Prompting for Policy: Forecasting Macroeconomic Scenarios with Synthetic LLM Personas

Giulia Iadisernia*
Banca d'Italia
Rome, Italy
giulia.iadisernia1@gmail.com

Carolina Camassa†
Banca d'Italia
Rome, Italy
Carolina.camassa@bancaditalia.it

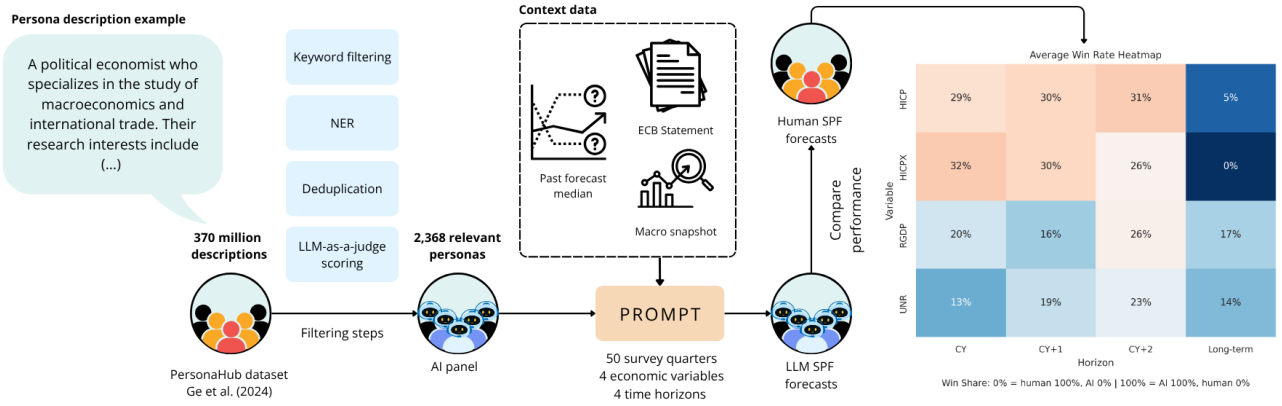


Figure 1: Experimental pipeline for evaluating synthetic LLM personas in macroeconomic forecasting. Starting from the PersonaHub corpus of 370M domain expert descriptions, we apply multi-stage filtering to extract 2,368 relevant personas. Each persona is then prompted to simulate responses to the ECB Survey of Professional Forecasters across 50 quarterly rounds (2013-2025), generating forecasts for four key macroeconomic variables (HICP inflation, core HICP, real GDP growth, unemployment) at multiple forecast horizons. The resulting 118,400 AI-generated forecasts are compared against human expert predictions to evaluate forecasting accuracy and the contribution of persona prompting to LLM performance in economic tasks.

Abstract

We evaluate whether persona-based prompting improves Large Language Model (LLM) performance on macroeconomic forecasting tasks. Using 2,368 economics-related personas from the PersonaHub corpus, we prompt GPT-4o to replicate the ECB Survey of Professional Forecasters across 50 quarterly rounds (2013-2025). We compare the persona-prompted forecasts against the human experts panel, across four target variables (HICP, core HICP, GDP growth, unemployment) and four forecast horizons. We also compare the results against 100 baseline forecasts without persona descriptions to isolate its effect. We report two main findings. Firstly, GPT-4o and human forecasters achieve remarkably similar accuracy levels, with differences that are statistically significant yet practically modest. Our out-of-sample evaluation on 2024-2025 data demonstrates that GPT-4o can maintain competitive forecasting performance on unseen events, though with notable differences compared to the in-sample period. Secondly, our ablation experiment reveals no measurable forecasting advantage from persona descriptions, suggesting these prompt components can be omitted to reduce computational costs without sacrificing accuracy. Our results provide evidence that GPT-4o can achieve competitive forecasting accuracy even on out-of-sample macroeconomic events, if provided with relevant context data, while revealing that diverse prompts produce remarkably homogeneous forecasts compared to human panels.

CCS Concepts

• **Computing methodologies** → **Natural language processing**; • **Applied computing** → **Economics**.

Keywords

large language models, prompt engineering, monetary policy, central bank communication, financial forecasting

ACM Reference Format:
Giulia Iadisernia and Carolina Camassa. 2025. Prompting for Policy: Forecasting Macroeconomic Scenarios with Synthetic LLM Personas. In *6th*

*Work done during an internship at Banca d'Italia. The views and opinions expressed are those of the authors and do not necessarily reflect the official policy or position of Banca d'Italia.

†Corresponding author.

ICAIF '25, Singapore, Singapore

ACM ISBN 979-8-4007-2220-2/25/11
https://doi.org/10.1145/3768292.3770385

ACM International Conference on AI in Finance (ICAIF '25), November 15–18, 2025, Singapore, Singapore. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3768292.3770385

1 Introduction

Macroeconomic forecasting has become increasingly critical for central bank communication and the transmission of monetary policy. The European Central Bank Survey of Professional Forecasters (ECB-SPF) [1], conducted quarterly since 1999, represents one of the most systematic efforts to capture expert expectations of inflation, GDP growth, and unemployment in the euro area. Because these forecasts directly influence policy decisions and market expectations, producing accurate and consistent forecasts is essential for economic stability.

Large Language Models (LLMs) have emerged as promising tools for economic forecasting tasks, offering the potential to simulate expert judgment at scale. Yet, current applications face a key methodological limitation: most studies using LLMs to simulate economic forecasts [10] or model inflation expectations [19] typically rely on one or few handcrafted "expert prompts". Despite the increasing adoption of LLMs in economic research [15], and the fact that LLM output is highly sensitive to prompt content and even formatting [17], there is a lack of empirical evidence on how prompt design affects performance in economics tasks.

