NBER WORKING PAPER SERIES

LLM SURVEY FRAMEWORK:
COVERAGE, REASONING, DYNAMICS, IDENTIFICATION

Jing Cynthia Wu
Jin Xi
Shihan Xie

Working Paper 34308
http://www.nber.org/papers/w34308

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2025

The authors received no external funding for this research and have no relevant financial relationships to disclose. We thank Drew Creal, Yuriy Gorodnichenko, Marcelo Medeiros, Pedro Simon, Haitian Xie, as well as seminar and conference participants at Midwest Econometrics Group Conference for helpful comments. An earlier version of this paper was circulated under the title “LLM Survey Framework: Coverage, Consistency, Identification.” The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

LLM Survey Framework: Coverage, Reasoning, Dynamics, Identification
Jing Cynthia Wu, Jin Xi, and Shihan Xie
NBER Working Paper No. 34308
October 2025
JEL No. C83, E31, E52

ABSTRACT

We propose a new LLM-based survey framework that enables retrospective coverage, economic reasoning, dynamic effects, and clean identification. We recover human-comparable treatment effects in a multi-wave randomized controlled trial of inflation expectations surveys, at 1/1000 the cost. To demonstrate the framework's full potential, we extend the benchmark human survey (10 waves, 2018–2023) to over 50 waves dating back to 1990. We further examine the economic mechanisms underlying agents' expectation formation, identifying the mean-reversion and individual-attention channels. Finally, we trace dynamic treatment effects and demonstrate clean identification. Together, these innovations demonstrate that LLM surveys enable research designs unattainable with human surveys.

Jing Cynthia Wu
University of Illinois Urbana-Champaign
Department of Economics
and NBER
j.cynthia.wu@gmail.com

Jin Xi
Chinese Academy of Sciences
Academy of Mathematics and
Systems Science
Center for Forecasting Science
xijin@amss.ac.cn

Shihan Xie
University of Illinois Urbana-Champaign
shihan.xie@gmail.com

1 Introduction

Surveys play a central role in economics, offering direct measures of expectations not observable in market or administrative data (Manski, 2004). They are widely used to study monetary policy communication (Coibion et al., 2022), firm price-setting behavior (Coibion et al., 2020), investor beliefs (Giglio et al., 2021), and household decisions in housing and labor markets (Armona et al., 2019; Faberman et al., 2022). However, traditional human surveys, while invaluable, face significant limitations. Most fundamentally, they cannot be fielded retrospectively, placing many counterfactuals and historical settings permanently out of reach. Moreover, human surveys usually record numerical or categorical answers without the reasoning behind them. In practice, they often rely on access to institutional channels such as funding, survey firms, or established panels, which creates barriers for many researchers and limits the scope and sample size of each survey. We propose a survey design built on large language models (LLMs) to address these limitations.

Our new LLM-based survey framework is broadly applicable for studying expectation formation: it replicates the essential features of human surveys while enabling the following capabilities that traditional methods cannot. First, it provides retrospective coverage, allowing surveys to extend systematically across decades and at higher frequencies. Second, it leverages LLMs' native capacity for reasoning to explain underlying economic mechanisms. Third, it traces the dynamic responses of the same agents over time. Fourth, it achieves clean identification by separating prior knowledge from subsequent factual treatments. Finally, it delivers these capabilities at roughly 0.5¢ per response, or 1/1000 the cost of a standard online survey platform. Collectively, these features create a survey framework that is ideal in design yet unattainable with human respondents.

A key challenge of LLMs is hindsight bias, a well-documented tendency of LLMs to incorporate future knowledge when asked about past states of the world (Roberts et al., 2024; Liu et al., 2025; Golchin and Surdeanu, 2024). Methodologically, our framework introduces a date-restrictive prompting design tailored to the survey setting. This design freezes the LLM agent's knowledge to information available up to any specified date. We then demonstrate that date restriction operates as intended by testing LLMs' awareness of salient events between 2001 and 2020, such as 9/11, the Iraq war, and the onset of COVID-19. As expected, LLMs report no awareness before these events and near-universal recognition afterward, confirming that our design prevents hindsight bias while preserving historically grounded awareness.

Another methodological foundation of our framework is internal consistency: LLMs allow researchers to fix respondent personas across control and treatment groups to create a balanced synthetic panel, follow the same individuals over time, and hold treatments word-for-word identical across waves. Human surveys, by contrast, often suffer from sample drift, respondent dropout, and evolving question wording. Economically, it ensures that changes in responses reflect genuine expectation updating. Econometrically, it reduces estimator variance and improves precision relative to simple randomization, a result we formally prove.

Building on the two methodological foundations, our survey framework extends beyond the limits of human surveys through the following key innovations and benefits. Date restriction enables retrospective coverage across decades and clean identification of treatment effects. Internal consistency allows us to trace dynamic responses of the same agents over time. In addition, the framework leverages LLMs' inherent reasoning capability to analyze the economic mechanisms underlying expectation formation, and their affordability to broaden access to a wider community of researchers. Together, these features create an ideal survey environment beyond the reach of human respondents.

To validate our LLM survey framework, we replicate the state-of-the-art multi-wave RCT of inflation expectations in Weber et al. (2025) within an LLM-based environment. We draw a random sample of 200 personas from the New York Fed's Survey of Consumer Expectations (SCE), holding them fixed across control/treatment arms and waves to ensure internal consistency. We impose date restrictions to reflect information available at the time of each wave and require each agent to articulate its reasoning before reporting a numerical forecast. We adopt the same treatments as in Weber et al. (2025): past inflation (T1), the Fed target (T2), and the Fed forecast (T3). The resulting LLM sample mirrors household demographics, and LLM priors closely align with those of human respondents. This setup enables a direct comparison between treatment effects in human and LLM surveys under parallel conditions.

The validation results show strong alignment: treatment effects are stronger during low-inflation episodes and weaker during high-inflation episodes. Correlations between treatment effects and inflation track or even exceed their human counterparts, ranging from 0.73 to 0.92 for LLM surveys and from 0.13 to 0.87 for human surveys. This contrast likely reflects reduced sampling noise from our internally consistent design. These findings underscore the value of our LLM survey framework as a credible complement to traditional survey experiments.

Next, we highlight the full potential of the LLM survey framework along four dimensions. First, we focus on retrospective coverage. To illustrate, we expand the benchmark human survey from 10 waves in 2018–23 to more than 50 waves dating back to 1990. Survey waves are fielded quarterly during turbulent episodes such as the Great Recession and the post-COVID inflation surge, and annually otherwise. The state dependence remains stable in the longer sample: the responsiveness of inflation expectations to new information declines during high-inflation periods and rises when inflation is low, although this correlation is somewhat attenuated.

Second, we leverage LLMs' native capacity to generate and classify reasoning in order to uncover the economic mechanisms underlying expectation formation. A mixed human–LLM coding strategy identifies three economic channels: mean reversion toward the long-run average, attention at both individual and aggregate levels, and business-cycle conditions. The mean-reversion and business-cycle channels emerge as new mechanisms revealed by our reasoning analysis. Although the attention channel has been discussed in prior work, our approach further decomposes it into individual and aggregate dimensions. Comparing high- and low-inflation environments, or examining dynamics over time, we find that two channels dominate: the mean-reversion and individual-attention channels. During high-inflation periods, agents perceive inflation as deviating from its long-run mean. Consequently, the mean-reversion channel places greater weight on priors that inflation will return toward that mean over time, thereby dampening responsiveness to informational treatments. At the same time, they rely more on personal price experiences and less on external information.

Third, we showcase dynamic treatment effects by implementing monthly follow-up surveys for up to twelve months after the initial treatment, asking the same agents to update their inflation expectations. We find that treatment effects decay as follows: after three months the treatment effects are about half the size of the initial response, become statistically insignificant by six months, and vanishes within a year. These results are consistent with human benchmarks where available, but our framework extends them systematically to horizons and frequencies that human surveys cannot feasibly cover.

Fourth, we incorporate clean identification by freezing knowledge at the survey date and using treatments released afterward. Implemented for the past-inflation treatment, although the overall economic conclusion remains unchanged in this experiment, clean and non-clean results are only moderately correlated at 0.54. This highlights that clean identification can introduce meaningful variation and becomes especially valuable when future information departs more sharply from past realizations.

The remainder of the paper, following a brief literature review, is organized as follows. Section 2 introduces our LLM-based survey framework. Section 3 validates the framework by replicating a benchmark inflation expectations experiment. Section 4 extends the analysis retrospectively over 35 years, demonstrating long-run coverage. Section 5 leveraging the LLM's reasoning capacity to uncover economic mechanisms behind expectation formation. Section 6 analyzes dynamic treatment effects and illustrates clean identification. Section 7 concludes.

1.1 Literature

Recent reviews highlight how generative AI is transforming research in economics (Korinek, 2024) and finance (Eisfeldt and Schubert, 2024). Our paper contributes to an emerging branch of this literature that uses LLMs as simulated agents in economic studies. In particular, Horton (2023) introduces the notion of *homo silicus*, showing that LLMs can reproduce canonical behavioral findings at low cost. Zarifhonarvar (2024) applies an LLM-based survey design to study household inflation expectations. Hansen et al. (2024) simulate the Survey of Professional Forecasters with synthetic LLM personas, which often outperform human forecasters. We advance this literature by moving beyond static LLM survey designs, introducing knowledge restrictions that align the model's information with each survey date and enable retrospective surveys across decades.

Beyond economics, a broader literature uses LLM agents to conduct experiments across the social sciences (Bail, 2024). Among them, Hewitt et al. (2024) demonstrate their ability to predict treatment effects in preregistered social science experiments, while Argyle et al. (2023) find that they reproduce realistic partisan and demographic patterns. Park et al. (2024) study group behaviors, highlighting the potential of LLMs as experimental participants at scale. At the same time, recent evaluations highlight limitations: Salecha et al. (2024) document social desirability bias, Bisbee et al. (2024) find that synthetic survey data understate variation across respondents, and Tjuatja et al. (2024) show that LLMs diverge from known human response biases in survey settings. Our contribution is both methodological and economic, developing new survey designs precisely to address questions of expectation formation central to macroeconomics.

Our paper also relates to a large literature using survey data to study how households form inflation expectations (see D'Acunto et al. (2023); Weber et al. (2022) for reviews).

Using the Nielsen scan panel, studies show that exposure to grocery prices influences household inflation expectations (D'Acunto et al., 2021) and establish causal effects of monetary policy communication (Coibion et al., 2022) and forward guidance (Coibion et al., 2023) on household beliefs. More broadly, various household survey panels—both long-running benchmarks (Michigan Survey of Consumers and New York Fed Survey of Consumer Expectations) and more recent innovations—have been widely used to understand inflation expectation formation (e.g., Malmendier and Nagel 2016; Cavallo et al. 2017; Binder 2017; Hajdini et al. 2024; Binder et al. 2024; Xie 2025). Recent work also leverages inflation expectations to measure attention (e.g., Bracha and Tang 2025; Weber et al. 2025), showing that households' responsiveness to information varies systematically with the inflation environment. Unlike human surveys, which can only be fielded once a questionnaire is developed, our LLM framework uses date-restricted knowledge sets to enable retrospective panels. Moreover, it delivers internal consistency across treatments and waves, something infeasible with human samples.

2 LLM-Based Survey Framework

In this section, we develop a new LLM-based survey framework built on two methodological foundations: Section 2.1 develops and validates the date-restriction design, and Section 2.2 introduces internal consistency.

Together, these two foundations power the key innovations detailed in Section 2.3: date restriction enables retrospective coverage and clean identification, while internal consistency enables dynamic responses of the same agents over time. In addition, the framework leverages LLMs' native reasoning capabilities, while offering greater affordability and accessibility. After validation in Section 3, Sections 4-6 illustrate how these innovations are implemented in practice and present their results.

2.1 Date Restriction

2.1.1 Design and Implementation

The first methodological foundation is date restriction. In practice, date restriction instructs the model to ignore any information released after the survey date and to role-play as if answering at that historical moment. This setup enables us to generate counterfactual survey waves at chosen points in the past and to vary survey frequency flexibly.

Our knowledge restriction is implemented through prompt design, specifically by instructing LLM-agents as follows:

> *You are responding to this survey in [*survey time*]. Do not reference or rely on any events or developments that occurred afterwards. Do not search for real data or base your answers on actual economic figures or official statistics. Instead, answer as a typical person might, based on their observations and general sense of the economy at the time.*

2.1.2 Validation of Date-Restriction Instructions

In this section, we validate whether the date-restriction setup functions as intended, addressing concerns about hindsight bias within the survey context. Specifically, we test the model's awareness of major historical events using the same set of LLM agents, and the same LLM model as in our main surveys detailed in Section 3.1. For each event, we pose the same awareness question both shortly before the event occurred and shortly after, and measure the share of personas answering "yes". For details on prompts, see Appendix A.1.

We include six major events: the September 11 attacks (2001), the U.S. invasion of Iraq (2003), the bankruptcy of Lehman Brothers (2008), and the elections of Barack Obama (2008) and Donald Trump (2016), as well as the onset of the COVID-19 pandemic (2019). These were all highly salient in U.S. public life, and a typical household would have become aware of them shortly after their occurrence. This makes them a natural benchmark for assessing whether our knowledge-restriction instructions prevent hindsight bias while preserving historically grounded awareness.

The results in Table 1 are fully consistent with our expectations. Before each event, LLM agents almost always respond "no", while immediately afterward the responses shift sharply to nearly universal recognition. In a few cases, such as the U.S. invasion of Iraq, the Lehman Brothers bankruptcy, and the COVID-19 pandemic, some "yes" responses appear before the event. These are plausibly explained by widespread public discussion and news coverage in the lead-up, making them less clear-cut than sudden shocks like 9/11 or presidential election outcomes. Likewise, the small number of "no" responses after Lehman Brothers' bankruptcy typically cite not following financial news, which is consistent with the behavior of some households. Overall, the exercise provides strong evidence that the knowledge restriction instructions succeed in their intended purpose: ensuring the model role-plays as if situated at the given experiment date, without drawing on future knowledge.

2.2 Internal Consistency

The second methodological foundation, internal consistency, operates along three dimensions. First, fixing personas across control and treatment arms and across waves ensures balance across groups. Second, following up with the same agent over time enables us to capture dynamic responses. Third, keeping treatment wording identical across waves ensures comparability without introducing wording effects.

We formalize the benefits of internal consistency using econometric theory. Assigning identical personas across arms is equivalent to complete stratification, in which treated and control units are exactly balanced within each stratum. Stratification promotes covariate balance, thereby improving the efficiency of the average treatment effect (ATE) estimator:

Table 1: Validation of knowledge restriction instructions

Event	**Event date**	**Before**		**After**	
		Survey date	%Yes	Survey date	%Yes
September 11 attacks, also known as 9/11	Sep 11, 2001	Aug 2001	0	Oct 2001	100
U.S.-led military invasion of Iraq	Mar 20, 2003	Jan 2003	0	Apr 2003	100
		Feb 2003	5		
Lehman Brothers declare bankruptcy	Sep 15, 2008	Jul 2008	0	Oct 2008	98
		Aug 2008	5		
Barack Obama has been elected U.S. President	Nov 4, 2008	Oct 2008	0	Dec 2008	100
Donald Trump has been elected U.S. President	Nov 8, 2016	Oct 2016	0	Dec 2016	100
COVID-19 declared as global pandemic	Mar 11, 2020	Dec 2019	0	Apr 2020	100
		Jan 2020	5		

Notes: This table reports the results of our knowledge restriction validation exercise (see Section 3.1 for LLM agents and model). For each event and experiment date, we prompted all personas to answer in character with knowledge restricted to that date, using the question format "Do you know of {event}?". The entries report the share of personas responding "yes".

