LLM–Based Survey Framework for Expectations: Retrospective Coverage and Consistency

Jing Cynthia Wu[*] Jin Xi[†] Shihan Xie[‡]

September 10, 2025

Abstract

We propose a new LLM-based survey framework with broad applications for studying expectations. It introduces two innovations beyond traditional surveys: (1) retrospective and systematic temporal coverage through date-restricted prompts, enabling analysis across decades and at higher frequency; and (2) perfect internal consistency by fixing personas across control and treatment arms as well as over survey waves, and by holding treatments constant over time. We validate the framework in a multi-wave randomized controlled trial of inflation expectations surveys, recovering human-comparable treatment effects. We extend the benchmark human survey (10 waves, 2018–2023) to over 50 waves dating back to 1990 using LLMs, demonstrating the scalability and broader potential of the framework.

Keywords: large language models (LLMs), survey RCT, inflation expectations
JEL codes: C83, E52, E31

[*]University of Illinois Urbana-Champaign and NBER. Email: wujing@illinois.edu
[†]Chinese Academy of Sciences. Email: xijin@amss.ac.cn
[‡]University of Illinois Urbana-Champaign. Email: shihanx@illinois.edu

1 Introduction

Surveys play a central role in economics, offering direct measures of expectations not observable from 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 housing and labor market decisions (Armona et al., 2019; Faberman et al., 2022). However, traditional human surveys, while invaluable, face significant limitations. First, they often depend on access to institutional channels—such as funding, survey firms, or established panels— which imposes barriers for many researchers and limits the scope and sample size of each survey. More fundamentally, they cannot be fielded retrospectively, makeing many counterfactuals and historical settings permanently out of reach. To address these limitations, we propose a survey design using LLMs that is expandable over time and accessible to a broader research community.

Our new LLM-based survey framework is broadly applicable to the study of expectations: it replicates the essential features of human surveys while enabling capabilities that traditional methods cannot. First, it provides retrospective and systematic temporal coverage through date-restricted prompts. Second, it ensures perfect internal consistency by reusing the same personas across control and treatment arms and repeated waves, and applying identical treatments over time. In doing so, our framework creates an ideal survey environment—retrospective and systematic coverage with internal consistency—opening avenues for research beyond the reach of human surveys.

A key limitation of human surveys is that past expectations cannot be reconstructed without archival data. Our framework overcomes this by introducing date restriction, which, to our knowledge, is the first application of this idea to LLM-based surveys. By restricting the model's knowledge to information available only up to each survey date, the LLM can recreate past surveys as if they had been fielded at the time. This design thereby

makes it possible to extend existing surveys backward in time or at higher frequency, enabling systematic retrospective analysis of expectations and the study of long-run dynamics.

We validate that these date-restriction instructions work as intended by testing LLM's awareness of salient events, which include a total of six major events between 2001 and 2020, such as 9/11, the Iraq war, and the onset of COVID-19. As expected, LLMs reported no awareness before these events and near-universal recognition afterward, confirming that our design prevents hindsight bias while preserving historically grounded awareness.

Another central feature of our framework is is perfect internal consistency: LLMs allow researchers to fix respondent personas across control and treatment groups and over time, and to hold treatments word-for-word identical across waves. Human surveys, by contrast, often suffer from sample drift, respondent dropout, and evolving question wording. Our design therefore creates a fully balanced synthetic panel, guaranteeing comparability across time and conditions and ensuring that changes in responses reflect genuine information updates rather than shifts in sample composition or survey design.

To validate our LLM survey framework, we benchmark it against a state-of-the-art multi-wave randomized controlled trial (RCT) survey of inflation expectations. Specifically, we replicate the experimental design in Weber et al. (2025), adapting their treatments—past inflation (T1), the Fed target (T2), and the Fed forecast (T3)—into our LLM-based survey environment. We construct a set of personas sampled from the New York Fed's Survey of Consumer Expectations (SCE). 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. LLM agents reproduce the cross-wave variation observed in human data. Treatment effects are stronger during low-inflation episodes, when households typically pay less attention to inflation. They are weaker during

high-inflation episodes, when inflation is already salient and additional information has less impact on expectations. Estimated responsiveness tracks human benchmarks but often exhibits stronger correlations with inflation. For LLM surveys, the correlations with inflation range from 0.73 to 0.92, whereas for human surveys they range from 0.13 to 0.87 across treatments. This contrast reflects cleaner patterns of state-dependent attention from LLMs, plausibly reflecting 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, extending it to applications beyond the reach of traditional surveys. Using date-restricted prompts, we field surveys "as if" at any point in the past, systematically reconstructing decades of expectations data into a long-run panel of treatment effects under consistent experimental conditions. This design scales the benchmark human survey from 10 waves in 2018–23 to more than 50 waves dating back to 1990, with quarterly coverage during turbulent episodes such as the Great Recession and the post-COVID inflation surge, and annual coverage otherwise. This retrospective design enables analysis of how agents would have responded to information treatments during historical episodes, thereby overcoming the temporal and logistical constraints that limit human surveys.

The extended panel uncovers stable patterns of state-dependent attention: the responsiveness of inflation expectations to new information declines sharply during high-inflation periods and rises when inflation is low. Treatment effects remain strongly correlated with inflation across decades: 0.61 for past inflation (T1) and 0.76 for Fed forecasts (T3), while weaker for the Fed target (T2), potentially reflecting that the Fed's 2% target is fixed over time. Unlike human surveys, which are constrained by when and where data were collected, the LLM survey framework delivers complete historical coverage, enabling systematic analysis of how expectation formation responds to information across diverse macroeconomic environments.

We demonstrate that our findings are robust to alternative implementation choices along three dimensions. First, results remain stable when the experiment is re-run with an entirely new set of personas. Second, they are consistent across LLM versions: GPT-4.1, our baseline fixed-cutoff model, and GPT-5, a larger and more recent model with expanded capabilities. Third, they are unaffected by income bin definitions, with nearly identical results when bins are indexed to inflation rather than fixed in nominal terms. Together, these exercises confirm that our framework's conclusions are not sensitive to design variations.

