Artificial Consensus: AI and Survey Measurement*

Thomas Cauthorn

First draft: July 2026

Abstract

Central banks, governments, industry, and investors use surveys to measure economic conditions that are otherwise difficult to observe. AI can distort survey measurement at three stages: by replacing respondents, assisting them, or analyzing the resulting data. I study all three channels by fielding an online survey based on the Survey of Consumer Expectations (SCE), benchmarking answers from fifteen AI models, and analyzing SCE microdata. A browser-side answer-origin measure distinguishes between AI-assisted humans, agentic tools, and autonomous bots, revealing that the dominant respondent-side problem is assistance rather than substitution. Among 1,294 respondents, 31.4% leave evidence of AI involvement, 4.8% disclose it, and only 0.4% are autonomous bots. AI-generated answers reproduce a plausible normal-times center of inflation expectations but collapse their dispersion, while AI-led cleaning disproportionately removes genuine upper-tail SCE responses. With independent respondents, tighter answer distributions generally signal more convergent beliefs; with a shared algorithmic source, the same pattern can arise without belief convergence. AI-induced homogenization can make surveys appear more precise while biasing inference.

Keywords: artificial intelligence; large language models; data quality; data contamination; survey methodology; inflation expectations; AI detection; paradata; behavioral finance.

JEL Classification: C83, C81, D84, E31, E58, G41, O33.

*Thomas Cauthorn: University of Mannheim, Finance Area, thomas.cauthorn@uni-mannheim.de. I am grateful for comments from Matti Keloharju and Adair Morse. Funding for this project was provided by the Mannheim Center for Corporate Behavior and Regulation Analysis (COBRA) and the Behavioral Finance e.V. research grant. The software for conducting the online survey was provided by SoSci Survey GmbH. SoSci Survey GmbH had no influence on the content of the study. The survey was approved by the Ethics Committee of the German Association for Experimental Economic Research.

1 Introduction

Surveys are a ubiquitous tool for central banks, governments, industry, and investors seeking to measure economic conditions that are otherwise difficult to observe. The value of those measurements rests on a simple assumption: each recorded answer reflects the sampled person's own belief about economic conditions. Generative AI weakens that link. It can complete a survey directly, help a human respondent formulate an answer, or clean and interpret the resulting data. Standard quality screens can reject implausible responses; they cannot establish whose information a plausible response contains. A dataset that appears to aggregate many independent beliefs may partly reflect a common algorithmic signal.

This measurement problem arrives as the Federal Reserve is reassessing the data on which policy decisions rely. At the June 17, 2026 FOMC press conference, Chair Kevin Warsh criticized "old-fashioned survey methods," citing inadequate response rates, questions designed for an earlier economy, and official statistics that are revised and may describe the economy only with a lag (Warsh, 2026). Those concerns involve whether surveys reach enough people, ask the right questions, and arrive in time. Generative AI adds a fourth concern: even a timely and numerically plausible answer to a relevant question may no longer measure the sampled person's belief.

I study this dimension of survey quality using the New York Federal Reserve's Survey of Consumer Expectations. Researchers and policymakers use the SCE to monitor the distribution of consumers' inflation expectations. I examine the three points at which AI can enter the survey pipeline: as a respondent, as an assistant to a respondent, and as an analyst of the resulting data. I ask three questions: whether respondents use AI to answer SCE-style questions, how AI involvement affects the statistics researchers and policymakers monitor, and whether an AI analyst introduces bias when analyzing SCE data.

First, to detect AI use, I introduce a novel browser-side measure focusing on how respondents answer rather than on what they answer. I find substantially more evidence of AI involvement than respondents disclose. Among the 1,294 respondents who complete the survey, only 4.8% self-report using AI, whereas 23.5% leave strong evidence of AI use and a further 7.9% leave some evidence. Agentic tools, by contrast, account for 1.3% of respondents and autonomous bots for a further 0.4%. The respondents who leave this evidence are not careless participants: they outperform the no-evidence group on financial-literacy, numeracy, and cognitive-reflection screens, so the quality filters surveys already run would retain them. The central measurement issue is not autonomous AI but AI-assisted humans: real respondents completing the survey while partly relying on AI to assist them.

Second, detected AI involvement is associated with apparent consensus: the center of

reported inflation expectations remains plausible, but the distribution around it becomes sharply compressed. At the five-year horizon, both the AI-evidence and no-evidence groups report a median inflation expectation of 3%. Yet the middle 80% of the distribution spans only 4 percentage points among respondents with AI evidence, compared with 25 points among respondents with no detected evidence. The distribution also exhibits pronounced concentration at a single focal answer. Across horizons, respondents with AI evidence are 20.9–26.6 percentage points more likely to report an inflation expectation of exactly 3%. The model benchmark points to a simple common-source mechanism: consumer AI models themselves return highly similar answers. Across fifteen models, one-year inflation expectations have a standard deviation of just 0.4 percentage points, compared with 12.7 points in the pre-AI SCE. Moreover, 81.8% of model answers equal 3.0%, 3.5%, or 4.0%, and none exceeds 5%. Model answers therefore reproduce a plausible normal-times center while eliminating most of the heterogeneity around it. This loss of heterogeneity weakens the signals policymakers monitor during high-inflation episodes. In simulations using 2022 SCE data, replacing 30% of household responses with model answers causes monitoring rules based on the interdecile spread, the ninetieth percentile, and the standard deviation to identify the episode in only 21%, 26%, and 37% of simulations, respectively. Rules based on the median, mean, and interquartile range continue to identify the episode, but portray it as materially milder. The reported median falls from 8.0% to 5.0%, and the share of respondents expecting inflation above 5% falls from 64.2% to 44.9%. AI-shaped answers can make a severe episode appear more moderate without making the center of the survey distribution obviously implausible. This distortion is economically important because growth in the upper tail can indicate that some households expect high inflation to persist rather than return toward the central bank's target (Reis, 2021; Natoli and Sigalotti, 2018; Grishchenko and Wilcox, 2024).

Third, I ask whether an AI analyst introduces a further distortion when cleaning and interpreting SCE data. When left to clean the raw survey, AI often mistakes observed high-inflation SCE responses for data errors and outliers. An AI analyst excludes 8.9% of the observations it receives, with 39% of the discarded observations coming from the highest decile of inflation expectations. After an AI analyst prepares the data, the share of respondents expecting inflation above 10% falls from 24.7% to 6.0%. In 56% of high-inflation SCE samples, the AI analyst's own data cleaning leads it to downgrade its assessment of unanchoring.

This paper contributes to four related literatures. The first literature studies AI involvement in online research. Westwood (2025) shows that autonomous AI respondents can maintain fairly coherent personas while passing almost all conventional attention checks, and Affonso (2026) finds that visual-cognitive traps distinguish agents from humans much more

effectively than conventional checks do. Yet the more difficult case is not AI substituting a respondent but AI assisting a respondent. One proposed active defense embeds prompt-injection traps intended to provoke a model into revealing itself (Wang et al., 2025). I instead take a hybrid answer-origin approach, combining passive paradata with active clipboard canaries. I extend this work with a multi-layered detection method for quantitative survey questions that infers from the response process how each answer was produced, rather than relying only on the answer itself. The method distinguishes autonomous bots and agentic tools from AI-assisted humans.

The second literature uses surveys to measure beliefs and reasoning that cannot be recovered from observed choices alone (Manski, 2004; Stantcheva, 2023). D'Acunto and Weber (2024) review evidence that these subjective expectations predict economically important decisions. In finance, belief surveys have become measurement instruments in their own right: expectations elicited from investors help explain portfolios and prices (Greenwood and Shleifer, 2014; Giglio et al., 2021). AI complicates this measurement by making it uncertain who produced an answer. Building on the premise that language models can be treated as simulated economic agents (Horton et al., 2023), a literature on "silicon respondents" asks whether they can stand in for human survey participants. Argyle et al. (2023) show that demographic conditioning allows a language model to approximate subgroup-level response patterns, whereas Bisbee et al. (2024) find that synthetic samples exhibit too little variation and unstable relationships between demographics and attitudes. More broadly, models reproduce plausible averages but struggle with human heterogeneity. They are unrepresentative of large population segments (Santurkar et al., 2023), their survey responses can be sensitive to prompt wording, answer ordering, and labeling (Y. Gao et al., 2025; Dominguez-Olmedo et al., 2024), and they fail to reproduce documented human response biases (Tjuatja et al., 2024). Closest to the inflation-expectations application, Zarifhonarvar (2026) finds that model architecture materially affects generated expectations, with proprietary models generally exhibiting less disagreement than open-source alternatives. I observe AI involvement among real human respondents and connect answer origin to the resulting survey distributions.

Third, the paper contributes to the macroeconomics of expectations literature. Mankiw et al. (2003) document substantial and time-varying disagreement in inflation expectations, while Coibion and Gorodnichenko (2012) show that disagreement and forecast errors reveal information rigidities. Andre et al. (2022) trace part of this heterogeneity to the different subjective models through which households, unlike experts, reason about macroeconomic shocks. More recent work treats changes in the upper tail as an early indication that inflation expectations are becoming unanchored (Natoli and Sigalotti, 2018; Reis, 2021; Grishchenko

and Wilcox, 2024). These tail observations also carry information about uncertainty. Rounding and coarse probability reporting can reveal how respondents express uncertainty rather than being merely measurement noise (Binder, 2017; Manski and Molinari, 2010). I show that detected involvement is associated with less of this cross-sectional information: the center of the distribution can remain plausible even as the dispersion and upper-tail statistics carrying information erode. What is lost is therefore not merely statistical spread but a human signal of uncertainty that the expectations literature has shown to be economically meaningful.