Proposition 1. *The variance of the ATE under complete stratification is no greater than that under simple randomization.*

Proof. See Appendix B.

Proposition 1 rests on the following key intuition. Under simple randomization, the variance of ATE can be decomposed into two parts: (1) a weighted average of within-stratum variation and (2) an additional component due to imbalances in strata across treatment arms. Under complete stratification, treatment assignment is perfectly balanced within each stratum, thereby eliminating variation due to imbalances across arms. Consequently, the variance is weakly smaller than under simple randomization. The two designs become equivalent only when the stratifying characteristics are uncorrelated with the potential outcomes or the sample size approaches infinity.

Athey and Imbens (2017) derive a similar result for the special case of two strata defined by a single covariate (gender). Proposition 1 generalizes their result in two dimensions: allowing any finite number of strata and permitting strata to be defined by multiple covariates.

2.3 Beyond Human Surveys

Building on the two methodological foundations, our framework delivers five key innovations and benefits. Date restriction underpins retrospective coverage and clean identification, while internal consistency enables dynamic treatment effects. Reasoning capability and affordability further complement these features. Together, these create an ideal survey environment unattainable by human respondents. We organize the discussion of the five capabilities in the order in which they appear in Sections 4-6.

Retrospective Coverage Expanding coverage is a key capability enabled by date restriction. Unlike human surveys, which can only be fielded contemporaneously, LLM-based surveys can be extended backward to construct long panels of expectation data, or run at higher frequency. This feature opens the door to systematic historical analysis of treatment effects across macroeconomic regimes. In Section 4, we illustrate its potential by extending a benchmark ten-wave survey in 2018–2023 to more than fifty waves dating back to 1990.

Reasoning LLMs are inherently capable of generating and analyzing reasoning. Our framework formalizes this capacity for survey analysis, offering insight into the economic mechanisms underlying expectation formation. Such direct elicitation of thought processes is infeasible with human surveys, which typically depend on follow-up questions or indirect inference.

Specifically, we first implement a mixed human–LLM coding strategy to identify a set of economically interpretable categories, and then instruct the LLM to assign each response to the relevant categories. Section 5 demonstrates this process to examine how agents update their inflation expectations.

Dynamic Treatment Effects Internal consistency, particularly the ability to follow up with the same agents after the initial survey date, enables dynamic treatment effects by tracing their answers at multiple horizons after treatment. This is another feature infeasible with human surveys. In Section 6.1, we illustrate this innovation.

Clean Identification Date restriction also enables survey designs that achieve clean identification of treatment effects. By this, we mean that factual treatments are excluded from the respondent's prior knowledge at the survey date. With this separation, subsequent developments affect beliefs only through the treatment itself.

In human surveys, by contrast, factual treatments researchers use must already be public at the survey date. As a result, this information is available to respondents and potentially embedded in priors, compromising identification. To preserve clean identification, some surveys rely on hypothetical treatments (see Armantier et al. 2022; Andre et al. 2022; Haaland et al. 2023; Jiang et al. 2024; Wu et al. 2025).

Our LLM-based survey framework allows clean identification with even factual treatments, which we illustrate in Section 6.2.

Affordability Finally, beyond methodological advances, our framework delivers a practical benefit: affordability and accessibility, enabling surveys to be conducted at a negligible fraction of the cost of traditional human surveys. With GPT-4.1 (the model used for our main results), an agent answering a pre- and post-treatment question costs only approximately 0.6¢, compared to typically $1-$5 per respondent in a standard Qualtrics panel survey (a widely used online survey platform). Moreover, because personas are fixed and reusable, much smaller samples can deliver comparable statistical power, further reducing costs. This affordability expands the range of feasible research designs and broadens access to a wider community of scholars to implement studies that would otherwise be out of reach.

3 Validating the LLM Survey Framework

We begin by validating our LLM-based survey framework against a state-of-the-art multi-wave inflation expectations experiment, showing that it recovers updating patterns comparable to those of human respondents. Specifically, we replicate the survey experiment of Weber et al. (2025), which examines how households update their inflation expectations when provided with new information in high- and low-inflation settings.

Weber et al. (2025) conduct ten survey waves between 2018 and 2023 with the Nielsen Homescan Panel, providing exceptional scale and continuity for household survey experiments on expectations. This makes their study a canonical benchmark for validating our framework.

Replicating this design with LLM surveys serves two purposes. First, it tests whether LLM responses reproduce the expectation-updating behavior observed in human data. Second, it demonstrates methodological credibility by situating our framework within canonical survey experiments central to the literature on household inflation expectations.

Section 3.1 describes the LLM implementation, Section 3.2 examines demographic composition and prior beliefs, and Section 3.3 reports the main results on expectation updating. Appendix C provides further details, while Appendix D presents robustness checks.

3.1 LLM Implementation

Our validation experiment employs large language models. Our main implementation uses GPT-4.1 via the OpenAI API, released in April 2025 with a fixed knowledge cutoff in June 2024. This model provides stable and reproducible outputs while retaining realistic sampling variation, making it a close analogue to human survey responses. As a robustness check, we replicate the validation experiment with GPT-5, released in August 2025, and find that the baseline results remain robust under this alternative; for details, see Appendix D.1.

Figure 1 presents the essential structure of the survey, which follows the original human RCT design but re-implements it through a prompting framework tailored to LLM respondents. Components specific to the LLM implementation are highlighted in blue, and the full prompt text is provided in Appendix A.2.

A survey begins by defining a persona, which specifies attributes such as age, gender, marital status, education, income category, and U.S. state. Personas are drawn from a random subsample of 200 respondents in the Federal Reserve Bank of New York's Survey of Consumer Expectations (SCE).[1, 2] This ensures that the demographic composition of the

[1]To verify that results are not driven by this particular draw, we re-run the full experiment using a new random sample of 200 personas, and the resulting estimates closely track the baseline, as documented in Appendix D.2.

[2]The SCE classifies households into three nominal income bins: 35.7% under $50k, 35.8% between $50k and $100k, and 28.5% above $100k. In our baseline (Sections 3 and 4), we adopt these fixed nominal bins for comparability. As a robustness check, we redefine income bins in real terms by anchoring the 2019 cutoffs and adjusting for inflation; results remain unchanged; see Appendix D.3.

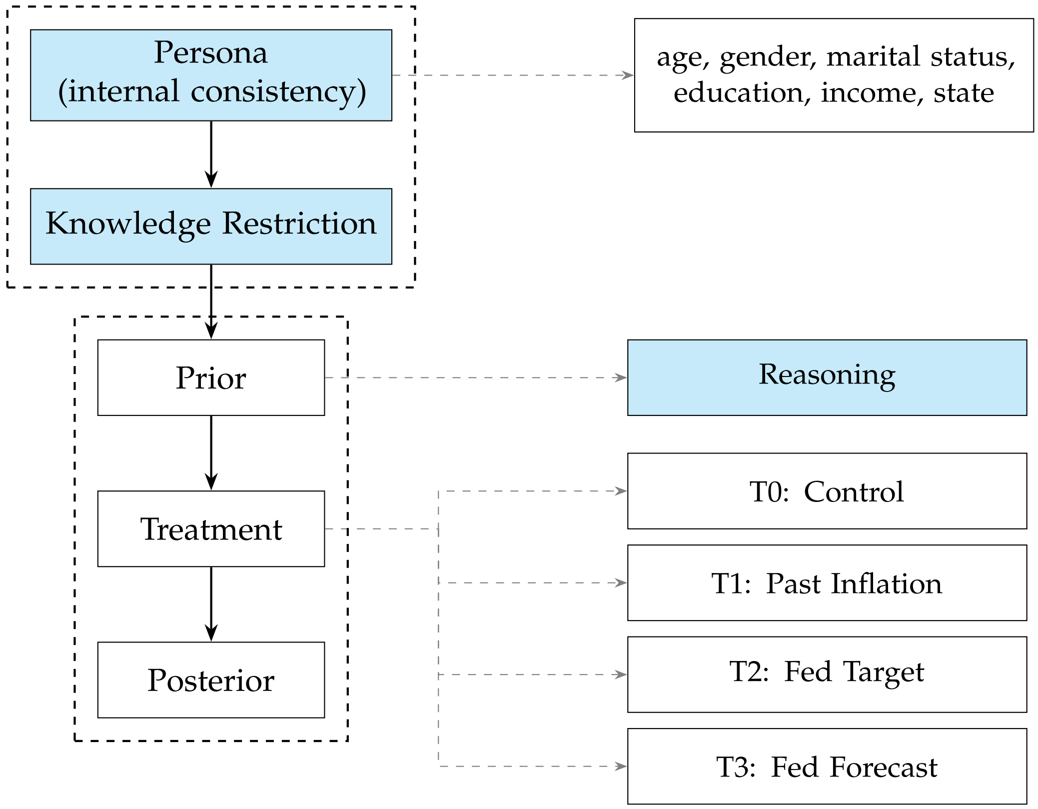

Figure 1: LLM SURVEY FLOW CHART

LLM sample mirrors that of the SCE.[3]

Once personas are created, they remain fixed across all survey waves and each group (control or treatment), creating a fully balanced synthetic panel. Although we use repeated personas, each LLM agent participates in only one wave and one group, ensuring the independence of responses. Using the same set of 200 personas achieves internal consistency (Section 2.2).

By contrast, the human benchmark of Weber et al. (2025) is based on a national panel of approximately 80,000 households, with typical sample sizes of 15,000–20,000 respondents per wave. The efficiency gain of obtaining credible estimates with just 1% of the human survey sample size makes our LLM framework both affordable and scalable.

To mimic the historical context of the original survey waves, we apply knowledge

[3]Data for SCE is available here: https://www.newyorkfed.org/microeconomics/sce.

restriction prompts per Section 2.1 so that responses reflect only the information available at the time of each wave.

Then, we follow the same sequence and wording of prior elicitation, informational treatments, and posterior forecasts as in Weber et al. (2025). In each survey, participants first report 12-month-ahead inflation expectations using a probabilistic distribution. We instruct LLM agents, unlike human respondents, to articulate their reasoning in full sentences before reporting a final numerical answer, enabling us to capture the underlying thought process behind each forecast.

Agents are then randomly assigned to either a control group (T0) or to a treatment group. Following Weber et al.'s (2025), there are three treatments: information on the most recent inflation rate (T1), the Federal Reserve's inflation target (T2), or the Federal Open Market Committee's forecast (T3). Survey waves and the corresponding treatment values are taken directly from Weber et al.'s (2025) replication files. Because not all treatments are implemented in every survey wave, the panel is unbalanced: T1 is fielded in 10 waves, T2 in 7, and T3 in 6. Table C.1 reports the extracted period coverage and numerical treatment values. After the intervention, all respondents provide a point forecast of 12-month-ahead inflation.

Additional implementation details are as follows. Agents retain memory of prior questions to replicate the sequential nature of human surveys, implemented through the full-memory setting. We only ask relevant questions, omitting the background and screening questions typical of human surveys.

3.2 Demographics and Priors

As a first step, we establish that our LLM-based survey reproduces the key features of the human benchmark, both in terms of demographic composition (aggregated from LLM personas) and in the priors reported before any informational treatment.

First, we show that demographic distributions of LLM agents are well aligned with those of the SCE. For instance, the share of college-educated respondents is 50.0% in the LLM sample and 55.3% in the SCE sample, and the gender shares are nearly identical. Differences are modest and largely confined to age and regional composition: the LLM demographics slightly over-represent younger respondents and those from the Midwest, whereas the SCE demographics include a larger share from the Northeast. For details, see Appendix C.2. While neither sample is nationally representative by design, both provide broad coverage across key demographic and geographic dimensions.

Next, we compare pre-treatment inflation forecasts from the LLM survey and the SCE, both based on the same distributional question. For each respondent, we compute the mean expectation as a weighted average of the midpoints of the reported bins, weighted by their assigned probabilities. In Section 3.3, we subsequently use this measure as the prior forecast of LLM agents.

Figure 2 compares the cross-sectional median inflation expectations of LLM agents with those of SCE respondents across survey waves. Regressing the LLM medians on the corresponding SCE medians yields a slope close to one (1.05) with a small negative intercept, indicating that the LLM priors closely track human priors over the inflation cycle.

3.3 Expectation Updating

After confirming that our LLM-based survey reproduces the baseline features of human benchmarks, both in demographic composition and in prior expectations, we now turn to the main validation exercise, examining whether LLM agents replicate the expectation-updating patterns documented in Weber et al. (2025).

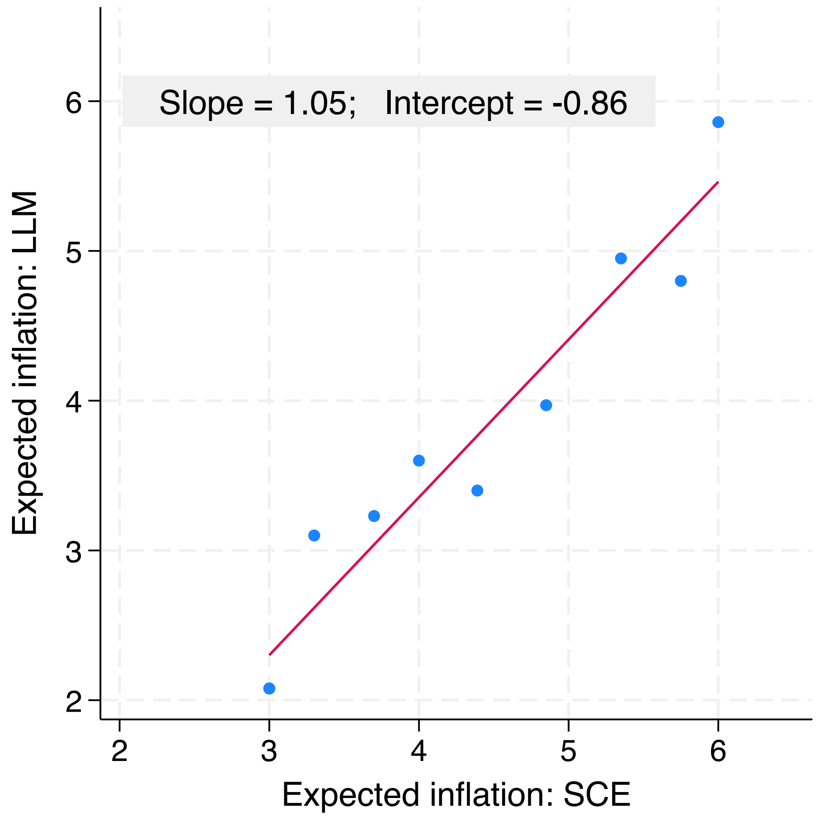

Figure 2: PRIOR INFLATION EXPECTATION: LLM VS. HUMAN

Notes: This figure compares 12-month-ahead inflation expectations from LLM-based priors with those from the New York Fed's Survey of Consumer Expectations (SCE). Each point represents the cross-sectional median of expectations within a given quarter.

