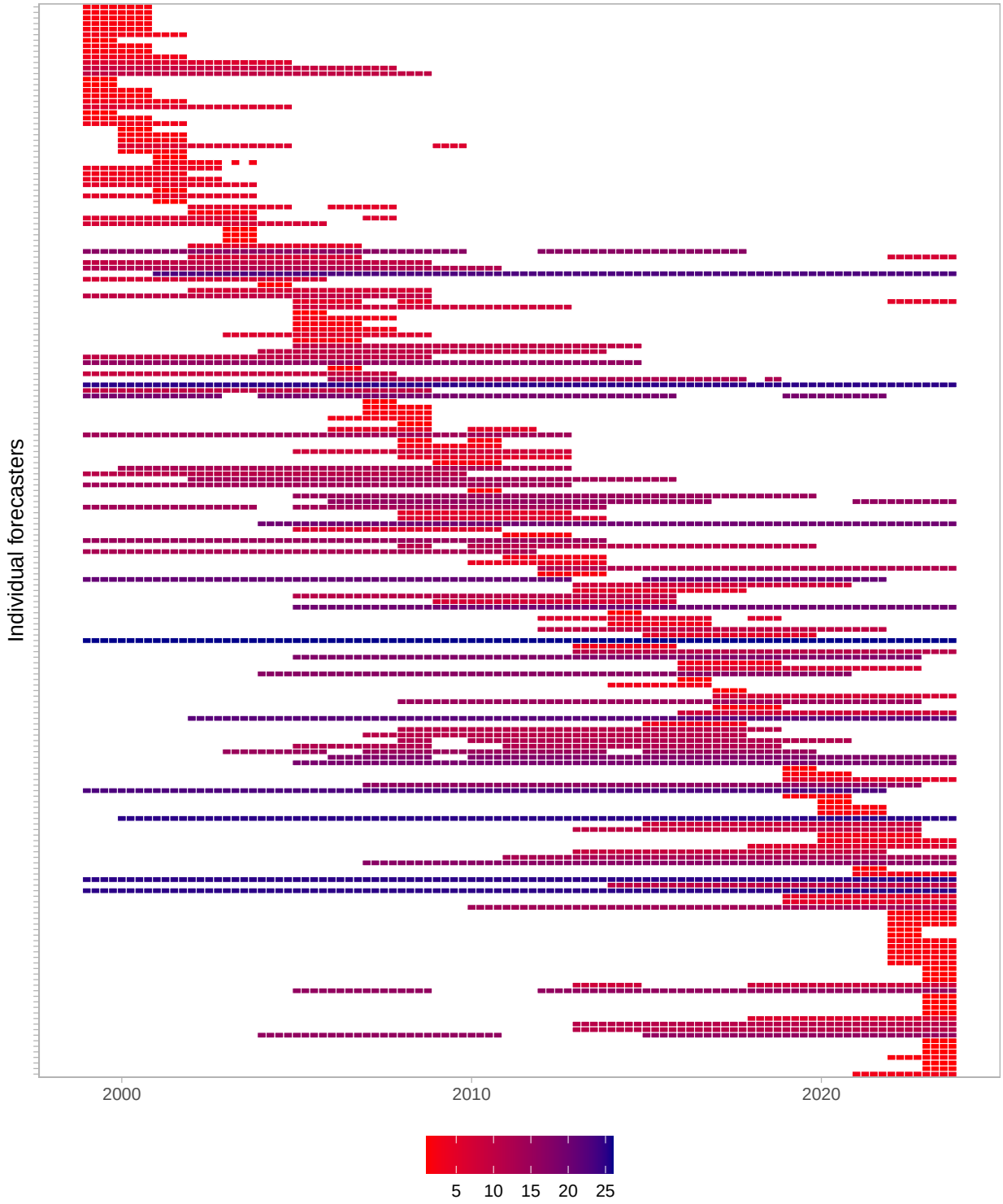
Prompt A.6: Number of forecasters in the SPF panel over time

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.



Prompt A.7: 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.



B Out of Sample Results

Tables B.12 and B.13 report MAEs for respectively median and best forecasts for the nowcasts and one- and two-quarter-ahead forecasts. At the time of writing, we cannot evaluate the performance of four-quarter forecasts due to lack of realized data values. 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. Overall, AI forecasters continue to perform well. 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.