We build on this premise by producing the first systematic replication of the ECB Survey of Professional Forecasters using LLMs. Our main research question is: **Do sophisticated persona descriptions—detailed biographical prompts designed to simulate specific expert types—improve LLM performance in a macroeconomic forecasting task?** We answer the question by extracting 2,368 economics-related synthetic biographies from the Persona Hub corpus[1] [8]. We evaluate the performance of these personas on 50 quarterly rounds (2013Q1-2025Q2) of the Survey of Professional Forecasters. This results in 118,400 AI-generated forecasts for four key macroeconomic variables at multiple forecast horizons. Our experimental design includes both *in-sample* evaluation (2013Q1-2023Q4) and *out-of-sample* testing on 2024-2025 data, which falls outside the model's training data.

Our study makes three main contributions. First, we conduct the first systematic replication of the ECB Survey of Professional Forecasters using LLMs, extending previous US-focused studies to European monetary policy contexts. Second, we implement a large-scale experimental design evaluating the macroeconomic forecasting performance of over 2,000 LLM-based synthetic personas. We compare these forecasts both to realized economic outcomes and, perhaps more interestingly, to the forecasting patterns of a panel of human experts. Third, we conduct a controlled ablation experiment to isolate the specific contribution of persona descriptions. We observe two main results: while LLMs can achieve competitive forecasting performance alongside human experts, the sophisticated persona descriptions contribute negligible improvements over simple baseline prompts. This result has significant implications for practitioners, suggesting that computational resources may be better allocated to ensemble methods or model improvements rather than elaborate persona engineering. These insights contribute to the growing understanding of LLMs as "synthetic forecasters" while providing practical guidance for central banks and financial institutions considering AI-augmented forecasting systems. Our results suggest that effective LLM-based forecasting may depend more on robust data integration and model architecture than on prompt engineering.

The remainder of this paper is organized as follows. Section 2 reviews related work on LLM forecasting and prompt engineering. Section 3 describes our data and experimental setup, including the ECB Survey of Professional Forecasters, and the persona dataset. Section 3.4 outlines our evaluation methodology. Section 4 presents results on the effectiveness of persona prompting, forecasting accuracy and human vs AI panel performance. Section 5 discusses future work, and Section 6 concludes.

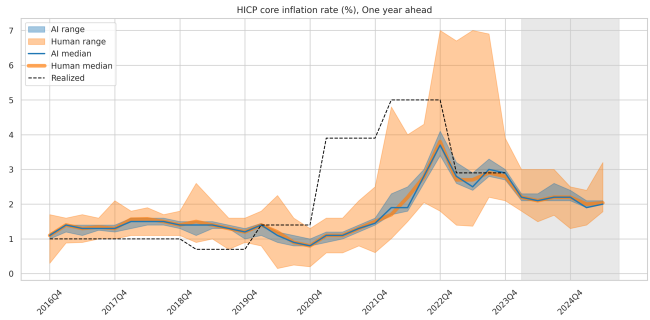


Figure 2: AI and human forecasters achieve remarkably similar accuracy across key macroeconomic variables. Time series comparison of realized outcomes (black), human expert forecasts from the ECB Survey of Professional Forecasters (orange), and AI forecasts using 2,368 synthetic personas (blue) for euro area core inflation (2016-2025). Despite using a variety of persona descriptions, LLM predictions converge to a much narrower forecast distribution compared to the human experts.

2 Related work

Macroeconomic forecasting with Large Language Models. The application of large language models to macroeconomic forecasting has emerged as a significant research area at the intersection of artificial intelligence and economics. Recent studies have explored direct applications of LLMs to a variety of forecasting tasks. Several studies share our focus on macroeconomic variables: Carriero et al. [4] examined LLM performance on macroeconomic time series, Bybee [3] fed Wall Street Journal articles to an LLM to predict financial and macroeconomic variables, while Faria-e Castro and Leibovici [6] demonstrated that Google's PaLM model could generate competitive inflation forecasts. Our work extends these existing studies primarily through our rigorous focus on persona prompting and the deployment of a panel of more than 2,000 "synthetic forecasters", which enables systematic comparison against both expert human panels and realized macroeconomic outcomes. Beyond evaluating LLM performance in specific forecasting applications, ongoing research has investigated general methodological

[1] https://huggingface.co/datasets/proj-persona/PersonaHub.

Table 1: The ECB Survey of Professional Forecasters: data description

(a) Main macroeconomic variables included in the ECB's Survey of Professional Forecasters

Variable	Definition
HICP	Harmonized Index of Consumer Prices (inflation)
HICPX*	Core HICP (excl. energy, food, alcohol, tobacco)
rGDP	Real GDP growth rate (annual %)
UNR	Unemployment rate (% of labor force)

*HICPX is included in SPF rounds since 2016Q4.

(b) Dataset overview: number of human forecasts from ECB data VS simulated forecasts with AI

Data source	SPF Rounds	Forecasters	Forecasts
Human SPF			
In-sample	44	56.2*	~2,473
Out-of-sample	6	58.4*	~350
AI personas (per model)			
With personas	50	2,368	118,400
No-persona baseline	50	100	5,000

*Average per round.

approaches to LLM-based forecasting. Lopez-Lira et al. [14] investigated memorization effects in LLM-based economic forecasts, while Tan et al. [18] compared language models to traditional time series methods. This methodological investigation connects to parallel efforts developing rigorous frameworks for LLM usage in economics research, which has become increasingly important as the field matures [13, 15].