As in Weber et al. (2025), we estimate the following regression specification:

$$\begin{aligned} \text{posterior}_i = & \alpha + \sum_{j=1,2,3} \delta_j \times \mathbb{I}_{\{i \in \text{Tj}\}} \\ & + \beta \times \text{prior}_i + \sum_{j=1,2,3} \gamma_j \times \mathbb{I}_{\{i \in \text{Tj}\}} \times \text{prior}_i + \text{error}_i, \end{aligned} \tag{3.1}$$

where the prior is elicited before information intervention (calculation described in Section 3.2). The posterior is the point forecast reported after the treatment. The indicator variable $\mathbb{I}_{\{i \in \text{Treat } j\}}$ equals one if respondent i is assigned to treatment arm Tj (T1 - past inflation, T2 - Fed target, or T3 - Fed forecast), and zero otherwise. Because the personas

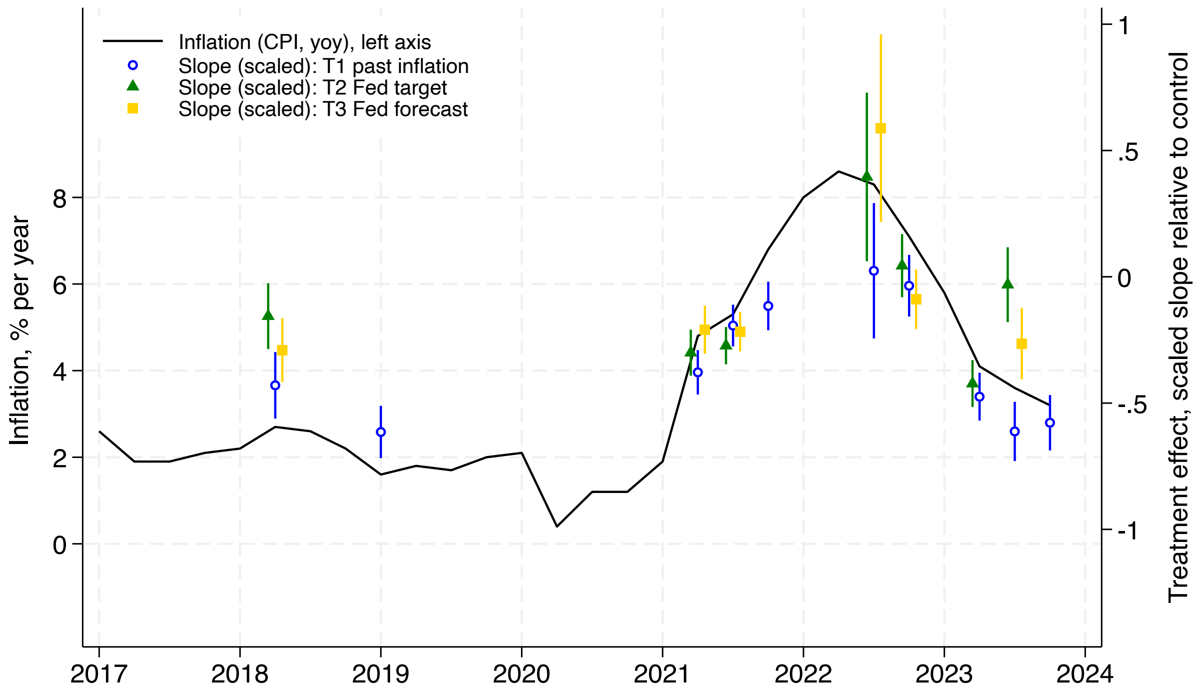

Figure 3: LLM REPRODUCTION OF WEBER ET AL. (2025) FIGURE 4A

Notes: This figure reproduces the scaled slope estimates from Figure 4A in Weber et al. (2025) using LLM-based survey responses, following the same experimental design. Inflation (black line, left axis) is year-over-year CPI in percentage points, taken from their replication files. Scaled slopes (right axis): T1 past inflation (blue dots), T2 Fed target (green triangles), T3 Fed forecast (yellow squares). Vertical bars indicate 90% confidence intervals, spanning 1.65 standard errors on each side.

are identical between treatment and control groups, we exclude demographic controls, which have no effect on the estimates.

In this setup, the coefficient β captures how closely posteriors track priors in the control group, and the interaction coefficients γ_j measure how this relationship changes across treatment arms. We follow Weber et al. (2025) and focus on the scaled slope γ_j/β. This ratio provides a unit-free measure of responsiveness: values close to -1 imply that agents ignore their priors entirely and fully rely on the treatment signal, whereas values near zero imply little or no responsiveness to the treatment relative to the control group. In what follows, we calculate the scaled slope across inflation environments to compare the LLM-based surveys with the human benchmarks in Weber et al. (2025).

Table 2: CORRELATION OF SCALED SLOPES WITH INFLATION

Agent type	Correlation with inflation			
	T1: Past inflation	T2: Fed target	T3: Fed forecast	Pooled
LLM	0.92	0.73	0.85	0.79
Human	0.69	0.87	0.13	0.57

Notes: This table reports the correlation between scaled slopes and current inflation. Results for human respondents are taken from Weber et al. (2025). Results for LLM respondents are from our replication under the identical treatment design. "Pooled" correlations are calculated by pooling estimates from all treatment types.

Figure 3 uses our LLM-based surveys to reproduce the human benchmark (Figure 4A in Weber et al. 2025), plotting scaled slope estimates for T1–T3 treatments with 90% confidence intervals (right axis) against year-over-year inflation (left axis). The figure shows that informational treatments have large effects on expectations during the low-inflation period of 2018–2019, but responsiveness declines sharply once inflation begins rising in 2021, converging toward zero by late 2022. Thus, our LLM results strikingly reproduce the central empirical finding of Weber et al. (2025).

To further assess alignment, Figure 4 plots LLM estimates of scaled slopes (y-axis) directly against their human counterparts (x-axis). The first three panels show one treatment arm per panel (T1: past inflation, T2: Fed target, T3: Fed forecast), and the fourth panel pools all three treatments. Each dot represents a single survey wave–treatment combination. Correlations are positive for T1 (0.62), T2 (0.71), and the pooled sample (0.31), but slightly negative for T3 (–0.24). Thus, while LLMs broadly reproduce the cross-wave variation observed in human data, the strength of alignment differs by treatment.

We further inspect what causes the negative correlation in T3 using Table 2, which evaluates Weber et al.'s (2025) main message by comparing treatment effects directly with inflation, rather than between LLMs and human surveys. In both cases, the correlations are positive across treatments: scaled slopes become less negative as inflation rises. The strength of correlation, however, differs. For LLM-based surveys, the correlation with inflation is consistently high across treatments—0.92 (T1), 0.73 (T2), 0.85 (T3), and 0.79

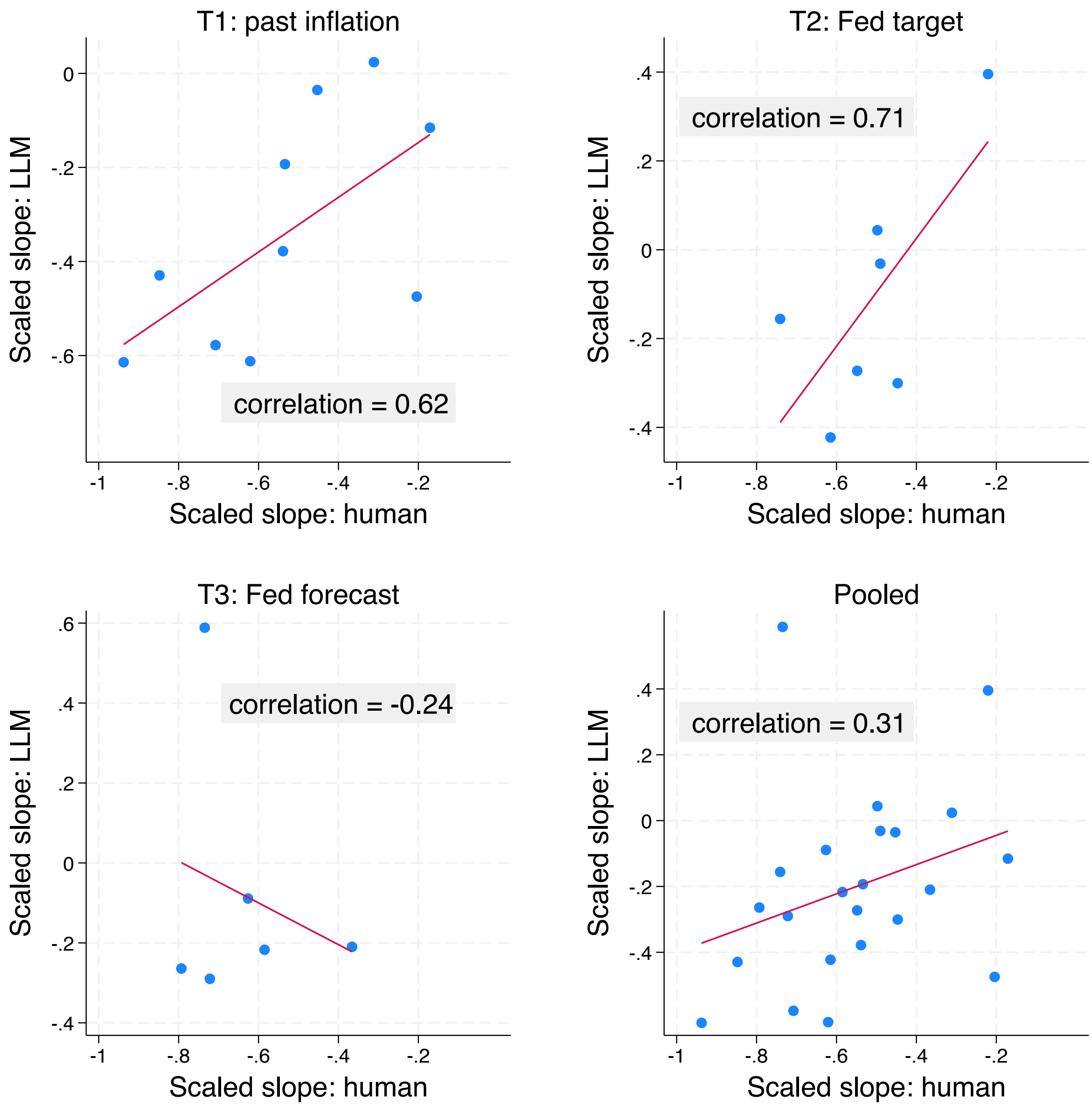

Figure 4: SCALED SLOPES: LLM VS. HUMAN

Notes: This figure compares the scaled slope estimated using human surveys and LLM-based surveys.

(pooled). Therefore, LLM-based surveys deliver a clear overall pattern of state-dependence of expectation updating.

For human surveys, the corresponding correlations are lower and more variable: 0.69 (T1), 0.87 (T2), 0.13 (T3), and 0.57 (pooled). The weaker human survey-inflation correlations, together with the negative LLM–human correlation, likely reflect greater sampling noise in conventional surveys.

Taken together, these findings show that LLM surveys successfully reproduce the core results of Weber et al. (2025). In particular, LLM agents recover the central empirical pattern: responsiveness to informational treatments diminishes as inflation becomes more salient. Moreover, the LLM-based estimates often appear cleaner and more consistent than their human counterparts, reflecting the absence of sampling noise and attrition. This underscores the value of LLM-based surveys as a credible and efficient complement to traditional survey experiments.

4 Beyond Human Surveys: Retrospective Coverage over 35 Years

We now turn to the broader potential unlocked by LLM-based surveys. In this section, we focus on retrospective coverage. Section 5 examines economic reasoning, while Section 6 explores dynamic treatment effects and clean identification. Together, these innovations realize the ideal form of the Weber et al. (2025) design, in a way that is unattainable with human surveys.

Specifically, we extend the survey design described in Section 3.1 retroactively to 1990, yielding over 50 survey waves across three and a half decades. For comparison, Weber et al. (2025) conduct 10 survey waves between 2018–2023, which already represent a major advance for human surveys.

4.1 Implementation Details

This section discusses implementation details. Table 3 summarizes the sample coverage, frequency, and the number of waves for each treatment. During rapidly shifting economic conditions, such as the Great Recession (2008–2009) and the COVID inflation surge (2021–2023), we implement the surveys quarterly to capture fast-moving dynamics. For

Table 3: SAMPLE COVERAGE

Coverage	Frequency	Number of waves		
		T1 (1990-2023)	T2 (2012-2023)	T3 (2015-2023)
1990 - 2007	Annual	18	N.A.	N.A.
2008 - 2009	Quarterly	8	N.A.	N.A.
2010 - 2020	Annual + replication	13	10	7
2021 - 2023	Quarterly	12	12	12

Notes: This table summarizes the sample coverage, frequency, and the number of waves across treatments in the extension experiment. In addition to the waves in Weber et al. (2025) (see Table C.1), the surveys are implemented quarterly during the Great Recession (2008 - 2009) and the COVID inflation surge (2021 - 2023), and annually for the rest.

more stable periods, we rely on annual frequency to balance coverage with computational cost, while always including the replication waves (see Table C.1). Our sample begins in 1990, as pre-1990 periods suffer from sparse digitization of contemporaneous news and statistical releases. It ends in 2023, given GPT-4.1's June 2024 knowledge cutoff.

For T1 (past inflation treatment), we begin in 1990, which marks the beginning of our sample. For treatments related to the Federal Reserve, the rollouts are anchored in institutional milestones. T2 (Fed target treatment) start in 2012, when the Federal Reserve first formally announced an inflation target of 2%.[4] T3 (Fed forecast treatment) is introduced in 2015, when the Federal Open Market Committee's (FOMC) Summary of Economic Projections (SEP) began reporting median forecasts for public release, thereby providing a clear numerical forecast that could be used in survey treatments.[5] This staggered design ensures that the informational treatments reflect the institutional reality of their historical period.

All experiments are conducted in the final month of each calendar quarter or year, giving us precise control over survey timing and ensuring consistency across periods. Unlike human surveys, which may face logistical delays or irregular field dates, LLM-

[4]See Federal Reserve Board Press Release on January 25, 2012: "Federal Reserve Issues FOMC Statement of Longer-Run Goals and Policy Strategy".

[5]See Federal Reserve Board, Review of Monetary Policy Strategy, Tools, and Communications: "Timelines of Policy Actions and Communications: Summary of Economic Projections".

based surveys can be implemented precisely as planned, keeping treatments tightly aligned with experiment dates.

To follow human survey timing, for T1 (past inflation treatment), we use the year-over-year CPI inflation rate (CPIAUCSL) of the first month of the corresponding quarter.[6] This convention accounts for the one-month release lag in CPI data.[7] For T2 (Fed target treatment), we provide the Federal Reserve's 2% inflation target. For T3 (Fed forecast treatment), we use the PCE inflation projection from the FOMC's Summary of Economic Projections (SEP) released at the end of the previous quarter. Following the practice in Weber et al. (2025), we use current-year forecasts in the first- and second-quarter experiments, and next-year forecasts in the third- and fourth-quarter experiments.