Fourth, the paper relates to the use of language models as research tools. Models perform well at tasks such as text classification (Gilardi et al., 2023) and can help analyze open-ended survey responses, although categorizing topics need not recover the underlying attitudes (Hobbs and Green, 2025). Their use for measurement still requires explicit validation (Ludwig et al., 2026), and historical applications may introduce look-ahead bias (Glasserman and Lin, 2024). In my setting, the output-side risk documented in the exploratory benchmark is different. When given discretion to clean survey data, the AI analyst disproportionately removes observed upper-tail SCE responses as errors. AI assistance therefore produces the same compression at both ends of the research pipeline, when respondents generate the data and when analysts interpret them.

Section 2 develops the detection framework. Section 3 applies it in the field. Section 4 calibrates the distribution of answers from fifteen AI models. Section 5 studies the consequences for monitored survey distributional statistics, Section 6 examines AI-assisted analysis, and Section 7 concludes.

2 AI in Surveys

Near-universal access to AI raises a question that existing survey-quality checks were not built to answer: who or what produced the recorded answer? This section introduces a method to classify survey respondents by degree of AI involvement.

2.1 The Challenge

Imagine a researcher reading a survey to learn about respondents' expectations regarding economic conditions. One respondent forecasts inflation at 4% and explains that grocery and housing costs still feel as though they are rising. Both answers are plausible and internally coherent. Neither answer, on its own, reveals whether the respondent answered unaided, consulted a chatbot, edited a model-generated response, used software that entered the

answer, or was replaced by an autonomous bot. The survey records what was submitted, not a complete account of how it was produced.

AI creates two challenges pertaining to respondents. The first is substitution. An autonomous bot can complete the survey without a person in the loop, or an agentic tool can enter one or more answers inside a session that a person otherwise navigates. Current systems can produce plausible responses and pass many conventional attention and quality checks (Westwood, 2025; Affonso, 2026). The second challenge is assistance. The sampled person fills out the survey but delegates part of one or more answers to AI. A respondent may copy a question into a chatbot, consult AI on another device, or use an AI-enabled browser or operating-system assistant. This is harder to detect because the account, browser, mouse movements, and much of the typing can all be genuinely human. I group these AI-involvement levels into five mutually exclusive respondent classifications: autonomous bots, agentic tools, AI-assisted humans, possibly AI-assisted humans, and respondents with no detected evidence of AI involvement. The distinction between an agentic tool and an AI-assisted human concerns how the answer enters the survey: software enters at least some answers in an agentic session, whereas the respondent enters the answer through ordinary input channels in an AI-assisted session.

It is useful to distinguish conventional survey-quality measures from measures of answer origin. The survey includes an instructed-response attention check; financial-literacy, numeracy, and cognitive-reflection items; and a set of visual-cognitive tasks.[1] The first four measure instruction following, knowledge, or reasoning performance. They do not reveal whether the submitted answer reflects the respondent's unaided judgment. Recent work suggests that visual-cognitive tasks may distinguish some autonomous agents from humans (Westwood, 2025; Affonso, 2026), but such items need not detect a human respondent who only consults AI for certain questions. I therefore use these measures primarily to ask whether conventional quality screens would retain respondents who leave evidence of AI involvement when answering other questions.

[1]For instructed-response attention checks, see Oppenheimer et al. (2009) and Berinsky et al. (2014); for the financial-literacy battery, see Lusardi and Mitchell (2014); for the numeracy items, see Lipkus et al. (2001); for the Cognitive Reflection Test, see Frederick (2005); and for visual-cognitive diagnostics of AI agents, see Affonso (2026).

2.2 A Simple Framework

To fix ideas, let H_i denote the answer respondent i would report unaided and let A_i denote the answer supplied by AI. I summarize the recorded answer as

$$Y_i^{\text{obs}} = \alpha_i H_i + (1 - \alpha_i) A_i, \tag{1}$$

where $\alpha_i = 1$ corresponds to an unaided response, $0 < \alpha_i < 1$ to an assisted response that retains some of the respondent's own information, and $\alpha_i = 0$ to complete replacement by a model answer.

Equation (1) motivates the paper's main prediction. If model answers are more concentrated than household beliefs, reducing α_i moves observations toward a common model signal and compresses the cross-section. This differs from idiosyncratic reporting noise, which tends to add dispersion. AI assistance can instead reduce measured disagreement and obscure economically meaningful heterogeneity.

2.3 Detecting AI Involvement

The detection method is an answer-origin measure: it uses evidence from the completion process to infer how an answer was produced, rather than classifying respondents from whether their answers look unusual. The procedure has two stages. First, I construct five families of evidence at the question, answer-field, and session levels: question routing, model processing, answer delivery, session operation, and supplementary evidence. Second, I apply an ordered rule that maps those signals into five mutually exclusive respondent classifications. Table 1 summarizes what each signal can, and cannot, establish.

2.3.1 Measurement Architecture and Observability

Respondents are recruited through Prolific and complete the survey in SoSci Survey, which presents the questions, implements the routing logic, and stores the submitted responses.[2] A separate paradata layer records timestamped completion-process events and links them to the response record by respondent, page, question, field, and time. A forthcoming internet appendix describes the infrastructure and event-processing rules in detail.

The paradata layer observes events at the survey boundary: text copied from a page, text pasted into an answer field, departures from and returns to the survey window, keyboard and other input events, page timing, and browser, device, rendering, and network characteristics.

[2]The online survey was conducted with SoSci Survey (Leiner, 2026); the data were collected on June 3–4, 2026, at `https://sosci.sowi.uni-mannheim.de/CE_survey/`.

It does not record the respondent's screen or activity inside another application.[3] This observability boundary is important. The method can identify several routes through which a question or answer crosses the survey boundary, but it cannot directly observe assistance obtained on a second device or manually retyped from an external source.

2.3.2 Constructing the Evidence

Question routing. For every clipboard event, I compare the copied text with the question visible on the page. I classify the selection as a complete or nearly complete question, a number-only selection, a short lexical fragment, or the respondent's own draft answer. The principal routing event is a complete or nearly complete question copy followed by a switch to another tab, window, or application before the respondent advances to the next survey page. The coding is deliberately conservative. Number-only copies may reflect calculator use; copying only "200" from a compound-interest item, for example, differs from copying the item's prose together with its dollar amount, interest rate, and time horizon. Short words and phrases receive little weight because they may reflect dictionary or translation use, and copies of a respondent's own draft answer are excluded.[4] A routing event therefore shows that a question left the survey, but not where it went or whether the external consultation affected the answer. It is consistent with AI use, but also with search, translation, or other assistance, and is not treated as proof of AI involvement in isolation.

Model processing. To distinguish external routing from model processing, I embed clipboard canaries in three survey questions. A canary is hidden material added only when the question is copied; it never appears on the survey page. For two questions, the copied text contains a base64-encoded instruction directing an instruction-following model to include an unusual marker phrase in its response.[5] For the third, the copied text contains a short planted phrase designed to be mentioned in the response. Because neither canary is visible in the survey, it can affect an answer only after the question has been copied. I record a canary round trip when text subsequently entered into the survey contains the designated marker or mentions the planted phrase. This event provides mechanism-specific evidence that an instruction-following language model processed the copied question. Its interpretation is intentionally narrower than "the model wrote the answer": a round trip does not establish that the respondent pasted the output verbatim, accepted its substantive content,

[3]Screen recording would have required more intrusive data collection and could itself have changed respondent behavior. I therefore rely on browser-side events that leave the survey experience unchanged.

[4]Eligibility was restricted to Prolific participants who reported fluency in English.

[5]The encoded string resembles an internal survey tag, reducing the likelihood that a respondent who inspects the copied text interprets it as part of the visible question.

or left it unedited.

Answer delivery. For each answer field, I reconcile the submitted value with its recorded input history, including keystrokes, paste actions, dictation, and other browser input events. I use two signatures of machine delivery. The first is *synthetic typing*: a sustained sequence of keyboard events in which most events report no key identity or otherwise differ from ordinary physical keystrokes. I evaluate this signature within respondent, comparing the suspect field with the respondent's named, conventional keystrokes elsewhere in the same session. The second is a *direct field write*: a completed field that cannot be matched to any recorded keystroke, paste, or dictation event.[6] These signatures identify how the answer entered the survey, not who generated its content. Machine entry is therefore decisive for distinguishing software delivery from ordinary input, but it does not by itself establish that an AI model produced the underlying answer.