Table B.12: Out-of-sample 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	725.96	**255.92**	0.03**	844.98	**157.99**	0.03**	789.59	**573.08**	0.09*
GDP Price	2.64	**2.62**	0.38	3.24	**3.13**	0.51	5.10	**5.06**	0.58
Corporate Profits	429.19	**225.32**	0.10	374.94	**335.66**	0.24	629.90	**569.84**	0.68
Unemployment Rate	0.17	**0.10**	0.63	0.19	**0.18**	0.01**	0.25	**0.17**	0.02**
Non-Farm Payroll	463.79	**334.69**	0.08*	1051.89	**530.01**	1.00	3477.57	**1220.57**	0.98
Industrial Production	0.62	**0.57**	0.32	**0.80**	0.81	0.08*	1.61	**0.76**	0.31
Housing Starts	0.07	**0.04**	0.01***	0.11	**0.06**	0.01***	0.22	**0.07**	0.01**
Treasury Bill Rate (3M)	0.45	**0.19**	0.58	0.53	**0.23**	0.75	0.56	**0.35**	0.02**
AAA Corp Bond Yield	0.47	**0.19**	0.33	**0.36**	0.40	0.35	**0.32**	0.34	0.92
BAA Corp Bond Yield	0.50	**0.20**	0.42	0.48	**0.35**	0.28	0.35	**0.29**	0.77
Treasury Bond Rate (10Y)	0.42	**0.26**	0.74	**0.36**	0.44	0.78	**0.31**	0.48	0.24
Section 2: Real GDP and Its Components									
Real GDP	1019.93	**289.94**	0.01***	1341.85	**557.81**	0.01***	1643.82	**1211.03**	0.01**
Real PCE	642.02	**173.32**	0.01***	789.10	**304.66**	0.01***	979.04	**686.76**	0.01**
Real Non-Res Fixed Inv	57.65	**40.80**	0.11	115.33	**78.73**	0.62	254.91	**235.59**	0.63
Real Res Fixed Inv	73.78	**25.20**	0.71	101.92	**53.79**	0.58	124.89	**98.07**	0.65
Real Federal C&GI	43.77	**14.23**	0.82	46.84	**22.36**	0.58	61.48	**47.24**	0.16
Real State/Local C&GI	135.51	**33.40**	0.01***	157.28	**58.33**	0.01***	269.73	**139.59**	0.01**
Real Change in Private Inv	30.23	**21.60**	0.44	47.05	**36.32**	0.21	53.89	**46.01**	0.73
Real Net Exports	167.51	**45.10**	0.27	263.69	**129.06**	0.26	**172.46**	195.93	0.48
Section 3: CPI and PCE Inflation									
CPI Inflation Rate	1.10	**1.07**	0.30	**0.74**	0.94	0.90	**0.65**	0.96	0.03**
Core CPI Inflation Rate	1.21	**0.72**	0.62	0.92	**0.71**	0.38	0.98	**0.82**	0.06*
PCE Inflation Rate	**0.68**	0.72	0.98	0.93	**0.91**	0.38	**0.61**	0.63	0.03**
Core PCE Inflation Rate	5.38	**5.23**	0.99	5.05	**4.89**	0.35	0.84	**0.65**	0.10

Table B.13: 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

C Reasoning Classification Details

C.1 Keyword Lists

Each LLM-generated reasoning response is classified using keyword-frequency rules. A response is tagged with a category if it contains at least one keyword from that category's list:

- **Mechanical:** "trend," "historical," "average," "mean reversion," "pattern," "previous," "past," "typical," "normal," "baseline," "extrapolat."
- **Policy:** "fed," "federal reserve," "monetary policy," "fiscal," "stimulus," "rate hike," "rate cut," "tightening," "easing," "central bank," "government," "policy response," "intervention."
- **Model:** "phillips curve," "okun," "gdp growth," "economic model," "structural," "relationship between," "economic theory," "multiplier," "elasticity," "economic indicator," "leading indicator."
- **Event-driven:** "covid," "pandemic," "crisis," "recession," "shock," "war," "collapse," "lockdown," "unprecedented," "disruption," "financial crisis," "lehman," "subprime," "outbreak," "emergency."

C.2 ECB-Style Aggregation

Following the ECB Survey of Professional Forecasters framework:

$$\text{Model-based} = \text{Model} \cup \text{Mechanical}$$

$$\text{Judgment-based} = \text{Policy} \cup \text{Event}$$

C.3 Crisis Period Definitions

- Financial Crisis: 2008 Q3 – 2009 Q4
- COVID-19 Pandemic: 2020 Q1 – 2020 Q4
- Inflation Shock: 2022 Q1 – 2022 Q4

All other quarters are classified as "Normal" periods.

D Robustness Checks

D.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 D.14 presents examples of personas generated by both methods:

Table D.14: 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.

D.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 D.1

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