4.2 Priors vs MSC

We begin our analysis by examining whether LLM agents produce credible prior inflation expectations across the extended sample period. While Section 3 benchmarks LLM priors against the SCE, its short post-2013 sample makes it unsuitable for our extended analysis. We therefore use the University of Michigan Survey of Consumers (MSC), the longest-running source of U.S. household inflation expectations.[8]

Figure 5 compares the LLM priors with those from the MSC. Each dot corresponds to the cross-sectional mean of expectations in a given survey wave, with the fitted line summarizing the linear relationship between the two series.[9] The alignment is striking: LLM priors and MSC expectations correlate at 0.84 over the extended sample. This high correlation is particularly notable given differences in the question design: our priors elicit probability distributions over inflation, whereas the MSC asks for point forecasts

[6] Consumer Price Index for All Urban Consumers: All Items in U.S. City Average (CPIAUCSL) is available on FRED here: https://fred.stlouisfed.org/series/CPIAUCSL.

[7] The Bureau of Labor Statistics releases monthly CPI data with a one-month delay; for example, August 2025 inflation is released on September 11, 2025.

[8] Data for MSC is available here: https://data.sca.isr.umich.edu/sda.php.

[9] We present the priors from the experiment in Section 4.3 for brevity. With the same set of personas, the cross-sectional mean of priors from the experiment in Section 6.2 is nearly identical.

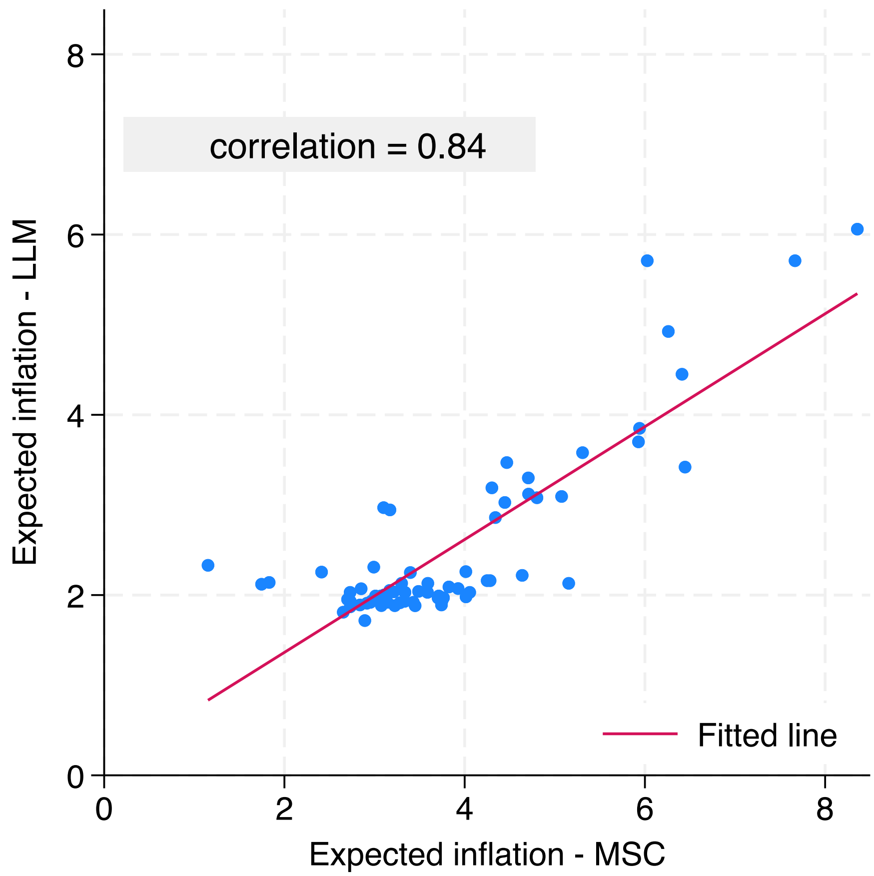

Figure 5: PRIOR INFLATION EXPECTATION: LLM VS. HUMAN (MSC)

Notes: This figure compares the 12-month ahead inflation expectations from LLM-based priors with those from the Michigan Survey of Consumers (MSC). Each point represents the cross-sectional average of expectations within a given survey month.

about price changes. This result makes the comparison conservative and underscores the robustness of the LLM-based expectations.

4.3 Retrospective Coverage

We next turn to retrospective coverage, examining how LLM agents update their expectations in response to informational treatments over 35 years. Figure 6 plots the scaled slope estimates for T1–T3 (right axis) against year-over-year CPI inflation (left axis) in the experiment month. Specifically, the top panel shows that treatment effects for T1 (past

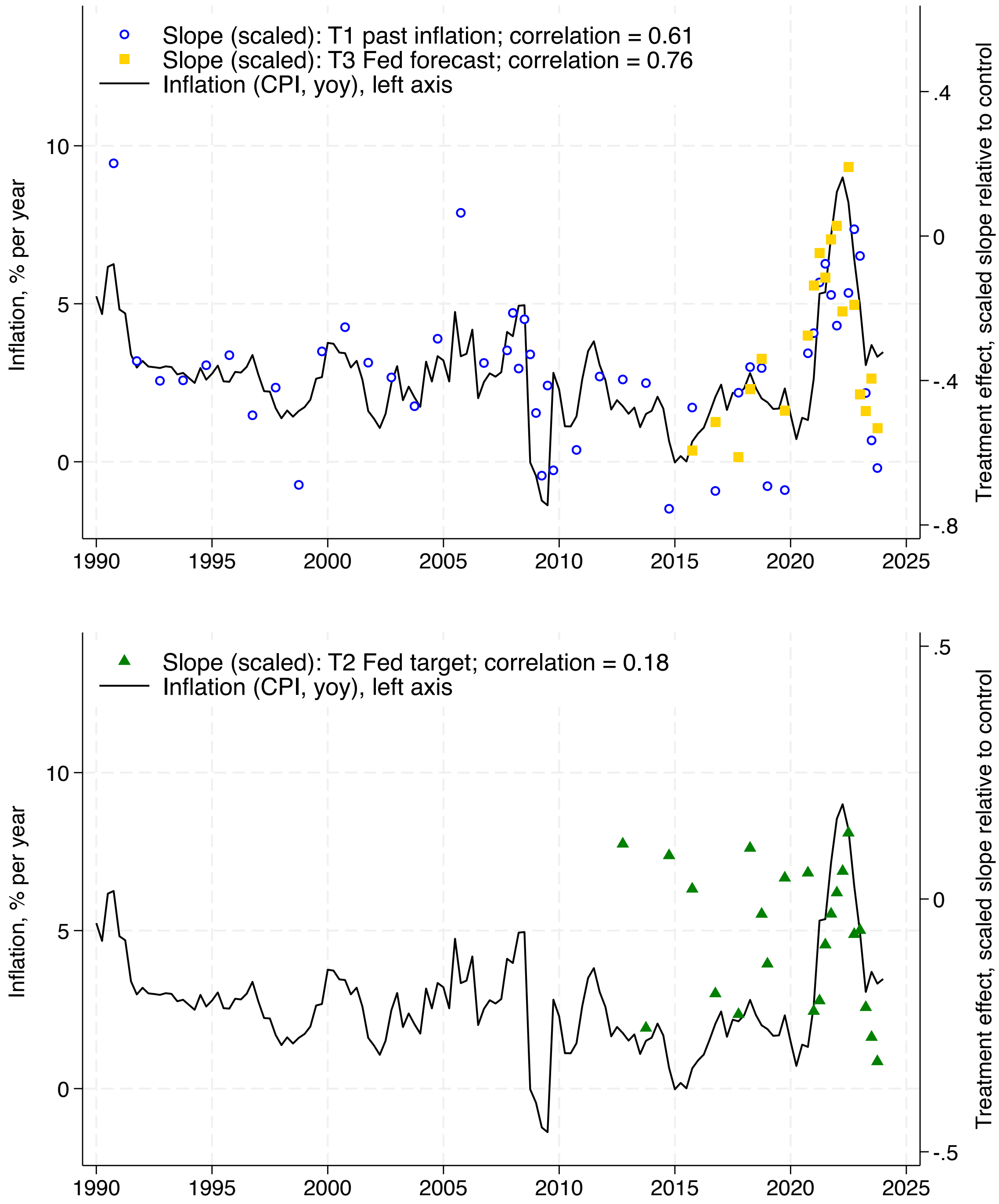

Figure 6: SCALED SLOPES: FULL SAMPLE

Notes: This figure plots the extension of scaled slopes using LLM-generated responses. Inflation (black line, left axis) is the year-over-year CPI of the experiment month in percentage points, downloaded from FRED. Scaled slopes (right axis): T1 past inflation (top panel, blue dots), T2 Fed target (top panel, green triangles), T3 Fed forecast (bottom panel, yellow squares).

inflation) are more negative during the low-inflation environment of the 1990s and early 2000s, but shift toward zero before the Great Recession and the post-COVID inflation surge, reflecting reduced responsiveness when inflation is high. The same pattern is stronger for T3 (Fed forecast). Overall, the correlations between scaled slopes and inflation remain positive: 0.61 for T1, 0.76 for T3 treatments. These positive correlations suggest that the main result in Weber et al. (2025) extends to our longer sample, though the effects are attenuated relative to the very recent period.

The bottom panel shows a weaker relationship for T2 (Fed target), with a correlation of 0.18 in the extended sample. One plausible explanation is that the Fed's 2% target is fixed through time, which LLM agents appear aware of. This conjecture also sheds light on why, in the replication sample summarized in Table 2, LLM correlations for T2 are lower than those of human respondents. Importantly, however, the same table shows that for the other treatments (T1 past inflation and T3 Fed forecasts), LLM correlations exceed the human benchmark.

To summarize, we demonstrate how our framework allows for extending the LLM-based survey design retrospectively over three decades, replicating key human survey behaviors where benchmarks exist, and delivering long-run coverage unattainable in traditional settings.

5 Beyond Human Surveys: Economic Mechanisms

Having established coverage, we now turn to the economic mechanisms underlying expectation formation. Specifically, this section leverages the LLMs' native reasoning capacity to generate and analyze the verbal explanations accompanying survey responses. In doing so, we evaluate the economic channels that shape belief updating across macroeconomic environments through agents' cognitive reasoning.

The literature has primarily attributed attention as the main economic channel behind state-dependent inflation expectations. Existing studies rely on indirect measures: Weber et al. (2025) draw on treatment effects in survey experiments, Bracha and Tang (2025) use survey-based inattention proxies, and Korenok et al. (2023) employ online search intensities.

By contrast, our LLM-based framework identifies economic channels directly through agents' explicitly articulated reasoning accompanying their final answers. Beyond providing direct evidence, our approach enables a comprehensive analysis of plausible economic mechanisms, identifying new channels and pinpointing specific forms of attention.

We focus on the reasoning underlying prior expectations, as these capture how agents conceptualize inflation before receiving new information and therefore reveal how they distribute weights between prior beliefs and informational treatment. Section 5.1 presents classification procedure and resulting categories and economic channels. Section 5.2 compares these patterns across high- and low-inflation environments, while Section 5.3 examines how they evolve over time.

5.1 Classification

This section explains how we construct economically interpretable categories from the reasoning responses, map them into distinct economic channels, and assign each response to the relevant categories.

For each survey question, the LLM agent first generates a brief textual explanation describing how it forms its expectation before producing its final numerical answer. We adopt a mixed human–LLM coding strategy following Haaland et al. (forthcoming). Specifically, we first manually review a sample of explanations to identify the salient reasoning categories. Independently, we prompt the LLM to detect common themes and propose an initial categorization scheme. We then integrate the researchers' judgment with the LLM's detailed reading to finalize a set of economically interpretable categories. Finally, we apply a GPT-based classifier to assign each explanation to the relevant categories. Details of the classification procedures appear in Appendix E.1.

This procedure yields three economically meaningful channels, encompassing five reasoning categories, summarized in the first three columns of Table 4. The top block captures the mean-reversion channel: whether agents view inflation as near its normal level or expect it to return toward that level. The middle block reflects attention, which operates along two dimensions: individual (references to personal price experiences) and aggregate (references to monetary policy). Finally, the third channel links inflation expectations to business-cycle conditions. Both the mean-reversion and business-cycle channels represent new mechanisms uncovered by our reasoning analysis, whereas the attention channel corresponds to a existing mechanism that we further decompose into individual and aggregate dimensions. Appendix E.2 provides illustrative examples for each category.

5.2 High- vs. Low-Inflation

Having established the different economic channels, we now examine how they vary across different inflation environments. Table 4 summarizes reasoning patterns for two representative periods: 2018Q2 (low inflation) and 2022Q4 (high inflation). We frame our discussion within a Bayesian framework, using these patterns to explain how the three channels shape how much weight agents place on prior versus treatment information.

First, for the mean-reversion channel: during the high-inflation period, no respondents view the current high inflation as representing its normal level or long-run mean. Most (68%) instead interpret current inflation conditions as deviations from normal times, limiting extrapolation from new information. By contrast, during the low-inflation period, reasoning is dominated by normal narratives (90% of responses). In this environment, agents place greater weight on treatments. The difference is substantial: from high- to low-inflation periods, the share of normal reasoning rises by 90 percentage points, while normalizing reasoning falls by 58.

Table 4: Reasoning across high- and low-inflation periods

Channel	Label	Definition	High	Low	Diff
Mean reversion	Normal	Predicts that inflation will stay close to its typical or target level, with gradual adjustments and low perceived risk of persistent deviations, reflecting a view that the current environment is normal and stable.	0.00	0.90	0.90
	Normalizing	Expects inflation to move back toward its normal or target range from unusually high (or low) levels, signaling stabilization or moderation.	0.68	0.10	-0.58
Attention	Personal observation	Refers to personally observed price experiences (e.g., groceries, gas, rent, wages).	0.54	0.30	-0.24
	Monetary policy	Mentions central banks, interest rates, or monetary policy actions (e.g., tightening, easing, QE, QT).	0.01	0.01	0.00
Business cycle	Business cycle	States that the economy is in expansion (recovering, rebounding, or improving) or recession (contracting, in a downturn, or facing a hard landing).	0.01	0.02	0.01

Notes: This table defines three economic channels along with their constituent reasoning categories, and reports the share of responses classified under each label during high- and low-inflation periods. Columns (4) and (5) show the proportions for 2018Q2 (low-inflation) and 2022Q4 (high-inflation), respectively, while Column (6) reports the difference between the two (low – high).

Second, under the individual attention channel, agents pay greater attention to personal experiences when inflation is high: 54% of responses reference personal observations such as grocery prices, gas, wages, or rent. This indicates that during high inflation, agents place greater weight on priors grounded in their own firsthand price experiences, rather than on new information or treatments. The prevalence of this reasoning declines by about 24 percentage points, from 54% in the high-inflation period to 30% in the low-inflation period.

By contrast, references to monetary policy (the aggregate attention channel) or to business-cycle conditions are rare in both periods, suggesting that prior expectations are largely shaped by whether households perceive inflation as normal and by their own price experiences.