Simulating surveys responses. A related strand of research examines LLMs' capacity to simulate human survey responses and expert judgment, similar to our comparison of human and AI panels in the Survey of Professional Forecasters. Zarifhonarvar [19] investigated inflation expectations formation using generative AI, finding that LLMs replicate key behavioral patterns including the tendency to predict higher inflation than realized rates. Horton [11] explored LLMs as "simulated economic agents", while Argyle et al. [2] demonstrated that language models can replicate human samples in political surveys. Geng et al. [9] and Fell [7] examined LLMs' ability to simulate social survey responses, though Dominguez-Olmedo et al. [5] identified important limitations regarding ordering and labeling biases in LLM survey responses. Most directly relevant to our investigation, Hansen et al. [10] conducted the first systematic replication of the U.S. Survey of Professional Forecasters using LLMs, demonstrating comparable accuracy between AI and human forecasts. Their approach employed manually crafted forecaster personas based on SPF participant characteristics, in contrast with our larger extraction of persona prompts from the PersonaHub dataset.

Persona prompting. A parallel line of research has examined the effects of prompt design and expert personas on LLM behavior. LLMs are highly sensitive to prompt content, structure, and formatting, as demonstrated by Sclar et al. [17], who show that even subtle variations in prompt phrasing can lead to large shifts in model output. This sensitivity complicates the interpretation of results in applied settings, where changes in tone, emphasis, or structure may unintentionally influence forecast quality. One strategy to improve LLM performance on reasoning tasks involves persona prompting or role-play. Kong et al. [12] introduces a role-play prompting method in which LLMs are instructed to assume the identity of domain experts. This approach improves zero-shot performance across multiple benchmarks and has inspired further exploration of expert simulation in applied domains, including economics. Persona prompting has also been used in macroeconomic forecasting experiments [10, 19]. Zarifhonarvar [19] prompts LLM to adopt distinct persona attributes, such as political orientation, background, and socioeconomic characteristics. This induces realistic behaviors like partisan bias in inflation expectations, which mirror behaviors observed in actual human surveys. Hansen et al. [10] construct detailed forecaster personas by manually gathering background data on SPF participants, including education, institutional affiliations, professional roles, and degrees. The study finds that removing these personas, i.e. replacing them with a generic forecaster prompt, leads to measurable drops in forecast accuracy, highlighting the value of role-based prompting. Interestingly, this is partially in contrast with our findings as presented in Section 4.

3 Data and experimental setup

3.1 Persona dataset

Starting from the Persona Hub corpus [8], which is publicly available and contains $\approx$ 370 million descriptions of domain experts (p_i), we implement a multi-stage filtering pipeline to extract only the items relevant to our study. The *Persona Hub* dataset includes highly heterogeneous personas—from lawyers to artists and policy analysts—thus requiring several layers of domain-specific filtering. The steps below were applied in the order presented:

(1) **Keyword search and domain filtering**: retain entries containing ≥ 2 tokens from a lexicon of terms related to monetary policy and the ECB (e.g., "central banking", "monetary policy", "Governing Council"). The dataset also includes four *domain* columns that were used as additional filtering dimensions. This step broadly excludes irrelevant personas while preserving those in macroeconomic or financial domains.
(2) **Name filtering**: using a named entity recognition (NER) algorithm, remove any description that directly mentions individuals (e.g., "Mario Draghi") to avoid role-play blurbs. After the first two steps, the dataset was reduced to $\approx 200,000$ personas.
(3) **Duplicate removal**: drop overly similar personas by computing vector embeddings[2] and discard those with cosine

[2] `sentence-transformers/all-mpnet-base-v2`.

Table 2: Examples of economics-related blurbs contained in the PersonaHub dataset, evaluated on the three dimensions relevant to our study: EU-centrality, neutrality and monetary policy depth. Only one meets all three criteria and was retained for the experiments.

Persona blurb (truncated)	EU centrality	Neutrality	Expertise
"A financial economist who specializes in the analysis of economic cycles and monetary policy. This person is interested in the degree of synchronisation of the euro area's economic cycle with that of the US, and how this affects the implementation of monetary policies. They are also interested in the factors that contribute to the degree of synchronisation and how they differ between the euro area and the US."	✓	✓	✓
"A global economist with Bank of America Merrill Lynch, with expertise in inflation and deflation, particularly in the context of the US and Europe. They are optimistic about the potential for economic growth in the US, but also recognize the potential for shocks that could trigger deflation."	✗	✗	✓
"A technologist who is skeptical of the effectiveness of information technology in stimulating economic growth. This persona believes that technology must be implemented and funded in order to generate economic growth. They also believe that the central bank's role in manipulating financial markets is a major impediment to economic growth."	✗	✗	✗

similarity scores ≥ 0.90, ensuring a diverse and representative subset of domain experts. This step reduced the dataset to $\approx 43,000$ personas.

(4) **Zero-shot relevance rating**: prompt an LLM (`gpt-4o-mini`) to evaluate each persona based on three independent binary criteria: *EU-centrality*, *neutrality*, and *monetary policy depth* (see Appendix A for the full prompt). Keep only those personas that satisfy all three (i.e., $score_{p_i} = 3$). For increased robustness, we run the evaluation three times with temperature `T = 1` and apply majority voting to determine the final decision. To validate the model's reliability, we randomly sampled 50 personas and had two human annotators rate them manually. Cohen's kappa scores between human ratings and GPT ratings ranged from 0.61 to 0.81 across the three criteria, indicating substantial agreement and confirming that the model selections are consistent with human judgment.

Given the size of the starting dataset, a multi-step pipeline is necessary to exclude candidate prompts that would pass the initial keyword-based screening but are ultimately unsuitable for our study. This strategy yields a candidate pool $P^\star$ containing **2368** biographies, representing a highly selective filter that retains approximately six personas per million from the initial dataset. Appendix B provides examples of economics-related persona descriptions from the PersonaHub dataset, illustrating how personas were evaluated and selected according to EU-centrality, neutrality, and monetary policy expertise.