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 the framework's ability to deliver long-run coverage. Section 5 provides robustness checks, while Section 6 concludes.

1.1 Literature

Recent reviews highlighting 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. Hewitt et al. (2024) demonstrate their ability to predict treatment effects in preregistered social science experiments. 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.

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 directly influences household inflation expectations (D'Acunto et al., 2021) and establish the causal effects of monetary policy communication (Coibion et al., 2022) and forward guidance (Coibion et al., 2023) on household beliefs. Recent work also leverages inflation expectations to measure attention (e.g., Bracha and Tang 2025; Weber et al. 2025; Pfäuti 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 with broad applications for studying expectations. The framework rests on two innovations. First, retrospective and systematic temporal coverage: by restricting the model's accessible knowledge to information available only up to each survey date, we can field credible surveys both retrospectively and at higher frequency. Second, perfect internal consistency: by reusing the same personas across treatment and control groups and over time, and repeating the same treatment arms and phrasing, we ensure comparability across time and conditions.

2.1 Date-Restricted Knowledge for Systematic Temporal Coverage

2.1.1 Design and Implementation

Our first design principle is date restriction. In practice, this means we instruct the model to ignore any information released after the designated survey date and to role-play as if

answering at that historical moment. This procedure allows us to generate counterfactual survey waves at arbitrary points in the past and to vary survey frequency flexibly, all at negligible marginal cost. The date-restriction setup thus provides the technical foundation for retrospective and systematic temporal coverage.

Our temporal 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.*

Notably, simply instructing the model to "answer as of [survey time]" is insufficient: LLMs tend to exhibit hindsight bias, drawing on knowledge of subsequent events unless explicitly prohibited. Our design adds such prohibitions against referencing future developments or ex post data, which are essential to enforce genuine date restriction.

2.1.2 Validation of Date-Restriction Instructions

Before proceeding to the second design principle, we first validate that the date-restriction setup functions as intended. Specifically, we test the model's awareness of major historical events using the same style of prompts and the same set of LLM agents as in our main surveys, anchored to specific experiment dates. For each event, we posed the same awareness question both shortly before the event occurred and shortly after, and measured the share of personas answering "yes".

The events included major, widely recognized developments such as 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). Each of these events was highly salient in U.S. public life, and a typical household would have been aware of them shortly after their occurrence. This makes them a natural benchmark for assessing whether our knowledge-restriction instructions prevent hindsight bias—i.e., the model's tendency to use information from the future—while still allowing it to reflect historically grounded awareness. For details on prompts, see Appendix A.1.

The results summarized in Table 1 align exactly with expectations. Before each event, LLM-agents almost always responded "no", while immediately afterward the responses shifted 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 appeared 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 cited 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 Perfect Internal Consistency

Another key advantage of using large language models (LLMs) for survey research is their ability to maintain perfect internal consistency, both on the respondent personas and the treatments.

In terms of personas, human surveys face inherent limitations. It is virtually impossible to guarantee that the same respondent profile appears in both treatment and control groups, or that the same individuals can be tracked consistently over long horizons. By contrast, LLMs allow researchers to define characteristics such as age, income, and education, and

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 Feb 2003	0 5	Apr 2003	100
Lehman Brothers declare bankruptcy	Sep 15, 2008	Jul 2008 Aug 2008	0 5	Oct 2008	98
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 Jan 2020	0 5	Apr 2020	100

Notes: This table reports the results of our knowledge restriction validation exercise. 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".

lock in the same set of personas across control and treatment arms and over all waves.

This design generates what is effectively a synthetic panel: a fixed set of respondents with constant demographic and background attributes. Because identical personas can be replicated across control and treatment arms and survey waves, the resulting panel is perfectly balanced and free of demographic drift or sample turnover. As a result, heterogeneity across waves or between treatment and control groups is greatly reduced relative to human panels. Moreover, by holding personas fixed, any shifts in responses over time can be more confidently attributed to evolving information rather than to changes in sample composition. Finally, personas can be re-run under different treatments or experimental conditions at any point, with no concerns about dropout or recruitment costs.

On the treatment side, the advantage of LLM agents lies in their ability to keep question phrasing perfectly consistent over time. In human surveys, wording often drifts across waves—even when addressing the same question—due to evolving research interests, survey redesigns, or efforts to improve clarity. By contrast, LLM-based surveys can administer identical treatments and phrasing repeatedly, ensuring comparability across time without introducing wording effects.

2.3 Ideal Survey Environment

Our framework delivers two methodological innovations that are unattainable with human surveys—retrospective and systematic temporal coverage and perfect internal consistency. In addition, it offers affordability and accessibility, enabling surveys to be scaled up and replicated at negligible marginal cost. Taken together, these features constitute an ideal survey environment.

To illustrate the affordability and accessibility, with GPT-4.1 (the LLM 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, smaller samples can deliver comparable statistical power, further reducing costs. This affordability expands the range of feasible research designs and broadens access to survey tools, allowing a wider community of scholars to implement studies that would otherwise be out of reach.

3 Validating the LLM Survey Approach

We begin by validating our LLM-based survey framework against a state-of-the-art multi-wave inflation expectations experiment, showing that it recovers information updating patterns comparable to those obtained from 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.

3.1 Benchmark Experiment: Weber et al. (2025)

Weber et al. (2025) roll out ten waves between 2018 and 2023 with the Nielsen Homescan Panel, which constitutes an exceptional scale and level of consistency for household survey experiments on expectations, making it a canonical benchmark for methodological validation.

Survey waves and the corresponding treatment information are taken directly from Weber et al.'s (2025) replication files. Respondents were exposed to 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). Because these treatments were not implemented in every survey wave, the panel is unbalanced: T1 was fielded in 10 waves, T2 in 7, and T3 in 6. For detailed treatment information, see Table B.1.

In each wave, participants first reported 12-month-ahead inflation expectations using

a probabilistic distributional question. They were then randomly assigned to a control group (T0) or to a treatment group. After the intervention, all respondents provided a point forecast of 12-month-ahead inflation.