Session operation and environment. For sessions with machine-entry evidence, I use the surrounding interaction record to determine whether software operated inside a human-run session or completed the survey autonomously. Machine entry on selected fields, together with ordinary navigation, mouse movement, and named keystrokes elsewhere, indicates an agentic tool used with a person in the loop. Machine entry throughout the survey, little or no ordinary human interaction, and corroborating browser or network evidence indicate autonomous completion. The environmental record includes browser-reported automation control, consistency across the reported browser, device, and operating system, graphics-rendering characteristics, and whether the connection originates from a residential network or a commercial hosting provider. These variables are useful mainly for separating agentic tools from autonomous bots. They reveal little about a respondent who consults AI in another tab or application while continuing to use an ordinary consumer browser. I therefore never classify a session from an environmental anomaly alone. Such anomalies enter only when they identify a specific automation mechanism or corroborate independent evidence from answer delivery and session behavior.[7]

Supplementary evidence and unobserved assistance. Two additional sources can capture assistance that does not generate a clear process trace. First, I examine open-ended

[6] Autofill can also generate incomplete input histories. I therefore apply these signatures only to survey-specific fields for which autofill has no plausible purpose and exclude cases with a credible benign explanation.

[7] For example, software rendering can result from disabled hardware acceleration, a virtual machine, or remote desktop; an unusual browser configuration can reflect privacy software; and a nonresidential connection can reflect a VPN or corporate network.

responses for linguistic patterns associated with model-generated text. Second, I ask respondents in the debrief whether they used an AI tool. Both measures are informative but imperfect: human writing can be formulaic, model output can be edited, and self-reports are subject to nondisclosure and reporting error. I therefore retain them as separate, supporting channels rather than treating either as a stand-alone behavioral proof. Some assistance remains outside the observable boundary altogether. A respondent may photograph a question, use a second device or voice assistant, and then enter the answer manually. Such a session can display ordinary navigation, mouse movement, and keystrokes even when AI shaped the substantive response. Ordinary human interaction therefore establishes human participation, not unaided authorship.

2.3.3 Classification and Interpretation

I aggregate the evidence in a fixed priority order, from the most specific machine-delivery mechanisms to the least conclusive forms of AI assistance. Once a session meets a higher-priority rule, it is not reconsidered for a lower category. The classification is outcome-blind: no reported inflation expectation or downstream distributional outcome enters any rule.

Autonomous bot. A session is classified as an autonomous bot when machine entry is pervasive, ordinary human interaction is absent, and the machine-delivery interpretation is corroborated by browser, rendering, automation, device, or network evidence.

Agentic tool. A session is classified as agentic when software enters answers on selected fields but the same session contains ordinary human navigation, mouse movement, or named keystrokes elsewhere.

AI-assisted human. A human-operated session is classified as AI-assisted when it contains a mechanism-specific trace of model processing, such as a canary round trip, or when question-routing evidence is corroborated by a mechanically distinct channel. The respondent may still enter the resulting answer through ordinary typing, paste, or dictation.

Possibly AI-assisted human. A session enters the possibly-AI-assisted category when an informative but ambiguous signal meets the operational threshold for retention without sufficient corroboration for a stronger category. This tier preserves evidence with a plausible non-AI explanation while keeping it separate from mechanism-specific or multi-channel cases.

No detected evidence. The residual category contains sessions for which no signal meets the threshold for one of the four positive-evidence categories.

Three points guide my interpretation. First, the labels describe observable production mechanisms, not answer plausibility. A returned canary establishes model processing; machine entry establishes software delivery; other strong cases require corroboration across mechanically distinct channels. Second, "no detected evidence" does not mean verified unaided completion. It is a residual group that can contain second-device use, voice consultation, manual retyping, and other assistance that does not cross an observable browser boundary. Third, the resulting misclassification is asymmetric for the comparisons below: undetected AI-assisted respondents remain in the residual group and attenuate differences between detected and undetected sessions. The reported shares should therefore be read as the prevalence of detected mechanisms, not as a complete estimate of true AI use.

3 AI Involvement in the Field

I now apply the detection method to an SCE-style field survey. I first describe the survey setting and sample, then report the detection layers, respondent classifications, and performance on conventional quality screens, and conclude with question-specific evidence on inflation expectations. Throughout, I use "AI involvement" to mean AI involvement detected by this method; because each classification rests on how an answer was produced, the field results below report associations between detected AI involvement and reported expectations rather than the causal effect of AI on any individual respondent's belief.

3.1 The Surveys

Household inflation expectations are economically relevant because households make decisions using the inflation they anticipate rather than the inflation that is ultimately realized. These expectations are associated with economically important choices, including consumption and saving, and may influence wage bargaining and aggregate demand (Weber et al., 2022; D'Acunto et al., 2022). Central banks, governments, industry, and investors therefore monitor household expectations not simply to evaluate forecast accuracy, but to assess whether the public continues to expect inflation near the target (Reis, 2021). The cross-sectional distribution is central to that exercise. A stable median can coexist with a growing group of households that expects persistently high inflation. Disagreement can reflect differences in information, personal experience, and exposure to prices; upper-tail mass captures the households furthest above the central bank's objective; and skewness helps distinguish

a broad upward shift from one concentrated among a smaller group. These statistics are not direct measures of subsequent behavior. They are warning indicators that economically consequential beliefs may be changing.[8]

One might ask why we do not simply glean this information from the market. Market prices provide complementary information, but they measure a different object. Nominal yields, inflation-indexed bonds, and inflation swaps combine inflation expectations with risk and liquidity premia (Gürkaynak et al., 2010; Ang et al., 2007; Chernov and Mueller, 2012). By construction, these market measures do not reveal the cross-sectional distribution of beliefs across the households whose decisions transmit expectations into the real economy.

The SCE (Armantier et al., 2017) is a valuable setting in which to study AI involvement. The survey is informative because independently sampled households bring different information, experiences, and perceptions to their answers. The value of the SCE is therefore not that it supplies another point forecast but that it directly elicits household beliefs about economic conditions and shows how those beliefs vary across people. A model may reproduce the center of that distribution without reproducing its heterogeneity. If respondents replace their own judgments with a common model answer, the median can remain plausible even as the survey loses part of the cross-sectional information it was designed to measure.

I field an SCE-style survey on Prolific, a recruitment platform whose data quality compares favorably with other online panels and crowdsourcing platforms (Peer et al., 2022), and obtain 1,294 survey completions.[9] The survey contains the SCE core module, additional expectations and demographic items, an instructed-response attention check, three financial-literacy items, four numeracy items, the three-item Cognitive Reflection Test, and four visual-illusion tasks (the checker-shadow, Müller-Lyer, café-wall, and reverse Müller-Lyer tasks). These modules serve distinct purposes. The attention check measures compliance with an explicit instruction; the financial-literacy, numeracy, and cognitive-reflection items measure knowledge and deliberative performance; and the visual items measure susceptibility to familiar perceptual illusions. I use them as benchmarks for conventional respondent-quality screening rather than as direct evidence of AI involvement. Alongside the submitted answers, I record the completion-process paradata described in Section 2.3.

[8]See, for example, Mankiw et al. (2003); Coibion and Gorodnichenko (2012); Malmendier and Nagel (2016); D'Acunto et al. (2021); Reis (2021); Natoli and Sigalotti (2018); Grishchenko and Wilcox (2024).

[9]The survey was fielded on June 3–4, 2026. The detection method was applied only to this independently fielded survey, not to the official SCE. The survey reproduces and modestly extends the SCE questions (Federal Reserve Bank of New York, 2024) for the present study. Some survey questions were taken or adapted from the Survey of Consumer Expectations, © 2013–2024 Federal Reserve Bank of New York (FRBNY); the SCE questions are available without charge at newyorkfed.org/microeconomics/sce and may be used subject to the license terms posted there, and FRBNY did not participate in or endorse this study. FRBNY disclaims any responsibility or legal liability for the administration of the survey and the analysis and interpretation of data collected. The complete survey instrument is available from the author on request.

3.2 Results

Before investigating the survey data from Prolific, I compare the Prolific and SCE panels. I then turn to the Prolific panel and report the individual detection layers and respondent classifications. Next, I examine conventional quality screens and the characteristics associated with AI involvement in my survey data. I conclude with question-specific evidence on reported inflation expectations.

3.2.1 The Prolific and SCE Samples

Table 2 compares the Prolific and SCE panels. Gender, regional, and marital-status shares are close across the two panels. The main differences concern age, education, income, employment, and homeownership. The Prolific panel is younger (40.7% under age 40, against 28.7% in the SCE), substantially more educated (70.9% in the college category, against 31.9%), more likely to be employed (77.0% against 62.2%), less likely to own a home (54.5% against 68.0%), and more concentrated in the middle and upper portions of the income distribution. The Prolific panel, while drawn to be representative of the United States, is therefore not a perfect substitute for the SCE panel. The field survey is designed to reproduce SCE-style measurement in a web-panel environment, not to estimate the official SCE's rate of AI involvement.

3.2.2 Detection Layers and AI Involvement Classification

Table 3 summarizes the detection evidence and resulting AI involvement classification. Panel A reports the individual detection layers and their overlap with respondent disclosure. Panel B maps the full evidence set into mutually exclusive AI involvement classifications. Because respondents can trigger multiple detection layers, the rows in Panel A overlap and do not sum to the totals in Panel B.