3.2 ECB Survey of Professional Forecasters

The European Central Bank's *Survey of Professional Forecasters* (SPF) is a quarterly survey that collects forecasts from a panel of experts on key euro area macroeconomic indicators. These include HICP (Harmonised Index of Consumer Prices) inflation, real GDP growth, and the unemployment rate. The survey covers multiple forecasting horizons: the current calendar year, the next year, two years ahead, rolling horizons, and a five-year outlook. Respondents provide both point forecasts and probability distributions. For this study, we consider a total of 50 SPF rounds from 2013Q1 to 2025Q2 and only focus on four of the macroeconomic variables tracked by the survey (see Table 1 for an overview of the dataset).

Training data leakage. When trying to simulate responses to past events, such as past inflation, it is essential to consider potential data leakage issues which could lead to the "memorization problem" [14, 15]. We build an out-of-sample training set including SPF rounds from 2024-Q1 to 2025-Q2. This time period contains six editions of the ECB Survey of Professional Forecasters. These surveys occur after the training cutoff date for the GPT-4o model[3], which is the object of this study [16], ensuring the absence of data leakage.

3.3 Prompt architecture

The central element of each LLM prompt in this study is the persona description (p_i), which we evaluate for its impact on task performance. In addition to the persona, each prompt includes a set of standardized components necessary to perform the two experimental tasks. For each experimental call, the LLM receives:

- **System rules**: date anchoring (T_s), instructions for output formatting.
- **Persona blurb** $p_i \in P^\star$, unedited except for the controlled experiment settings—see below.
- **Monetary policy context**: the full text of the latest ECB press release at time T_s, plus a macro snapshot (π^{HICP}, gdp^{now}, ...).
- **Task instruction**:
  "Provide your forecasts for euro area HICP inflation, real GDP growth, and unemployment rate for the current quarter (t) the following time horizons t+1, t+2, t+3, t+4 (...) Format your response as numerical values: 'HICP (t): X.X%, HICP (t+1): X.X%, ...' etc.".

To isolate the effect of the persona selection on performance, we test two variations of the persona blurb: (i) *raw text*: unedited, full p_i; (ii) *empty persona*: the persona description block is omitted, providing

[3] Specifically, we prompt **gpt-4o-2024-11-20**.

Table 3: Intra-panel dispersion comparison between AI persona-based and human forecasters. Values represent the median dispersion, measured with standard deviation (SD) and inter-quartile range (IQR), across all survey rounds. AI forecasts consistently exhibit substantially lower dispersion than human forecasters across all variables and horizons, while human forecasters display broader disagreement patterns typical of professional survey panels.

(a) HICP and HICPX

Variable	Horizon	SD		IQR	
		AI	Human	AI	Human
HICP	CY	0.030	0.164	0.000	0.200
	CY+1	0.041	0.243	0.000	0.300
	CY+2	0.039	0.255	0.000	0.270
	LT	0.009	0.210	0.000	0.215
HICPX	CY	0.045	0.170	0.006	0.200
	CY+1	0.045	0.242	0.000	0.275
	CY+2	0.040	0.253	0.025	0.300
	LT	0.012	0.257	0.000	0.285

(b) rGDP and UNR

Variable	Horizon	SD		IQR	
		AI	Human	AI	Human
rGDP	CY	0.040	0.183	0.000	0.200
	CY+1	0.048	0.257	0.000	0.300
	CY+2	0.042	0.262	0.000	0.300
	LT	0.018	0.300	0.000	0.331
UNR	CY	0.047	0.194	0.050	0.177
	CY+1	0.046	0.296	0.000	0.300
	CY+2	0.043	0.417	0.000	0.419
	LT	0.024	0.582	0.000	0.700

only the monetary policy context and the task instructions. An example of the full prompt can be found in Appendix C.

3.4 Scoring metrics

Evaluation is carried out *match-by-match*, where a *match* is the unique combination of survey round r, macro variable $v \in \{\text{HICP, HICPX, rGDP, UNR}\}$, and forecast horizon $h \in \{\text{t0, t1, t2, lt}\}$. For each match we collapse the ∼2 000 persona completions of a given LLM to a single forecast—its cross-sectional *median*—and compare it with the published SPF-panel median. With the first real-time annual average that becomes available after the reference year closes, we form the absolute errors

$$e^{\text{AI}}_{rvh} = |\hat{y}^{\text{AI}}_{rvh} - y_{rvh}|, \qquad e^{\text{H}}_{rvh} = |\hat{y}^{\text{SPF}}_{rvh} - y_{rvh}|.$$

Point accuracy. For each variable and horizon, we measure forecast accuracy against the realized yearly data with the mean-absolute error (MAE)

$$\text{MAE}^{\text{panel}}_{vh} = \frac{1}{n_{vh}} \sum_{r=1}^{n_{vh}} e^{\text{panel}}_{rvh},$$

with n_{vh} being the number of available rounds. The score is reported in Table 4. The lower of the two numbers is bold-faced.

Panel disagreement. For each round we measure cross-sectional dispersion with the inter-quartile range $\text{IQR}_{rvh} = q_{75}(\hat{y}_{\bullet rvh}) - q_{25}(\hat{y}_{\bullet rvh})$, computed separately for personas and for human respondents. Table 3 reports the *median* IQR and variance across rounds.

Relative performance (win–share). For every match, we record $\text{win}_{rvh} = \mathbf{1}\{e^{\text{AI}}_{rvh} < e^{\text{H}}_{rvh}\}$, so that win = 1 denotes an AI victory over the SPF median. Let $w_{vh} = \sum_r \text{win}_{rvh}$ and n_{vh} be the number of matches for variable v and horizon h. The *win-share*

$$\bar{w}_{vh} = w_{vh}/n_{vh}$$

is reported alongside two p-values that address distinct questions:

- **One-tailed (H_0: Pr(win) = 0.5 vs. H_A: Pr(win) > 0.5).** Answers "is the AI panel *strictly better* than humans?".
- **Two-tailed** tests the symmetric alternative H_A: Pr(win) ≠ 0.5 and answers "is there *any* systematic difference in accuracy?".