The key insight from these experiments is that the responsiveness of expectations to information treatments varies systematically with the inflation environment. During periods of low and stable inflation, households placed significant weight on the provided signals, indicating limited prior knowledge. As inflation rose after 2021, however, households became markedly less responsive, suggesting that they had already internalized much of the publicly available information. Weber et al. (2025) interpret this pattern as evidence of endogenous attention to inflation: when inflation is high and salient, agents update less when presented with exogenous information interventions.

Replicating this design with LLM surveys serves two purposes. First, it tests whether LLM responses capture the same information 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.

3.2 LLM Implementation

We replicate the experiment in Weber et al. (2025) using large language models. Our main implementation uses GPT-4.1 via the OpenAI API.

To mimic the historical context of the original survey waves, we applied knowledge restriction prompts, restricting responses to the information that would have been available at the time of each survey wave (see Section 2.1).

Using repeated personas, LLM agents yield reliable estimates from far smaller samples than typical household surveys. Each group — control or treatment — consists of 200 LLM agents with the same persona (see Section 2.2). The same set of 200 personas is used consistently across all survey waves. Each persona is assigned attributes for age, gender, marital status, education, income category, and U.S. state, and is drawn from a random

subsample of 200 respondents in the Federal Reserve Bank of New York's Survey of Consumer Expectations (SCE). This ensures that the demographic composition of the LLM sample mirrors that of the SCE.[1] By contrast, the human benchmark of Weber et al. (2025) is based on a nationally representative panel of approximately 80,000 households, with typical sample sizes of 15,000–20,000 respondents per wave. The efficiency of obtaining credible estimates from just 1% of the human survey sample size reflects the internal consistency of the LLM framework, making it both practical and scalable.

The essential structure of the survey mirrors the original human RCT design, with adaptation to a prompting framework suitable for LLM-based respondents (for full prompts see Appendix A.2). We implement the same sequence of prior elicitation, treatments, and posterior forecasts as in Weber et al. (2025), using their survey waves and treatment information summarized in Table B.1. To enhance transparency, we instruct LLM agents to generate their reasoning in full sentences before providing a final answer, and agents retain memory of prior questions to replicate the sequential nature of human surveys. We only ask relevant questions abstracts from the additional background and screening questions typical of human surveys. Therefore, our survey is relatively short, and we adopt the full-memory setting.

3.3 Validation: Demographics and Priors

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

First, we compare demographic distributions and find that the LLM demographics 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:

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

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 B.2. Although neither sample is designed to be nationally representative, both samples provide broad demographic and geographic coverage.

Next, we compare the pre-treatment inflation forecasts generated by LLM agents with those from the SCE, which employs the same distributional question adopted by the Nielsen Homescan Panel. For both the LLM survey and the SCE, we compute the mean expectation for each respondent using the midpoint of the reported probability bins. For the LLM survey, this constitutes the prior forecast of each agent.

Figure 1 compares the median inflation expectations of LLM agents with those of SCE respondents, calculated across respondents within a given quarter. The regression slope is close to one (1.05) with a small negative intercept, indicating that the LLM priors closely track human priors over the inflation cycle.

3.4 Validation: Information 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 its central step: whether LLM agents replicate the information-updating patterns documented in Weber et al. (2025).

Following Weber et al. (2025), we estimate the following regression specification,

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

where the prior is the mean inflation expectation calculated based on the probability distribution, elicited before information intervention; 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

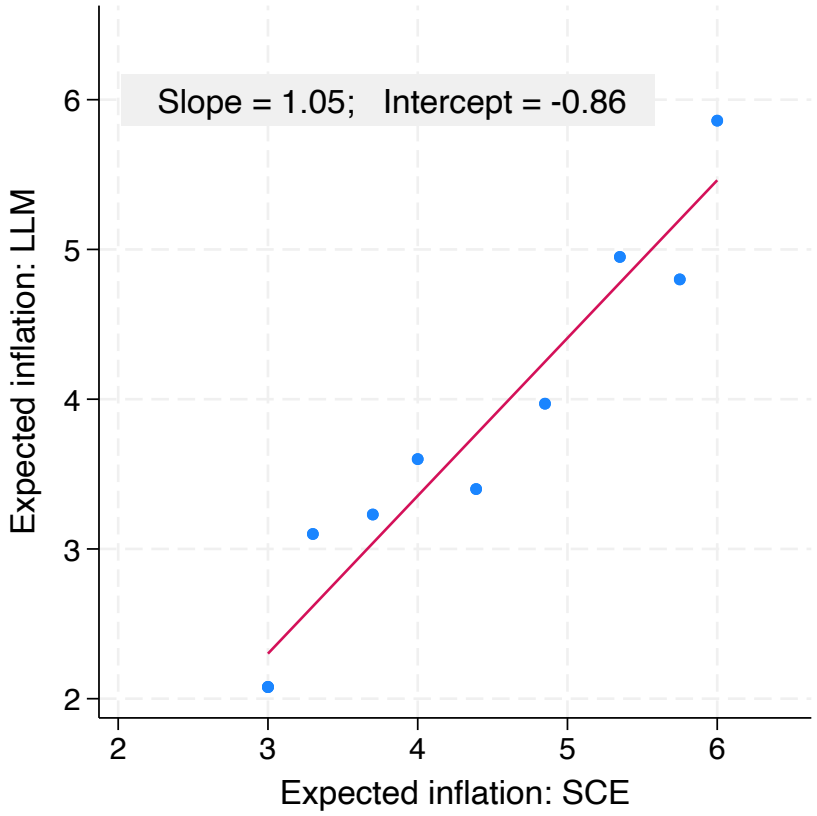


Figure 1: PRIOR INFLATION EXPECTATION: LLM VS. HUMAN

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.

was assigned to treatment arm Tj (T1 - past inflation, T2 - Fed target, or T3 - Fed forecast), and zero otherwise. Given that personas are identical between treatment and control groups, including demographic controls or not has no effect on the estimates.

