November 2026

# Generative AI for Economic Simulation: Heterogeneity in Subjective Models of Inflation

**Yusuke Takahashi[a] and Kazuki Otaka[a,*]**

[a] Institute for Monetary and Economic Studies, Bank of Japan, 2-1-1 Nihonbashi-Hongokucho, Chuo-ku, Tokyo 103-0021, Japan

[*] Corresponding author. E-mail address: kazuki.ootaka@boj.or.jp

## Abstract

Personas generated by large language models (LLMs) are candidate heterogeneous agents for economic simulation. We match each persona one-to-one with a respondent in the Bank of Japan's household survey and measure how far its inflation expectations track that respondent's answers. Correspondence is separated into dispersion, attribute-level, and individual-level parts, the last read against the benchmark that attribute-only personas imply. Short-horizon conditional means are partially recovered; dispersion and the five-year average are not. Describing a respondent's awareness of monetary policy can more than double the degree of tracking for one-year expectations, so information frictions can enter a simulation as persona text rather than calibrated parameters.

**Keywords:** Large language models, Inflation expectations, Agent-based simulation, Information frictions, Heterogeneous agents, Persona prompting

**JEL classification:** C63, D84, E31, E37

## 1. Introduction

Agent-based models (ABMs) allow researchers to analyze how aggregate phenomena emerge from interactions among heterogeneous agents (Farmer & Foley, 2009; Tesfatsion, 2006). Modeling that heterogeneity matters: household diversity alters policy transmission (Kaplan et al., 2018), and inflation expectations vary systematically across demographic groups (Weber et al., 2022). HANK models capture distributional dynamics but, for tractability, abstract from the unstructured diversity of household behavior. An ABM need not impose that abstraction, but its flexibility shifts the specification burden onto the researcher.

As Farmer (2003, 2024) characterizes it, departing from full rationality opens a "wilderness of bounded rationality" in which infinitely many behavioral specifications are possible.[1] Sargent's (1993) lectures had already named that burden. The modeler decides what agents know in advance, what they must learn, and how they learn; Sargent grouped those decisions under the heading of "prompting." The number of free parameters and the ad-hoc character of behavioral rules remain the main hurdles to the credibility of ABM-based policy analysis (Fagiolo & Roventini, 2017).

Large language models (LLMs) may offer a way through. Horton et al. (2023) introduced Homo silicus: computational agents that simulate human economic behavior through statistical regularities learned from text. Instead of hand-coding behavioral rules, the researcher describes an agent in natural language—a persona—and lets the model generate that agent's responses. Prompting in Sargent's sense is thereby relocated rather than removed. What the persona leaves unsaid is supplied by the model rather than by the modeler's equations, so part of the specification burden turns from a modeling choice into an empirical property of the model, one that can be measured against human data. Economics (Li et al., 2024; Hansen et al., 2025), political science (Argyle et al., 2023), and sociology (Gao et al., 2024) have all taken up the idea, and central banks have begun to explore LLM agents for macroeconomic simulation (Takahashi et al., 2025) and for expectation formation (Anesti et al., 2026).

Whether an LLM can fill this role depends on whether it reproduces not only the average behavior of households but the heterogeneity across them, and the validation literature is mixed on that point. LLMs display "algorithmic fidelity" in replicating population-level correlations

[1] Farmer (2003) attributes the phrase to Sims (1980), who writes of the "wilderness of 'disequilibrium economics'" and never uses the term bounded rationality. The substitution originates with Sargent (1993), who has economists venture into "what Christopher Sims (1980) characterized as the 'wilderness' of irrational expectations and bounded rationality."

(Argyle et al., 2023), yet they also exhibit the homogeneity problem: response distributions with less variation than human data (Bisbee et al., 2024; Lin et al., 2026).

This paper therefore asks how far LLM personas can serve as heterogeneous subjective models of household inflation expectations, and builds the apparatus needed to answer the question quantitatively. Each LLM agent receives the attribute profile of a real respondent to the Bank of Japan's household survey, and the human and LLM response vectors are compared one-to-one, with correspondence at the attribute level separated from correspondence at the individual level. Where the correspondence falls short, we test a repair suggested by information friction theory (Mankiw & Reis, 2002; Sims, 2003), conditioning personas on cognitive variables—awareness of monetary policy and price attention—elicited from the same respondents. Scenario experiments following Andre et al. (2022) establish beforehand that LLM personas hold attribute-dependent subjective models at all.

The answer is quantitative rather than binary, which is why the measurement device is the contribution. LLM personas reproduce the direction of most attribute effects on short-horizon expectations, and conditioning on cognitive variables improves that correspondence further. But dispersion remains a fraction of the human level in every configuration, and correspondence weakens as the horizon extends, to nothing at five years. Conditional mean structure is partially recoverable at short horizons; the dispersion of expectations is not.

The paper proceeds as follows. Section 2 reviews the literature and states what we add to it. Section 3 describes the persona design, the language model, the survey data, and the evaluation framework, and decomposes the variance of generated expectations to locate the homogeneity problem. Section 4 compares demographics-only personas with their matched respondents, and Section 5 repeats the comparison with the cognitive variables added. Section 6 interprets the results and discusses their implications and limitations, and Section 7 concludes.

## 2. Related Literature and Contributions

Recent work has explored the capacity of LLMs to simulate economic expectations. Bybee (2023) showed that LLMs can generate aggregate expectations of macroeconomic and financial variables from financial news, treating the model as a text processor. Zarifhonarvar (2026) built persona prompts from Survey of Consumer Expectations respondents and found attribute gradients in the generated expectations that match those documented for households in some models but differ in sign across models, and a range of expectations narrower than in human surveys. Lin et al. (2026) developed a simulation framework for macroeconomic

expectations—including a replication of the Andre et al. (2022) scenario design—and found that personas alone leave LLM agents far more homogeneous than humans, while supplying individual characteristics, prior expectations, and external information restores much of the heterogeneity. Bisbee et al. (2024) established the coefficient-comparison approach we adapt, generating synthetic counterparts for thousands of election-survey respondents and showing that the synthetic data understate variance and diverge at the coefficient level. Wu et al. (2025) re-implemented information-provision experiments on inflation expectations with synthetic panels and set out when the resulting treatment effects are identified. Anesti et al. (2026) is the closest central-bank study: they assign LLM personas the demographic profiles of actual respondents to a national household survey, estimate parallel demographic regressions on both sides, and report respondent-level correlations across horizons. Much of this work uses public survey data or replicates published experiments, which leaves open the concern that observed correspondence reflects patterns memorized from the training corpus (Larooij & Törnberg, 2026).

This paper makes three contributions. The first is a validation framework whose measurements are free of record-level contamination by construction. The whole pipeline runs on confidential microdata, on premises and offline, and because the survey records are never released, no training corpus can contain the individual answers against which the personas are scored. Within that setting the framework separates attribute-level from individual-level correspondence through one-to-one persona matching. Respondent-matched designs already exist: Anesti et al. (2026) evaluate LLM personas at aggregate, demographic, and respondent levels, Bisbee et al. (2024) generate synthetic counterparts for individual respondents, and Huang et al. (2026) benchmark LLM predictions of individual U.S. households' survey answers, including inflation expectations predicted from each respondent's earlier answers. What we add is the separate measurement of dispersion, attribute-level correspondence, and individual-level correspondence, with the last read against the benchmark that attribute-only personas imply.

The second contribution uses the framework to locate where LLM personas stop being usable as heterogeneous agents. The framework makes precise a pattern that aggregate distributional comparisons blur: attribute coefficients can match while dispersion falls short and individual-level correspondence is absent. In the model we validate, this is the pattern that appears: the direction of most short-horizon attribute effects is reproduced, dispersion never exceeds a third of the human level in any configuration, and attribute-level correspondence is absent at five years. Together these delimit what the validated model can be used for:

simulations of short-run expectation formation, but not of the five-year average or of anything that requires the distribution rather than its center.

The theoretical foundation for expectation heterogeneity comes from information friction models. Mankiw and Reis (2002) proposed the sticky information model, in which agents update their information sets at heterogeneous frequencies; Sims (2003) introduced rational inattention, in which agents with finite processing capacity optimally allocate attention. Empirically, cognitive factors drive expectation heterogeneity beyond demographic proxies: measured cognitive abilities help explain inflation expectations even after demographics are controlled for (D'Acunto et al., 2019), and information rigidities that differ across types of agents explain the sluggish adjustment of expectations to shocks (Coibion & Gorodnichenko, 2015). Existing LLM persona designs, however, condition predominantly on demographic characteristics, which carry little of this information: two individuals with identical demographic profiles may differ widely in their attention to economic information.

The third contribution measures what happens when personas are conditioned on exactly the characteristics information friction theory singles out, elicited from the same respondents whose expectations are the outcome. Describing a respondent's awareness of monetary policy in the persona, with both awareness items entered together, more than doubles the degree of tracking for one-year-ahead expectations, although the implied correlation remains small. Conditioning a persona on characteristics measured from the respondent is established practice (Bisbee et al., 2024; Lin et al., 2026). The design measures each configuration against a benchmark that expands with the conditioning set. The vocabulary of information frictions can therefore enter a simulation as persona text rather than as calibrated parameters.

The attribute–expectation relationships that Section 4 tests against are well documented: women, lower-income individuals, and the less educated tend to hold higher inflation expectations across countries (Bryan & Venkatu, 2001; D'Acunto et al., 2023), and Takahashi and Tamanyu (2022) confirm these patterns for perceived inflation in Japan, where households unaware of the 2% price stability target perceive inflation to be higher. Since LLMs are trained predominantly on English-language and US-centric corpora, Japanese data provide a demanding test of whether a model reproduces country-specific expectation patterns, shaped by decades of deflation, rather than defaulting to the patterns prevalent in its training data.

## 3. Data and Methods

### 3.1 LLM Agent Design

This section describes the survey-matched design of Sections 4 and 5; the scenario experiments of Appendix B use a separate set of US personas and are described there. The experiments use persona-based prompting to simulate heterogeneous economic agents. A persona is a natural-language description of an individual's demographic and socioeconomic characteristics, supplied to the LLM as context for generating responses. The prompt follows a system–user format: the system prompt instructs the model to embody the persona, and the user prompt supplies the persona description together with the survey questions. The personas carry age, gender, region, employment status, household income, liabilities, and household composition, and in Section 5 additionally the respondent's own answers to three cognitive items. The experiments are conducted entirely in Japanese, using the recorded attributes of actual respondents to the Bank of Japan survey. Appendix A reproduces the prompts.

Each persona receives its questions within a single session, so earlier answers form part of the context for later questions within a persona, while no context is shared across personas; responses are therefore independent across personas by construction. The questions follow the survey's own fixed order for humans and personas alike, so any anchoring induced by question position is common in kind to both sides of the comparison.

### 3.2 Model Selection and Configuration

The experiments use non-public survey microdata containing confidential respondent attributes, so they run exclusively on locally deployed Gemma-3 models (Google) in an environment disconnected from the internet; no commercial API-based model is used. We deliberately work with lightweight models rather than the largest available ones, because deploying thousands of heterogeneous agents requires models that can be run at that scale. Within the Gemma-3 family we designate Gemma-3-12B-it as the model to be measured, on the grounds that it offers the best balance between computational cost and general benchmark capability among the locally deployable sizes given our GPU constraints; the choice was made before any correspondence statistic was computed. Appendix C reports the same measurement for the 4B and 27B sizes, for one-year expectations in the demographics-only configuration. Every result in Sections 4 and 5 is a statement about the model measured; what carries over to another candidate model is the framework, not the values.

All three Gemma-3 sizes are run under 4-bit NF4 quantization because of local GPU memory constraints, so numerical precision is held fixed across sizes. Local deployment of open-weight models keeps the confidential survey microdata within the computing environment and allows the experiments to be re-run with identical weights.

The temperature parameter, which controls the stochasticity of LLM outputs, is set to the models' default sampling temperature of 1.0. Section 3.5 examines the temperature–variance relationship in this design and shows that no temperature setting between 0 and 2 resolves the homogeneity problem. At the default, the valid-response rate stays at 100% and about half of the between-attribute variance present at temperature 0 is retained for one-year expectations. Temperature 0 would leave no draw-to-draw variation for the repeated-generation analysis of Section 5.3 to measure, and because deployed simulations sample stochastically, the measured object is the one a simulation would face.