The same procedure is applied both to the *panel-median* AI forecast and to every *individual persona*:

(i) **In-sample rounds** (2013Q1–2023Q4; up to $n_{vh} = 44$). We approximate the null distribution by Monte-Carlo: $N = 10{,}000$ artificial panels $W^*_j \sim \text{Binom}(n_{vh}, 0.5)$.

$$p^{(1)}_{vh} = N^{-1} \sum_j \mathbf{1}\{W^*_j \geq w_{vh}\}, \tag{1}$$

$$p^{(2)}_{vh} = 2 \min\left(p^{(1)}_{vh}, N^{-1} \sum_j \mathbf{1}\{W^*_j \leq w_{vh}\}\right). \tag{2}$$

With $N = 10{,}000$ the Monte-Carlo error never exceeds 0.005.

(ii) **Out-of-sample rounds** (2024Q1–2025Q2; $n_{vh} \leq 6$). We use the exact binomial:

$$p^{(1)}_{vh} = \Pr\{W \geq w_{vh}\}, \tag{3}$$

$$p^{(2)}_{vh} = 2 \min\{\Pr(W \geq w_{vh}), \Pr(W \leq w_{vh})\}, \tag{4}$$

where $W \sim \text{Binom}(n_{vh}, 0.5)$.

Stars in every table refer to the *one-tailed* p-value and mark $^{*}\, p \leq 0.10$, $^{**}\, p \leq 0.05$, $^{***}\, p \leq 0.01$.

4 Results

We report results obtained with GPT-4o[4] using temperature T = 1 for stochasticity. For out-of-sample survey rounds (2024Q1 to 2025Q2), we report accuracy and win-share results only for horizons where realized data is available: current year (CY) and next year (CY+1). For the HICPX variables, only 34 rounds are available as it was only introduced to the survey in 2016Q4.

[4]Preliminary experiments using GPT-4O-MINI and O3-MINI showed qualitatively similar behavior.

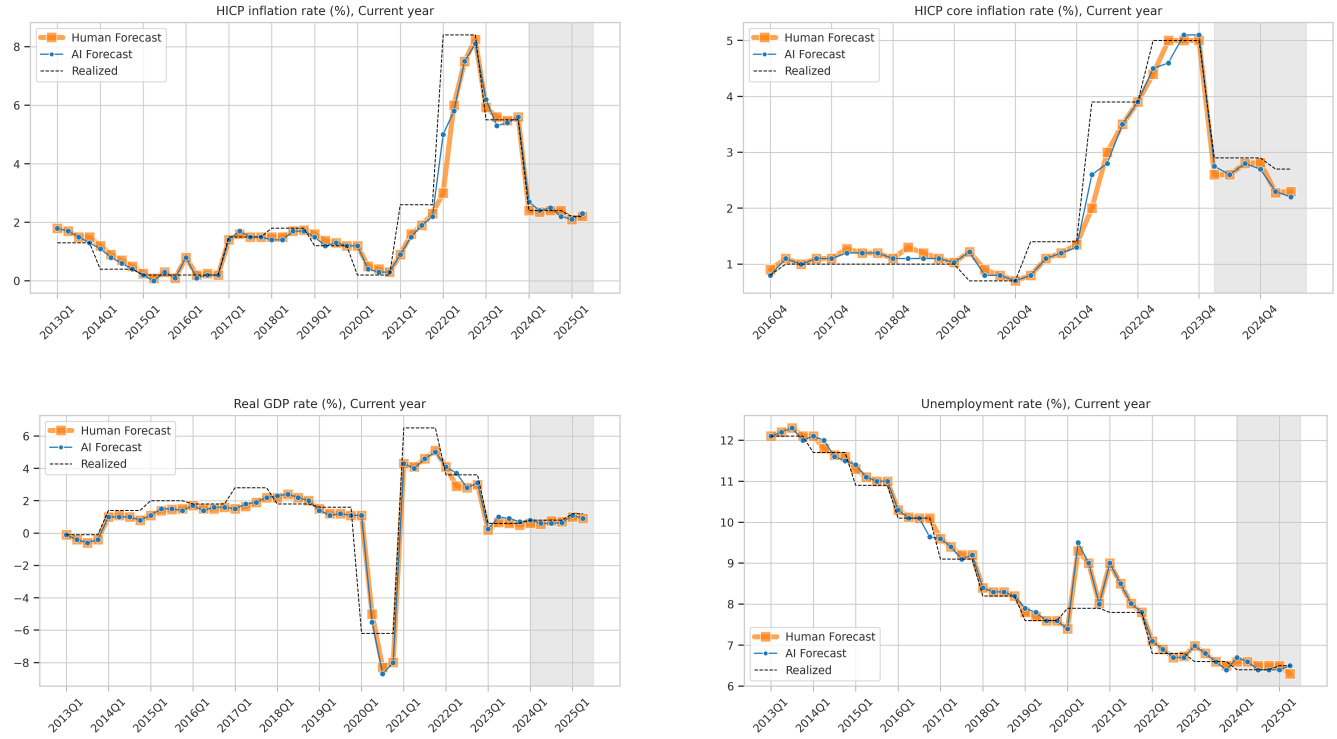


Figure 3: Comparison of AI persona-based and human forecasts for current-year horizon across four ECB-SPF variables (2013-2025): (a) HICP inflation, (b) HICP core inflation, (c) Real GDP growth, and (d) Unemployment rate. Gray shaded regions indicate out-of-sample evaluation period. AI-generated median forecasts often, but not always, match human forecasts; this occurs both in the in-sample and out-of-sample surveys.