In this setup, the coefficient β captures how closely posteriors track priors for the control group, which measures the baseline weight agents place on their own prior beliefs in the absence of new information. The interaction coefficients γ_j measure how this relationship changes in each treatment arm. A negative γ_j implies that treated respondents rely less on their priors and instead shift weight toward the provided information.

We follow Weber et al. (2025) by normalizing the slope γ_j by the control group slope β and focusing 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 compare the scaled slope (γ_j/β) across inflation environments and between LLM-based surveys and the human benchmarks in Weber et al. (2025).

Figure 2 reproduces the human benchmark (Figure 4A in Weber et al. 2025) using LLM surveys, 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 information treatments had large effects on expectations during the low-inflation period of 2018–2019, but responsiveness declined sharply once inflation began rising in 2021, converging toward zero by late 2022. Thus, our LLM results strikingly reproduce the central empirical finding of Weber et al. (2025): attention to inflation is endogenous to the macroeconomic environment.

To further assess alignment, Figure 3 plots LLM estimates of scaled slopes directly against their human counterparts. The first three panels show one treatment arm per panel (T1: past inflation, T2: Fed target, T3: Fed forecast). The fourth panel pools all three treatments. In each plot, every dot corresponds to a single survey wave, comparing the human scaled slope on the x-axis with the corresponding LLM scaled slope on the y-axis. 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. This naturally leads to the question of whether the two surveys capture a similar underlying link between treatment effects and inflation.

We address this question in Table 2, which evaluates Weber et al.'s (2025) main message by comparing treatment effects directly with inflation, rather than between LLM and human surveys. In both cases, the correlations are positive across different treatments: scaled slopes become less negative as inflation rises. The strength of correlation, however,

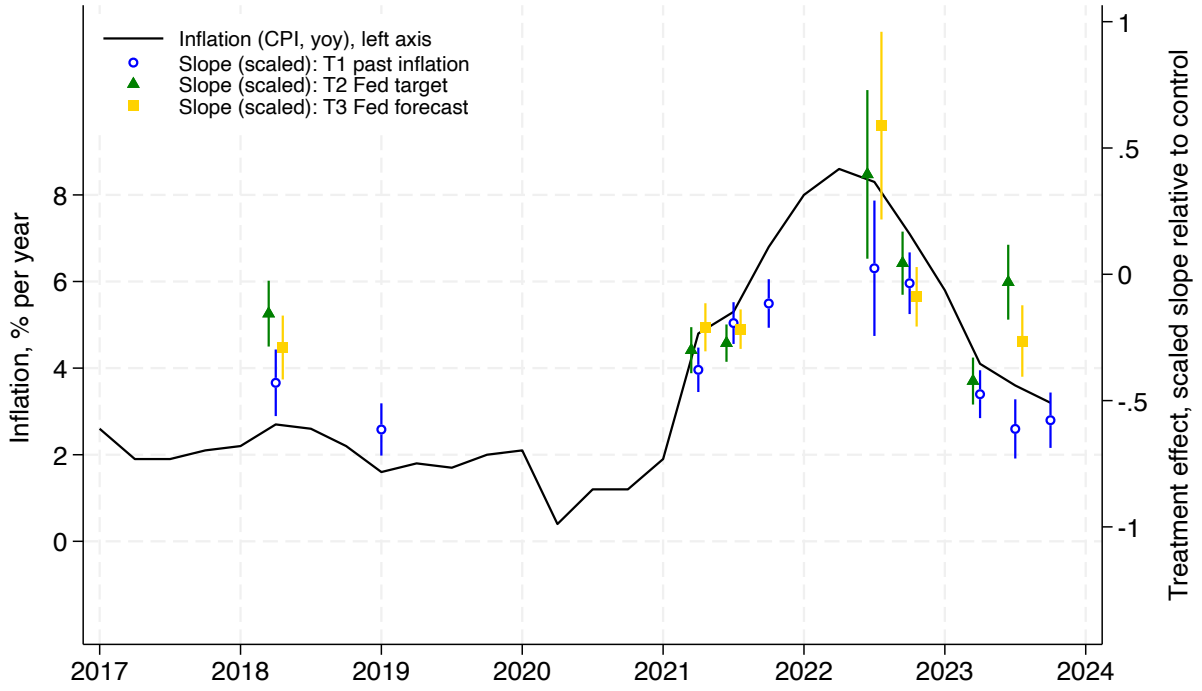


Figure 2: LLM REPRODUCTION OF WEBER ET AL. (2025) FIGURE 4A

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.

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 (pooled). For human surveys, the corresponding correlations are lower and more variable: 0.69 (T1), 0.87 (T2), 0.13 (T3), and 0.57 (pooled). These differences suggest that the LLM-based surveys deliver a clearer overall pattern of state-dependent attention to inflation, while some of the weaker human correlations may reflect greater sampling noise.

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 information treatments diminishes as inflation becomes more salient. Moreover, the LLM-based estimates often appear cleaner and more consistent