### 3.3 Data Sources

The experiments use non-public microdata from the Bank of Japan's "Opinion Survey on the General Public's Views and Behavior" (100th wave, December 2024 survey), conducted between November 7 and December 3, 2024, with results published on January 17, 2025. This quarterly survey collects inflation expectations and related economic attitudes from individuals aged 20 and over; the questionnaire also asks respondents their gender, age, employment status, income, borrowing, and household composition, and the region of residence is recorded by the survey. Personas are constructed directly from actual respondent attributes, which enables the one-to-one comparison at the core of our framework.

The design blocks contamination at two levels. The individual responses are never released in any form, so no training corpus can contain them, whatever a model's cutoff; and this wave's published aggregates appeared after the survey window, which itself postdates the reported knowledge cutoff of the models employed, August 2024. Together the two levels rule out direct retrieval of respondent-level records and, given the reported cutoff, of this wave's statistics as an explanation for any observed correspondence—the possibility that LLM outputs merely reproduce memorized data, raised by Lopez-Lira et al. (2025) for LLM forecasting.

The model may still have absorbed published findings from earlier waves, and this possibility fixes how each component of correspondence is read. Aggregate results of earlier waves are public and serially correlated with the present wave, and published research on Japanese households documents how perceived inflation varies with demographics in earlier

waves of this survey (Takahashi & Tamanyu, 2022) and how expectations rise with age (Diamond et al., 2020). Attribute-level correspondence is therefore read throughout as evidence that the model has internalized such regularities and can apply them to new profiles, and the same reading applies to the cognitive variables of Section 5. Individual-level correspondence is the component published sources cannot supply: attribute profiles here are close to unique, and published material does not report cross-tabulations of expectations by these attribute combinations. This is what the one-to-one design makes measurable.

We use the wave immediately following the reported cutoffs: an earlier wave cannot be cleansed of what the models have memorized, and a later one would require supplying the intervening macroeconomic information in full. The gap between the cutoffs and the survey window is roughly two to three months, over which the macroeconomic information bearing on the survey questions changed comparatively little, so the information sets available to the respondents and to the models are close in calendar terms.

For the cognitive-variable experiments we additionally use three survey items that capture information-processing characteristics beyond standard demographics: awareness of the Bank of Japan's price stability mandate, awareness of the 2% inflation target, and whether the respondent has a specific figure in mind for CPI inflation.

The survey drew a sample of 4,000 and obtained 2,099 valid responses (52.5%); the analysis sample of 1,764 respondents reflects listwise exclusion of respondents with missing attribute or outcome items.

### 3.4 Evaluation Framework

We evaluate dispersion, attribute-level correspondence, and individual-level correspondence; Appendix E gives the estimation details.

**Coefficient comparison.** For both human and LLM data we estimate a linear regression of inflation expectations on persona attributes:

$$y_i = \alpha + \sum_k \beta_k W_{ik} + \varepsilon_i, \tag{1}$$

where $y_i$ is the inflation expectation (or perception) of individual $i$ and $W_{ik}$ are dummy-coded demographic attributes. We then compare the estimated coefficient vectors $\hat{\beta}^H$ (human) and $\hat{\beta}^L$ (LLM). This approach, adapted from Bisbee et al. (2024), moves beyond aggregate distribution comparison to evaluate whether LLMs capture the direction and

magnitude of attribute effects. Sign agreement rates are tested by a permutation test that shuffles the LLM responses.

**Individual-level statistics.** The survey-matched design assigns each LLM agent the attribute profile of a real respondent, so each generated expectation can be compared with that respondent's actual answer. We report two primary statistics. The first is the standard deviation ratio,

$$\text{SD Ratio} = \frac{\text{SD}(y^L)}{\text{SD}(y^H)}, \tag{2}$$

which measures relative dispersion; values below 1 indicate the homogeneity problem. The ratio compares total dispersions, generation noise included; Section 3.5 defines a systematic SD ratio net of that noise. The second is the tracking coefficient

$$\theta_1 = \frac{\text{Cov}(y^H, y^L)}{\text{Var}(y^H)}, \tag{3}$$

the OLS slope from regressing LLM expectations on human expectations across matched pairs. Its systematic part is mediated by attributes: since the persona carries only $W$, even a model that had learned nothing beyond published attribute-level regularities could generate a positive $\theta_1$. We also report the residual association between the two sides after additive adjustment for $W$, which diagnoses whether anything beyond the additive attribute structure is tracked and which we test with a within-cell permutation check (Section 4.4). Two further quantities are $\theta_2$ from the reverse regression and $R^2 = \theta_1\theta_2$. These statistics are computed from one generation per persona and then averaged over 30 generations. Generation noise attenuates the reverse slope and the product while leaving $\theta_1$ unbiased; Appendix E measures the attenuation on persona means and reports the draw-to-draw variability of the tracking coefficient.

Since the LLM observes only $W$ and not $y^H$, its systematic output can covary with $y^H$ only through $W$; by the Cauchy–Schwarz inequality, in expectation

$$|\theta_1| \leq \sqrt{\tilde{R}^2_{H|W}} \times \text{SD Ratio}, \tag{4}$$

where $\tilde{R}^2_{H|W} = \text{Var}(E[y^H \mid W])/\text{Var}(y^H)$ is the explanatory power of attributes for human expectations. The right-hand side is the scale on which individual-level correspondence is read: no model can track within-cell variation it is given no information about, so $\theta_1$ is expressed relative to this attribute-implied reference rather than to 1, which places outcomes with different attribute structure on a common footing. Because the conditional mean is not

estimable cell by cell in this sample, we substitute the additive attribute regression for it throughout; the resulting figure is an additive benchmark rather than a bound, and it serves as a unit of measurement, not as a level the model is expected to reach.

**Persona utilization efficiency.** Let $\hat{y}^H$ denote fitted values from regressing human expectations on $W$, and define the benchmark $\theta_{OLS} = \mathrm{Cov}(y^H, \hat{y}^H)/\mathrm{Var}(y^H) = R^2_{H|W}$. Persona utilization efficiency is

$$\text{Efficiency} = \frac{\theta_1}{\theta_{OLS}} = \frac{\mathrm{Cov}(y^H, y^L)}{\mathrm{Cov}(y^H, \hat{y}^H)}, \tag{5}$$

a ratio of covariances that asks how much of the linearly attribute-predictable covariance with human responses the LLM recovers. A simulator built from the human data themselves, the attribute regression plus draws from its residual distribution, attains an efficiency of 100% and an SD ratio of 1 in expectation, so both statistics ask how far a model that never saw those data comes toward that reference.

Because the benchmark $R^2$ in the denominator is small (Section 4.4), we lead throughout with the benchmark-free quantities: $\theta_1$ and the implied raw human–LLM correlation $\theta_1/\text{SD Ratio}$. The latter is the tracking coefficient divided by the SD ratio, so it is the ordinary correlation coefficient between the two response vectors and, unlike efficiency, does not divide by an estimated benchmark.

### 3.5 The Structure of the Homogeneity Problem: Variance Decomposition

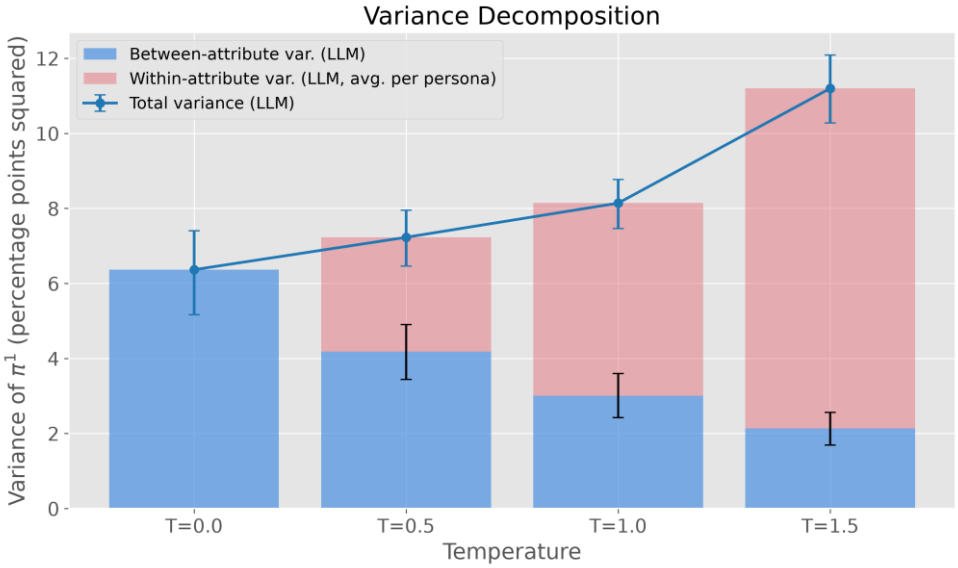
We defined the homogeneity problem, following the literature, as the tendency of LLMs to produce response distributions with less total variation than human data. For simulation purposes not all of that variation is equal: for a persona defined only by its attributes, variation that tracks those attributes is the systematic heterogeneity a simulation relies on, while draw-to-draw variation at fixed attributes is generation noise by construction. To identify which component is missing we take the demographics-only personas of the survey-matched design, generate their answers repeatedly at temperatures from 0 to 2, and decompose the total variance of the generated expectations at each temperature into between-attribute and within-attribute components:

$$\mathrm{Var}(y) = \underbrace{\mathrm{Var}(E[y \mid W])}_{\text{between-attribute}} + \underbrace{E[\mathrm{Var}(y \mid W)]}_{\text{within-attribute}}, \tag{6}$$

where y is the generated expectation and $W$ denotes persona attributes. For a random subsample of respondents' personas we generate the same number of responses per persona,

estimate the within-attribute component as the average of the persona-specific sample variances, obtain the between-attribute component as the residual from total variance, and attach bootstrap intervals over personas to the between-attribute component (Appendix E). The square root of the between-attribute component, divided by the human standard deviation, is the systematic SD ratio used in Section 5.3: the SD ratio of equation (2) with generation noise removed from the numerator. Figure 1 plots the decomposition for one-year expectations; Appendix E reports the underlying values for all three outcomes.

Figure 1: Variance Decomposition Across Temperature Settings




Note: One-year expectations; demographics-only personas for 274 respondents drawn at random from the trimmed sample of Section 4, with 10 generations per persona at temperatures 0.5 to 1.5 and two generations at temperature 0, where decoding is deterministic. Temperature 2.0 is reported in Appendix E only, because a few extreme answers dominate its total variance.

At temperature 0 decoding is deterministic, so all variance is between-attribute. Raising temperature to 1.5 raises total variance by about three quarters, but the added variance is entirely within-attribute noise: between-attribute variance, the component a persona-based simulation can control and rely on, falls monotonically over this range, to about one third of its value at temperature 0 for one-year expectations (Figure 1). The model therefore lacks between-attribute variation, and higher temperature does not restore it: what it adds is generation noise, not heterogeneity a simulation could use. In what follows, "homogeneity problem" refers to this deficit in between-attribute variation, and "variance compression" or "dispersion gap" to the total-dispersion symptom; the two come apart as temperature rises, since total dispersion grows while meaningful differentiation shrinks.

Even between temperatures 0 and 1.0, within the range practitioners use, between-attribute variance for one-year expectations falls by about half. Temperature is therefore not an instrument for manufacturing agent heterogeneity, and the operating range has to be chosen against this loss rather than for maximal output variance.

Survey evidence indicates that human heterogeneity in household inflation expectations likewise runs beyond demographic attributes: exposure to everyday price signals, lifetime inflation experiences, and cognitive abilities all shape expectations, and the gender gap, for instance, disappears once responsibility for grocery shopping is accounted for (D'Acunto et al., 2023). Much of the heterogeneity therefore lies within attribute cells rather than between them. An attribute-only design cannot recover variation within attribute cells, so it is held to the between-attribute component, the one temperature erodes. The natural hypothesis is that what is lacking is the information in the persona, and Sections 4 and 5 pursue it with data that bear on it.