Table 4: Forecast accuracy (MAE) comparing AI persona-based forecasts (median of 2,368 personas) versus human SPF medians across 50 ECB-SPF rounds (2013-2025). In-sample: 2013-2023, out-of-sample: 2024-2025. Bold indicates better performance. All errors in percentage points.

Horizon	In-sample								Out-of-sample			
	CY		CY+1		CY+2		Long-term		CY		CY+1	
	AI	Human	AI	Human	AI	Human	AI	Human	AI	Human	AI	Human
HICP	0.20	**0.19**	**0.95**	1.00	**1.02**	1.03	0.75	**0.74**	0.10	**0.01**	**0.13**	0.19
HICPX	**0.10**	**0.10**	**0.50**	**0.50**	**0.70**	0.75	**1.30**	**1.30**	**0.25**	0.30	**0.50**	**0.50**
rGDP	**0.50**	**0.50**	**0.50**	**0.50**	**0.60**	0.90	**0.85**	**0.85**	**0.17**	0.20	**0.20**	**0.20**
UNR	0.20	**0.10**	0.47	**0.42**	**0.70**	**0.70**	**1.00**	**1.00**	**0.05**	0.15	0.05	**0.02**

4.1 Persona ablation effect

Our primary methodological contribution examines whether detailed persona descriptions improve forecasting accuracy. We conduct a controlled ablation experiment comparing the results of our original prompt against 100 baselines in which we remove the persona description from the prompt. This results in a total of 5,000 forecasts which we can compare to the persona-enhanced ones. The results show no statistically significant difference in forecasting performance. A paired t-test on match-level median absolute errors yields a mean difference of 0.01 percentage points (personas minus no-personas), with $t = -1.02$, $p = 0.31$. This finding is corroborated by a size-matched Kolmogorov-Smirnov test ($D = 0.05$, $p = 0.28$) showing that error distributions are statistically indistinguishable (see Figure 4). This null result has significant practical implications: sophisticated persona engineering provides no measurable forecasting advantage and can be omitted to reduce computational costs without sacrificing accuracy. The finding suggests that model performance depends primarily on data quality and task framing rather than prompt elaboration.

Table 5: Win share of SPF and AI-generated forecasts (%). Win shares calculated as percentage of forecasting rounds where AI (or human) median strictly outperformed the other. Remaining percentage represents ties. *** $p \leq 0.01$

Horizon	CY			CY+1			CY+2			Long-term		
	AI wins	Human wins	P-val	AI wins	Human wins	P-val	AI wins	Human wins	P-val	AI wins	Human wins	P-val
In-sample												
HICP	29	**28**	***	30	**31**	***	**31**	26	***	5	**11**	***
HICPX	**31**	20	***	**30**	17	***	**26**	13	***	1	**5**	***
rGDP	20	**30**	***	16	**34**	***	26	**28**	***	**17**	9	***
UNR	13	**32**	***	**19**	17	***	**23**	15	***	**14**	9	***
Out-of-sample												
HICP	28	**68**	***	**70**	25	***						
HICPX	32	**40**	***	**25**	0	***						
rGDP	**53**	25	***	10	**50**	***						
UNR	**47**	37	***	3	**40**	***						

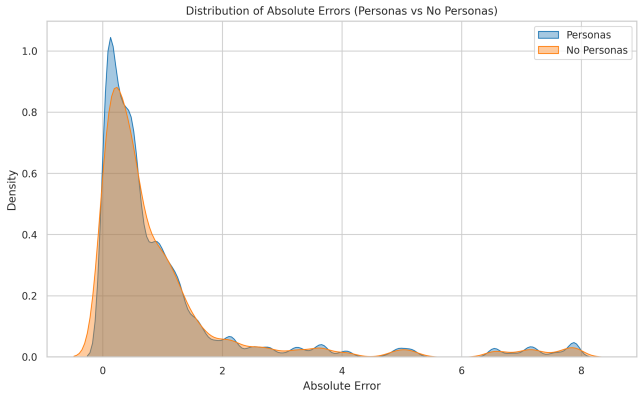


Figure 4: Persona prompting yields statistically indistinguishable error distributions. Kernel density estimates of absolute forecast errors for GPT-4o with persona descriptions (blue) versus baseline prompts without personas (orange) across all variable-horizon-round combinations. The near-perfect overlap supports our null hypothesis ($t = -1.02$, $p = 0.31$; Kolmogorov-Smirnov $D = 0.05$, $p = 0.28$).

4.2 Panel disagreement

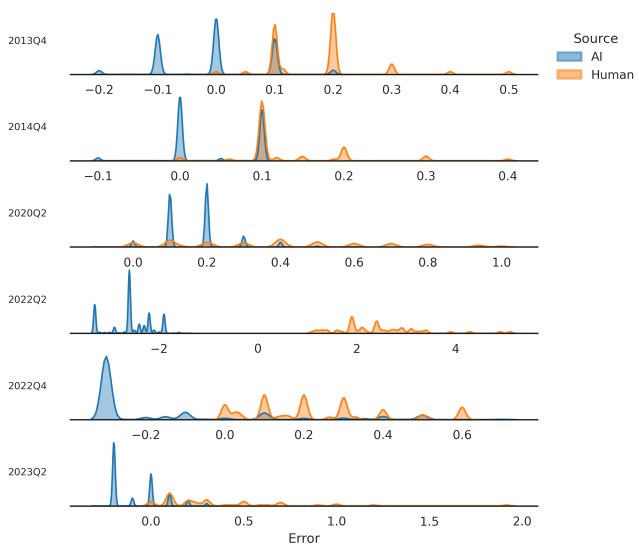
The dispersion of forecasts within each panel reveals a significant difference between the AI and human panels of forecasters, as shown in Table 3. AI personas exhibit near-zero disagreement, with median inter-quantile ranges (IQRs) mostly below 0.001 percentage points across all variables and horizons, roughly two orders of magnitude lower than human forecasters. Human forecasters, display substantially higher dispersion, with higher median IQRs ranging from 0.17 percentage points (UNR current-year) to slightly higher values at longer time horizons. The disparity is equally pronounced when measured by standard deviation, with AI personas exhibiting roughly one order of magnitude less variation than human forecasters across all variables and horizons. Both dispersion measures confirm that despite the variety of persona prompts, the model converges on mostly homogeneous forecasts, suggesting limited sensitivity to prompt variations in this task domain.