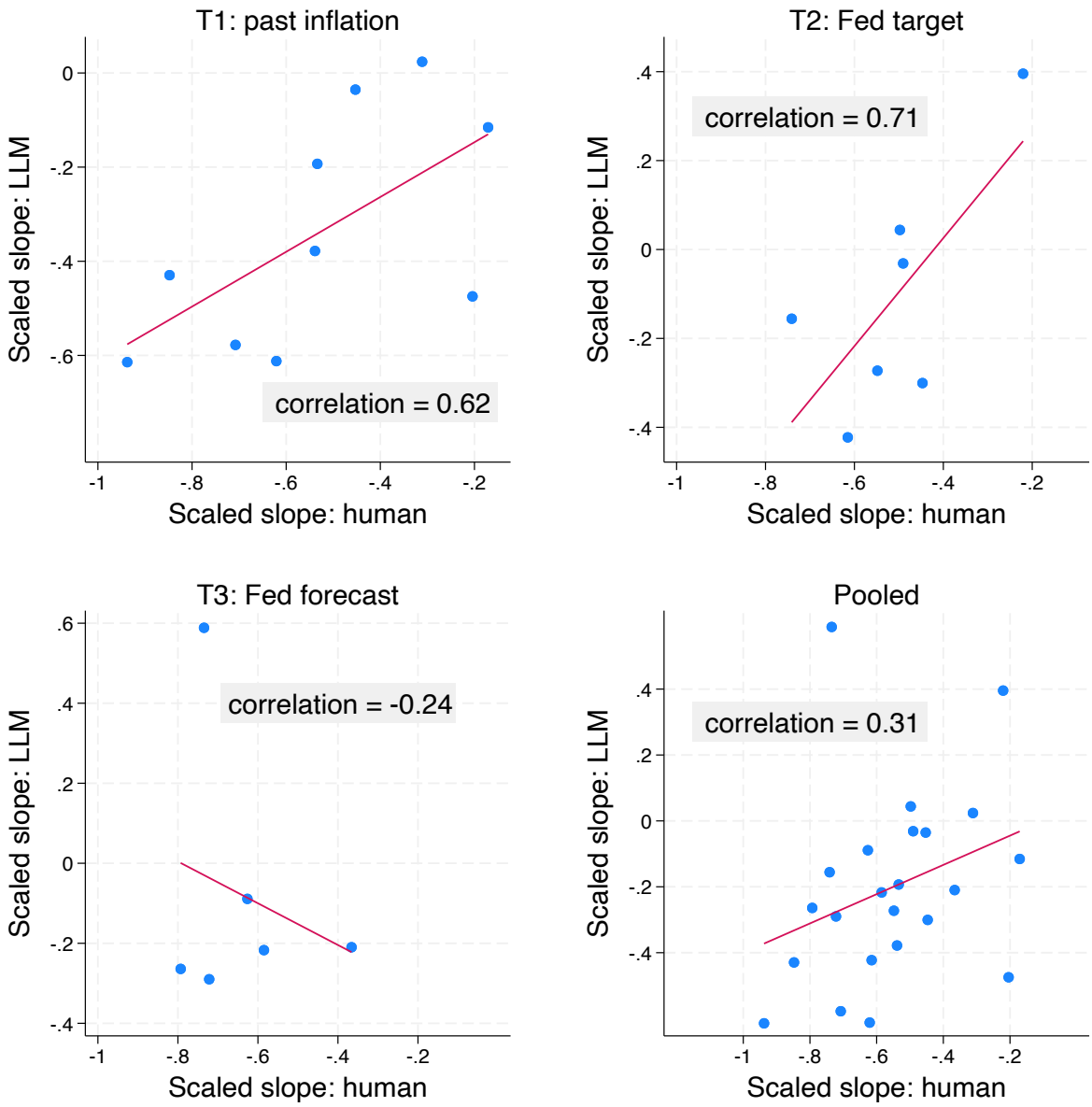


Figure 3: SCALED SLOPES: LLM VS. HUMAN

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

than their human counterparts, reflecting the absence of sampling noise and attrition. This underscores the value of LLM-based surveys as a credible complement to traditional survey experiments.

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.

4 LLM-Based Surveys over 35 Years: Retrospective Coverage and Internal Consistency

We now turn to the broader potential of LLM-based surveys, highlighting their ability to serve as an ideal experimental setting. By systematically extending surveys backward over 35 years, our framework demonstrates how LLMs can generate long-run, internally consistent series unattainable in traditional settings.

4.1 Retrospective Coverage and Internal Consistency

We extend the survey design described in Section 3.2 systematically from 1990. By comparison, Weber et al. (2025) conducted 10 survey waves between 2018–2023, which represented a major contribution for human surveys. Our framework scales this to three and a half decades of coverage, comprising over 50 survey waves.

Beyond coverage, our framework ensures perfect internal consistency by holding respondent personas fixed across control and treatment arms, across all waves, while rolling out the same treatments consistently over time. This design yields a consistent survey design and a synthetic survey panel that spans diverse inflation regimes and policy environments, allowing us to trace responses under comparable conditions across history.

In sum, the combination of retrospective coverage and internal consistency realizes

Table 3: 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 B.1), the surveys are implemented quarterly during the Great Recession (2008 - 2009) and the COVID inflation surge (2021 - 2023), and annually for the rest.

the ideal form of the Weber et al. (2025) design, in a way that is unattainable with human surveys. In this ideal survey environment, we can assess whether the state-dependent patterns documented in the benchmark setting persist across three decades—an experiment that human surveys cannot deliver.

4.2 Implementation Details

Having outlined the scope, we now turn to implementation details. Specifically, Table Table 3 summarizes the sample coverage, frequency, and the number of waves across treatments, including periods, number of waves, and survey frequency. 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 more stable periods, we rely on annual frequency to balance coverage with computational cost, unless additional waves are covered in the validation sample (see Table B.1). Our sample begins in 1990, as earlier periods suffer from sparse digitization of contemporaneous news and statistical releases. It ends in 2023, given GPT-4.1's June 2024 knowledge cutoff. We can extend to 2024 and beyond, once newer LLM models with full-year training data become available.

For T1 (past inflation treatment), we began 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) started in 2012, when the Federal Reserve first formally announced an inflation target of 2%.[2] T3 (Fed forecast treatment) was introduced in 2015, when the Federal Open Market Committee's (FOMC) Summary of Economic Projections (SEP) began reporting median forecasts for public release (in both the SEP summary and in the addendum to the FOMC minutes), thereby providing a clear numerical forecast that could be used in survey treatments.[3] 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-based surveys can be implemented precisely as planned, keeping treatments tightly aligned with experiment dates. For T1 (past inflation treatment), we use the year-over-year CPI inflation rate (CPIAUCSL) of the first month of the corresponding quarter.[4] This convention accounts for the one-month release lag in CPI data,[5] ensuring that the relevant figure is available when the experiment is fielded.

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. This design ensures that treatments reflect the information realistically available to survey respondents at the time.