Panel A shows that question-text copying is by far the most common recorded trace. In total, 319 respondents (24.7%) copy a question. Within this group, 77 respondents (6.0%) complete a returned-canary round trip, providing more specific evidence that an instruction-following model processed the copied text. The AI-prose register flags 40 respondents (3.1%), synthetic keystrokes 13 (1.0%), machine fingerprints 8 (0.6%), and direct field writes and datacenter origin 5 each (0.4%). In the debrief, 62 respondents (4.8%) disclose AI use, and my detection method independently classifies 31 of them as showing AI involvement. In summary, the dominant evidence is question-text copying, the routing of questions out of the survey; direct machine-entry and infrastructure traces are comparatively rare.

Panel B reports the AI involvement classification. In total, 406 of the 1,294 completers

(31.4%) show evidence of AI involvement. Of these, 282 respondents (21.8%) trigger multiple detection layers and 102 (7.9%) trigger one. A further 17 sessions (1.3%) are classified as agentic tools based on the layers they trigger and 5 (0.4%) as autonomous bots. Excluding respondents who trigger only a single detection layer leaves 304 (23.5%) strong cases of AI involvement. The dominant pattern is therefore human respondents completing the survey while routing at least some content through AI, not bots replacing respondents. The residual group contains 888 respondents (68.6%) with no detected evidence of AI involvement.

The accounting that links the two panels is as follows. The six core behavioral layers in Panel A fire for 331 unique respondents. Of these, 315 are corroborated and enter the positive-evidence set, while 16 isolated, uncorroborated firings are not promoted and remain in the no-evidence group. The 406 respondents with positive evidence therefore comprise 315 corroborated core-layer cases, 31 cases identified only by self-report, and 60 cases identified only by auxiliary paradata, chiefly window focus-loss; 58 of the 60 enter the possibly-AI-assisted tier. Self-report and behavioral detection also capture largely different people: of the 331 behaviorally flagged respondents, only 31 also self-report, and of the 62 self-reporters, 31 leave no behavioral trace. A McNemar test (31 self-report-only against 300 behavioral-only) rejects equality of the two channels' coverage ($\chi^2(1) = 216.99$, $p < 0.001$). A survey relying on disclosure alone would therefore miss most of the evidence of AI involvement.

3.2.3 Conventional Quality Screens

Conventional quality screens are designed to identify inattentive or low-effort responses. If respondents who leave AI evidence were simply poor-quality participants, they should perform worse on these measures. Table 4 tests this implication. The category rows report average performance within each block, while the indented rows show which individual items account for the aggregate differences.

I begin with financial literacy: respondents with evidence of AI involvement answer 87.4% of questions correctly, compared with 77.9% among respondents with no detected evidence, a 9.5-percentage-point difference that is statistically significant at the 1% level and driven mainly by the compound-interest question, on which the positive-evidence group performs 21.2 points better. The numeracy questions follow the same pattern. The positive-evidence group answers 91.9% of the four items correctly, compared with 85.5% in the no-evidence group, with the largest item-level gaps on the lottery-probability question (8.2 points) and the infection-risk question (16.0 points). Both groups perform very well on the numeracy items. The largest differences appear on the Cognitive Reflection Test. Respondents with positive evidence answer 69.5% of the three questions correctly, compared with 48.3% among respondents with no detected evidence, a 21.3-percentage-point difference to which

each item contributes roughly equally. The visual-illusion tasks are the exception to the previous pattern. Respondents with positive evidence answer 27.3% of the four items correctly, compared with 29.4% among respondents with no detected evidence, and the difference is not statistically significant. The item-level estimates do not point in a common direction: the positive-evidence group performs 7.6 percentage points better on the café-wall item but 7.3 points worse on the reverse Müller-Lyer item.

Taken together, the results are inconsistent with the view that respondents who use AI are simply inattentive or low-effort participants. They submit substantially more correct answers on financial-literacy, numeracy, and cognitive-reflection items, while showing no systematic advantage on the visual tasks. Several mechanisms could produce this pattern. Some respondents may consult AI or other external tools, experienced survey participants may recognize commonly used questions, and respondents who seek outside assistance may also be positively selected on education, digital fluency, effort, or willingness to investigate a difficult item rather than guess. Because question copying is itself one detection layer, the positive-evidence group may disproportionately capture this last form of help-seeking. Conventional quality screens would therefore tend to retain these respondents.

Table 5 asks whether the same pattern survives demographic adjustment. I report average marginal effects from a logit model for the probability that a respondent shows evidence of AI involvement. Column (1) contains the demographic baseline. Columns (2) through (4) add, one at a time, the comprehension score, the visual-illusion score, and survey duration; column (5) includes all three. The comprehension score counts correct answers across the ten financial-literacy, numeracy, and cognitive-reflection items in Table 4. The comprehension score, visual score, age, household income, and duration enter the model in standard-deviation units. All columns use the 1,294 observations from the Prolific sample.

Turning to column (1), a one-standard-deviation increase in age is associated with a 7.5-percentage-point lower probability of showing AI evidence. A college degree is associated with a 10.0-point higher probability, a one-standard-deviation increase in household income with a 2.9-point increase, and being employed with a 6.9-point increase. Women and White respondents are 4.7 and 5.0 points less likely to show positive evidence, respectively, but both are marginally significant and do not survive the full specification in column (5).

A one-standard-deviation increase in comprehension is associated with a 10.6-point higher probability of showing AI evidence (column (2)), and a one-standard-deviation increase in duration with a 7.5-point increase (column (4)). The visual-illusion score is insignificant in column (3) but negative and significant in column (5). In general, the coefficients are slightly attenuated in the joint specification of column (5). The McFadden pseudo-R^2 rises from 0.052 in the demographic baseline in column (1) to 0.109 in the joint specification in

column (5).

In summary, the evidence runs counter to the conventional screening logic. Respondents with positive evidence are younger, more educated, more successful on the comprehension battery, and slower to complete the survey; they do not differ systematically in visual perception accuracy. Standard screens can identify careful participation, but they cannot establish whether AI was used to answer the question.

3.2.4 Inflation-Question Routing and Reported Expectations

A useful feature of my method is that it records not only whether a respondent used AI but also where: the specific question on which the evidence occurs. A respondent may consult a model on a numeracy question or an open-text item and answer the inflation block using their own beliefs about future economic conditions. I therefore turn to question-specific evidence and ask whether respondents with AI evidence on the inflation questions report a different distribution of expectations than those who leave no evidence anywhere.

For this analysis, I focus on the 133 respondents (10.3% of the sample) with evidence of AI involvement during the inflation-question block. The group contains 111 AI-assisted respondents, 17 agentic-tool sessions, and 5 autonomous bots. The comparison group, which I call the no-evidence group, contains the 888 respondents with no detected evidence anywhere in the survey. I exclude a further 273 respondents whose AI use occurs only on other questions.

Figure 1 depicts how each group answers the inflation expectation question at the one-, three-, and five-year horizons. Each horizon shows the 133 respondents with AI evidence on the inflation block next to the 888 respondents with no detected evidence anywhere in the survey. The box spans the interquartile range, the horizontal line marks the median, and the whiskers show the tenth and ninetieth percentiles. The vertical axis is split above 10% so that the central mass near 3–4% and the upper-tail differences can be seen in the same figure.

The figure makes two points. First, the centers are close. At one year, the AI-involved group has a median of 3%, compared with 4% in the no-evidence group; at five years, both medians are 3%. Second, the surrounding distribution is much narrower for respondents with AI evidence on the inflation questions. The no-evidence group has long upper whiskers at every horizon and, at the three- and five-year horizons, a visibly wider lower tail as well. The AI-involved group instead clusters tightly around the center. A two-sample Kolmogorov–Smirnov test confirms that the full distributions differ at the one-, three-, and five-year horizons ($p < 0.001$ in each case). The difference is not a uniform downward shift in expected inflation. It is an inward movement from both tails toward a focal center.

Table 6 presents more descriptive statistics to complement Figure 1 and uses the same 1,021-respondent sample. Each panel corresponds to a forecast horizon. The first two rows compare the no-evidence and AI-involved groups; the indented rows split the AI-involved group into AI-assisted humans, agentic-tool sessions, and autonomous bots. The mean, median, standard deviation, and the tenth and ninetieth percentiles are provided.

Panel A presents the one-year inflation expectation result. The mean is 7.08% in the AI-involved group and 7.96% in the no-evidence group, while the corresponding medians are 3% and 4%. The standard deviations are nearly identical, at 16.63 and 16.79 percentage points. The percentile columns nevertheless show a narrower middle of the distribution among AI-involved respondents: their tenth-to-ninetieth-percentile range runs from 1% to 15.6%, compared with 0% to 25% in the no-evidence group. Thus, even where the standard deviation is almost unchanged, the upper part of the middle 80% is thinner among respondents with AI evidence on the inflation block. The differences become substantially larger at the three-year horizon in Panel B. The AI-involved group has a mean of 3.20%, a median of 3%, and a standard deviation of 8.15 percentage points. The corresponding values in the no-evidence group are 6.55%, 4%, and 16.92 points. The tenth-to-ninetieth-percentile range is 1.2% to 5% for AI-involved respondents, compared with −3% to 20% for the no-evidence group. Both sides of the distribution move inward: the AI-involved group has less mass in the high-inflation upper tail and less mass in the zero-or-negative lower tail. Panel C shows the same pattern at five years. Both groups report a median of 3%, so the center alone would suggest little difference. The dispersion measures tell a different story. The standard deviation is 7.90 percentage points in the AI-involved group and 18.31 points in the no-evidence group. The tenth-to-ninetieth-percentile range is 2% to 6% among AI-involved respondents, compared with −5% to 20% among respondents with no detected evidence. The five-year median is unchanged even though the distribution around it is substantially narrower. The indented rows show that this is not only a bot result. Autonomous bots account for five observations and agentic-tool sessions for seventeen. The bulk of the AI-involved group consists of 111 AI-assisted human sessions, and that subgroup also shows compression at every horizon. At one year its tenth and ninetieth percentiles are 2% and 20%, at three years they are 2% and 5%, and at five years they are 2% and 6%.