4.3 Point forecast accuracy

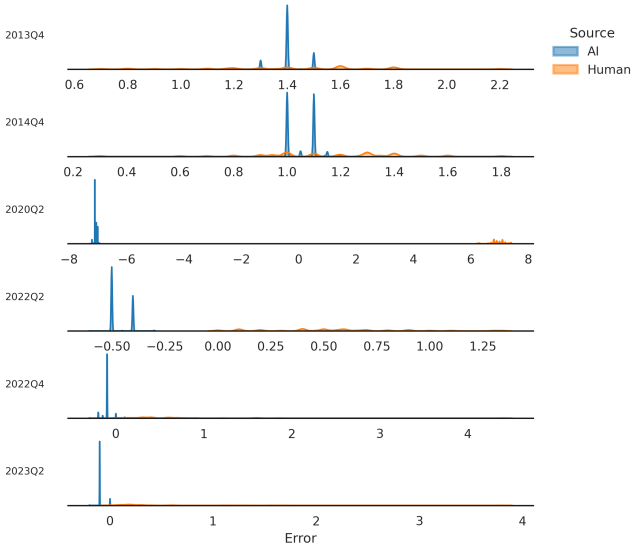
The mean absolute error results, reported in Table 4, show that AI and human forecasters often perform at remarkably similar levels, with identical errors observed in seven of sixteen in-sample comparisons and numerous cases where differences are minimal (within 0.05-0.10 percentage points). The largest performance gaps emerge in specific variable-horizon combinations: AI substantially outperforms humans for GDP growth at the CY+2 horizon (0.60 vs 0.90) and unemployment at current-year out-of-sample forecasts (0.05 vs 0.15), while humans show clear advantages for unemployment at current-year in-sample (0.10 vs 0.20) and HICP current-year out-of-sample (0.01 vs 0.10). The transition from in-sample to out-of-sample periods shows no systematic performance degradation. Despite the model forecasting economic conditions entirely absent from its training data (October 2023 cutoff), accuracy levels remain broadly comparable to the in-sample period, suggesting the model effectively utilizes the real-time economic context provided in prompts rather than relying purely on memorized patterns.

4.4 Win-share analysis

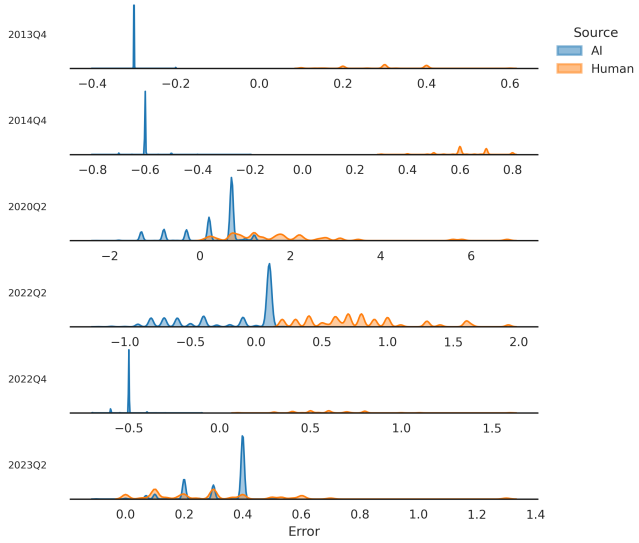
In addition to evaluating point accuracy, we compute win-share scores to compare AI and human performance net of ties. The aggregated results are shown in Table 5, both for in-sample and out-of-sample rounds. Appendix D additionally reports the win shares for each survey round by horizon and variable. The results demonstrate statistically significant yet practically modest differences in forecasting accuracy, with performance patterns varying systematically across variables and horizons. Despite uniform statistical significance at the 1% level, many win rate differentials are relatively narrow—particularly for inflation forecasts where margins often fall within 1-5 percentage points. The data reveal variable-specific comparative advantages: AI consistently outperforms on core inflation (HICPX) across most horizons, while humans maintain advantages in short-term GDP and unemployment forecasting



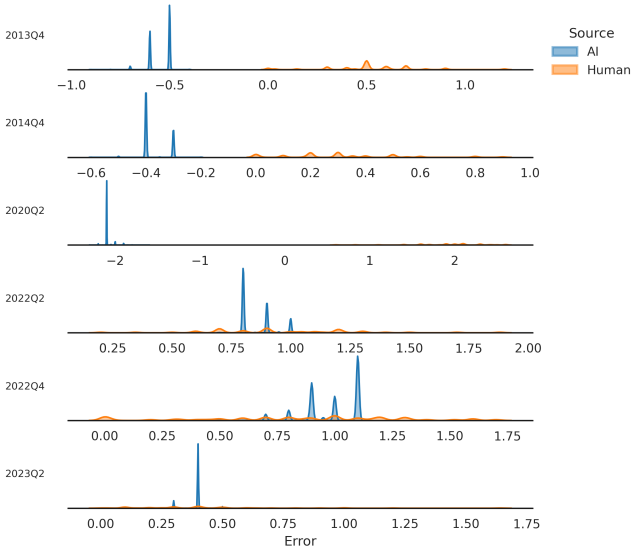
(a) HICP inflation, Current year



(b) HICP inflation, Two years ahead



(c) Real GDP growth, Current year



(d) Real GDP growth, Two years ahead

Figure 5: Distribution of forecast errors by variable and horizon. Each panel shows kernel density estimates of errors for AI forecasts (blue) and human SPF forecasts (orange) across selected survey rounds. Top row compares HICP inflation errors at current-year (t0) and two-years (t2) horizons. Bottom row shows the same comparison for real GDP growth. AI forecasts consistently exhibit lower dispersion and more concentrated error distributions than human forecasters across both inflation measures and forecast horizons.

that gradually erode at longer horizons. The out-of-sample results show more unstable results, with some outcome reversals compared to the in-sample. The limited out-of-sample observations (N=4-6) make it difficult to determine whether these reversals reflect genuine performance differences, structural breaks in the post-2021 period, or simply small-sample volatility.