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

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

[4]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.

[5]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.

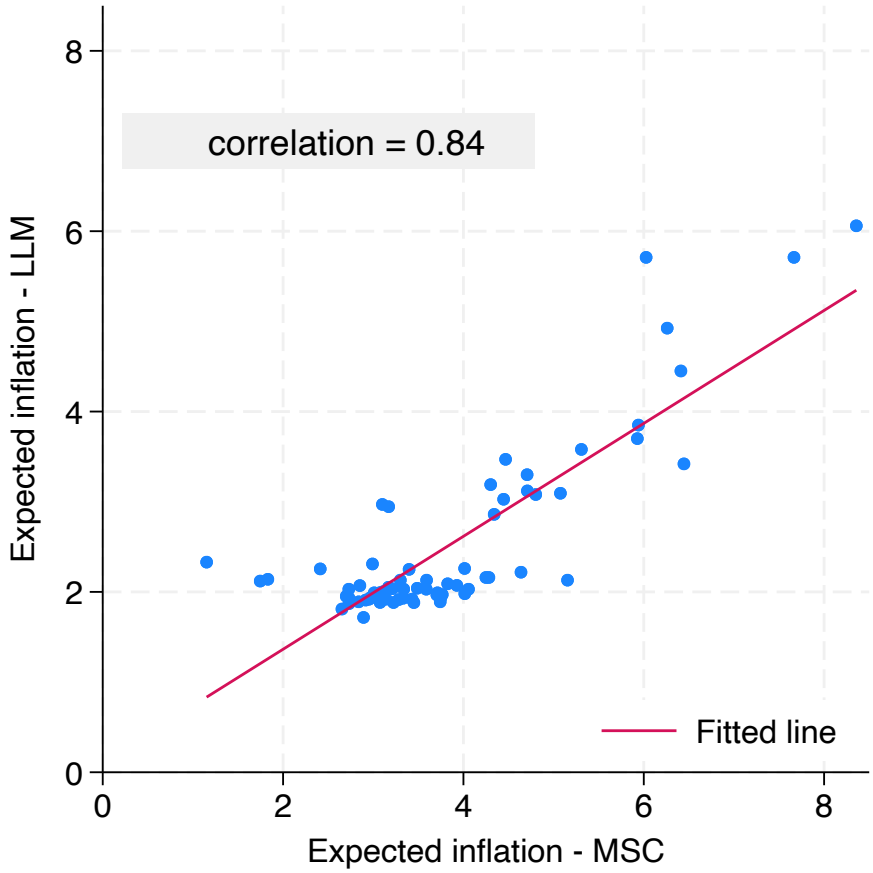


Figure 4: EXPECTED INFLATION: LLM VS. HUMAN

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.

4.3 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 benchmarked 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.[6]

Figure 4 compares the LLM priors with those from the MSC. Each dot corresponds

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

to the cross-sectional mean of expectations in a given survey wave, with the fitted line summarizing the linear relationship between the two series. 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 about price changes—making the comparison conservative and underscoring the robustness of the LLM-based expectations.

4.4 Information Updating over 35 Years

We next turn to how LLM agents update their expectations in response to information treatments over 35 years. A key test is whether the extended panel reproduces the state-dependent attention patterns observed in the human benchmark.

Figure 5 plots the scaled slope estimates for T1–T3 (right axis) against year-over-year CPI inflation (CPIAUCSL, see Footnote 4) in the experiment month (left axis). Specifically, the top panel shows that treatment effects for T1 (past 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 even stronger for T3 (Fed forecast). Overall, the correlations between scaled slopes and inflation remain strongly positive: 0.61 for T1, 0.76 for T3 treatments. The reasonably high correlations suggest that the main result in Weber et al. (2025) extends robustly to our longer sample, though the effects are somewhat 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, effectively giving them an informational edge over the average U.S. household. This conjecture also sheds light on why, in the replication sample summarized in Table 2, LLM correlations for T2 are lower