In summary, inflation-question evidence is associated with substantially less dispersion, particularly at the three- and five-year horizons. At one year, the middle of the distribution is narrower even though the standard deviations are nearly identical. At longer horizons, both the lower and upper percentiles move toward the center. The five-year result is especially informative: the median remains at 3%, while the middle 80% contracts from a 25-point range in the no-evidence group to a 4-point range in the AI-involved group. A focus on the

median alone would therefore miss most of the difference between the groups.

To identify which regions of the distribution account for this compression, Table 7 examines five probabilities. The first outcome, $y \leq 0$, captures the lower tail of the distribution. The next two, $y > 5\%$ and $y > 10\%$, measure progressively more extreme upper-tail mass. The final two, $y = 3\%$ and $y = 4\%$, measure concentration at adjacent focal values near the center.[10]

For each horizon h and distributional outcome c, I estimate the linear probability model

$$1\{Y_{ih} \in \mathcal{C}_c\} = \alpha_{ch} + \beta_{ch}\, AIInvolved_i + X_i'\gamma_{ch} + \varepsilon_{ich}, \tag{2}$$

where $\mathcal{C}_c \in \{(-\infty, 0], (5, \infty), (10, \infty), \{3\}, \{4\}\}$ and $h \in \{1, 3, 5\}$. The indicator $AIInvolved_i$ equals one for the 133 respondents in the AI-involved group and zero for the 888 respondents in the no-evidence group. The vector X_i contains controls for age, sex, education, household income, race, and employment. Each cell in Table 7 reports a separate estimate of β_{ch} with standard errors in parentheses.[11] Because the dependent variable is binary, the coefficients are adjusted probability differences. For example, the coefficient of 0.252 in the three-year $y = 3\%$ row means that AI-involved respondents are 25.2 percentage points more likely to report exactly 3%, conditional on the included covariates.

I present the one-year, three-year, and five-year inflation expectation horizons in columns (1), (2), and (3), respectively. The first row describes the lower tail of the distribution. At one year, the estimated difference in the probability of a nonpositive report is only -1.0 percentage point and is statistically insignificant. At three and five years, however, AI-involved respondents are 9.3 and 20.1 points less likely to report a zero or negative expectation. The next two rows show upper-tail contraction at every horizon. AI involvement is associated with reductions of 19.9, 28.9, and 19.3 percentage points in the probability of reporting inflation above 5% at the one-, three-, and five-year horizons. The probability of reporting inflation above 10% falls by 8.4, 15.7, and 10.8 points. The final two rows show where part of the mass concentrates. AI-involved respondents are 20.9 percentage points more likely to report exactly 3% at one year, 25.2 points more likely at three years, and 26.6 points more likely at five years. By contrast, the exact-4% coefficient is significant only at the five-year horizon, and only marginally, at -3.6 points.[12]

[10] The coefficients describe selected regions of the distribution and should not be added together.

[11] The bracketed third line in each cell reports the mean of the outcome in the no-evidence group, the base rate against which each coefficient should be read. The 8.4-point reduction in the probability of reporting one-year inflation above 10% is measured against a no-evidence base rate of 19.1%, so AI involvement roughly halves that tail.

[12] Direct tests reject equality of the exact-3 and exact-4 coefficients at every horizon, with differences of 16.9, 26.3, and 30.2 percentage points.

In summary, the estimates show upper-tail thinning at every horizon, lower-tail thinning at the three- and five-year horizons, and a large increase in the probability of reporting exactly 3%. The pattern is consistent with answers moving inward from the outer regions of the distribution and concentrating at a focal center. It is not simply a common downward shift: the longer-horizon distributions contract on both sides, and the five-year median remains unchanged.[13]

These differences are useful for diagnosing measurement error, not for screening individual respondents. The two distributions still overlap substantially, so a respondent's inflation forecasts alone cannot reliably reveal whether the answers were produced unaided or with AI. The classification therefore continues to rest on completion-process evidence rather than on reported expectations.[14]

These field results motivate a direct model calibration. The field comparison is observational: respondents who route inflation questions may differ from other respondents in unobserved ways. I therefore ask fifteen consumer AI models the same SCE questions. Because the source of every answer is known, this exercise tests whether model responses themselves reproduce the plausible center, thin tails, and focal-value concentration observed among routed respondents, and whether the pattern holds across a broad model panel.

4 Calibrating How AI Answers the SCE

The field evidence in Section 3 shows that AI involvement is associated with a distinctive pattern in reported inflation expectations, but it cannot fully separate the mechanism from selection. Respondents who route inflation questions outside the survey may differ from other respondents in uncertainty, effort, digital fluency, or willingness to seek outside help. I therefore turn from detected respondent behavior to the model-answer distribution itself. The question is whether a consumer AI model, when asked an SCE inflation question, returns something resembling a draw from the household cross-section or instead supplies a more centralized answer. This section calibrates that object directly by asking fifteen consumer AI models the same SCE inflation items and comparing their answers with public SCE microdata.

[13] A table in a forthcoming internet appendix shows that the dispersion gap remains after controlling for age, sex, education, household income, race, and employment. Eleven of the fifteen distributional estimates in Table 7, including every tail-mass and exact-3-percent coefficient, remain significant at the 5% level after a Benjamini–Hochberg correction across the fifteen cells.

[14] A forthcoming internet appendix reports an answer-based benchmark. The score is each respondent's distance from the model-pool consensus across the one-, three-, and five-year point expectations. It reaches an AUC of 0.73, where 0.5 corresponds to random sorting and 1 to perfect separation, but identifies only 29% of AI-involved respondents when the false-positive rate is limited to 10%.

4.1 Calibration Design

The human benchmark is the public SCE microdata.[15] The main comparison uses the pre-AI period from June 2013 through November 2022. I also investigate the normal-inflation period of 2019 and a high-inflation period from 2022. The model benchmark contains fifteen consumer AI models queried in May 2026. Each model receives the SCE inflation items in two formats. The single-question call presents the inflation item in isolation, mimicking a respondent who copies only that question into a chatbot. The complete-survey call embeds the same item in the full survey instrument, mimicking a respondent who routes the broader survey context. Together, the two formats produce 886 valid one-year answers.

Because the model calls were made in 2026 while the human benchmarks come from earlier SCE waves,[16] I do not interpret level differences as evidence that AI forecasts are too high or too low. Calendar time can move the center of inflation expectations. It cannot, however, explain why the model distributions are nearly degenerate relative to both calm and high-inflation human benchmarks. A contemporaneous comparison makes the same point: field respondents with no detected AI evidence, answering three weeks after the model calls, report a one-year standard deviation of 16.8 percentage points (Table 6). I therefore use the calibration to compare distributional shape rather than contemporaneous forecasting performance. I report the human series without winsorization and drop only the 21 one-year and 17 three-year observations that fall outside ±100 percentage points; no model answer lies outside that range.[17]

4.2 The Distribution of AI Inflation Expectations

Table 8 presents the main calibration. Panel A reports one-year expectations and Panel B reports three-year expectations. Within each panel, the first three columns provide the public SCE microdata results and the last two columns provide the AI model results pooled across all models. I report the mean, median, standard deviation, interquartile range, tenth percentile, ninetieth percentile, the share of answers equal to 3.0, 3.5, or 4.0%, and the observation count in rows.

I now turn to Panel A, where the one-year inflation expectations are presented. When

[15] Source: Survey of Consumer Expectations, © 2013–2024 Federal Reserve Bank of New York (FRBNY). The SCE data are available without charge at newyorkfed.org/microeconomics/sce and may be used subject to the license terms posted there. FRBNY disclaims any responsibility for this analysis and interpretation of Survey of Consumer Expectations data.

[16] Once the data become available, I plan to compare the SCE data from May 2026 to the AI answers from May 2026.