## 4. Validating LLM Expectations: A One-to-One Microdata Approach

### 4.1 Design

This section answers the question of the paper directly on human microdata, with the evaluation framework of Section 3.4: each LLM agent receives the attribute profile of a respondent to the Bank of Japan's household survey, and its generated responses are compared with that respondent's actual answers, one-to-one. Attribute-level correspondence is interpreted as consistency with human patterns.

We examine three outcome variables: perceived inflation over the past year ($\pi_p$), expected inflation one year ahead ($\pi_1$), and expected average annual inflation over the next five years ($\pi_5$), which is the quantity the survey itself elicits at that horizon. Since the recorded demographic attributes explain only a modest fraction of the variance of human expectations, the question is not whether the LLM explains the full heterogeneity, but how much of the attribute-predictable structure it recovers, and where the individual-level statistics sit relative to the additive benchmark.

### 4.2 Descriptive Overview

Table 1: Descriptive Statistics by Outcome Variable

| Variable | Source | Mean | SD | N |
|---|---|---|---|---|
| Perceived inflation ($\pi_p$) | Human | 17.3 | 17.0 | 1,764 |
| | LLM | 17.1 | 3.2 | 1,764 |
| One-year-ahead expectation ($\pi_1$) | Human | 11.8 | 16.0 | 1,764 |
| | LLM | 10.4 | 3.0 | 1,764 |
| Five-year average, annual rate ($\pi_5$) | Human | 9.7 | 17.8 | 1,764 |
| | LLM | 4.3 | 1.0 | 1,764 |

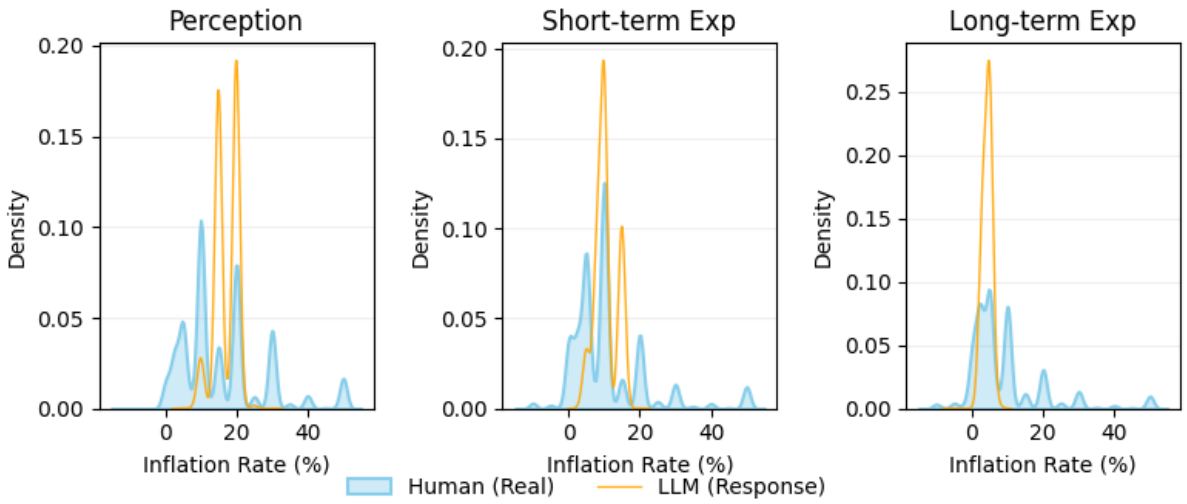
Note: One generation per persona.

Three features of Table 1 stand out. First, LLM means are close to human means for perceived inflation and one-year expectations.

Second, the LLM's mean for the five-year average (4.3) is less than half the human mean (9.7), and its dispersion nearly vanishes: the model's answers at this horizon lie within a few points of that mean, whereas the human mean is carried by answers of 10% and above, of which the model produces almost none.

Third—and central to our question—LLM standard deviations are a fraction of human values throughout, and the ratio collapses further for the five-year average. To the extent the noise share is larger on the LLM side—responses are generated once per persona—the table understates the gap in systematic, attribute-driven dispersion.

Figure 2: Distribution Comparison



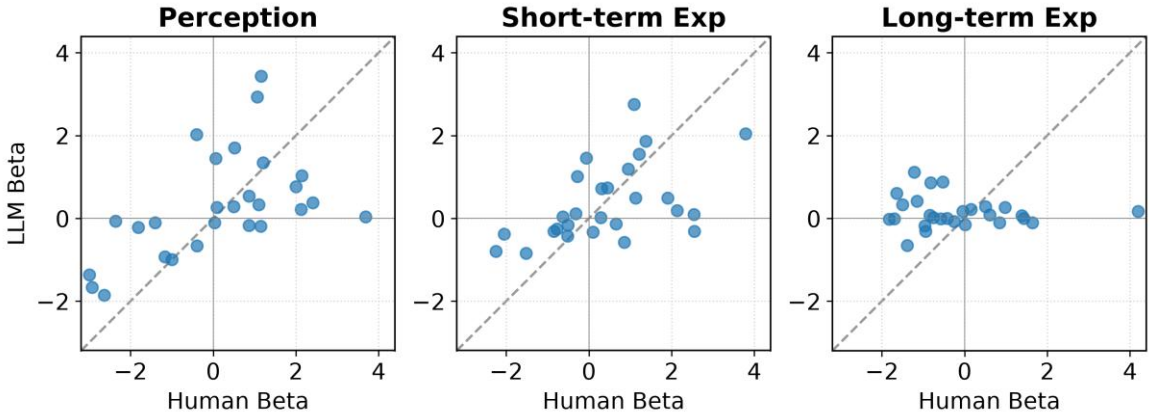

Note: Kernel density estimates of the responses summarized in Table 1, for perceived inflation, one-year expectations, and the five-year average.

Figure 2 visualizes the compression: at every horizon the LLM's answers sit at a few values within a narrow band and leave the tails empty, and the band narrows from perceived inflation to the five-year average, while the human answers spread across the axis, with visible mass at 20% and beyond.

### 4.3 Attribute-Level Correspondence: Coefficient Comparison

We estimate parallel attribute regressions for human and LLM data and compare coefficients.

Figure 3: Coefficient Comparison Scatter Plot




Note: From this figure on, human respondents below the 0.5th or above the 99.5th percentile of any of the three outcome variables are excluded (N = 1,744). LLM coefficients are estimated on each persona's mean response over 30 generations, as are the sign agreement rates in the text.

Sign agreement rates are 85.2% for perceived inflation, 70.4% for one-year expectations, and 51.9% for the five-year average (23, 19, and 14 of the 27 coefficients of the parallel regressions, with each persona's LLM response averaged over 30 generations before the regression is estimated). Under the permutation test the first two rates are significant while the five-year rate is not distinguishable from chance.[2] For perceived inflation and one-year expectations the LLM reproduces the direction of most attribute effects, but with attenuated amplitude: on the coefficient scatter of Figure 3 the points in the two short-horizon panels lie along a line markedly flatter than the 45-degree line. The one-year correspondence is thus positive but fragile. The expected directions of the age, gender, and income effects rest on documented mechanisms: experience-based learning, differential exposure to frequently

[2] The test shuffles the LLM responses (2,000 draws, one-sided); p = 0.001, 0.02, and 0.49 for perceived inflation, one-year, and five-year expectations respectively. Because sign agreement is computed over dummy coefficients, it also depends on the choice of reference categories: with one generation per persona, across 100 alternative sets of reference categories the permutation p-value falls below 0.05 in 92% of specifications for perceived inflation, in 38% for one-year expectations, and in none at five years.

purchased goods, and income-related differences in budget shares and informational resources (Bryan & Venkatu, 2001; Hobijn & Lagakos, 2005; Malmendier & Nagel, 2016; D'Acunto et al., 2019, 2021; Diamond et al., 2020, for Japan). Appendix E reports the coefficients of both regressions.

### 4.4 Individual-Level Correspondence and the Additive Benchmark

Table 2: Summary Statistics for Human–LLM Correspondence

| Metric | $\pi_p$ | $\pi_1$ | $\pi_5$ |
|---|---|---|---|
| $\theta_1$ (LLM on human) | 0.023 | 0.011 | −0.002 |
| $\theta_1$ SD over 30 generations | 0.003 | 0.004 | 0.001 |
| $\theta_2$ (human on LLM) | 0.497 | 0.228 | −0.366 |
| $R^2$ $(= \theta_1\theta_2)$ | 1.2% | 0.3% | 0.1% |
| SD Ratio | 0.22 | 0.22 | 0.08 |
| $\theta_{OLS}$ $(R^2_{H\|W})$ | 0.039 | 0.023 | 0.019 |
| Efficiency $(\theta_1/\theta_{OLS})$ | 59% | 47% | −12% |
| $\theta_1$ 95% CI | [0.015, 0.032] | [0.004, 0.017] | [−0.0052, 0.0005] |

Note: $\theta_{OLS}$ is the in-sample additive attribute benchmark (Appendix E). For $\pi_5$ both slopes are negative ($\theta_1, \theta_2 < 0$), so the positive $R^2$ does not indicate positive correspondence. Except for the SD and confidence-interval rows, entries are means of the statistics computed separately for each of the 30 generations per persona; the product row averages the per-generation products, and the SD row is the standard deviation of the tracking coefficient across the generations. The sample is trimmed as in Figure 3 (N = 1,744), so the SD ratio is not the ratio of the standard deviations in Table 1. The confidence interval for $\theta_1$ is over respondents, conditional on the generated responses, from the regression on per-persona means, with heteroskedasticity-robust (HC1) standard errors; the benchmark and efficiency carry no interval.

The individual-level statistics in Table 2 are read on the scale of Section 3.4. For attribute-only personas the reference value is the product of two factors: the square root of $\sqrt{R^2_{H|W}}$, the attribute information in the human data, and the SD ratio, the model's own dispersion. Estimating the first with the additive attribute regression gives reference values of about 0.043 for $\pi_p$ and 0.033 for $\pi_1$; the two factors are of similar size in both cases (0.20 against 0.22, and 0.15 against 0.22), so the model's own compression and the attribute information contribute about equally to the scale. The observed $\theta_1$ is positive for perceived inflation and one-year expectations, with confidence intervals that exclude zero, and amounts to about half and a third of the respective reference value. The reference values are rough scales, because the benchmark includes the roughly 0.016 that fitting the 27 dummies to noise alone produces (Appendix E). The benchmarks for the one-year and five-year horizons are of similar size (0.023 and 0.019); what separates the two horizons is the tracking coefficient, which for the five-year average is negative, close to zero, and has an interval that includes zero.

The remaining rows of Table 2 qualify this reading. The two kinds of uncertainty reported for the tracking coefficient, the confidence interval over respondents given the generated responses and the standard deviation across the 30 generations, are of similar size once the latter is scaled to a 95% band. The $R^2$ values show that human and LLM responses share almost no individual-level variation. The residual correlation after additive demographic adjustment, which measures correspondence beyond the additive attribute structure, is −0.006, −0.011, and 0.033 for the three outcomes, and none is detectable under the within-cell permutation check[3] (p = 0.79, 0.63, and 0.18; Appendix E).

The diagnosis is thus established on human data: attribute-level correspondence is real but partial, individual-level correspondence stays close to the low value the additive benchmark implies, and both dispersion and attribute-level correspondence deteriorate with horizon. If missing persona information is part of the cause of the attribute-level shortfall, richer conditioning should help, and information friction theory says what information to add.

## 5. Can Cognitive Variables Mitigate the Homogeneity Problem?

### 5.1 Design

Sections 3.5 and 4 identified insufficient between-attribute differentiation and modest persona utilization as the central failures. This section tests a theory-guided repair: conditioning personas on cognitive variables that proxy for information frictions. Mankiw and Reis (2002) attribute expectation heterogeneity to heterogeneous updating frequencies, Sims (2003) to heterogeneous attention allocation under capacity constraints. If part of the missing heterogeneity reflects the model's ignorance of how each agent processes information, then telling it—in natural language—should recover some of the shortfall.

Three quantities are tracked separately in this section: the alignment of the LLM's attribute-driven variation with human patterns, the quantity of between-attribute variance, and total dispersion. Appendix D reports a further, exploratory experiment on whether described awareness also shapes how strongly personas revise expectations in response to news.