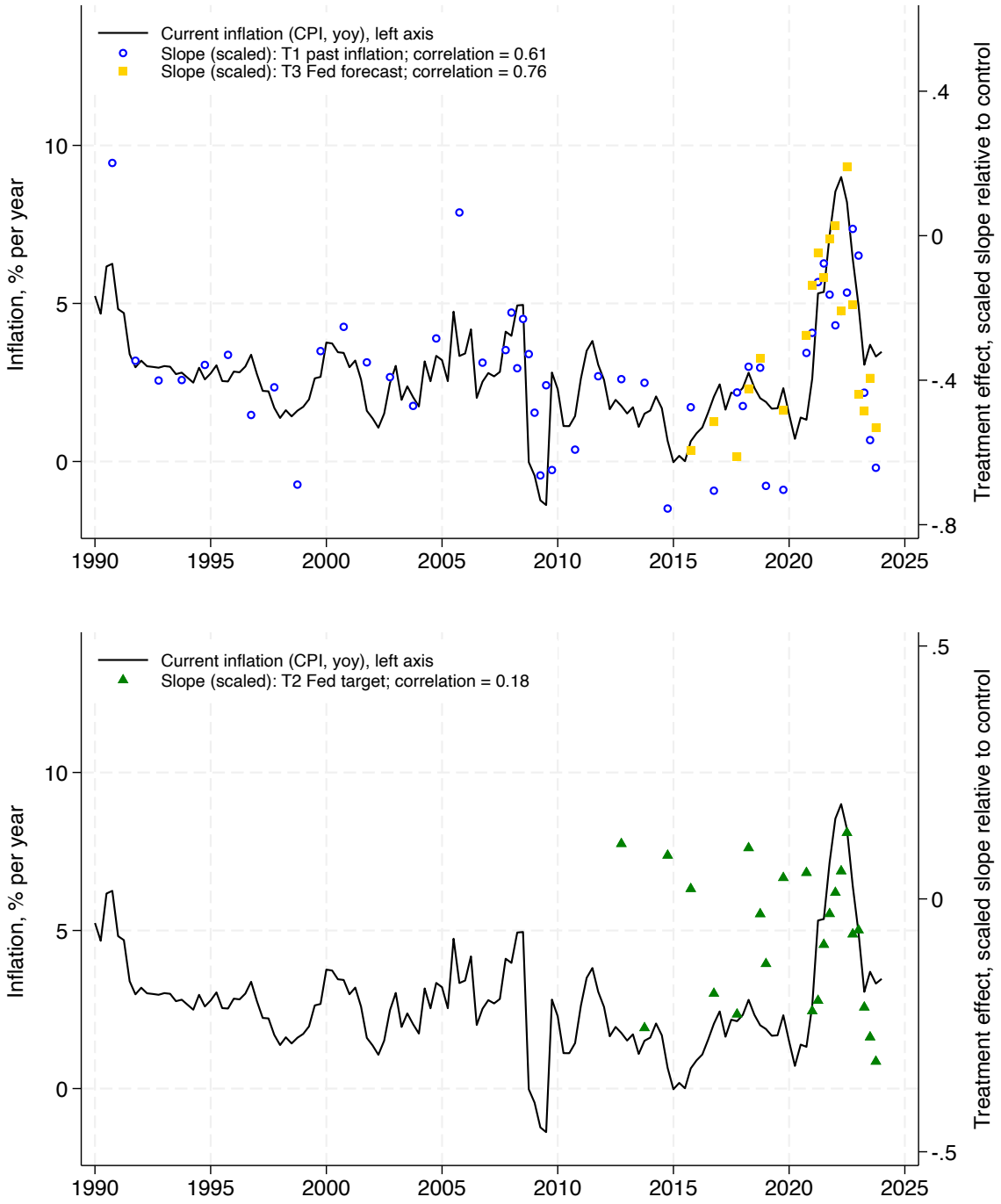


Figure 5: EXTENSION OF EXPERIMENTS

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).

than those of human respondents. Importantly, however, the same table shows that for the other treatments—past inflation (T1) and Fed forecasts (T3), both of which are time varying—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 and internal consistency unattainable in traditional settings.

5 Discussions and Robustness

Having established the broader potential of our framework, we next examine whether the results are robust to alternative implementation choices. Specifically, we examine three dimensions. First, we assess sensitivity to the draw of personas by repeating the experiment with a new set of respondents. Second, we evaluate model dependence by comparing results from our baseline GPT-4.1 with those from the updated GPT-5. Finally, we check specification robustness by re-estimating results with income bins indexed to inflation rather than fixed in nominal terms.

5.1 Personas

Our baseline relies on 200 fixed personas. As a robustness check, we draw a new set of respondents and re-run the experiment to test whether results are sensitive to the initial persona sample.

We repeat our main validation exercises using a fresh draw of 200 personas from the SCE. 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 3

remains similar, with all but one positive correlation. For details, see Panel A of Table C.3. We therefore conclude that our results are robust to alternative persona samples.

5.2 LLM Model Versions

Our baseline is implemented using GPT-4.1 released in April 2025, which is the latest stable model with a fixed knowledge cutoff in June 2024. In August 2025, OpenAI launched GPT-5, the most recent model. There are several key differences. First, unlike GPT-4.1's fixed cutoff, GPT-5's knowledge boundary may shift in future releases, which makes the model less predictable and less easily replicable across experiments. Second, GPT-5 reduces sampling variation, yielding smoother outputs but making GPT-4.1 a closer analogue to human surveys. Third, GPT-5's output involves substantially more reasoning and longer context windows, making the cost roughly six to seven times that of GPT-4.1 for our survey design. For these reasons, we use GPT-4.1 as our baseline.

However, as the newer model, GPT-5 has its own important advances: it is larger than GPT-4.1, trained on more recent data, and equipped with more parameters and longer context windows. We therefore check the robustness of our framework using GPT-5. The fact that our findings remain consistent across both GPT-4.1 and GPT-5 underscores that our results are robust to model choice.

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. For details, see Panel B of Table C.3.

5.3 Income Bins

In our baseline in Section 3 and Section 4, we follow the SCE and classify households into fixed nominal income bins across survey years. Specifically, we divide income into three nominal buckets, targeting the SCE composition: 35.7% under 50k, 35.8% between 50k

and 100k, and 28.5% over 100k. This facilitates comparability but raises a natural concern: over time, fixed nominal cutoffs capture households at very different real income levels, especially for the longer sample in Section 4.

To address this, we redefined 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 focus on the T1 past-inflation treatment, which spans the longest sample period and where the nominal–real distinction matters most, reporting results at annual frequency.

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, 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. For figures, see Appendix C.2.

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.

6 Conclusion

This paper proposes a new LLM-based survey framework that expands the frontier of expectation measurement. By combining retrospective coverage through date-restricted prompts with perfect internal consistency across respondents and over time, 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 shows that LLM surveys reproduce key patterns of information

updating, while often yielding clearer correlations with inflation, reflecting reduced sampling noise. Extending the design retrospectively to over three decades, we demonstrate that state-dependent attention to inflation is a stable feature of expectations formation across macroeconomic environments.

Taken together, these results establish LLMs as a credible and scalable complement to traditional survey methods. Beyond inflation, the framework opens avenues for systematically reconstructing expectations in domains such as monetary policy, housing, labor markets, and financial decision making. In this way, LLM-based surveys have the potential to transform the way economists design and implement survey experiments on expectations.

References

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.

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

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.

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).

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

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

Pfäuti, Oliver, "The inflation attention threshold and inflation surges," *working paper*, 2025.

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.

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

We employ the following prompt to test the effectiveness of our knowledge restriction:

> 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 }. 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. { Insert question here } Please provide your answer in yes or no.

The following set of questions is used on survey dates approximately one to two months before each event and again one month 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 }. You are responding to this survey in { survey date }.

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.

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 Validation Details

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

Table B.1: PERIOD 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.

B.2 Demographic Distribution

Table B.2: DEMOGRAPHIC DISTRIBUTION

	LLM (1)	SCE (2)
Education		
College	50.0%	55.3%
High School	13.5%	12.2%
Some College	36.5%	32.5%
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).