[17] In an internet appendix, I will provide replication materials with model access dates, sampling parameters, extraction rules, and invalid-response counts.

reading the mean and median rows together, the AI answers look plausible at the center but already differ from the human distribution. The single-question and complete-survey means are 3.3% and 3.4%, below the SCE means of 5.9% in the full pre-AI sample, 4.7% in 2019, and 9.4% in the high-inflation 2022 period. The pre-AI and 2019 SCE medians are both 3.0%, while the model medians are 3.2% and 3.5%. Only in the high-inflation 2022 period, where the SCE median rises to 7.0%, is the model center far from the human center. The dispersion rows show that this central plausibility does not hold for other statistics. The SCE standard deviation is 12.7 percentage points in the full pre-AI sample, 11.8 points in 2019, and 15.2 points in 2022. The AI standard deviation is 0.4 percentage points in both prompt formats. The interquartile-range row shows the same pattern: 5.0, 3.0, and 6.0 points in the SCE columns, compared with 0.5 points in each AI column. The percentile rows make the compression even clearer. In the SCE, the tenth-to-ninetieth-percentile range runs from 1% to 15% in the pre-AI period, from 1% to 12% in 2019, and from −2% to 20% in the high-inflation period. In both AI formats, the same range is only 3% to 4%. The normal period narrows the human distribution and the high-inflation period widens it, but neither human benchmark resembles the model distribution. The focal-value row shows that the compression is not only the absence of extreme forecasts. Human respondents do report round values, but the share answering exactly 3.0, 3.5, or 4.0% is 22.4% in the full pre-AI sample and 23.7% in 2019. In the high-inflation 2022 period, the share falls to 9.4% as the human distribution moves outward. The AI shares are much larger: 76.0% in the single-question calls and 87.5% in the complete-survey calls. The model answers therefore do not merely avoid the tails. Most of them collapse onto a small set of adjacent central values.

Panel B shows that the same pattern holds for the three-year inflation expectation. The mean row again places the AI answers below the SCE means: 2.9% and 2.8% for the two model formats, compared with 5.3% in the full pre-AI sample, 4.7% in 2019, and 4.2% in 2022. The median row is closer: the SCE medians are 3.0%, 3.0%, and 4.0%, while the AI medians are 3.0% and 2.8%. Thus, as in Panel A, the model answers are not implausible as central forecasts. The problem is that the rest of the row structure diverges. For three-year expectations, the SCE standard deviations are 13.1, 12.1, and 16.0 percentage points across the three human benchmarks. The AI standard deviation is 0.3 points in both formats. The SCE interquartile ranges are 4.0, 3.0, and 10.0 points; the AI interquartile range is again only 0.5 points. The percentile rows show the same collapse. The SCE tenth-to-ninetieth-percentile range is −2% to 15% in the pre-AI period, −1% to 12% in 2019, and −7% to 15% in the high-inflation period. The AI ranges are only 2.5% to 3.5% in the single-question calls and 2.5% to 3.0% in the complete-survey calls. Even the calm 2019 SCE distribution

contains far more lower-tail and upper-tail variation than the model answers. The focal-value row again points to concentration rather than replication of the cross section. In the SCE, the share at 3.0, 3.5, or 4.0% is 23.7% in the pre-AI period, 24.8% in 2019, and 15.0% in 2022. The corresponding AI shares are 63.6% and 46.0%. The complete-survey three-year calls have a lower focal-value share than the single-question calls, but this does not mean that they recover human heterogeneity: their tenth-to-ninetieth-percentile range is only 2.5% to 3.0%. They remain tightly concentrated, just around a slightly different set of central answers.

In summary, the row-by-row comparison paints a picture of AI homogenization. Across horizons, prompt formats, and human benchmark periods, AI produces something that looks like a reasonable central forecast but not like a draw from the household cross-section. AI answers preserve a plausible normal-times center but do not reproduce the state-dependent movement of that center or the cross-sectional heterogeneity of the household distribution. In the high-inflation 2022 period, the SCE median rises to 7.0%, while the model center stays near 3–4%. The SCE distribution remains dispersed even in a normal inflation-expectation year, and it becomes more dispersed in the high-inflation episode. The AI distribution is narrow. The human mean exceeds the human median because a relatively small set of households reports high inflation. Those observations may be inaccurate as forecasts, but their prevalence carries information about disagreement, uncertainty, and the risk of unanchoring. A model distribution that reproduces the median while eliminating those observations does not reproduce the object the SCE is designed to measure. It supplies a centralized forecast rather than an independent household belief.

4.3 Across Models, Prompts, and Personas

The data reported in Table 8 are generated using one of two prompts. The model receives either the inflation question alone or the full survey instrument, but it is not assigned a demographic identity. This is intentional. The object of interest in the field setting is not whether a researcher can use language models to construct a synthetic population. It is what a respondent is likely to receive when using AI as a quick aid while completing a survey. A respondent who copies an inflation question into a chatbot is unlikely to add a carefully specified age, income, education, region, and numeracy profile, let alone sample across a population of such profiles. I therefore design my prompts to match the natural use case most relevant for AI-assisted humans.

This design also raises a natural objection. If every model is asked the same question without a persona, perhaps the answers should be expected to look similar. In that view, the

compression in Table 8 might say less about AI-assisted survey response and more about the absence of demographic conditioning. I address that concern in two steps. I first show that my results are not driven by a particular model, by unequal call counts, or by whether the inflation item is asked alone or inside the full survey. I then ask whether explicitly supplying demographic personas restores the missing heterogeneity. The persona exercise is therefore a robustness check against the prompt-design concern. Persona prompting may be relevant when bots answer a survey or for a researcher building silicon respondents, but the dominant field concern in this paper is AI-assisted humans.

Table 9 reports the one-year inflation expectation separately for every model-by-prompt group. Across the fifteen AI models and two prompt formats, means range from 2.7% to 4.1%. No model has a standard deviation above 0.8 percentage points, and no answer exceeds 5%. In the call-weighted pool, the mean is 3.4%, the median is 3.5%, the standard deviation is 0.4 points, the interdecile spread is 1.0 percentage point, and 81.8% of answers fall exactly on 3.0, 3.5, or 4.0%. Giving each model equal weight leaves the mean and median at 3.5%, the standard deviation at 0.4 points, and the focal-value share at 84.3%. The pooled result is therefore not an artifact of one model, one prompt format, or unequal numbers of calls across systems.

The comparison between the two prompt formats is also informative. The single-question call mimics a respondent who routes only the inflation item to a chatbot. The complete-survey call gives the model more context by embedding the item inside the full survey instrument. If the narrow distribution were mainly a consequence of presenting the question in isolation, the complete-survey call should generate much more variation. It does not. The complete-survey answers are as concentrated as the single-question answers in Table 8. Thus, adding ordinary survey context does not make the model behave like a heterogeneous cross-section of households.

These estimates are still conditional on the prompt wording. The main analysis calls use a temperature of 0.7 and do not assign demographic identities. Lower-temperature decoding would ordinarily reduce stochastic variation further, while persona prompts can increase it. My claim is not that every conceivable prompt produces the same distribution but rather that under ordinary prompts that resemble quick respondent assistance, every model in the panel supplies a highly centralized answer distribution. Table 10 then asks whether demographic persona prompting restores the missing variation. This is a demanding robustness exercise because it gives the models information that ordinary routed respondents usually would not provide. Each call conditions the model on a demographic persona drawn from SCE sampling-frame margins, and six models[18] answer the full survey instrument for

[18]The six models are OpenAI's gpt-chat-latest and gpt-5.4-mini, Anthropic's claude-haiku-4.5, Google's

those personas. If the compression in the canonical calls were mainly due to the absence of demographic conditioning, the persona pool should move substantially toward the human SCE distribution.

Persona prompting widens the AI distribution, but it does not make it look like the SCE. Panel A shows that the one-year persona pool has a mean inflation expectation of 4.4%, a median of 3.8%, a standard deviation of 3.3 points, an interquartile range of 1.0 point, and a tenth-to-ninetieth-percentile range from 3.0% to 6.0%. This is substantially wider than the main analysis in Table 8, where the one-year standard deviation is about 0.4 points. But it remains far inside the pre-AI human distribution where the one-year standard deviation is 12.7 points and the tenth-to-ninetieth-percentile range runs from 1% to 15%. The same pattern appears at longer horizons. At three years, persona-prompted AI has an inflation expectation standard deviation of 1.1 points, an interquartile range of 0.4 points, and a tenth-to-ninetieth-percentile range from 2.4% to 4.0%. The corresponding pre-AI SCE range in Table 8 runs from -2% to 15%. At five years, the persona pool remains narrow as well, with a standard deviation of 2.0 points, an interquartile range of 0.5 points, and a tenth-to-ninetieth-percentile range from 2.3% to 4.0%. Focal-value concentration also remains high: 45.7% of one-year answers, 42.3% of three-year answers, and 31.3% of five-year answers land exactly on 3.0, 3.5, or 4.0%.