[3] Two-sided p-values, 5,000 permutations, with cells below 15 respondents excluded.

### 5.2 Cognitive Variables

The Bank of Japan survey includes items that can proxy for otherwise unobservable cognitive states. We use three (Table 3): awareness of the BOJ's price stability mandate (Q21(1)), awareness of the 2% inflation target (Q21(2)), and whether the respondent has a specific figure in mind for CPI inflation (Q22), which we call price attention. For each persona, the respondent's actual answers are incorporated into the prompt as natural-language descriptions (e.g., "You are aware that the Bank of Japan has set a 'price stability target' of 2%…" or "You have never heard of the 2% inflation target").

Table 3: Cognitive Variables Used in the Persona Descriptions

| Variable | Survey item | Response options |
|---|---|---|
| Q21(1) | Awareness of BOJ's price stability mandate | Know / Heard but do not know well / Never heard |
| Q21(2) | Awareness of 2% inflation target | Know / Heard but do not know well / Never heard |
| Q22 | Has a specific figure in mind for CPI inflation | Yes / No |

The three items are reasonable proxies for the cognitive states that the theory points to. Awareness of the mandate and of the target is informative about how much monetary policy information the respondent has acquired, and having a figure in mind for inflation is informative about attention to prices. Each item may also capture stable individual traits beyond the theoretical parameter, such as a disposition toward numerical reasoning, which does not detract from its use in describing an individual. The results are read as conditional correspondence: whether adding the information moves generated expectations in the direction observed in the human data.

The human benchmark is conditioned on the same information as the persona. The benchmark regression includes the cognitive variables as well as demographics, so the denominator of the efficiency metric expands with the conditioning set. Through omitted-variable channels the regression also absorbs unobserved heterogeneity correlated with the cognitive items, whereas the LLM receives only the descriptor text.

### 5.3 The Impact of Cognitive Variables on Expectation Alignment

To isolate the marginal contribution of each cognitive variable we test six persona configurations: demographics only (1), each cognitive item added singly (2) to (4), the two policy awareness items together (5), and all three together (6).

Table 4 reports the correspondence statistics for the six configurations across the three outcome variables, each averaged over 30 generations per persona, with the draw-to-draw standard deviation of the tracking coefficient in parentheses; Figure 4 plots these means, with the range over the 30 generations shown for the tracking coefficient.

Table 4: Persona Utilization by Configuration and Outcome

*Panel A: Perceived inflation ($\pi_p$)*

| Config. | Cognitive variables | $\theta_1$ (draw SD) | p vs (1) | $\theta_{OLS}$ | Efficiency | SD Ratio |
|---|---|---|---|---|---|---|
| (1) | – | 0.023 (0.003) | – | 0.039 | 59% | 0.22 |
| (2) | +Q21(1) | 0.028 (0.003) | 0.16 | 0.049 | 57% | 0.26 |
| (3) | +Q21(2) | 0.025 (0.003) | 0.63 | 0.044 | 56% | 0.26 |
| (4) | +Q22 | 0.035 (0.005) | 0.04 | 0.041 | 85% | 0.33 |
| (5) | +Q21(1,2) | 0.033 (0.003) | < 0.01 | 0.050 | 67% | 0.28 |
| (6) | Full | 0.033 (0.005) | 0.04 | 0.050 | 66% | 0.30 |

*Panel B: One-year-ahead expectation ($\pi_1$)*

| Config. | Cognitive variables | $\theta_1$ (draw SD) | p vs (1) | $\theta_{OLS}$ | Efficiency | SD Ratio |
|---|---|---|---|---|---|---|
| (1) | – | 0.011 (0.004) | – | 0.023 | 47% | 0.22 |
| (2) | +Q21(1) | 0.018 (0.003) | 0.04 | 0.033 | 54% | 0.23 |
| (3) | +Q21(2) | 0.018 (0.004) | 0.08 | 0.027 | 65% | 0.28 |
| (4) | +Q22 | 0.012 (0.004) | 0.82 | 0.023 | 50% | 0.22 |
| (5) | +Q21(1,2) | 0.026 (0.005) | < 0.01 | 0.033 | 79% | 0.28 |
| (6) | Full | 0.022 (0.005) | 0.01 | 0.033 | 67% | 0.27 |

*Panel C: Five-year average, annual rate ($\pi_5$)*

| Config. | Cognitive variables | $\theta_1$ (draw SD) | p vs (1) | $\theta_{OLS}$ | Efficiency | SD Ratio |
|---|---|---|---|---|---|---|
| (1) | – | −0.002 (0.001) | – | 0.019 | −12% | 0.08 |
| (2) | +Q21(1) | −0.001 (0.001) | 0.23 | 0.030 | −3% | 0.09 |
| (3) | +Q21(2) | −0.002 (0.001) | 1.00 | 0.020 | −12% | 0.10 |
| (4) | +Q22 | −0.003 (0.001) | 0.72 | 0.019 | −15% | 0.08 |
| (5) | +Q21(1,2) | −0.001 (0.002) | 0.32 | 0.031 | −4% | 0.10 |
| (6) | Full | −0.001 (0.002) | 0.20 | 0.031 | −2% | 0.10 |

Note: Efficiency in each row is computed against the benchmark regression including that row's conditioning set, so the benchmark grows as variables are added. The benchmarks shown are in-sample (Appendix E). Entries are means over 30 generations as in Table 2; the parenthetical after $\theta_1$ is its standard deviation across the 30 generations, not a standard error. The p-value column reports the two-sided p-value for the difference in $\theta_1$ from configuration (1), from a paired bootstrap over respondents (5,000 resamples) on the personas common to both configurations, conditional on the generated responses and unadjusted for multiple comparisons; after Holm adjustment over the five comparisons per outcome, configuration (5) remains below 0.05 at both short horizons and configuration (6) at one year is 0.05. The sample is the trimmed sample of Table 2, less personas with fewer than 30 parsable generations (N = 1,738 to 1,744).

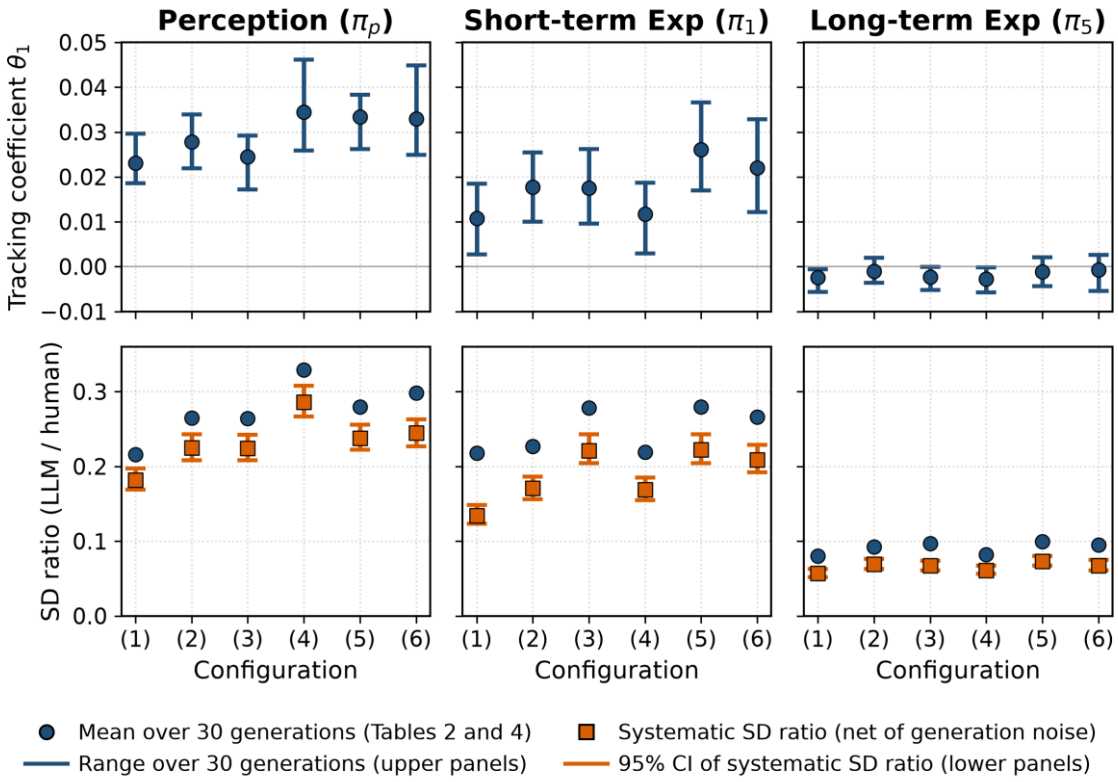
Three patterns emerge.

First, adding cognitive information raises the tracking coefficient, and with both awareness items included, in configuration (5), the rise is large relative to sampling uncertainty over

respondents and to variation across generations. The tracking coefficient $\theta_1$ for $\pi_1$ rises from 0.011 at baseline to 0.026 in configuration (5), and the implied raw human–LLM correlation from about 0.05 to about 0.09. The difference from baseline is statistically detectable in configurations (2), (5) and (6), marginal in (3) and absent with price attention alone, (4); for configuration (5) it is also several times the draw-to-draw standard deviation of the coefficient (Table 4). For $\pi_p$ every configuration improves on the baseline, with price attention (Q22) the strongest single addition, and the differences from baseline are detectable for (4), (5) and (6). After Holm adjustment over the five comparisons per outcome, configuration (5) remains detectable at both short horizons, configuration (6) is borderline for one-year expectations (adjusted p-value 0.05), and no other configuration remains detectable. SD ratios improve modestly and unevenly and reach at most a third of the human level (0.33 for perceived inflation with price attention).

The tracking coefficient $\theta_1$ scales with the amplitude of the model's attribute-driven response, so a larger response to the same attributes would raise it even if that response had not come closer to the human pattern. Figure 4 separates the two. Averaged over 30 generations, the tracking coefficient for one-year expectations more than doubles from baseline to configuration (5). The systematic SD ratio, the dispersion of the attribute-driven component relative to the total human standard deviation, rises by about two thirds over the same step. Dividing the tracking coefficient by this ratio gives the correlation between the attribute-driven component and human responses, which rises from about 0.08 to about 0.12. The improvement therefore comes from both sides: the attribute effects the model generates line up more closely with the human ones, and the dispersion of those effects moves closer to the human level.

Figure 4: Tracking Coefficient and SD Ratio by Configuration (30 Generations per Persona)




Note: Circles are means over 30 generations per persona (Tables 2 and 4); bars in the upper panels show the range of the 30 values. Squares in the lower panels are the systematic SD ratio, the attribute-driven component of Section 3.5 relative to the human standard deviation, with bootstrap 95% intervals over respondents (Appendix E).

Second, improvements are non-monotonic across configurations. The best-performing configuration differs by outcome—(4) for $\pi_p$, (5) for $\pi_1$—and the full configuration (6) does not improve on the two-variable configuration (5) for either outcome; for one-year expectations it falls short by less than one draw-to-draw standard deviation of the coefficient. For perceived inflation the $\theta_1$ differences among configurations (4) to (6) are within one draw-to-draw standard deviation and well inside the uncertainty of the paired bootstrap. For one-year expectations, configuration (4), which adds price attention alone, falls short of configurations (5) and (6), which include the awareness items, while the gap between (5) and (6) is not detectable. What survives is therefore the direction and rough magnitude of the improvement at short horizons rather than a ranking among the cognitive configurations.

Third, no configuration produces appreciable positive correspondence for the five-year average. Conditioning raises the benchmark by more than half, from 0.019 to about 0.031, yet for $\pi_5$ the tracking coefficient stays in a narrow band below zero in every configuration, between $-0.003$ and $-0.001$; no difference from baseline is statistically detectable, the

coefficient itself stays economically indistinguishable from zero, and the SD ratio never leaves 0.08–0.10.