C Robustness

C.1 Perosna and model versions

Table C.3: CORRELATION OF SCALED SLOPES

	(1) T1: Past inflation	(2) T2: Fed target	(3) T3: Fed forecast	(4) Pooled
Panel A: new personas				
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
Panle B: GPT-5				
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 a different set of personas (Panel A) and GPT-5 model (Panel B). The correlations are with current inflation, scaled slopes based on baseline results from Section 3.4, and scaled slopes based on the human survey from Weber et al. (2025). "Pooled" correlations are calculated by pooling estimates from all treatment types.

C.2 Inflation-adjusted Income

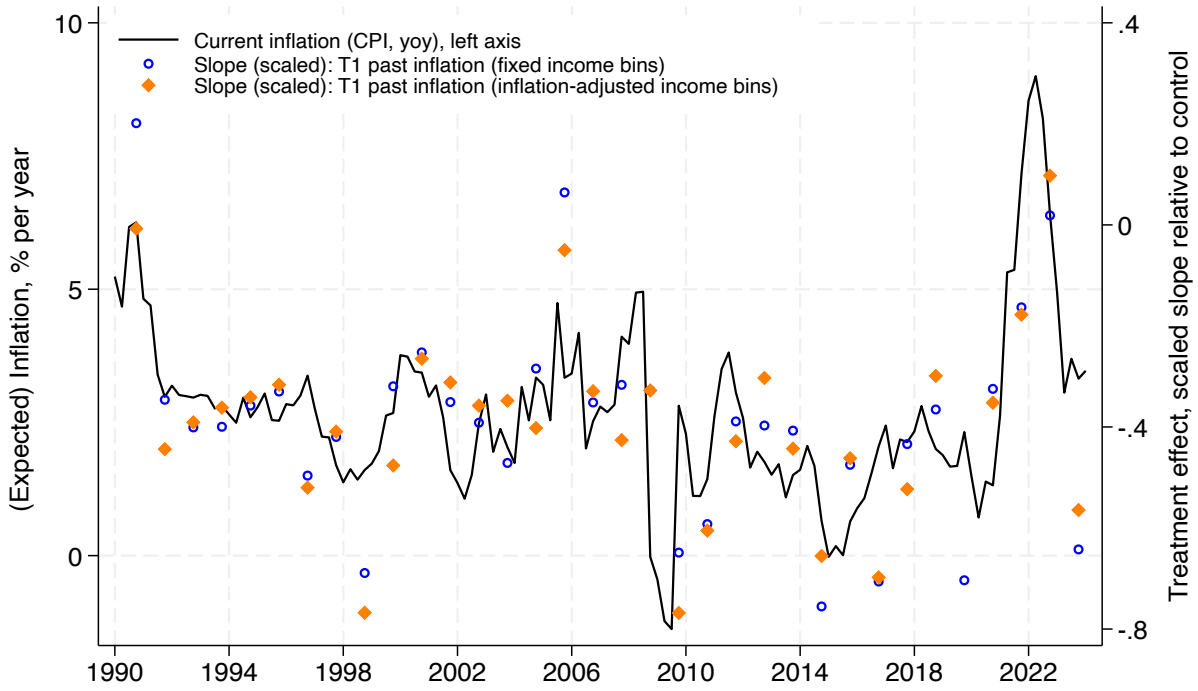


Figure C.1: SCALED SLOPE: INCOME BINS

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

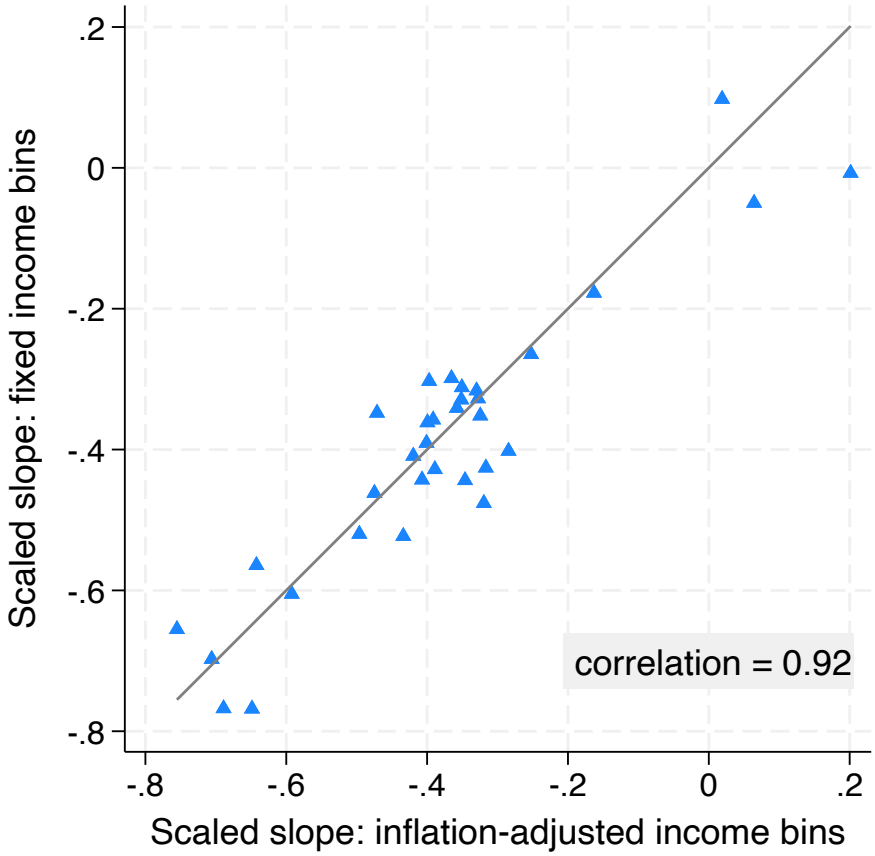


Figure C.2: SCALED SLOPE: INCOME BINS

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.