5.3 Dynamics over Time

We next extend the reasoning analysis to the full time series starting from 1990, using the same sample as in Section 4.[10] This section focuses on the mean-reversion and individual attention channels, which display strong variation across high- and low- inflation states (Table 4).

Figure 7 shows how the share of each reasoning category (right scale) varies with inflation (left scale). The top, middle, and bottom panels plot responses classified as "normal," "normalizing," and "personal observation," respectively, in each period.

Two patterns emerge clearly, consistent with Table 4. First, the mean-reversion channel remains strong. Specifically, the top panel of Figure 7 shows that the share of responses describing inflation expectations as "normal" is consistently high (close to 100%), except during three episodes: the high inflation of the early-1990s recession, the temporary rise during the Great Recession, and the pronounced surge during COVID-19. Consistent with this pattern, the second panel, showing the "normalizing" label, features visibly higher

[10] Section 5.2 examines two representative periods from this sample for illustration.

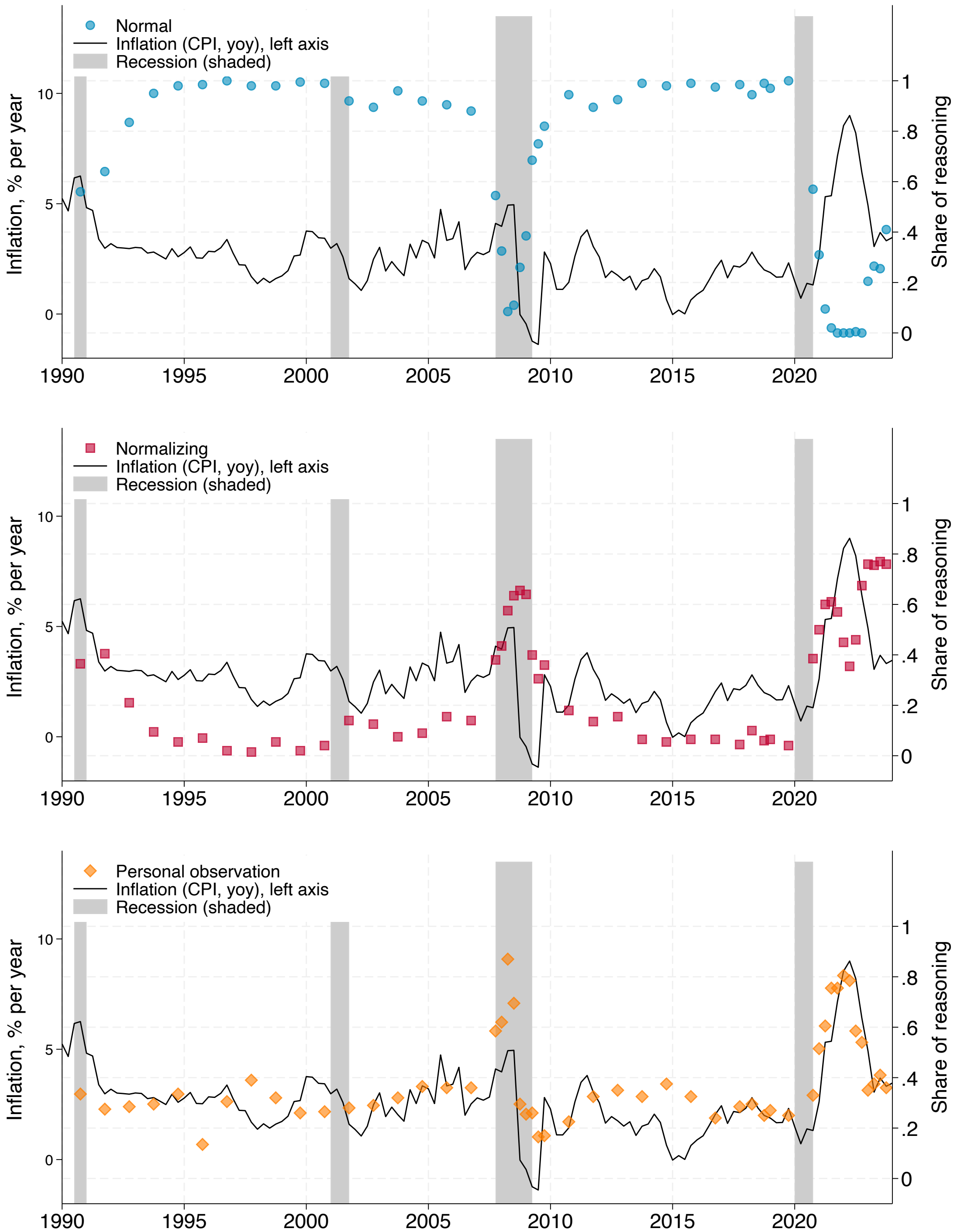

Figure 7: REASONING SHARES OVER TIME

Notes: This figure plots the time series of reasoning shares alongside current CPI-inflation (left axis). The upper, middle, and bottom panels plot the share of reasoning classified as "normal" (blue circles), "normalizing" (red squares), and "personal observation" (orange diamonds), respectively. All reasoning shares are based on the reasoning of LLM agents regarding prior inflation expectations.

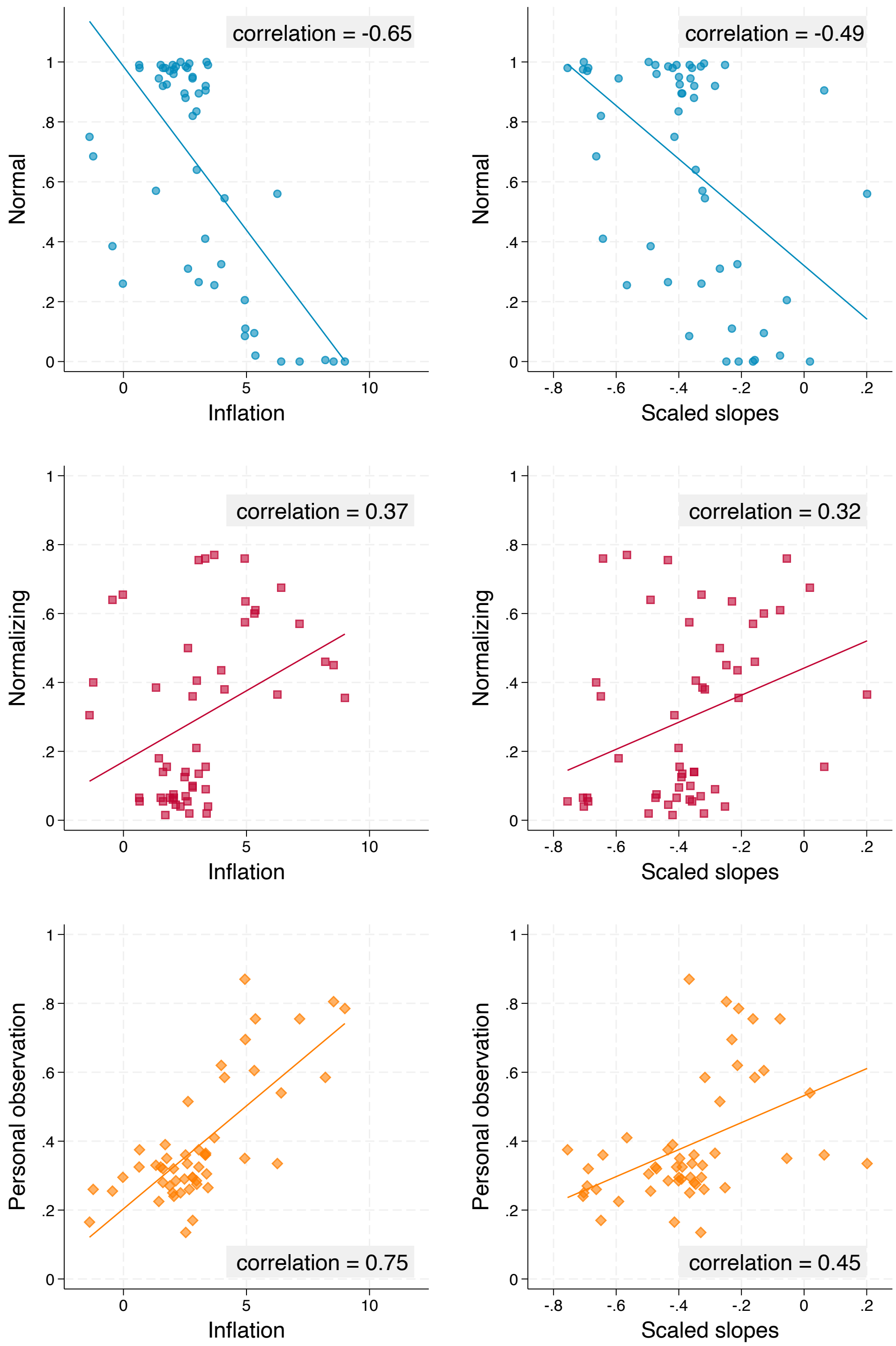

Figure 8: REASONING SHARES VS. INFLATION AND SCALED SLOPES

Notes: This figure plots the share of reasoning classification versus inflation (left column) and scaled slopes of T1 (right column). The top, middle, bottom rows plot the shares of reasoning classified as "normal" (blue circles), "normalizing" (red squares), and "personal observation" (orange diamonds), respectively.

shares during these three episodes.

The top two plots in the left panel of Figure 8 corroborate this relationship using scatter plots, which show correlations of –0.65 for the "normal" label and 0.37 for the "normalizing" label. This pattern reinforces the interpretation from the cross-period comparison: when inflation is high, respondents mainly view it as a temporary deviation from the long-run mean and therefore place less weight on the informational treatment.

Second, the individual attention channel also persists over time. The bottom panel of Figure 7 shows that mentions of personal observation increase during two periods: the Great Recession and the COVID-19 inflation surge. Consistently, the bottom-left scatter plot in Figure 8 shows a positive correlation (0.75) between the prevalence of personal observation and inflation.

Having discussed the relationship between economic channels and inflation, we now examine their relationship with the scaled slopes. The right panel of Figure 8 shows these correlations, focusing on the scaled slopes calculated using the past-inflation treatment (T1) described in Section 4.3. The results mirror those in the left panel using inflation: the correlation with "normal" remains negative, while those with "normalizing" and "personal observation" are positive. The strongest correlations are observed for "normal," followed by "personal observation."

Taken together, the cognitive reasoning accompanying the final answers reveals two mechanisms behind expectation formation about inflation: a new mean-reversion channel and an attention channel, where we identify personal observation as its key mechanism.

6 Beyond Human Surveys: Dynamics and Identification

This section further advances the framework along two dimensions that go beyond the reach of human surveys: tracing agents' expectation updating over time and establishing clean identification of treatment effects.

6.1 Dynamic Treatment Effects

LLMs enable us to quantify how beliefs evolve after exposure to information over time. To illustrate this, we elicit follow-up posteriors after completing the initial survey. The same respondents receive a sequence of follow-up prompts framed as:

> *It is now **[follow-up month]**. What do you expect the rate of inflation to be over the next 12 months? Please give your best guess.*

Similar to human surveys, we avoid referencing the initial survey to minimize experimenter demand effects.

We implement the follow-up survey monthly for up to twelve months. Such regular follow-ups are infeasible in human surveys, where repeatedly re-interviewing the same individuals at short intervals is prohibitively costly. Even Weber et al. (2025), who conducted a single follow-up three months after the initial survey, requires substantial resources. In contrast, LLM surveys make this design straightforward, enabling us to trace the dynamic treatment effects over time.

To quantify the dynamic treatment effects, we modify (3.1) as follows:

$$\begin{aligned} \text{posterior}_{ih} =& \alpha_h + \sum_{j=1,2,3} \delta_{jh} \times \mathbb{I}_{\{i \in \text{Tj}\}} \\ & + \beta_h \times \text{prior}_i + \sum_{j=1,2,3} \gamma_{jh} \times \mathbb{I}_{\{i \in \text{Tj}\}} \times \text{prior}_i + \text{error}_{ih}, \end{aligned} \tag{6.1}$$

where h indexes the horizon between the initial survey, when we elicit priors and administer treatments, and the follow-up survey, when we elicit the posterior. When $h = 0$, it collapses to the baseline regression in (3.1).

We plot dynamic treatment effects in Figure 9 together with confidence intervals in shaded areas. Specifically, we focus on 2018Q2 (a period of relatively low inflation where instantaneous treatment effects are strong) and T1 (past inflation). We find that after three months, the scaled slope is about half the size of the instantaneous response, consistent

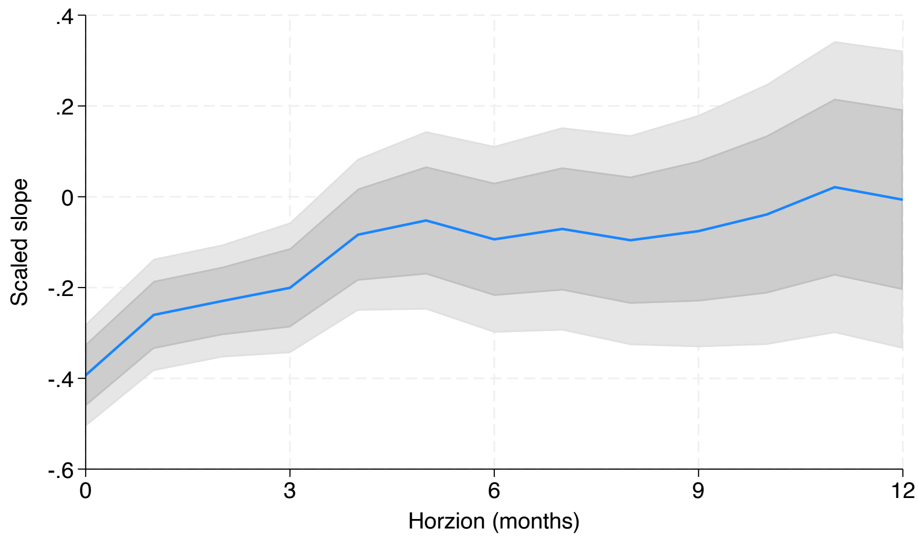

Figure 9: DYNAMIC TREATMENT EFFECTS

Notes: This figure plots the estimates of the scaled slope from regression (6.1). The horizontal axis shows the number of months h between the initial survey (when priors are elicited and treatments are implemented) and the follow-up survey (when posteriors are measured). The shaded areas are 68% and 90% confidence intervals.

with Weber et al. (2025) and Coibion et al. (2023). By six months, the effect becomes statistically insignificant, which aligns with Coibion et al. (2023). Finally, it completely vanishes in about a year.

Because these human survey benchmarks come from different studies, they may not be fully comparable. By contrast, our framework ensures consistency across horizons, reproduces human-survey benchmarks at the horizons where they exist, and enables us to trace dynamics flexibly at any horizon.

6.2 Clean Identification

While the timing of treatments used in Section 4 is designed to mimic the information available in human surveys, where factual treatments must already be public prior to the survey date and thus available to respondents. However, this constraint does not bind LLM surveys. This section implements clean identification by ensuring that factual treatments are excluded from the prior knowledge set.