Cognitive conditioning affects between-attribute variance and total dispersion, the second and third quantities of Section 5.1, differently. Applying the variance-component estimator of Section 3.5 with 30 generations per persona, we find that every cognitive configuration raises the between-attribute variance for all three outcomes, and the systematic SD ratio rises with it (Figure 4, lower panels). The attribute-driven component remains small relative to total human dispersion. Against the part of human dispersion that attributes explain, however, measured by the square root of the additive benchmark in Table 4 (0.20 to 0.22 for perceived inflation and 0.15 to 0.18 for one-year expectations), the systematic SD ratio matches or exceeds that part in most cognitive configurations at the two short horizons, while at five years it stays well below it. Total dispersion moves much less, as the SD ratios in Table 4 show. What the cognitive variables add is therefore differentiation across attribute profiles inside a total distribution that remains compressed.

## 6. Discussion

### 6.1 What the Framework Establishes

Throughout the survey-matched design the model reproduces the direction of inflation expectations better than their dispersion: it reproduces the sign of most attribute effects at short horizons (Section 4), and cognitive conditioning moves correspondence in the right direction (Section 5), while the ratio of LLM to human standard deviations never exceeds a third in any configuration. For the uses that need the distribution this is the decisive statistic, and the near-absence of extreme values in the LLM output means that the missing part is precisely the tails on which those uses depend. The differentiation of agents rests on the between-attribute component, which cognitive conditioning raises to the level that attributes explain in the human data at short horizons (Section 5). A single distributional distance would collapse these separations into one number and give no guidance on what to repair. The failure is structured rather than uniform, and the sharpest structure is the horizon gradient: attribute-level correspondence is partial and improvable at short horizons but absent at five years, and the individual-level statistics stay at the low level that attribute-only personas imply, with no detectable correspondence beyond the additive attribute structure (Section 4).

The value of LLM agents lies where matched data end: in counterfactual scenarios, novel policy environments, interaction and feedback among agents, and conditioning richer than a regression can absorb. Precisely there, little direct validation is possible ex ante. The discipline that remains is the one adopted here: validate the model where it can be tested, understand what it reproduces and what it does not, and carry into the untestable settings only what has passed.

### 6.2 Implications

The constructive finding—that natural-language descriptions of policy awareness and price attention improve attribute-level correspondence at short horizons—bears on information friction theory in a specific way. Conditioning on reported policy awareness moves generated expectations in a direction that increases covariance with human responses, without any numerical parameterization. For model building this suggests a workflow in which heterogeneity in information processing is specified descriptively and validated against survey microdata, rather than calibrated as free parameters. This replaces the step at which agent-based practice has been most ad hoc. Adding all three items together did not improve on the best two-variable configuration (Section 5); the estimates are too close to establish that descriptors interfere with one another, but for persona design they suggest that choosing relevant descriptors matters more than adding them.

The horizon gradient is informative in its own right. The model reports a low, near-constant five-year average while human answers at that horizon retain their level, though they carry little attribute structure for the model to track (Section 4). The human pattern may reflect experience effects (Malmendier & Nagel, 2016; Diamond et al., 2020), imperfect credibility, or both. Although the pattern is documented mostly in publications the model has plausibly seen, the model does not reproduce it: what the model lacks is not the fact but the weight it places on it. The nearest respondent-matched study finds the opposite horizon pattern: for the United Kingdom, at temperature zero and with a different model family and outcome definition, Anesti et al. (2026) report respondent-level correlations that strengthen with the horizon, though only from 0.03 to 0.07. The gradient measured here is therefore a property of this model, this measurement, and this country rather than of LLM personas in general. Japan, where household long-horizon expectations have stayed persistently above the target and carry little attribute structure even in the human benchmark, may make the five-year failure visible where other countries' data would not.

For central-bank and research applications the results support a differentiated verdict. Directional and short-horizon uses—screening the sign of expectation responses to scenarios, prototyping heterogeneous-agent mechanisms, generating hypotheses about which population groups react most—are the uses consistent with the observed attribute-level correspondence. They are supported for a model that has first been validated in this way, not for LLM personas in general. Uses that require the distribution of expectations—welfare analysis across households, tail risks of de-anchoring, quantitative calibration of heterogeneous-agent models—are not currently supported, because of the unresolved dispersion gap and the five-year failure.

All survey-matched analyses ran on locally deployed open-weight models in an internet-disconnected environment, so confidential microdata can be used for LLM validation without any data leaving the premises, a requirement most central banks share. The protocol—one-to-one matching, decomposition statistics, configuration experiments—is portable to other institutions' confidential surveys with minimal adaptation, provided the survey carries attribute profiles and, for the conditioning experiments, items that capture how respondents acquire and process economic information.

### 6.3 Limitations

Two limitations affect the interpretation of the results. The first is scope: we study one economic variable in one country, and the survey-matched evidence comes from one model family, with the conditioning experiments run on a single model. A persona is also not a theoretically consistent agent: it optimizes nothing and faces no budget constraint, so consistency across its answers is an empirical property to be checked rather than a guarantee (Appendix D records one failure).

The second concerns the benchmark the framework validates against. The attribute structure in the human survey data is thin—demographic attributes explain a few percent of the variance of human expectations (Section 4)—so the five-year result is a joint statement about the model and the benchmark rather than a verdict on the model alone. Dispersion on the human side also mixes genuine heterogeneity with survey response noise. Repeated generations split the LLM side into systematic and generation components (Sections 3.5 and 5.3), but a single wave cannot do the same on the human side, so the systematic SD ratio has a noise-inclusive denominator and the total SD ratios compare noise-inclusive totals on both sides. Sharper

verdicts would require panel structure, additional survey items, or information-provision experiments that supply exogenous variation.

## 7. Conclusion and Future Directions

This paper asked how far LLM personas can serve as heterogeneous subjective models of household inflation expectations, and answered it with a validation framework built on non-public microdata and one-to-one persona matching. In the model validated here, short-horizon attribute-level structure is partially reproduced, and conditioning the personas on cognitive variables motivated by information friction theory improves it, with persona text taking on part of the work that ABMs assign to calibrated parameters. Correspondence at the five-year horizon is absent under every conditioning we tried, dispersion stays well below the human level, and at the short horizons the individual-level statistics stay at the low level an attribute-only design implies.

The results open four research directions. Dynamics: embedding validated personas in settings with sequential information arrival and expectation updating, ultimately closing the loop in which macroeconomic outcomes feed back into agent states, as in heterogeneous-agent New Keynesian environments. Information provision: varying the content and precision of the macroeconomic information personas receive implements the common-signal comparative statics of information friction theory and turns cognitive conditioning from an observed correlate into an experimental treatment. A post-cutoff shock episode observed prospectively would extend it into strict real-time validation. Richer personas and models: the homogeneity problem can be addressed from both sides of the persona–model pair. On the persona side, individual information beyond attributes—free-text responses, answers to other survey items—would raise the benchmark on individual-level correspondence and make it a live diagnostic. On the model side, heterogeneity need not come from prompts alone: model-family ensembles and fine-tuned variants offer ways to build it into the agents themselves. Other agents: the same matched design applies wherever an institution holds microdata that pair attributes with an economic decision, and the natural next cases are firms' price setting and wage setting, whose subjective models bear on inflation dynamics as much as households' do.

With confidential microdata, one-to-one matching, and a decomposition of correspondence into its attribute-level and individual-level parts, it is now possible to say what a given LLM reproduces of human expectation heterogeneity and what it does not. The boundary drawn here is a measurement, not a verdict: as models and persona designs advance, the same yardstick

will show how far it moves. The wilderness of bounded rationality is vast, but how humanlike a simulated agent is, and where it stops being so, can now be measured against the human population the agent is meant to represent.

## Appendix A: Prompt Structure

This appendix reproduces the prompts used in the two designs. Figure A.1 shows the survey-matched experiments of Sections 4 and 5, in the baseline configuration and with the cognitive descriptors added; Figure A.2 shows the scenario experiment of Appendix B. The survey-matched experiments were administered in Japanese and the scenario experiment in English; the Japanese prompts are shown here in translation, and the originals are available from the authors on request.

Figure A.1: Example Prompt Structure for the Survey-Matched Experiments

**System Prompt**

You are tasked with answering based on the personal, social, and economic circumstances of the profile provided, fully embodying that persona.

**User Prompt (baseline: demographic attributes only)**

You are a male in your 60s living in the Kinki region. Your employment status is "Other (e.g., full-time homemaker, student, pensioner, or unemployed)". Your total household income for the last year falls into the category of "3 to 5 million yen". Your household has liabilities including mortgage. Your household composition is "one-generation household (married couple only)".

It is currently November 2024. Based on the current economic environment, please answer the following questions.

Q1. In your own perception, how have "prices" changed compared to one year ago? ("Prices" refers to the overall prices of goods and services you purchase.) Please answer with the option number only. 1. Rose considerably 2. Rose slightly 3. Almost unchanged 4. Fell slightly 5. Fell considerably

Q2. By what percentage do you think "prices" have changed compared to one year ago? Please provide a number. Put + before the number if you think prices rose, or - if you think prices fell. If you think 0%, answer "0".

[Q3–Q4 ask about one-year-ahead expectations in the same format; Q5–Q6 ask, following the wording of the source survey, about the average annual rate of price change over the next five years]

**User Prompt (extended: demographic attributes + cognitive variables)**

[Same as above, plus:] You are aware that the Bank of Japan has "price stability" as one of its objectives. You are aware that the Bank of Japan has set a "price stability target" of 2% year-on-year increase in consumer prices. You have a concrete numerical image of the year-on-year change in the Consumer Price Index.

Figure A.2: Example Prompt Structure for the Scenario Experiment

**System Prompt**

You are tasked with answering based on the personal, social, and economic circumstances of the profile provided, fully embodying that persona.

**User Prompt 1 (Persona + Scenario)**

You are a 45-year-old Female living in the state of California. Your highest level of education is "Bachelor's degree" and your primary job/status is "Employed". Your marital status is "Married" and your race is "White". Your income is in the range of "$75,000 to $99,999".

Imagine that the average price of oil unexpectedly rises due to a problem with the local production technology issues in the Middle East. On average, the price will be $30 higher for the next 12 months than the current price.

**User Prompt 2 (Question + Options)**

Please account for the actions of policymakers that you would expect in this scenario and include them in your predictions. Under this scenario, what do you think the US inflation rate will be over the next 12 months? Note: The current inflation rate in the US is 1.8%.

Answering method: 1. Select the range with the highest probability from the options. 2. State the range string exactly as it is (e.g., above +1.75% and below or equal to +2.25%). 3. Do not include any reasons or other text.

— Please select the range with the highest probability — above -2.25% and below or equal to -1.75% / above -1.75% and below or equal to -1.25% / … / above +1.75% and below or equal to +2.25% / …

## Appendix B: Scenario Experiments

This appendix establishes that LLM personas hold attribute-dependent subjective models of the macroeconomy, which is the premise the survey-matched validation of Section 4 builds on.

### B.1 Design

Following the experimental paradigm of Andre et al. (2022), LLM agents with different personas respond to hypothetical oil price scenarios. The experiment presents two oil price conditions (stable versus a $30 shock) and elicits two outcome variables (inflation and unemployment expectations) under each. In the baseline condition, respondents are told that oil prices will remain at $54 per barrel over the next 12 months; in the shock condition, a supply disruption raises prices by $30 to $84 per barrel. For each condition, personas provide 12-month-ahead forecasts for both inflation and unemployment; the expectation change $\Delta\pi$ ($\Delta u$) for a persona is the difference between its shock-condition and baseline-condition responses,

elicited in separate sessions. Throughout, we use "subjective model" operationally: a mapping from scenario descriptions to expectation responses whose direction and attribute dependence can be compared with human counterparts.

The scenario experiments use English prompts with US personas, so that the results can be set beside the US findings of Andre et al. (2022). The 999 personas are derived from actual respondent profiles in the Survey of Consumer Expectations (SCE, January 2025 wave), retaining respondents for whom all required persona attributes were available, and carry age, gender, state, education, employment status, marital status, race, and income; building demographic personas from SCE respondent microdata follows Zarifhonarvar (2026). The models are GPT-4o-mini and GPT-4.1-nano/mini (OpenAI), the Gemma-3 series with 4B, 12B, and 27B parameters (Google), and gpt-oss-20b (OpenAI), an open-weight reasoning-specialized model; all are run at their default sampling temperature. Figure A.2 reproduces the prompt.