Panel B shows which imposed characteristics generate the additional spread. Age produces almost no gradient. The one-year means are 4.4%, 4.3%, and 4.4% for young, middle-aged, and older personas, respectively; the tenth percentile is 3.0% in all three groups, and the ninetieth percentile ranges only from 5.6% to 6.2%. Education produces a slightly larger but still modest difference. College personas have a mean of 4.1%, a median of 3.5%, and a ninetieth percentile of 5.0%, while non-college personas have a mean of 4.5%, a median of 3.8%, and a ninetieth percentile of 6.0%. Income generates the clearest gradient. Low-income personas have a mean of 5.2%, a median of 4.0%, a standard deviation of 4.7 points, and a ninetieth percentile of 8.0%. High-income personas have a mean of 3.9%, a median of 3.5%, a standard deviation of 1.4 points, and a ninetieth percentile of 5.0%. These differences are sensible, but they are not large enough to recover household heterogeneity. Across all demographic splits, interquartile ranges remain between 0.5 and 1.3 points, tenth percentiles

gemini-3-flash-preview, xAI's grok-4.3, and DeepSeek's deepseek-v4-pro, accessed through OpenRouter on 11–13 May 2026. I construct 231 personas from New York Fed SCE sampling-frame margins tabulated from 177,000 SCE respondent-months. Each persona is a cell defined by age group, education, household-income group, census region, and numeracy. Each model answers the full survey instrument once for each persona, giving an intended $6 \times 231 = 1{,}386$ full-survey calls, and therefore about 1,386 observations at each inflation horizon. The usable counts after invalid responses are 1,375, 1,381, and 1,384 at the one-, three-, and five-year horizons. This design uses many calls because the relevant variation is across demographic profiles and model families.

cluster between 3.0% and 3.2%, and large shares of answers still fall on the same three focal values. Persona prompting therefore adds some demographic structure, especially by income, but the resulting distributions remain concentrated around a narrow set of central forecasts.

Demographic conditioning can push model answers away from the tight distribution from the main analysis, but it does not recover the heterogeneity observed among households. Household expectations reflect personal price experiences, income shocks, local conditions, news exposure, attention, uncertainty, and subjective views of the economy. A demographic label supplied in a prompt is only a coarse representation of that information. It assigns heterogeneity to the model; it does not elicit beliefs from independently sampled people. The robustness results support the same conclusion as Table 8: consumer AI systems can supply plausible central forecasts, but they do not reproduce the cross-sectional heterogeneity that the SCE is designed to measure.

4.4 Density Forecasts

The main calibration uses point forecasts of inflation expectations. I now ask whether the same measurement problem appears in the SCE density questions where respondents can assess the probability of multiple future inflation scenarios. This check matters because the density instrument gives respondents a way to express uncertainty. A possible interpretation of the point-forecast results is that AI supplies a conventional central forecast but still conveys uncertainty through the probability distribution. In that case, the compression of point estimates would be less damaging for surveys that also collect densities.

Table 11 rejects that interpretation in a specific way.[19] The AI density forecasts are not invalid or mechanically degenerate. They are smooth. Their modal-bin shares are lower than those of SCE respondents, their entropy is higher, and their implied within-forecast standard deviations are comparable to those of human density forecasts. A standard response-quality screen would therefore have little reason to discard them. The loss appears across responses rather than within a single response. Among SCE respondents, the cross-sectional standard deviation of density means is 3.7 to 3.8 percentage points; among no-evidence field respondents, it is 3.4 to 4.0 points. Across model calls, it is below one percentage point at both horizons and in both prompt formats. The density instrument therefore shows the same homogenization as the point forecasts: individual AI responses can look reasonable while the cross-section loses disagreement.

The table also shows that AI removes a form of human probability reporting that is

[19]The SCE density statistics are computed from public microdata running from June 2013 through December 2021, a slightly shorter window than the June 2013–November 2022 period used for the point-forecast calibration in Section 4.

itself informative. About 18% of SCE density forecasts put at least 90% of the probability in a single bin. None of the model density forecasts does so. Prior work interprets coarse probability reporting and rounding as evidence about how respondents express uncertainty, not simply as noise (Binder, 2017; Manski and Molinari, 2010). AI therefore not only smooths the answer, but also replaces a human measurement object with a model-generated one that is valid-looking, centralized, and behaviorally different.

5 Consequences for Survey Measurement

The previous section calibrates the distribution of answers supplied by consumer AI models. The main result is that model answers can look reasonable at the center while containing little of the dispersion, tail mass, or focal-value behavior observed in the SCE. I now ask what this kind of answer distribution does to the statistics that researchers, central banks, and market participants monitor. I take human SCE answers from a high-inflation episode and replace a share of them with draws from the model-answer pool in Section 4. I then track the distributional statistics used to monitor disagreement and anchoring (Mankiw et al., 2003; Coibion and Gorodnichenko, 2012, 2015; Reis, 2021). The question is not whether the official SCE is currently contaminated at any particular rate. The question is what happens to monitored distributional statistics when a centralized AI answer distribution is inserted into a dispersed human cross-section.

It is useful to clarify what it means for a monitor to "fire." The word does not mean that the rule detects AI involvement. It means that the survey statistic still flags the high-inflation episode. Imagine a reader with a dashboard of inflation-expectations distributional statistics: the median, the mean, the standard deviation, the interquartile range, the interdecile spread, upper percentiles, and shares above fixed thresholds. Before seeing the high-inflation period, the reader uses normal-period data to learn what usual monthly values of each statistic look like. A monitor fires when the statistic in a simulated survey wave crosses its normal-period threshold in the direction associated with elevated or unanchored inflation expectations. The firing rate is the fraction of simulated survey waves in which that alarm would still go off after some human answers are replaced with AI answers.

This distinction separates the value of a statistic from its ability to send a warning. A statistic can keep firing even though its reported value is materially distorted. For example, the median may remain high enough to signal an elevated episode even after AI replacement has pulled it down from the clean 2022 value. Conversely, a statistic can stop firing when AI replacement pushes it back inside its normal-period range. That is the loss-of-signal case. The underlying clean episode has not changed; only the recorded survey answers have. The

monitor no longer warns the reader because the observations carrying the warning have been compressed toward the center.

This is why I report both distributional statistics and firing probabilities. The statistic values show the distortion in the reported statistic. The firing probabilities show whether a reader using that statistic as a monitoring rule would still recognize the high-inflation episode. The central concern is that these two can diverge. Center-based monitors can continue to fire while reporting a milder episode, whereas dispersion and upper-tail monitors can stop firing altogether. A dashboard can therefore look partly reassuring even when the survey has lost the disagreement and tail information it was designed to measure.

5.1 Design: Replacing Human Answers with AI Answers

The human benchmark is the SCE during the 2022 high-inflation episode. This period is useful because inflation expectations were elevated not only at the center but also in the upper tail. For each simulation, I draw a monthly sample of 1,000 observations using the official SCE respondent weights. I then replace a random fraction λ of the human answers with draws from the model-answer pool described in Section 4. I repeat this exercise 3,000 times at each value of λ.

Table 12 reports the average value of each monitored distributional statistic after replacement in columns (1) to (3) and the probability that the corresponding monitoring rule still identifies the high-inflation episode in columns (4) to (6). The statistic values show how much each reported statistic changes. The monitor-power column shows whether a rule based on that statistic still fires. A rule can continue to fire even when the statistic it reports has become economically distorted.

The monitoring thresholds are estimated from the 2014–2019 normal-period months. Calibration and evaluation use a disjoint calibrate-and-read split, so a simulated wave never both sets and tests its own threshold. In the uncontaminated 2022 benchmark, the main monitoring rules fire in essentially every simulation. The power columns therefore measure how often each rule continues to recognize the same high-inflation episode after AI replacement. I report three replacement shares, $\lambda \in \{0.1, 0.3, 0.5\}$.[20] The parameter λ is not directly observed: it is the share of answers actually replaced by AI content. I therefore treat λ as a stress scenario that traces how the monitored distributional statistics respond as AI replacement rises, not as an estimate of the rate of AI contamination in the official SCE.

[20] In the field survey, 133 of 1,294 respondents, or 10.3%, leave detected evidence during the inflation block, while 31.4% leave positive evidence of AI involvement somewhere in the survey.

One feature of the baseline is important for interpretation. The injected model answers come from May 2026, when the model pool is centered near 3–4%. The human benchmark comes from 2022, when the median one-year expectation is 8%. The baseline therefore combines two forces: replacement by a lower central forecast and replacement by a much less dispersed distribution. To isolate compression from this level shift, I repeat the exercise after shifting the AI pool so that its median equals the 2022 episode median. Table 13 reports that median-preserving exercise.

5.2 Results

Table 12 presents the main replacement results. I begin with the original SCE data in column (1). In the clean 2022 benchmark, the median one-year inflation expectation is 8.0%, the mean is 10.7%, the standard deviation is 17.5 percentage points, and the interquartile range is 8.6 points. The upper-tail rows show why this is a useful stress episode. The interdecile spread is 30.3 points, the ninetieth percentile is 28.1%, the ninety-fifth percentile is 47.9%, and 29.1% of respondents expect inflation above 10%. The episode is therefore not only high at the center. It has substantial disagreement and a large upper tail.