We implement a "clean" version of the experiment using T1, which offers the longest sample and provides a natural benchmark for comparison. Specifically, respondents are provided with the factual statement:

> *Over the last twelve months, the inflation rate in the U.S. (as measured by the Consumer Price Index) was [treatment]%.* ***This information is provided to you for the purpose of this survey and has not yet been publicly released.***

The first sentence is identical to the baseline survey (detailed in Appendix A.2), and we added a clause clarifying that the information had not yet been publicly released, ensuring it entered only through the experiment and thereby implementing a "clean" design. Moreover, the treatment corresponds to the CPI inflation rate for the survey month, which would only be released in the subsequent month. Such a design is impossible in real-time human surveys but feasible in LLM surveys.

Figure 10 plots the scaled slope estimates (right axis) using treatments with clean identification (pink diamonds) and from our baseline (blue circles) against year-over-year CPI inflation in the experiment month (left axis). The blue circles are identical to the blue circles in the upper panel of Figure 6. The correlation between the clean and non-clean series is 0.54, implying that clean identification makes some difference. Moreover, this exercise highlights a unique advantage of LLM surveys: they render clean identification feasible.

However, in this experiment, clean versus non-clean treatments do not alter the economic conclusion. The correlation of scaled slopes with inflation is 0.61 for both treatments. Similarly, in the replication sample, the scaled slope estimates remain strongly correlated with actual inflation (0.85), with those of Weber et al. (2025) (0.68), and with our baseline treatment (0.88). This outcome reflects the unusual persistence of inflation, as upcoming CPI releases are mechanically close to past values.

While the benefits of clean identification are not realized in this setting, the broader principle remains valuable, especially in contexts where future information departs more

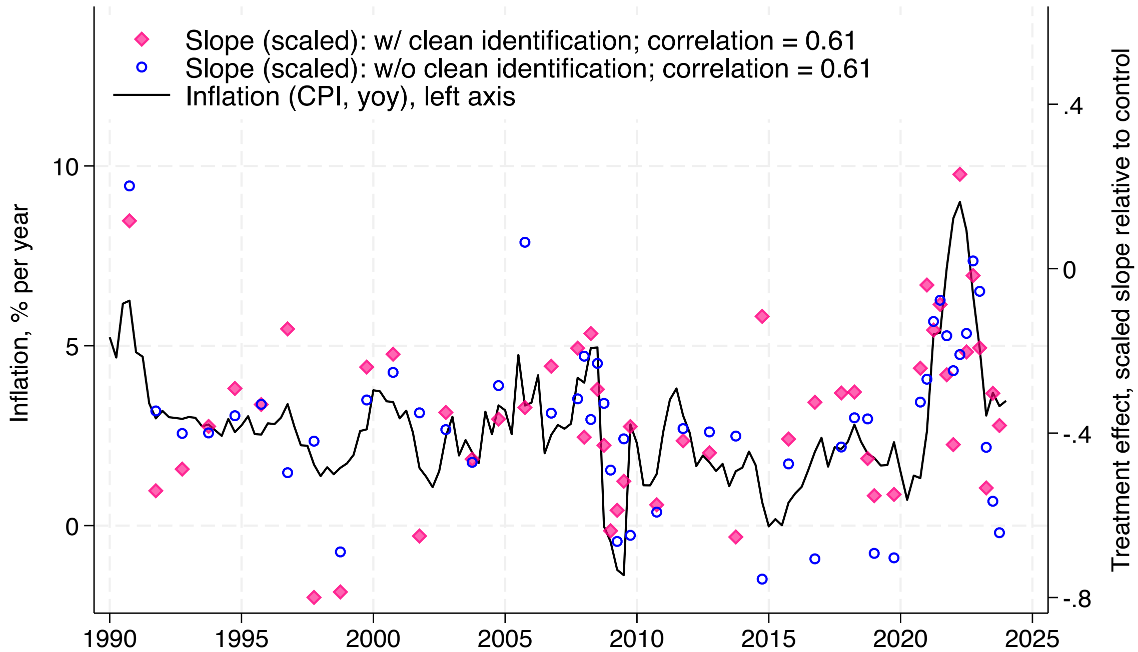

Figure 10: CLEAN VS. NON-CLEAN IDENTIFICATION

Notes: This figure plots the extension of scaled slopes under T1 past inflation with LLM-generated responses. Inflation (black line, left axis) is the year-over-year CPI of the experiment month in percentage points, downloaded from FRED. Scaled slopes (right axis): with clean identification (pink diamond), without clean identification (blue circle).

sharply from prior realizations.

7 Conclusion

This paper proposes a new LLM-based survey framework that expands the frontier of expectation measurement. By combining retrospective coverage, economic reasoning, dynamic treatment effects, clean identification, as well as affordability and accessibility, the framework delivers an ideal survey environment that traditional human surveys cannot match.

We validate our LLM framework against a canonical household experiment (Weber et al., 2025). Our validation exercise shows that LLM surveys reproduce key patterns of expectation updating, while often yielding clearer correlations with inflation, reflecting reduced sampling noise. To showcase the capabilities of our framework, we extend the experiment retrospectively from 10 waves (2018–2023) to more than 50 waves dating back to 1990. The extended panel shows that treatment responsiveness (scaled slope) co-moves with inflation, though attenuated. Our framework uncovers the mechanisms behind this relationship, identifying the mean-reversion and individual-attention channels as the main drivers. Finally, it traces the dynamic responses of the same agents over time and achieves clean identification by separating prior knowledge from subsequent factual treatments.

Taken together, these results establish our LLM framework as a credible complement to traditional surveys, offering advantages that human surveys cannot deliver. Beyond inflation, the framework opens avenues for systematically reconstructing expectations to study monetary policy, housing, and labor markets, among other topics beyond economics.

References

Andre, Peter, Carlo Pizzinelli, Christopher Roth, and Johannes Wohlfart, "Subjective models of the macroeconomy: Evidence from experts and representative samples," *The Review of Economic Studies*, 2022, *89* (6), 2958–2991.

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

Armantier, Olivier, Argia Sbordone, Giorgio Topa, Wilbert Van der Klaauw, and John C Williams, "A new approach to assess inflation expectations anchoring using strategic surveys," *Journal of Monetary Economics*, 2022, *129*, S82–S101.

Armona, Luis, Andreas Fuster, and Basit Zafar, "Home price expectations and behaviour: Evidence from a randomized information experiment," *The Review of Economic Studies*, 2019, *86* (4), 1371–1410.

Athey, Susan and Guido W Imbens, "The econometrics of randomized experiments," in "Handbook of economic field experiments," Vol. 1, Elsevier, 2017, pp. 73–140.

Bail, Christopher A, "Can Generative AI improve social science?," *Proceedings of the National Academy of Sciences*, 2024, *121* (21), e2314021121.

Binder, Carola C, "Measuring uncertainty based on rounding: New method and application to inflation expectations," *Journal of Monetary Economics*, 2017, *90*, 1–12.

Binder, Carola Conces, Jeffrey R Campbell, and Jane M Ryngaert, "Consumer inflation expectations: Daily dynamics," *Journal of Monetary Economics*, 2024, *145*, 103613.

Bisbee, James, Joshua D Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M Larson, "Synthetic replacements for human survey data? the perils of large language models," *Political Analysis*, 2024, *32* (4), 401–416.

Bracha, Anat and Jenny Tang, "Inflation levels and (in) attention," *Review of Economic Studies*, 2025, *92* (3), 1564–1594.

Cavallo, Alberto, Guillermo Cruces, and Ricardo Perez-Truglia, "Inflation expectations, learning, and supermarket prices: Evidence from survey experiments," *American Economic Journal: Macroeconomics*, 2017, *9* (3), 1–35.

Coibion, Olivier, Dimitris Georgarakos, Yuriy Gorodnichenko, and Michael Weber, "Forward guidance and household expectations," *Journal of the European Economic Association*, 2023, *21* (5), 2131–2171.

_, Yuriy Gorodnichenko, and Michael Weber, "Monetary policy communications and their effects on household inflation expectations," *Journal of Political Economy*, 2022, *130* (6), 1537–1584.

_, _, and Tiziano Ropele, "Inflation expectations and firm decisions: New causal evidence," *The Quarterly Journal of Economics*, 2020, *135* (1), 165–219.

D'Acunto, Francesco, Ulrike Malmendier, and Michael Weber, "What do the data tell us about inflation expectations?," in "Handbook of economic expectations," Elsevier, 2023, pp. 133–161.

D'Acunto, Francesco, Ulrike Malmendier, Juan Ospina, and Michael Weber, "Exposure to grocery prices and inflation expectations," *Journal of Political Economy*, 2021, *129* (5), 1615–1639.

Eisfeldt, Andrea L and Gregor Schubert, "AI and Finance," *NBER working paper*, 2024, (No. w33076).

Faberman, R Jason, Andreas I Mueller, Ayşegül Şahin, and Giorgio Topa, "Job search behavior among the employed and non-employed," *Econometrica*, 2022, *90* (4), 1743–1779.

Giglio, Stefano, Matteo Maggiori, Johannes Stroebel, and Stephen Utkus, "Five facts about beliefs and portfolios," *American Economic Review*, 2021, *111* (5), 1481–1522.

Golchin, Shahriar and Mihai Surdeanu, "Time Travel in LLMs: Tracing Data Contamination in Large Language Models," in "The Twelfth International Conference on Learning Representations" 2024.

Haaland, Ingar, Christopher Roth, and Johannes Wohlfart, "Designing information provision experiments," *Journal of economic literature*, 2023, *61* (1), 3–40.

_, _, Stefanie Stantcheva, and Johannes Wohlfart, "Understanding Economic Behavior Using Open-Ended Survey Data," *Journal of Economic Literature*, forthcoming.

Hajdini, Ina, Edward S Knotek II, John Leer, Mathieu Pedemonte, Robert Rich, and Raphael Schoenle, "Indirect consumer inflation expectations: Theory and evidence," *Journal of monetary economics*, 2024, *145*, 103568.

Hansen, Anne Lundgaard, John J Horton, Sophia Kazinnik, Daniela Puzzello, and Ali Zarifhonarvar, "Simulating the survey of professional forecasters," *Available at SSRN*, 2024.

Hewitt, Luke, Ashwini Ashokkumar, Isaias Ghezae, and Robb Willer, "Predicting results of social science experiments using large language models," *Preprint*, 2024.

Horton, John J, "Large language models as simulated economic agents: What can we learn from homo silicus?," *NBER working paper*, 2023, (No. w31122).

Jiang, Janet Hua, Rupal Kamdar, Kelin Lu, and Daniela Puzzello, "How do households respond to expected inflation? an investigation of transmission mechanisms," *Bank of Canada Staff Working Paper*, 2024.

Korenok, Oleg, David Munro, and Jiayi Chen, "Inflation and attention thresholds," *Review of Economics and Statistics*, 2023, pp. 1–28.

Korinek, Anton, "Generative AI for economic research: LLMs learn to collaborate and reason," *NBER working paper*, 2024, (No. w33198).

Liu, Yachuan, Xiaochun Wei, Lin Shi, Xinnuo Li, Bohan Zhang, Paramveer Dhillon, and Qiaozhu Mei, "ExAnte: A Benchmark for Ex-Ante Inference in Large Language Models," *arXiv preprint arXiv:2505.19533*, 2025.

Malmendier, Ulrike and Stefan Nagel, "Learning from inflation experiences," *The Quarterly Journal of Economics*, 2016, *131* (1), 53–87.

Manski, Charles F, "Measuring expectations," *Econometrica*, 2004, *72* (5), 1329–1376.

Park, Joon Sung, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein, "Generative agent simulations of 1,000 people," *arXiv preprint arXiv:2411.10109*, 2024.

Roberts, Manley, Himanshu Thakur, Christine Herlihy, Colin White, and Samuel Dooley, "To the Cutoff... and Beyond? A Longitudinal Perspective on LLM Data Contamination," in "ICLR" 2024.

Salecha, Aadesh, Molly E Ireland, Shashanka Subrahmanya, João Sedoc, Lyle H Ungar, and Johannes C Eichstaedt, "Large language models show human-like social desirability biases in survey responses," *arXiv preprint arXiv:2405.06058*, 2024.

Tjuatja, Lindia, Valerie Chen, Tongshuang Wu, Ameet Talwalkwar, and Graham Neubig, "Do llms exhibit human-like response biases? a case study in survey design," *Transactions of the Association for Computational Linguistics*, 2024, *12*, 1011–1026.

Weber, Michael, Bernardo Candia, Hassan Afrouzi, Tiziano Ropele, Rodrigo Lluberas, Serafin Frache, Brent Meyer, Saten Kumar, Yuriy Gorodnichenko, Dimitris Georgarakos, Olivier Coibion, Geoff Kenny, and Jorge Ponce, "Tell Me Something I Don't Already Know: Learning in Low-and High-Inflation Settings," *Econometrica*, 2025, *93* (1), 229–264.

_ , Francesco d'Acunto, Yuriy Gorodnichenko, and Olivier Coibion, "The subjective inflation expectations of households and firms: Measurement, determinants, and implications," *Journal of Economic Perspectives*, 2022, *36* (3), 157–184.

Wu, Jing Cynthia, Shihan Xie, Yinxi Xie, and Ji Zhang, "Fiscal Policy: Financing and Indebtedness," *Available at SSRN*, 2025.

Xie, Shihan, "An Estimated Model of Household Inflation Expectations: Information Frictions and Implications," *The Review of Economics and Statistics*, 07 2025, *107* (4), 1042–1058.

Zarifhonarvar, Ali, "Evidence on Inflation Expectations Formation Using Large Language Models," *Available at SSRN 4825076*, 2024.

Appendices

A LLM-based survey prompts

A.1 Prompts for testing knowledge restriction

Persona You are answering questions as if you were a human. Do not break character. You are a { age } year old { gender } who is { marital status } with an education level of { education } degree and income category of { income } who lives in the state of { state }.

Knowledge restriction Do not reference or rely on any events or developments that occurred afterwards. Do not search for real data or base your answers on actual economic figures or official statistics. Instead, answer as a typical person might, based on their observations and general sense of the economy at the time.

Even awareness { Event question } Please provide your answer in yes or no.

List of events The following set of questions is asked on survey dates both before and after the event:

- Do you know of the September 11 attacks, also known as 9/11?
- Do you know of a U.S.-led military invasion of Iraq recently?
- Do you know of Lehman Brothers declaring bankruptcy?
- Do you know if Barack Obama has been elected U.S. President?
- Do you know if Donald Trump has been elected U.S. President?
- Do you know of a global pandemic called COVID-19?

A.2 Full survey prompts

Persona You are answering questions as if you were a human. Do not break character. You are a { age } year old { gender } who is { marital status } with an education level of { education } degree and income category of { income } who lives in the state of { state }.

Knowledge restriction You are responding to this survey in { survey date }. Do not reference or rely on any events or developments that occurred afterwards. Do not search for real data or base your answers on actual economic figures or official statistics. Instead, answer as a typical person might, based on their observations and general sense of the economy at the time.

Prior Please estimate the probability (as a percentage) for each of the following inflation/deflation scenarios over the next 12 months. Each probability must be between 0% and 100%. You may use up to 2 decimal points (e.g., 7.25%). The sum of all probabilities must equal exactly 100%. Return only a list of the numbers (i.e., 50 instead of '50%').