The shock condition is a compound manipulation, changing the oil price level, its unexpectedness, its stated cause, and its salience together, so what is identified is the response to the scenario as a bundle. The scenario text, the response scale, the human results of Andre et al. (2022) and the textbook direction of a cost-push shock all predate the models' training cutoffs, so directional agreement establishes that the models apply published relationships; the test of reasoning beyond published material is the matched design of Section 4.

### B.2 Directional Consistency

Table B.1 reports, for each model, the proportion of "Rise" responses and the SD ratio for inflation and unemployment expectations following the oil price shock, beside the human data of Andre et al. (2022).

Table B.1: Proportion of Responses Indicating "Rise" and SD Ratio

| Source | $\Delta\pi > 0$ (%) | SD Ratio ($\Delta\pi$) | $\Delta u > 0$ (%) | SD Ratio ($\Delta u$) | N |
|---|---|---|---|---|---|
| *Human (Andre et al., 2022)* | | | | | |
| General public | 78 | – | 63 | – | ~560 |
| Experts | 86 | – | 64 | – | ~90 |
| *LLM* | | | | | |
| GPT-4o-mini | 89.5 | 1.00 | 18.9 | 0.31 | 901 |
| GPT-4.1-nano | 100.0 | 0.46 | 69.0 | 0.38 | 826 |
| GPT-4.1-mini | 100.0 | 0.93 | 100.0 | 0.33 | 901 |
| gpt-oss-20b | 98.6 | 0.63 | 83.3 | 0.49 | 735 |
| Gemma-3-4B-it† | 92.2 | 0.43 | 0.9 | 0.29 | 115 |
| Gemma-3-12B-it | 83.7 | 0.62 | 0.0 | 0.29 | 901 |
| Gemma-3-27B-it | 100.0 | 0.18 | 0.0 | 0.00 | 901 |

Note: Human shares and N refer to the oil price scenario in Wave 3 of Andre et al. (2022), which elicited rise scenarios only; the human SD in the denominator of the SD ratio is that of their general-public sample in Waves 1 and 2. For each model and variable, the lowest and highest 5% of valid responses are dropped, as Andre et al. (2022) do for their dispersion statistics; N is the number of responses that remain, and all LLM statistics are computed on this sample. †Gemma-3-4B-it returned valid responses for only a small fraction of the 999 personas; its row is reported for completeness and is not comparable with the others.

For inflation expectations, LLM agents show comparable or higher rates of "Rise" responses than the general public. The models reproduce the textbook direction—oil price shocks generate cost-push inflation—at least as consistently as human survey respondents. In three of them every persona answers "Rise": complete agreement is not a closer match to the human pattern but an absence of the directional heterogeneity a simulation would use. For unemployment expectations the models differ widely: some show comparable or higher "Rise" rates than humans, while others produce far fewer, or almost none. Different model families embed different subjective models of the relationship between supply shocks and unemployment, and the measurement separates them. Where the unemployment responses have no variance at all, as for the 27B model, what the measurement records is an absence of response rather than a distinct subjective model.

The SD ratio column quantifies dispersion relative to the human benchmark. For inflation expectations it ranges from 0.18 to 1.00 across models: some models achieve human-level variance, while others show severe homogeneity. For unemployment expectations it is uniformly low.

### B.3 Heterogeneity Across Sociodemographic Attributes

To assess whether LLMs generate attribute-dependent response patterns, we regress inflation expectation changes ($\Delta\pi$) on persona attributes. Table B.2 presents results from Gemma-3-12B-it, the model designated as primary for the survey-matched experiments.

Table B.2: Regression of $\Delta\pi$ on Persona Attributes (Gemma-3-12B-it)

| Variable | Coefficient | Std. Error |
|---|---|---|
| Age 20s (vs. 40s) | −0.17** | (0.07) |
| Age 30s | −0.08 | (0.06) |
| Age 50s | 0.14** | (0.06) |
| Age 60s | 0.18** | (0.08) |
| Age 70+ | 0.28* | (0.15) |
| Female | 0.06 | (0.04) |
| Married | 0.22*** | (0.05) |
| Income \$20–30k (vs. <\$10k) | −0.58*** | (0.13) |
| Income \$60–75k | −0.65*** | (0.14) |
| Income \$100–150k | −0.59*** | (0.14) |
| Controls | Education, Race, Employment, State | |
| $R^2$ | 0.23 | |
| N | 999 | |

Note: ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively. Standard errors are heteroskedasticity-robust (HC1) and reflect variation across generated responses, so the significance levels indicate generation-level precision, not sampling uncertainty with respect to a human population. Unlike Table B.1, the regressions here use the untrimmed sample to preserve cell sizes across attribute categories; their N therefore differs from Table B.1.

The results reveal a clear age gradient: younger personas (20s) show significantly lower $\Delta\pi$ relative to the 40s reference category, while older personas (50s, 60s, 70+) show progressively higher values. This pattern is directionally consistent with the literature documenting that older individuals tend to hold higher inflation expectations, an effect often attributed to their experience of past inflationary episodes (Malmendier & Nagel, 2016). The income coefficients are negative relative to the lowest income category across the categories shown: higher-income personas expect smaller inflation increases following the oil shock, consistent with prior findings (Bryan & Venkatu, 2001). The gender effect is positive but not statistically significant, qualitatively consistent with the gender gap documented in human surveys (D'Acunto et al., 2021).

## Appendix C: The Same Measurement at Other Model Sizes

Table C.1 reports the key metrics for the three Gemma-3 sizes under the demographics-only configuration; the relationship between model size and performance is not monotonic: the

12B model attains the highest tracking coefficient and SD ratio, while the 27B model shows more severe variance compression. Larger scale within a family does not by itself improve the reproduction of human attribute–expectation relationships. Two qualitative conclusions of Section 4—variance compression and weak individual-level correspondence—hold at all three sizes. The comparison covers the demographics-only configuration for one-year-ahead expectations.

Table C.1: Key Metrics by Model Size ($\pi_1$, demographics-only configuration)

| Metric | Gemma-3-4B-it | Gemma-3-12B-it | Gemma-3-27B-it |
|---|---|---|---|
| $\theta_1$ (LLM on human) | 0.006 | 0.012 | 0.002 |
| $R^2$ | 0.1% | 0.2% | 0.2% |
| SD Ratio | 0.21 | 0.23 | 0.04 |
| Efficiency | 27% | 50% | 7% |

Note: Valid-response rates differ across model sizes, so the matched sample, and with it the human-side benchmark $\theta_{OLS}$, differs by column; the efficiency column depends on that benchmark and is not compared across columns. All three columns use one generation per persona, so the 12B column differs from the 30-generation averages of Table 2.

## Appendix D: Shock Responses and Information Frictions

This appendix explores whether cognitive conditioning shapes responses to economic news, adapting the oil price scenario of Appendix B to a survey-respondent framing. The shock scenario states: "Today, a sudden production disruption occurred in the Middle East, and news reports indicate that the prices of all petroleum products will rise significantly from tomorrow onward. Furthermore, the news indicates that the problem will continue for at least the next 12 months." The baseline scenario describes stable oil production with prices expected to remain at current levels; the scenarios were administered in Japanese and are translated here. We use configuration (6), not the best-performing configuration but the only one in which all three cognitive variables enter the persona together, and regress shock responses on cognitive-variable dummies with demographic controls (Table D.1):

$$\Delta y_i = \alpha + \sum_j \gamma_j C_{ij} + \sum_k \beta_k W_{ik} + \varepsilon_i, \tag{D.1}$$

where $C_{ij}$ are the cognitive variables and $W_{ik}$ the demographic attributes. No matched human experiment exists for these scenarios, so the analysis establishes what the model does rather than whether humans do the same.

The average responses are $\Delta\pi_p$ = 1.81 (SD = 3.85), $\Delta\pi_1$ = 6.53 (SD = 3.68) and $\Delta\pi_5$ = 0.28 (SD = 1.53); the largest is at one year, consistent with the shock's stated duration. A one-

year effect enters the five-year average annual rate at about a fifth of its weight, but the scenario states only that the disruption lasts at least 12 months, so this is a reference point rather than an implied value; the observed five-year response is below it in either case. Because the news framing does not quantify the price increase, the differentials across cognitive groups rather than the levels are the quantities of interest, and because each $\Delta$ is the difference of two generated responses, only group means and their differentials, not individual values, are interpretable, because averaging over hundreds of personas per cognitive category shrinks the noise.

Table D.1: Shock Response by Cognitive Variable (Config. 6)

| Variable | Category (ref.: Know / Yes) | $\Delta\pi_p$ | $\Delta\pi_1$ | $\Delta\pi_5$ |
|---|---|---|---|---|
| Q21(1): BOJ mandate | Heard but do not know well | −0.12 | +1.00*** | +0.24*** |
| Q21(1): BOJ mandate | Never heard | +3.03*** | +1.19*** | +0.90*** |
| Q21(2): 2% target | Heard but do not know well | +0.50** | +1.36*** | +0.13 |
| Q21(2): 2% target | Never heard | +0.46** | +1.23*** | −0.15 |
| Q22: CPI image | No image | −0.57** | +0.37 | −0.18* |

Note: Reference categories are "Know" for Q21(1) and Q21(2) and "Yes" for Q22. Demographic controls included; the sample consists of the 1,939 respondents with complete attribute and cognitive items, whether or not they answered the expectation questions, since no human responses enter the comparison; one generation per persona and condition. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively. Standard errors are heteroskedasticity-robust (HC1) and reflect variation across generated responses, so the significance levels indicate generation-level precision, not sampling uncertainty with respect to a human population. Each Δ is the difference of two single generations, so generation noise enters twice.

Personas described as less aware of monetary policy revise one-year expectations more strongly: for Q21(1), "never heard" personas respond about 1.2 percentage points more than "know" personas, and Q21(2) shows the same broad pattern, though not monotonically. The direction is consistent with information friction logic, in that better-informed agents may anticipate a stabilizing policy response, but a role-playing reading is also available: a model told that a persona is uninformed may generate more extreme numbers.

One result bears on the internal consistency of the simulated agent. The shock begins "tomorrow onward" and cannot affect inflation over the past year, yet the average revision of perceived inflation is positive ($\Delta\pi_p$ = 1.81) and "never heard" personas revise it by 3.03 percentage points more than "know" personas. This departs from the logic that news about the future should leave a backward-looking report unrevised, though salient news can spill over onto reports about the past for human respondents as well.

## Appendix E: Estimation Details

This appendix collects the estimation detail underlying the statistics defined in Sections 3.4 and 3.5.

Temperature sweep. Table E.1 gives the variance decomposition of Section 3.5 for the three outcomes. The demographics-only personas of 274 respondents drawn at random from the trimmed sample of Table 2 were generated 10 times each at temperatures 0.5, 1.0, 1.5, and 2.0, and twice at temperature 0, where decoding is deterministic and the within-attribute component is zero by construction; the temperature-1.0 rows use the first 10 of the 30 generations of the main run for the same personas, and on the full sample with all 30 generations they reproduce the between-attribute variances of Table E.3. At temperature 2.0 a few extreme answers dominate total variance, and for perceived inflation the within-attribute estimate exceeds the total, so the between-attribute component is set to zero there. Because attribute profiles are almost unique in this sample, the within-attribute component coincides with the within-persona component.