5 Future work

Having established that persona descriptions are expendable, future work should systematically ablate other prompt components—ECB policy communications, past SPF medians, and real-time macro data—to identify which contextual information genuinely drives forecasting performance versus merely increasing token costs. Several other extensions merit investigation. First, evaluating density

forecasts—which are included in the ECB SPF—rather than just point estimates would test whether LLMs can meaningfully quantify uncertainty. Second, alternative prompting strategies beyond personas, such as explicit chain-of-thought reasoning or adversarial perspectives, may prove more effective at generating forecast diversity. Finally, extending the out-of-sample evaluation beyond our limited six rounds would provide more robust evidence of generalization.

6 Conclusion

We present the first systematic replication of the ECB Survey of Professional Forecasters using LLMs, evaluating over 2,000 synthetic personas extracted from the PersonaHub corpus across 50 quarterly rounds. Our controlled ablation experiment reveals that adding these descriptions to the prompt provides no measurable forecasting advantage, with statistical tests showing no significant difference between persona-enhanced and baseline approaches. However, we find that LLMs can achieve competitive accuracy with human forecasters, even on out-of-sample data from 2024-2025 that was entirely absent from model training. These results have practical implications for AI-assisted forecasting systems. Rather than investing computational resources in elaborate persona engineering, practitioners should focus on robust data integration and model improvements. Our findings also reveal behavioral differences between AI and human forecasting panels: despite diverse prompting, LLMs exhibit very low dispersion and consensus-seeking behavior, in contrast with the heterogeneity observed in human expert panels. Future research should explore density forecasting capabilities and scenario coherence across multiple variables, while investigating whether alternative prompt engineering approaches beyond persona descriptions can enhance LLM forecasting performance in economic applications.

References

[1] Juan Angel, García. 2003. *An introduction to the ECB's survey of professional forecasters.* Occasional Paper Series 8. European Central Bank. https://EconPapers.repec.org/RePEc:ecb:ecbops:20038

[2] Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis* 31, 3 (2023), 337–351.

[3] Leland Bybee. 2023. Surveying Generative AI's Economic Expectations. doi:10.48550/arXiv.2305.02823 arXiv:2305.02823 [econ].

[4] Andrea Carriero, Davide Pettenuzzo, and Shubhranshu Shekhar. 2024. Macroeconomic Forecasting with Large Language Models. *CoRR* (Jan. 2024). https://openreview.net/forum?id=hNU5kFeo9r

[5] Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. 2024. Questioning the Survey Responses of Large Language Models. https://openreview.net/forum?id=Oo7dlLgqQX

[6] Miguel Faria-e Castro and Fernando Leibovici. 2024. Artificial Intelligence and Inflation Forecasts. *Review* (2024). doi:10.20955/r.2024.12

[7] Michael J Fell. 2024. Energy social surveys replicated with Large Language Model agents. *Available at SSRN 4686345* (2024).

[8] Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling Synthetic Data Creation with 1,000,000,000 Personas. doi:10.48550/arXiv.2406.20094 arXiv:2406.20094 [cs].

[9] Mingmeng Geng, Sihong He, and Roberto Trotta. 2024. Are large language models chameleons? An attempt to simulate social surveys. *arXiv preprint arXiv:2405.19323* (2024).

[10] Anne Lundgaard Hansen, John J. Horton, Sophia Kazinnik, Daniela Puzzello, and Ali Zarifhonarvar. 2025. Simulating the Survey of Professional Forecasters. doi:10.2139/ssrn.5066286

[11] John J. Horton. 2023. Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus? doi:10.3386/w31122

[12] Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2024. Better Zero-Shot Reasoning with Role-Play Prompting. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 4099–4113. doi:10.18653/v1/2024.naacl-long.228

[13] Anton Korinek. 2023. Language Models and Cognitive Automation for Economic Research. doi:10.3386/w30957

[14] Alejandro Lopez-Lira, Yuehua Tang, and Mingyin Zhu. 2025. The Memorization Problem: Can We Trust LLMs' Economic Forecasts? doi:10.48550/arXiv.2504.14765 arXiv:2504.14765 [q-fin].

[15] Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. 2025. *Large Language Models: An Applied Econometric Framework.* NBER Working Papers 33344. National Bureau of Economic Research, Inc. https://ideas.repec.org/p/nbr/nberwo/33344.html

[16] OpenAI. 2024. GPT-4o Model Documentation. https://platform.openai.com/docs/models/gpt-4o Model released May 13, 2024. Knowledge cutoff: October 1, 2023.

[17] Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations.* https://openreview.net/forum?id=RIu5lyNXjT

[18] Mingtian Tan, Mike Merrill, Vinayak Gupta, Tim Althoff, and Tom Hartvigsen. 2024. Are language models actually useful for time series forecasting? *Advances in Neural Information Processing Systems* 37 (2024), 60162–60191.

[19] Ali Zarifhonarvar. 2024. Evidence on Inflation Expectations Formation Using Large Language Models. doi:10.2139/ssrn.4825076 Place: Rochester, NY Type: SSRN Scholarly Paper.