- Inflation of 12% or more: ______________%
- Inflation between 8% and 12%: ______________%
- Inflation between 4% and 8%: ______________%
- Inflation between 2% and 4%: ______________%
- Inflation between 0% and 2%: ______________%
- Deflation between 2% and 4%: ______________%
- Deflation between 4% and 8%: ______________%
- Deflation between 8% and 12%: ______________%
- Deflation of 12% or more: ______________%

Treatments

T0 Control group: No information provided

T1 Past inflation: *" Over the last twelve months, the inflation rate in the U.S. (as measured by the Consumer Price Index) was { treatment } %. "*

T2 Fed target: *" The inflation target of the Federal Reserve is 2% per year. "*

T3 Fed forecast: *" The U.S. Federal Open Market Committee (which sets short-term interest rates) forecasts { treatment }% inflation rate in { year }. "*

Posterior What do you expect the rate of inflation to be over the next 12 months? Please give your best guess. This question requires a numerical response in the form of an integer or decimal (e.g., -12, 0, 1, 2, 3.45, ...). Respond with just your number on a single line. If your response is equivalent to zero, report '0'.

B Proof of Proposition 1

Before presenting the final proof, we first introduce the notation and setup, then state and prove two lemmas, and finally present the main result.

B.1 Notation and Setup

Let $\hat{\tau} = \bar{Y}t - \bar{Y}c$ denote the ATE estimator, where $\bar{Y}t$ and $\bar{Y}c$ are the sample means of the outcomes for the treated and control groups, respectively. Suppose the full sample contains N units, partitioned into H strata for $h \in \mathcal{H} := 1, 2, \ldots, H$. Let n_h denote the sample size of stratum h, and let w_h denote the probability that a participant belongs to stratum h. In a simple randomization design, each participant in both the control and treatment groups has a probability w_h of being in stratum h. Under a complete stratification design, by contrast, exactly a fraction w_h of the control group and a fraction w_h of the treatment group belong to stratum h. For each stratum h, let σ_{ht}^2 and σ_{hc}^2 denote the variances of Y for the treated and control groups. Define the treatment indicator $g_i \in \{t, c\}$, where $g_i = t$ if unit i is assigned to the treatment group and $g_i = c$ if assigned to the control group. Finally, define $\mu_{ht} := \mathbb{E}[Y_i | i \in h, g_i = t]$ and $\mu_{hc} := \mathbb{E}[Y_i | i \in h, g_i = c]$ as the expected potential outcomes in stratum h for treated and control groups.

For simplicity, we assume the treated and control groups contain equal numbers of participants. The results can be easily generalized to settings with unequal treatment probabilities.

B.2 Lemmas

We derive the variances of simple randomization and complete stratification in Lemmas 1 and 2, respectively.

Lemma 1. *Under simple randomization, the variance of $\hat{\tau}$ is*

$$var^r(\hat{\tau}) = \underbrace{\frac{2}{N} \sum_{h \in \mathcal{H}} w_h \left(\sigma_{ht}^2 + \sigma_{hc}^2\right)}_{\textit{within-stratum}} + \underbrace{\frac{1}{N} \sum_{h \in \mathcal{H}} w_h \left[(\mu_{ht} - \mu_t)^2 + (\mu_{hc} - \mu_c)^2 \right]}_{\textit{between-stratum imbalance}}. \tag{B.1}$$

Proof of Lemma 1. We first rewrite the sample means $\bar{Y}_t$ and $\bar{Y}_c$ as weighted averages of sample means within each stratum and arms:

$$\bar{Y}_t = \sum_h r_{ht} \, \bar{Y}_{h1}, \qquad \bar{Y}_c = \sum_h r_{hc} \, \bar{Y}_{h0},$$

where $r_{ht} := n_{ht}/n_t$ and $r_{hc} := n_{hc}/n_c$ are the arm compositions across strata. With balanced arms, $n_t = n_c = N/2$ and $\sum_{h\in\mathcal{H}} r_{ht} = \sum_{h\in\mathcal{H}} r_{hc} = 1$. We also define the strata composition $r_t := (r_{1t}, r_{2t}, \ldots, r_{Ht})'$, $r_c := (r_{1c}, r_{2c}, \ldots, r_{Hc})'$ and $w := (w_1, w_2, \ldots, w_H)'$, where $r_t \perp r_c$, and $\mathbb{E}[r_{ht}] = \mathbb{E}[r_{hc}] = w_h$. Decompose each stratum-by-arm mean as a population mean and a noise term:

$$\bar{Y}_{ht} = \mu_{ht} + \bar{\varepsilon}_{ht}, \qquad \bar{Y}_{hc} = \mu_{hc} + \bar{\varepsilon}_{hc},$$

where $\bar{\varepsilon}_{h\cdot}$ denotes the average noise of treated/control units within strata h. With independent survey samples, $\mathbb{E}[\bar{\varepsilon}_{ht}] = \mathbb{E}[\bar{\varepsilon}_{hc}] = 0$, $\mathrm{var}(\bar{\varepsilon}_{ht}) = \frac{\sigma_{ht}^2}{n_{ht}}$, and $\mathrm{var}(\bar{\varepsilon}_{hc}) = \frac{\sigma_{hc}^2}{n_{hc}}$. Rewrite $\hat{\tau}$ as follows:

$$\begin{aligned}\hat{\tau} &= \sum_{h\in\mathcal{H}} \{r_{ht}\mu_{ht} - r_{hc}\mu_{hc}\} + \sum_{h\in\mathcal{H}} \{r_{ht}\varepsilon_{h1} - r_{hc}\varepsilon_{hc}\} \\ &= \underbrace{\sum_{h\in\mathcal{H}} w_h(\mu_{ht} - \mu_{hc})}_{\text{ATE under balanced strata}} + \underbrace{\sum_{h\in\mathcal{H}} [(r_{ht} - w_h)\mu_{ht} + (r_{hc} - w_h)\mu_{hc}]}_{\text{between-stratum imbalance}} + \underbrace{\sum_{h\in\mathcal{H}} (r_{ht}\bar{\varepsilon}_{ht} - r_{hc}\bar{\varepsilon}_{hc})}_{\text{within-stratum noise}}.\end{aligned}$$

The ATE estimator has three components. The first component is the ATE when both arms have the same strata composition, i.e., $r_{hc} = r_{ht} = w_h$ for all h. The second term characterizes the imbalance of strata composition in the two treatment arms. The last term represents the within-stratum noise. Condition on the composition r_t, we decompose the variance:

$$\mathrm{var}^r(\hat{\tau}) = \underbrace{\mathbb{E}\big[\mathrm{var}(\hat{\tau}|r_t)\big]}_{\text{within-stratum}} + \underbrace{\mathrm{var}\big(\mathbb{E}[\hat{\tau}|r_t]\big)}_{\text{between-stratum}}.$$

Define the composition term $C := \sum_{h\in\mathcal{H}} [(r_{ht} - w_h)\mu_{ht} + (r_{hc} - w_h)\mu_{hc}]$ and the noise term $U := \sum_{h\in\mathcal{H}}\{r_{ht}\varepsilon_{h1} - r_{hc}\varepsilon_{h0}\}$. As we will show shortly, $\mathbb{E}\big[\mathrm{var}(\hat{\tau}|r_t)\big] = \mathbb{E}\big[\mathrm{var}(U|r_t)\big]$, which carries the within-stratum variation; $\mathrm{var}\big(\mathbb{E}[\hat{\tau}|r_t]\big) = \mathrm{var}(\mathbb{E}[C|r_t])$ is the between-stratum variation.

We start with analyzing $\mathbb{E}\big[\mathrm{var}(\hat{\tau}|r_t)\big]$. First, it is clear that the first term $\sum_{h\in\mathcal{H}} w_h(\mu_{ht} - \mu_{hc})$ is a constant, and hence itself does not have variation. Conditioning on r_t, the second term's variance is zero, i.e. $\mathrm{var}(\sum_{h\in\mathcal{H}}(r_{ht} - w_h)\,(\mu_{ht} + \mu_{hc})|r_t) = 0$. Thus $\mathbb{E}\big[\mathrm{var}(\hat{\tau}|r_t)\big] = \mathbb{E}\big[\mathrm{var}(U|r_t)\big]$, and

$$\begin{aligned}\mathrm{var}(U|r_t) =& \mathrm{var}(r_{ht}\bar{\varepsilon}_{ht} - r_{hc}\bar{\varepsilon}_{hc}) \\ =& \mathrm{var}(r_{ht}\bar{\varepsilon}_{ht}) + \mathrm{var}(r_{hc}\bar{\varepsilon}_{hc})\end{aligned}$$

$$
\begin{aligned}
&= \sum_{h\in\mathcal{H}} r_{ht}^2 \frac{\sigma_{ht}^2}{n_{ht}} + \sum_{h\in\mathcal{H}} r_{hc}^2 \frac{\sigma_{hc}^2}{n_{hc}} \\
&= \frac{2}{N} \sum_{h\in\mathcal{H}} (r_{ht}\sigma_{ht}^2 + r_{hc}\sigma_{hc}^2).
\end{aligned}
$$

Since $\mathbb{E}[r_{ht}] = \mathbb{E}[r_{hc}] = w_h$, we have

$$
\mathbb{E}[\mathrm{var}(\hat{\tau}|r_t)] = \mathbb{E}\left[\frac{2}{N} \sum_{h\in\mathcal{H}} (r_{ht}\sigma_{ht}^2 + r_{hc}\sigma_{hc}^2)\right] = \frac{2}{N} \sum_{h\in\mathcal{H}} w_h(\sigma_{ht}^2 + \sigma_{hc}^2).
$$

Next, we analyze $\mathrm{var}\big(\mathbb{E}[\hat{\tau}|r_t]\big)$. Again, the first term in $\hat{\tau}$ is a constant and has zero variance. The third term has zero conditional expectation, because

$$
\mathbb{E}[\sum_{h\in\mathcal{H}} (r_{ht}\bar{\varepsilon}_{ht} - r_{hc}\bar{\varepsilon}_{hc})|r_{ht}] = \sum_{h\in\mathcal{H}} (r_{ht}\mathbb{E}[\bar{\varepsilon}_{ht}] - r_{hc}\mathbb{E}[\bar{\varepsilon}_{hc}]) = 0.
$$

Hence, $\mathrm{var}\big(\mathbb{E}[\hat{\tau}|r_t]\big) = \mathrm{var}(\mathbb{E}[C|r_t]) = \mathrm{var}(C)$, with

$$
C = \sum_{h\in\mathcal{H}} \left[(r_{ht} - w_h)\mu_{ht} + (r_{hc} - w_h)\mu_{hc}\right].
$$

Notice that

$$
\mathrm{var}(C) = \mathrm{var}\Bigg(\underbrace{\sum_h (r_{ht} - w_h)\,\mu_{ht}}_{\mathrm{A}}\Bigg) + \mathrm{var}\Bigg(\underbrace{\sum_h (r_{hc} - w_h)\,\mu_{hc}}_{\mathrm{B}}\Bigg)
$$

under $r_t \perp r_c$. Moreover, $\mathbb{E}[\sum_h (r_{ht} - w_h)] = \mathbb{E}[\sum_h (r_{hc} - w_h)] = 0$, and

$$
\mathrm{cov}(r_t) = \mathrm{cov}(r_c) = \frac{1}{N}(Diag(w) - ww^T),
$$

where $Diag(w)$ denotes an $H \times H$ diagonal matrix with the entries of w on its diagonal. We have

$$
\begin{aligned}
\mathrm{var}(A) &= \frac{1}{N}\left[\sum_{h\in\mathcal{H}} w_h\mu_{ht}^2 - (\sum_{h\in\mathcal{H}} w_h\mu_{ht})^2\right] = \frac{1}{N}\left[\sum_{h\in\mathcal{H}} w_h(\mu_{ht} - \mu_t)^2\right], \\
\mathrm{var}(B) &= \frac{1}{N}\left[\sum_{h\in\mathcal{H}} w_h\mu_{hc}^2 - (\sum_{h\in\mathcal{H}} w_h\mu_{hc})^2\right] = \frac{1}{N}\left[\sum_{h\in\mathcal{H}} w_h(\mu_{hc} - \mu_c)^2\right],
\end{aligned}
$$

which concludes the proof. $\square$

Lemma 2. *Under complete stratification, the variance of $\hat{\tau}$ is:*

$$var^s(\hat{\tau}) = \frac{2}{N} \sum_{h \in \mathcal{H}} w_h(\sigma^2_{ht} + \sigma^2_{hc}). \tag{B.2}$$

Proof of Lemma 2. Notice that $r_{ht} = r_{hc} = w_h$, therefore

$$\begin{aligned}
\hat{\tau} &= \sum_{h \in \mathcal{H}} \{r_{ht}\mu_{ht} - r_{hc}\mu_{hc}\} + \sum_{h \in \mathcal{H}} \{r_{ht}\varepsilon_{h1} - r_{hc}\varepsilon_{hc}\} \\
&= \underbrace{\sum_{h \in \mathcal{H}} w_h(\mu_{ht} - \mu_{hc})}_{\text{ATE under balanced strata}} + \underbrace{\sum_{h \in \mathcal{H}} w_h(\bar{\varepsilon}_{ht} - \bar{\varepsilon}_{hc})}_{\text{within-stratum noise}},
\end{aligned}$$

which implies that $\text{var}\big(\mathbb{E}[\hat{\tau}|r_t]\big) = 0$. Thus, the variance is simply

$$\text{var}^s(\hat{\tau}) = \mathbb{E}\big[\text{var}(\hat{\tau}|r_t)\big] = \frac{2}{N} \sum_{h \in \mathcal{H}} w_h(\sigma^2_{ht} + \sigma^2_{hc}).$$

□

B.3 Proof

Proof of Proposition 1. Comparing results in Lemmas 1 - 2,

$$\text{var}^r(\hat{\tau}) - \text{var}^s(\hat{\tau}) = \frac{1}{N} \sum_{h \in \mathcal{H}} w_h \left[(\mu_{ht} - \mu_t)^2 + (\mu_{hc} - \mu_c)^2\right] \geq 0. \tag{B.3}$$

Therefore, we conclude that $\text{var}^s(\hat{\tau}) \leq \text{var}^r(\hat{\tau})$. □

C Validation Details

C.1 Treatment Information of Weber et al. (2025)

Table C.1: SAMPLE COVERAGE AND TREATMENT INFORMATION

Experiment date	**Treatment information**		
	T1: Past inflation	T2: Fed target	T3: Fed forecast
2018Q2	2.3	2	1.9
2019Q1	1.8		
2021Q2	2.6	2	2.3
2021Q3	5.8	2	2.1
2021Q4	6.2		
2022Q3	8.5	2	2.6
2022Q4	7.8	2	2.8
2023Q2	6.0	2	
2023Q3	3.0	2	2.5
2023Q4	3.2		

Notes: This table summarizes the RCTs conducted with U.S. households in Weber et al. (2025). Each experiment date corresponds to a survey wave in which respondents were randomly assigned to a control group or to one of three treatment groups: T1 - past inflation, T2 - Fed target, T3 - Fed forecast. The values shown in the table are the exact numerical information provided to treated respondents in each survey wave.