Table E.1: Variance Decomposition Across Temperature Settings (Survey-Matched Design, Configuration (1))

| Outcome | Temp. | Total Var. | Within-Attr. Var. | Between-Attr. Var. [95% CI] | Systematic SD Ratio [95% CI] |
|---|---|---|---|---|---|
| $\pi_p$ | 0.0 | 7.87 | 0.00 | 7.87 [6.87, 8.92] | 0.200 [0.187, 0.213] |
| $\pi_p$ | 0.5 | 7.82 | 0.83 | 6.99 [6.06, 7.91] | 0.189 [0.176, 0.201] |
| $\pi_p$ | 1.0 | 8.53 | 2.39 | 6.13 [5.28, 6.98] | 0.177 [0.164, 0.189] |
| $\pi_p$ | 1.5 | 11.59 | 5.99 | 5.61 [4.77, 6.37] | 0.169 [0.156, 0.180] |
| $\pi_p$ | 2.0 | 152.43 | 169.13 | 0.00 [0.00, 8.24] | 0.000 [0.000, 0.205] |
| $\pi_1$ | 0.0 | 6.37 | 0.00 | 6.37 [5.17, 7.41] | 0.192 [0.173, 0.207] |
| $\pi_1$ | 0.5 | 7.23 | 3.05 | 4.18 [3.44, 4.90] | 0.156 [0.141, 0.169] |
| $\pi_1$ | 1.0 | 8.14 | 5.14 | 3.00 [2.42, 3.60] | 0.132 [0.119, 0.145] |
| $\pi_1$ | 1.5 | 11.20 | 9.07 | 2.13 [1.69, 2.56] | 0.111 [0.099, 0.122] |
| $\pi_1$ | 2.0 | 39.07 | 36.89 | 2.18 [1.24, 3.12] | 0.113 [0.085, 0.135] |
| $\pi_5$ | 0.0 | 0.95 | 0.00 | 0.95 [0.88, 0.99] | 0.076 [0.074, 0.078] |
| $\pi_5$ | 0.5 | 0.94 | 0.30 | 0.64 [0.58, 0.70] | 0.063 [0.060, 0.065] |
| $\pi_5$ | 1.0 | 1.05 | 0.57 | 0.48 [0.43, 0.53] | 0.054 [0.051, 0.057] |
| $\pi_5$ | 1.5 | 2.04 | 1.67 | 0.36 [0.28, 0.44] | 0.047 [0.042, 0.052] |
| $\pi_5$ | 2.0 | 4.54 | 4.24 | 0.30 [0.18, 0.47] | 0.043 [0.033, 0.053] |

Note: Variances in squared percentage points, 274 personas, 10 generations per persona (two at temperature 0). Bootstrap 95% intervals over personas in brackets. The systematic SD ratio is the square root of the between-attribute component divided by the human standard deviation of the trimmed sample of Table 2 (14.0, 13.1, and 12.8 for the three outcomes). The valid-response rate is at least 99.9% up to temperature 1.5 and 97.6% at temperature 2.0.

Sign agreement and coefficient equality. Because coefficients from the same regression are dependent, sign agreement rates are tested by a permutation test that shuffles the LLM responses. Two limitations of the statistic are worth naming. Agreement on coefficients that are themselves noise is uninformative, so the rate is read together with the significance of the human coefficients rather than on its own; and because dummy-coefficient signs depend on the

choice of reference categories, the rate is sensitive to that choice, which Section 4.3 quantifies by reporting how often significance survives across 100 alternative base-category specifications. Because the comparison is between coefficient vectors, the sign-agreement statistic is computed once on the LLM responses averaged over the 30 generations rather than averaged over per-generation regressions. Table E.2 reports the coefficients of the parallel attribute regressions of Section 4.3 for configuration (1), from which Figure 3 and the sign agreement rates are computed.

Table E.2: Attribute Coefficients of the Parallel Regressions (Configuration (1))

| Attribute | $\pi_p$ Human | $\pi_p$ LLM | $\pi_1$ Human | $\pi_1$ LLM | $\pi_5$ Human | $\pi_5$ LLM |
|---|---|---|---|---|---|---|
| Female | 2.141 | 1.030 | 0.451 | 0.739 | −1.503 | 0.336† |
| | (0.723) | (0.047) | (0.695) | (0.040) | (0.671) | (0.013) |
| Age 20s | −2.992 | −1.365 | −1.515 | −0.845 | −0.965 | −0.184 |
| | (1.380) | (0.144) | (1.315) | (0.101) | (1.328) | (0.027) |
| Age 30s | −0.392 | −0.660 | 0.856 | −0.579† | 1.641 | −0.100† |
| | (1.155) | (0.091) | (1.071) | (0.066) | (1.251) | (0.020) |
| Age 50s | 2.406 | 0.377 | 1.909 | 0.490 | −0.041 | 0.174† |
| | (1.015) | (0.065) | (1.019) | (0.054) | (0.955) | (0.018) |
| Age 60s | 2.004 | 0.762 | 0.305 | 0.715 | 0.976 | 0.261 |
| | (1.222) | (0.077) | (1.055) | (0.063) | (1.095) | (0.020) |
| Age 70+ | 0.062 | 1.447 | −0.282 | 1.012† | −1.153 | 0.415† |
| | (1.355) | (0.091) | (1.244) | (0.077) | (1.209) | (0.025) |
| Self-employed | 0.511 | 1.706 | −0.066 | 1.455† | −1.641 | 0.606† |
| | (1.088) | (0.096) | (1.186) | (0.081) | (1.125) | (0.029) |
| Non-regular employee | 1.155 | 3.437 | 1.088 | 2.753 | −1.221 | 1.120† |
| | (1.058) | (0.071) | (1.032) | (0.059) | (0.861) | (0.018) |
| Agriculture, forestry, fishery | −0.405 | 2.022† | 1.217 | 1.549 | −0.815 | 0.865† |
| | (3.029) | (0.243) | (3.023) | (0.161) | (1.563) | (0.055) |
| Not employed | 1.068 | 2.928 | 1.378 | 1.867 | −0.528 | 0.880† |
| | (1.163) | (0.081) | (1.091) | (0.065) | (1.091) | (0.021) |
| Income ¥3–5M | −1.002 | −0.992 | −0.322 | 0.114† | 0.010 | −0.148† |
| | (0.920) | (0.069) | (0.879) | (0.055) | (0.804) | (0.016) |
| Income ¥5–10M | −2.925 | −1.665 | −2.046 | −0.379 | −0.940 | −0.313 |
| | (0.971) | (0.080) | (0.873) | (0.059) | (0.836) | (0.019) |
| Income ¥10M+ | −2.633 | −1.858 | −2.249 | −0.797 | −1.387 | −0.654 |
| | (1.432) | (0.123) | (1.250) | (0.092) | (1.201) | (0.029) |
| No income | 3.682 | 0.038 | 3.792 | 2.046 | 4.193 | 0.170 |
| | (4.158) | (0.200) | (4.415) | (0.211) | (3.798) | (0.040) |
| Has borrowing | 1.205 | 1.342 | 0.946 | 1.196 | 0.508 | 0.287 |
| | (0.799) | (0.055) | (0.747) | (0.044) | (0.717) | (0.013) |
| Couple only | 0.860 | −0.172† | 0.289 | 0.017 | −1.824 | −0.020 |
| | (1.116) | (0.080) | (1.068) | (0.069) | (1.094) | (0.019) |
| Two generations | 0.033 | −0.102† | −0.510 | −0.159 | −1.696 | −0.011 |
| | (1.037) | (0.080) | (1.033) | (0.064) | (1.095) | (0.017) |
| Three generations | 2.119 | 0.214 | 0.653 | −0.132† | −0.749 | 0.020† |
| | (1.638) | (0.098) | (1.641) | (0.092) | (1.791) | (0.026) |
| Other household | 1.147 | −0.190† | 2.550 | −0.308† | −0.569 | −0.005 |
| | (2.164) | (0.167) | (2.683) | (0.129) | (2.481) | (0.040) |
| Hokkaido | 0.088 | 0.262 | 2.538 | 0.092 | 1.383 | 0.068 |
| | (1.581) | (0.093) | (1.627) | (0.087) | (1.386) | (0.028) |
| Tohoku | 0.866 | 0.535 | 1.126 | 0.495 | 0.156 | 0.216 |
| | (1.439) | (0.082) | (1.363) | (0.075) | (1.168) | (0.025) |
| Hokuriku | −2.362 | −0.064 | 0.102 | −0.330† | −0.250 | −0.079 |
| | (1.567) | (0.105) | (1.511) | (0.085) | (1.337) | (0.024) |
| Chubu | −1.808 | −0.220 | −0.769 | −0.255 | −0.420 | 0.004† |
| | (0.951) | (0.073) | (0.896) | (0.060) | (0.883) | (0.016) |
| Kinki | −1.409 | −0.104 | −0.521 | −0.424 | 0.843 | −0.104† |
| | (0.921) | (0.069) | (0.854) | (0.056) | (0.992) | (0.017) |
| Chugoku | −1.174 | −0.927 | −0.847 | −0.316 | 1.432 | 0.000 |
| | (1.874) | (0.142) | (1.235) | (0.122) | (1.571) | (0.025) |
| Shikoku | 0.499 | 0.282 | −0.631 | 0.040† | −0.834 | 0.081† |
| | (2.230) | (0.156) | (2.018) | (0.112) | (1.649) | (0.039) |
| Kyushu | 1.101 | 0.335 | 2.136 | 0.186 | 0.611 | 0.085 |
| | (1.278) | (0.082) | (1.391) | (0.065) | (1.099) | (0.021) |

Note: The human regression uses the trimmed sample of Table 2 (N = 1,744); the LLM regression uses each persona's mean response over 30 generations, on the same attributes. Heteroskedasticity-robust (HC1) standard errors in parentheses; for the LLM they are conditional on the averaged generations. Reference categories: male, age 40s, regular employee, income below ¥3 million, no borrowing, single-person household, and Kanto. † marks an LLM coefficient whose sign differs from the human coefficient; the unmarked coefficients are the 23, 19, and 14 sign agreements of Section 4.3.

Residual association. The within-cell permutation check of Section 4.4 shuffles the LLM residuals within age-by-gender cells, because cells defined by the full attribute vector are too sparse; since the exchangeability this requires does not hold across income, region, and employment, the procedure is read as a check rather than as a formal test of person-specific correspondence.

The additive benchmark. $\widetilde{R}^2_{H|W}$ is the share of human variance explained by the conditional mean of expectations given an attribute profile. With seven attribute dimensions most cells contain a single respondent, so a saturated-cell fit would approach one mechanically and cannot serve as a benchmark, and the reference values of Section 4.4 substitute the $R^2$ of the additive attribute regression for $\widetilde{R}^2_{H|W}$. The substitution errs in both directions: omitting interactions between dimensions understates the reference value, while in-sample fitting optimism overstates it, so the figures are read as additive benchmarks rather than as bounds.

The efficiency benchmark and its estimation. The denominators are in-sample $R^2$ values from a regression with roughly 27 dummy regressors on about 1,700 observations. Under the null of no attribute structure such a regression returns an $R^2$ of about $k/N \approx 0.016$ from fitting noise alone, the same order as the observed benchmarks (0.039, 0.023, 0.019); the conventional degrees-of-freedom adjustment gives roughly 0.024, 0.008, and 0.004. Efficiency measures the share of attribute-predictable covariance recovered, not coefficient equality or distributional similarity, and rescaling LLM outputs changes it without changing their information content; the substantive statements therefore rest on the benchmark-free quantities.

Reverse regression and attenuation. We report $\theta_2$, the slope from regressing human on LLM expectations, and $R^2 = \theta_1\theta_2$. Generation noise inflates $\mathrm{Var}(y^L)$ without contributing covariance, which attenuates $\theta_2$ and $R^2$ as classical measurement error would, whereas $\theta_1$ and hence efficiency are unbiased by it. A second, opposing force is mechanical and dominates at the dispersion levels observed: the reverse regression divides by the compressed $\mathrm{Var}(y^L)$ ($\theta_2 = \theta_1/\text{SD Ratio}^2$), so $\theta_2$ exceeds $\theta_1$ by an order of magnitude in the tables because the LLM's variance is compressed, not because the model tracks humans more strongly in that direction. Averaging the LLM side over the 30 generations before computing the statistics shows the size of the attenuation (Table E.4): at baseline, for perceived inflation, the reverse slope rises from 0.50 in Table 2 to 0.69 on persona means and the product of the two slopes from 1.2% to 1.6%, while the tracking coefficient is unchanged, necessarily, since it is linear in the generated responses.

Generation-level variability and the multi-draw run. With one draw per persona an estimate of $\theta_1$ is one realization of a Monte Carlo experiment whose variability a bootstrap over respondents does not capture. Each of the six configurations of Section 5 was therefore generated 30 times per persona with the same prompts, model, and temperature setting as the analyses with one generation per persona; personas with fewer than 30 parsable responses (0 to 6 per configuration) are dropped so that every persona contributes the same number of draws. Tables 2 and 4 report means over the 30 generations with the draw-to-draw standard deviation, which is 0.003–0.005 for perceived inflation and one-year expectations and 0.001–0.002 for the five-year average; the baseline-to-configuration-(5) movement of Section 5.3, 0.011 to 0.026 for one-year expectations, is several times this standard deviation, whereas efficiency ranges from 0.12 to 0.80 across draws for one-year expectations at baseline (not tabulated). Table E.3 reports the range of the tracking coefficient over generations and the variance decomposition by configuration, and Table E.4 the correspondence statistics on per-persona means for the baseline configuration.