C.2 Demographic Distribution

Table C.2: DEMOGRAPHIC DISTRIBUTION

	LLM (1)	SCE (2)
Education		
College	50.0%	55.3%
Some College	36.5%	32.5%
High School	13.5%	12.2%
Income		
Under 50k	32.0%	35.7%
50k to 100k	34.0%	35.8%
Over 100k	34.0%	28.5%
Age		
Under 40	34.0%	26.6%
40 to 60	41.5%	38.0%
Over 60	24.5%	35.4%
Gender		
Female	50.5%	51.8%
Male	49.5%	48.2%
Marital Status		
Married	60.0%	63.3%
Not Married	40.0%	36.7%
Region		
Midwest	27.5%	23.2%
Northeast	12.5%	20.3%
South	33.5%	33.7%
West	26.5%	22.8%

Notes: This table reports the distribution of LLM personas in Column (1) and the distribution of demographics from the Federal Reserve Bank of New York's Survey of Consumer Expectations (SCE) in Column (2).

D Robustness

D.1 GPT-5

In August 2025, OpenAI launched GPT-5, which costs roughly 10 times the cost of GPT-4.1 for our survey design due to substantially more reasoning and longer context windows.

We repeat our main validation exercises in Section 3.3 with GPT-5, and Table D.1 reports the results. Using GPT-5, correlations with inflation range from 0.38 to 0.83, with a pooled value of 0.59, comparable to our baseline. Correlations with the GPT-4.1 baseline are even higher, between 0.72 and 0.85 across treatments. The correlations with human surveys also display similar patterns. The fact that our findings remain consistent across both GPT-4.1 and GPT-5 underscores that our results are robust to model choice.

Table D.1: ROBUSTNESS: GPT-5

	(1) T1: Past inflation	(2) T2: Fed target	(3) T3: Fed forecast	(4) Pooled
Inflation	0.83	0.38	0.42	0.59
Baseline	0.85	0.72	0.72	0.79
Human	0.57	0.47	-0.44	0.26

Notes: This table reports the correlation between scaled slopes generated by using the GPT-5 model. The correlations are with current inflation, scaled slopes based on baseline results from Section 3.3, and scaled slopes based on the human survey from Weber et al. (2025). "Pooled" correlations are calculated by pooling estimates from all treatment types.

D.2 Personas

We repeat our main validation exercises in Section 3.3 using a fresh draw of 200 personas from the SCE, and Table D.2 reports the results. Despite replacing the entire respondent pool, the correlations of treatment effects with inflation remain strong, ranging between 0.52 and 0.9, comparable to our baseline in Table 2, which ranges from 0.73 to 0.92. Correlations with the baseline scaled slopes are high, between 0.63 and 0.90, and the alignment with human survey estimates in Figure 4 remains similar, with all but one positive correlation. We therefore conclude that our results are not driven by the choice of initial personas, but are robust to alternative respondent samples.

Table D.2: ROBUSTNESS: PERSONAS

	(1) T1: Past inflation	(2) T2: Fed target	(3) T3: Fed forecast	(4) Pooled
Inflation	0.52	0.59	0.90	0.53
Baseline	0.63	0.79	0.90	0.77
Human	0.54	0.38	-0.28	0.22

Notes: This table reports the correlation between scaled slopes generated by using a different set of personas. The correlations are with current inflation, scaled slopes based on baseline results from Section 3.3, and scaled slopes based on the human survey from Weber et al. (2025). "Pooled" correlations are calculated by pooling estimates from all treatment types.

D.3 Inflation-adjusted Income Bins

To adjust income bins to inflation, we redefine bins in real terms: we anchor cutoffs in 2019 and adjust for inflation in all other years while keeping the bin shares fixed at their baseline levels. For this robustness check, we use the extended sample in Section 4.3 and focus on the T1 past-inflation treatment, which spans the longest period and is most sensitive to the nominal–real distinction. Results are reported at an annual frequency.

Figure D.1 shows that using inflation-adjusted bins delivers results that closely mirror our baseline with fixed bins: both approaches yield nearly identical patterns of state-dependent attention to inflation. The correlation with realized inflation is 0.55 when using inflation-adjusted bins, which is very close to the 0.62 correlation obtained when using fixed bins. Furthermore, Figure D.2 shows that the scaled slopes under the two binning methods line up closely along the 45-degree line, with a correlation of 0.92 across the survey waves. This confirms that our main findings are robust to the choice of binning method.

Overall, our findings are not sensitive to the binning method: the estimated slopes remain highly correlated and lead to the same substantive conclusions, whether income bins are fixed in nominal terms or indexed to inflation.

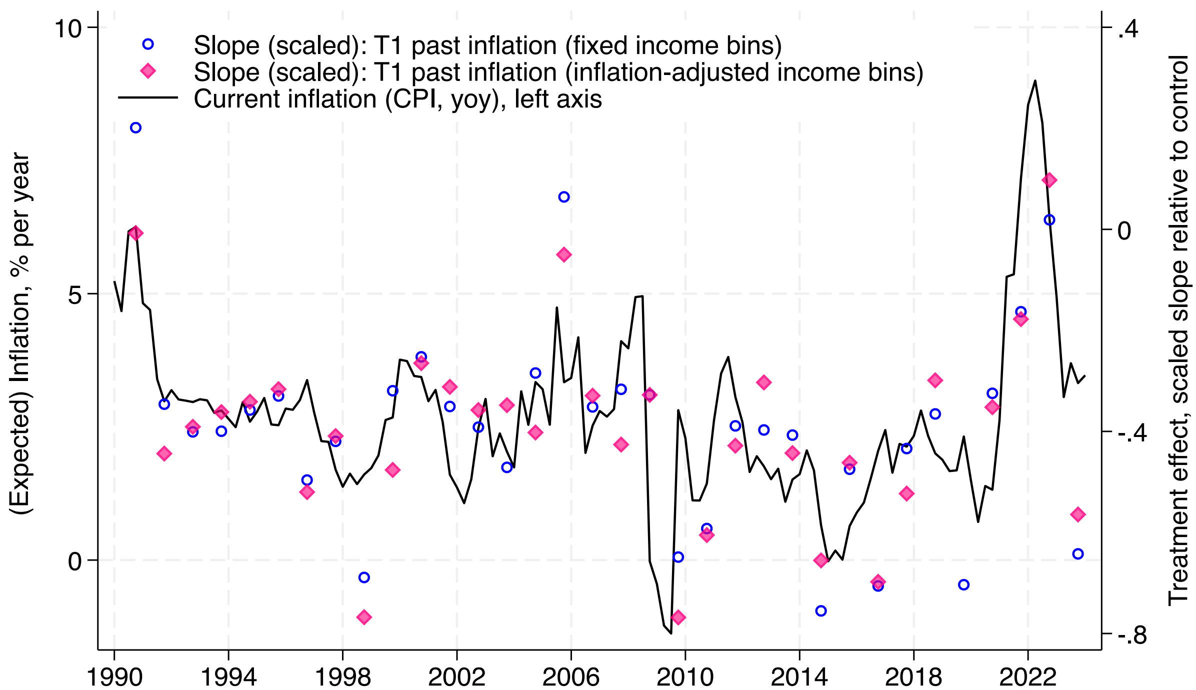

Figure D.1: ROBUSTNESS: INCOME BINS (OVER TIME)

Notes: This figure plots the comparison of scaled slopes from fixed (blue circle) and inflation-adjusted (pink diamond) income bins, based on past inflation treatment using LLM-generated responses.

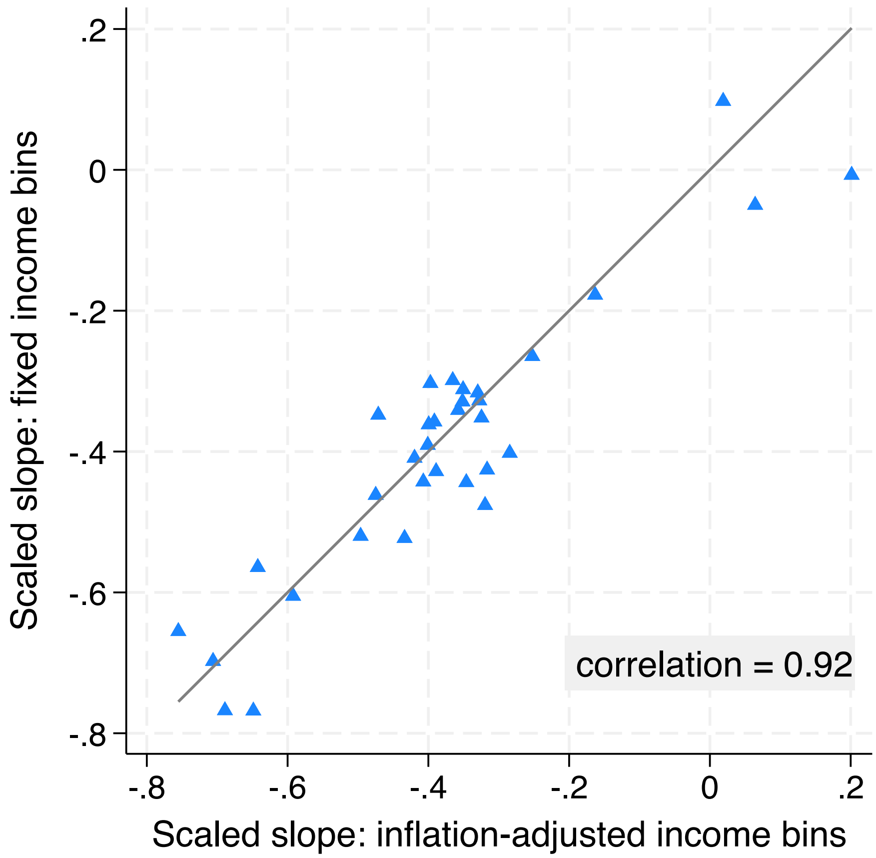

Figure D.2: Robustness: income bins (scatter plot)

Notes: This figure plots the comparison between the scaled slope estimated using fixed income bins and inflation-adjusted income bins, based on past inflation treatment. The solid gray line represents the 45-degree line.

E Reasoning

E.1 Classification

To analyze reasoning patterns using a GPT-based classifier, we first develop a coding scheme that translates open-ended text into analyzable data. The scheme consists of well-defined categories, definitions, and illustrative examples. We remain agnostic about the model's internal process for forming inflation expectations and focus instead on identifying recurring themes that emerge in respondents' explanations. Although qualitative analysis ultimately relies on human judgment, LLMs can complement this process by detecting subtle or diffuse themes that human readers may overlook. Accordingly, we construct the coding scheme through a hybrid process that combines human analysis with LLM-assisted reading.

The procedure unfolds as follows:

1. **Human review:** We manually read through explanations across survey waves to identify salient reasoning categories.

2. **LLM-assisted exploration:** In parallel, we prompt the LLM to analyze all explanations collectively and group them into at most ten reasoning categories, each illustrated with examples.

3. **Comparison:** We compare the LLM-generated classification with the human-derived scheme to identify categories or themes missed by human coders.

4. **Refinement:** For each additional category suggested by the LLM, we review its supporting examples and determine whether it should be incorporated into the final coding scheme.

Once the coding scheme is finalized, we prompt the LLM to evaluate each category independently to determine whether a given explanation exhibits that form of reasoning. The classifier is implemented through detailed instructions that specify the label, its definition, coding guidelines, and illustrative examples. Classifications are non-exclusive: a single explanation may be assigned to multiple categories or to none. Crucially, this LLM-based classification scheme does not rely on predefined keywords but interprets the full textual context, distinguishing this approach from traditional keyword-based text analysis.

E.2 Examples

Normal This category is consistently prevalent except during the Great Recession and COVID inflation surge, we select one example from before the Great Recession and one from after, to illustrate its stability across normal times.

1996Q4 *"I don't think we'll see anything crazy like big inflation or deflation.* ***Prices seem fairly steady right now, and the economy feels stable enough.*** *There could be a little uptick or downturn but probably just in that 0–2% range, which seems to be how it's been going."*

2017Q4 *"I think we'll* ***most likely continue to see mild inflation or prices staying about the same****, since things haven't been changing a whole lot lately. I don't expect big jumps in prices or any serious deflation, just more of what we've been experiencing—slow and* ***steady****."*

Normalizing This category appears primarily during the early 1990-recession, the Great Recession, and COVID inflation surge. We include two examples from those episodes.

1990Q4 *"I think prices are still rising, but not as fast as they did back in the '70s.* ***Inflation seems to be slowing down*** *but I doubt it will drop super low or go negative. I don't expect deflation at all. Most likely, inflation will stay in the 2-4% or 0-2% range, maybe a little higher, but not like double digits."*

2023Q1 *"I think we're unlikely to see either extreme inflation or real deflation this year, though prices still feel high and rising, just* ***not at the crazy levels of a year ago****. Most likely, inflation will stay in that somewhat elevated low-single-digit range,* ***with a chance it dips closer to normal****, but not enough to see actual deflation."*

Personal Observation This category remains relatively common across all periods. We include one example from a crisis period and one from a normal period to capture its continuity and variation in context.

1992Q4 *"I just don't see prices getting out of control with inflation way up above 8%, but prices keep creeping up—****grocery store, gas, property taxes****. So I put most of my odds in the 2-4% and 0-2% range, with a bit higher for the 4-8% just to be careful. Almost no chance of real deflation, that just doesn't seem likely these days."*

2023Q1 *"I'm seeing prices still climbing in a lot of areas, **especially groceries and insurance**, but not as wild as before—it feels like things are slowing down a bit. I don't expect huge inflation (over 12%), but I don't think we'll get back to really low inflation or any serious deflation soon. Florida seems to keep getting more expensive overall, but not at a crazy rate."*

Monetary Policy References to monetary policy are rare and arise almost exclusively during the Great Recession and COVID inflation surge. We therefore select one example from each major crisis to reflect the contexts in which this reasoning typically emerges.

2009Q2 *"Given everything that's going on with the economy in 2009—coming off the financial crisis, recession, and **the Fed lowering interest rates**—I think inflation is not our main concern over the next year. Prices seem pretty stable, and there are even worries about deflation, but I doubt it'll be severe. Most likely, inflation will be low, maybe between 0% and 2%, or we could see mild deflation. High inflation seems very unlikely to me right now. "*

2022Q4 *"I think inflation will stay a bit elevated compared to normal, but not spike out of control. I don't see deflation as likely at all, given all the supply chain issues, energy prices, and the way things seem to still be costing more at the grocery store and gas pump. **The Fed is raising rates**, but I don't think that's enough to turn us toward deflation soon. Most likely we'll land somewhere between 4% and 8%, maybe a bit higher for a while."*

Business Cycle This category is concentrated almost entirely in the Great Recession period, so we select examples from that time.

2009Q1 *"I think we're most likely to see low inflation or even a bit of deflation in the next year because of how **the economy has been struggling lately—lots of layoffs, businesses closing**, and people being cautious about spending. High inflation seems pretty unlikely right now, but I do think there's a decent chance of mild deflation, just because demand has really dropped off."*

2011Q4 *"I think we'll see low inflation overall, with prices rising only slightly since **the economy is still recovering from the recession**. There's more chance of mild inflation than deflation, but some chance that prices could barely move or even fall a bit, especially with so many foreclosures and people struggling with work here in Nevada. Wild swings like very high inflation or deflation seem unlikely to me."*