Table E.3: Generation-Level Variability and Variance Decomposition by Configuration (30 generations per persona)

| Outcome | Config. | $\theta_1$: range | Systematic SD Ratio [95% CI] | Between-attr. var. | Δ vs (1) [95% CI] |
|---|---|---|---|---|---|
| $\pi_p$ | (1) | 0.0186 to 0.0296 | 0.182 [0.169, 0.197] | 6.54 | – |
| $\pi_p$ | (2) | 0.0219 to 0.0339 | 0.225 [0.208, 0.243] | 9.92 | 3.38 [2.96, 3.80] |
| $\pi_p$ | (3) | 0.0172 to 0.0292 | 0.224 [0.208, 0.242] | 9.89 | 3.35 [2.95, 3.79] |
| $\pi_p$ | (4) | 0.0259 to 0.0461 | 0.286 [0.267, 0.308] | 16.07 | 9.54 [8.88, 10.17] |
| $\pi_p$ | (5) | 0.0262 to 0.0383 | 0.238 [0.222, 0.256] | 11.09 | 4.55 [4.10, 5.02] |
| $\pi_p$ | (6) | 0.0249 to 0.0449 | 0.245 [0.227, 0.263] | 11.78 | 5.24 [4.60, 5.84] |
| $\pi_1$ | (1) | 0.0027 to 0.0185 | 0.134 [0.123, 0.148] | 3.11 | – |
| $\pi_1$ | (2) | 0.0100 to 0.0254 | 0.171 [0.156, 0.186] | 5.01 | 1.89 [1.61, 2.18] |
| $\pi_1$ | (3) | 0.0096 to 0.0262 | 0.221 [0.204, 0.243] | 8.42 | 5.31 [4.96, 5.67] |
| $\pi_1$ | (4) | 0.0029 to 0.0187 | 0.169 [0.155, 0.185] | 4.94 | 1.83 [1.56, 2.09] |
| $\pi_1$ | (5) | 0.0170 to 0.0366 | 0.222 [0.204, 0.243] | 8.46 | 5.34 [4.94, 5.73] |
| $\pi_1$ | (6) | 0.0121 to 0.0328 | 0.209 [0.192, 0.229] | 7.54 | 4.42 [3.97, 4.85] |
| $\pi_5$ | (1) | −0.0057 to −0.0006 | 0.057 [0.052, 0.063] | 0.52 | – |
| $\pi_5$ | (2) | −0.0036 to 0.0020 | 0.069 [0.063, 0.076] | 0.78 | 0.25 [0.22, 0.29] |
| $\pi_5$ | (3) | −0.0052 to −0.0001 | 0.067 [0.061, 0.074] | 0.73 | 0.21 [0.18, 0.23] |
| $\pi_5$ | (4) | −0.0058 to −0.0002 | 0.061 [0.056, 0.067] | 0.60 | 0.08 [0.06, 0.11] |
| $\pi_5$ | (5) | −0.0044 to 0.0021 | 0.073 [0.067, 0.080] | 0.86 | 0.34 [0.31, 0.38] |
| $\pi_5$ | (6) | −0.0054 to 0.0026 | 0.067 [0.061, 0.075] | 0.74 | 0.22 [0.18, 0.26] |

Note: Column 3 gives the range over the 30 generations of the tracking coefficient of Tables 2 and 4. The systematic SD ratio has the square root of the between-attribute variance of Section 3.5 in the numerator and the total human standard deviation in the denominator, with bootstrap 95% intervals over respondents. The last two columns give the between-attribute variance in squared percentage points and its paired-bootstrap difference from configuration (1) on personas common to both configurations; all fifteen intervals exclude zero. Sample as in Table 4.

Table E.4: Correspondence Statistics on Persona Means over 30 Generations (Configuration (1))

| Outcome | Config. | $\theta_1$ (SE) | $\theta_2$ | $R^2$ | Sign agreement (p) | Residualized corr. (p) |
|---|---|---|---|---|---|---|
| $\pi_p$ | (1) | 0.0231 (0.0044) | 0.685 | 1.6% | 23/27 (0.001) | −0.006 (0.79) |
| $\pi_1$ | (1) | 0.0108 (0.0033) | 0.566 | 0.6% | 19/27 (0.019) | −0.011 (0.63) |
| $\pi_5$ | (1) | −0.0024 (0.0015) | −0.712 | 0.2% | 14/27 (0.494) | 0.033 (0.18) |

Note: The LLM side is the mean of 30 generations per persona; the human side and the sample are as in Table E.3. Standard errors are heteroskedasticity-robust (HC1), over respondents, conditional on the averaged generations. Sign agreement is the number of attribute coefficients with the same sign in the human and LLM regressions over the number compared; p-values are from the permutation test of Section 4.3. Residualized correlation is the in-sample correlation after additive adjustment for the row's conditioning set, with p-values from the within-cell permutation check of Section 4.4.

## Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the authors used ChatGPT and Codex (OpenAI), Claude (Anthropic), and Gemini (Google) in order to improve the language and readability of the manuscript, check consistency between the text, tables, and references, and assist with writing and checking analysis and figure-generating code. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

## Acknowledgements

The authors would like to thank participants at the 2025 Banque de France FFJ LAB Winter Conference, the 2026 BOK/ERI-BOJ/IMES Research Workshop, WEHIA 2026, CEBRA Annual Meeting 2026, and seminars at the Bank of Japan for many helpful comments and suggestions. The views expressed in this paper are those of the authors and do not necessarily reflect the official views of the Bank of Japan. Both authors are employees of the Bank of Japan, and this research was conducted as part of their work there. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The household survey microdata were accessed in the authors' capacity as Bank of Japan staff. The authors have no other financial relationships or potential conflicts of interest relevant to this paper.

## References

Andre, P., Pizzinelli, C., Roth, C., & Wohlfart, J. (2022). Subjective models of the macroeconomy: Evidence from experts and representative samples. *Review of Economic Studies, 89*(6), 2958–2991.

Anesti, N., Hill, E., & Joseph, A. (2026). Inflation attitudes of large language models. Bank of England Staff Working Paper No. 1,190.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis, 31*(3), 337–351.

Bisbee, J., Clinton, J., Dorff, C., Kenkel, B., & Larson, J. (2024). Synthetic replacements for human survey data? The perils of large language models. *Political Analysis, 32*(4), 401–416.

Bryan, M. F., & Venkatu, G. (2001). The demographics of inflation opinion surveys. *Federal Reserve Bank of Cleveland Economic Commentary*.

Bybee, L. (2023). Surveying generative AI's economic expectations. SSRN Working Paper No. 4430515.

Coibion, O., & Gorodnichenko, Y. (2015). Information rigidity and the expectations formation process: A simple framework and new facts. *American Economic Review, 105*(8), 2644–2678.

D'Acunto, F., Hoang, D., Paloviita, M., & Weber, M. (2019). Cognitive abilities and inflation expectations. *AEA Papers and Proceedings, 109*, 562–566.

D'Acunto, F., Malmendier, U., Ospina, J., & Weber, M. (2021). Exposure to grocery prices and inflation expectations. *Journal of Political Economy, 129*(5), 1615–1639.

D'Acunto, F., Malmendier, U., & Weber, M. (2023). What do the data tell us about inflation expectations? In R. Bachmann, G. Topa, & W. van der Klaauw (Eds.), Handbook of Economic Expectations (pp. 133–161). Elsevier.

Diamond, J., Watanabe, K., & Watanabe, T. (2020). The formation of consumer inflation expectations: New evidence from Japan's deflation experience. *International Economic Review, 61*(1), 241–281.

Fagiolo, G., & Roventini, A. (2017). Macroeconomic policy in DSGE and agent-based models redux: New developments and challenges ahead. *Journal of Artificial Societies and Social Simulation, 20*(1), 1.

Farmer, J. D. (2003). Avoiding getting lost in the wilderness of bounded rationality: The path from zero intelligence to no arbitrage. Quantitative Finance, 3(4), C64–C65.

Farmer, J. D. (2024). *Making sense of chaos: A better economics for a better world*. Yale University Press.

Farmer, J. D., & Foley, D. (2009). The economy needs agent-based modelling. *Nature, 460*(7256), 685–686.

Gao, C., Lan, X., Li, N., Yuan, Y., Ding, J., Zhou, Z., Xu, F., & Li, Y. (2024). Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications, 11*, 1259.

Hansen, A. L., Horton, J. J., Kazinnik, S., Puzzello, D., & Zarifhonarvar, A. (2025). Simulating the Survey of Professional Forecasters. SSRN Working Paper No. 5066286.

Hobijn, B., & Lagakos, D. (2005). Inflation inequality in the United States. *Review of Income and Wealth, 51*(4), 581–606.

Horton, J. J., Filippas, A., & Manning, B. S. (2023). Large language models as simulated economic agents: What can we learn from Homo silicus? NBER Working Paper No. 31122.

Huang, J., Ferreras Garrucho, D., Xie, Y., Yuan, W. M., Mei, Q., Lian, C., & Hazell, J. (2026). HouseholdBench: Evaluating large language models as predictors of household economic behavior. arXiv:2610.07563.

Kaplan, G., Moll, B., & Violante, G. L. (2018). Monetary policy according to HANK. *American Economic Review, 108*(3), 697–743.

Larooij, M., & Törnberg, P. (2026). Validation is the central challenge for generative social simulation: A critical review of LLMs in agent-based modeling. *Artificial Intelligence Review, 59*, 15.

Li, N., Gao, C., Li, M., Li, Y., & Liao, Q. (2024). EconAgent: Large language model-empowered agents for simulating macroeconomic activities. In *Proceedings of the 62nd*

*Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 15523–15536). Association for Computational Linguistics.

Lin, J., Sun, L., & Yan, Y. (2026). Simulating macroeconomic expectations in survey experiments with LLM-based economic agents. arXiv:2505.17648v5 (first circulated May 2025).

Lopez-Lira, A., Tang, Y., & Zhu, M. (2025). The memorization problem: Can we trust LLMs' economic forecasts? arXiv:2504.14765.

Malmendier, U., & Nagel, S. (2016). Learning from inflation experiences. *Quarterly Journal of Economics, 131*(1), 53–87.

Mankiw, N. G., & Reis, R. (2002). Sticky information versus sticky prices: A proposal to replace the New Keynesian Phillips curve. *Quarterly Journal of Economics, 117*(4), 1295–1328.

Sargent, T. J. (1993). *Bounded rationality in macroeconomics: The Arne Ryde memorial lectures*. Oxford University Press.

Sims, C. A. (1980). Macroeconomics and reality. Econometrica, 48(1), 1–48.

Sims, C. A. (2003). Implications of rational inattention. *Journal of Monetary Economics, 50*(3), 665–690.

Takahashi, Y., & Tamanyu, Y. (2022). Households' perceived inflation and CPI inflation: The case of Japan. Bank of Japan Working Paper Series No. 22-E-1.

Takahashi, Y., Otaka, K., & Kato, N. (2025). Potential applications of generative AI in economic simulations. Bank of Japan Research LAB No. 25-E-1.

Tesfatsion, L. (2006). Agent-based computational economics: A constructive approach to economic theory. In L. Tesfatsion & K. L. Judd (Eds.), *Handbook of computational economics* (Vol. 2, pp. 831–880). Elsevier.

Weber, M., D'Acunto, F., Gorodnichenko, Y., & Coibion, O. (2022). The subjective inflation expectations of households and firms: Measurement, determinants, and implications. *Journal of Economic Perspectives, 36*(3), 157–184.

Wu, J. C., Xi, J., & Xie, S. (2025). LLM survey framework: Coverage, reasoning, dynamics, identification. NBER Working Paper No. 34308.

Zarifhonarvar, A. (2026). Generating inflation expectations with large language models. *Journal of Monetary Economics, 157*, 103859.