The first replacement share, $\lambda = 0.1$, already moves the reported distribution, but most monitoring rules still fire. The median falls from 8.0% to 7.1%, the mean falls from 10.7% to 10.0%, and the interdecile spread falls from 30.3 to 27.5 points. The upper-tail statistics also decline: the ninetieth percentile falls from 28.1% to 25.9%, the ninety-fifth percentile from 47.9% to 45.1%, and the share above 10% from 29.1% to 26.2%. These are meaningful movements, but the right half of the table shows that the alarms mostly remain intact at this replacement rate. The median, mean, interquartile range, interdecile spread, ninetieth percentile, ninety-fifth percentile, and threshold-share rules all fire with probabilities close to one. The main exception is skewness.[21]

At $\lambda = 0.3$ the monitoring rules split. The median falls to 5.0%, and the mean falls to 8.5%. The interquartile range remains 6.8 points, and the share above 5% remains 44.9%. These center and threshold statistics are still elevated enough that their monitoring rules fire in every draw. But the dispersion and upper-tail rows look different. The standard deviation falls to 15.0 points, the interdecile spread to 19.5 points, and the ninetieth percentile to 21.0%. The corresponding monitor powers fall sharply. The standard-deviation alarm fires in 37% of draws, the interdecile-spread alarm in 21%, and the ninetieth-percentile alarm in 26%. The

[21] Skewness is the one non-monotone monitor: its alarm fires in 1% of draws at $\lambda = 0.1$, 17% at $\lambda = 0.3$, and 58% at $\lambda = 0.5$, because replacement first moves the statistic toward its normal-period range before pushing it beyond. I report it in Table 12 but do not use it in the main comparison. The kurtosis alarm's firing rate declines from 98% at $\lambda = 0.1$ to 34% at $\lambda = 0.3$ and zero at $\lambda = 0.5$.

ninety-fifth-percentile alarm is more resilient, firing in 78% of draws, but it too loses power relative to the clean benchmark. The rules that survive are not unaffected. They fail more quietly. The median rule fires in every draw even though the median it reports has fallen from 8.0% to 5.0%. The share-above-5% rule also never misses even though the reported share falls from 64.2% to 44.9%. Thus, the rules that continue to signal a high-inflation episode report a materially milder one. The rules that would reveal the loss of disagreement and upper-tail mass are the first to stop firing.

At the $\lambda = 0.5$ level, the compression becomes severe. The median falls to 3.5%, the mean to 7.0%, and the interquartile range to 5.0 points. The interdecile spread is only 13.4 points, compared with 30.3 points in the clean benchmark. The ninetieth percentile falls to 16.1%, the ninety-fifth percentile to 28.0%, and the share above 10% to 14.6%. At this point, every dispersion and upper-tail rule is at or near zero power. The median rule still fires in every draw, and the mean rule still fires in 89% of draws, but the distribution has lost much of the tail information that made the episode informative.

Figure 2 shows the same pattern over the full grid of λ. The statistic values move smoothly as replacement rises. The alarm probabilities do not all move smoothly. The central alarms remain high, while the interdecile-spread and ninetieth-percentile alarms collapse between $\lambda = 0.1$ and $\lambda = 0.3$. This is the monitoring problem: the center can remain informative enough to identify that something is wrong, while the tail statistics that describe how wrong it is lose power first.

Table 13 shows that this pattern is not merely a level effect. I shift the AI pool so that its median equals the 2022 episode median. The contaminated median is therefore fixed at 8.0% by construction. The mean changes little: at $\lambda = 0.3$, it falls only from 10.7% to 9.9%. If the baseline results were driven only by replacing 2022 households with lower 2026 AI forecasts, the recentered exercise should remove the problem. It does not. At $\lambda = 0.3$, the standard deviation alarm fires in only 23% of draws, the interdecile-spread alarm in 22%, and the ninetieth-percentile alarm in 27%. These values are close to the baseline of 37%, 21%, and 26% in Table 12. The interquartile-range alarm now fails completely because the recentered model draws cluster near the median itself. The share-above-5% rule survives for a mechanical reason: after recentering, the AI answers lie above the 5% cutoff, so the reported share is distorted upward rather than downward. The lesson is unchanged. Tail erasure is not an artifact of comparing 2022 households with 2026 models. It is what happens when a low-variance answer distribution is inserted into a high-variance human cross-section.

5.3 Loss of Signal Mechanism

The baseline simulation treats an AI answer as complete replacement of the respondent's belief. Table 14 relaxes that assumption using Equation (1). I hold $\lambda = 0.3$ fixed and allow the recorded answer to retain a fraction α of the human answer, plus a small idiosyncratic noise term so that blended answers are not exact linear combinations. The parameter α therefore measures how much of the respondent's own belief survives AI influence. When $\alpha = 0$, the answer is pure replacement. When α is higher, the recorded answer remains closer to the original human response.

The rows of Table 14 show that retention is decisive. At $\alpha = 0$, the interdecile spread is 19.5 points and the interdecile-spread alarm fires in only 20% of draws, close to the baseline value in Table 12. At $\alpha = 0.3$, the interdecile spread rises to 22.5 points and the alarm fires in 59% of draws. At $\alpha = 0.5$, the interdecile spread rises to 25.7 points and the alarm fires in 96% of draws. At $\alpha = 0.7$, the interdecile spread is 27.7 points and the alarm never misses. The ninety-fifth-percentile monitor follows the same logic, although it is more resilient: its power rises from 78% at $\alpha = 0$ to 82% at $\alpha = 0.3$, 96% at $\alpha = 0.5$, and 100% at $\alpha = 0.7$.

The table separates two quantities that are otherwise bundled together: how many respondents consult AI and how much of their own belief survives the consultation. The damage is large when many routed answers are mostly model answers. It is much smaller when respondents use AI lightly and retain most of their original beliefs. If the covariate-adjusted contraction in long-horizon field answers, reported in a forthcoming internet appendix, were attributed entirely to blending rather than to selection into AI use, the implied retention would be low, roughly $\alpha \in [0, 0.3]$. This mapping is a calibration, not structural identification, because routing is not randomized and the field estimates come from different horizons. Its implication is nevertheless direct: mostly-human blends would not produce the degree of contraction observed in the field.

Table 15 asks a second question: which human answers get replaced? The baseline assumes random replacement. A more damaging alternative is tail-selected replacement, where replacement begins with the most extreme observations. This is the pattern one would expect if respondents with uncertain beliefs were more likely to seek assistance from AI. The field data cannot test this channel because pre-consultation beliefs are unobserved. The exercise therefore brackets the damage between random and tail-selected replacement.

The tail-selected rows show how quickly the signal can disappear when replacement targets the observations that carry it. At $\lambda = 0.1$, the interdecile spread is already down to 18.1 points, and the interdecile-spread alarm fires in only 10% of draws. Under random replacement in Table 12, the same alarm fires in 99% of draws at $\lambda = 0.1$. The share-above-10% alarm still fires in 97% of tail-selected draws at $\lambda = 0.1$, because enough upper-tail mass

remains. By $\lambda = 0.3$, however, the interdecile spread is only 12.2 points, the share above 10% is 0.2%, and both alarms are dead. When replacement is concentrated among extreme-belief respondents, it removes exactly the observations that contain the warning signal.

The remaining rows of Table 15 separate what is mechanical from what is specific to AI. The persona-prompted pool is wider than the baseline pool, but it still produces similar erosion. At $\lambda = 0.3$, contamination from the persona pool leaves a median of 5.1%, an interdecile spread of 19.8 points, and a share above 10% of 20.5%. The interdecile-spread alarm fires in 26% of draws, close to the 21% baseline power in Table 12. At $\lambda = 0.5$, the interdecile-spread alarm is again at zero and the share-above-10% alarm fires in only 7% of draws. Persona prompting makes the contaminant wider, but not wide enough to restore the human tail.

Finally, I turn to the generic-Gaussian rows. I draw contamination from a non-AI Gaussian distribution matched to the AI pool's mean and variance. At $\lambda = 0.3$, the median is 5.0%, the interdecile spread is 19.6 points, and the share above 10% is 20.4%. The interdecile-spread alarm fires in 21% of draws, essentially the same as in the model-pool baseline. Tail erasure is therefore a property of any low-variance contaminant inserted near the center of a dispersed distribution. It is not a unique property of large language models.

Two checks confirm that the simulation mechanics are not driving the result. The ordering of surviving and failing rules is unchanged without survey weights (reported in a forthcoming internet appendix). The erosion is also smooth in λ, rather than an artifact of the three reported points: the full grid in the same appendix shows the interdecile spread passing through each intermediate value.

In summary, what makes AI empirically relevant is the combination established earlier in the paper. Consumer models supply a distribution that is centralized enough to generate the effect, and respondents route expectation questions to those models at material rates. Expectation questions may be especially exposed. The routing pattern suggests that some respondents treat them like questions with a discoverable correct answer. That is precisely the situation in which a model's confident central estimate can displace private information. For expectations surveys, the lost information is often the object of interest: the gap between a household's belief and the model consensus is part of what the survey is meant to measure.

6 AI-Assisted Analysis of Survey Data

The preceding sections investigate how AI can influence survey data during the collection process. I now turn to the other end of the pipeline and present an exploratory benchmark of how the data are cleaned, summarized, and interpreted. This stage matters because

many survey statistics are not read directly from raw responses. Researchers often inspect outliers, decide which observations to exclude, summarize the distribution, and write an interpretation based on their findings. Furthermore, AI is increasingly used to label text, summarize responses, and automate measurement from survey data (Gilardi et al., 2023; Hobbs and Green, 2025; Ludwig et al., 2026).

If a survey is designed to measure disagreement and tail risk, then the observations that look unusual may be the observations that carry valuable information. An analyst that treats those observations as errors can create the same compression documented above, even when the underlying respondent data are clean. An analyst can also inherit compression if the input data have already been affected by AI-assisted responding. A third risk, emphasized in finance and machine learning, is look-ahead or contamination: a model may use information that would not have been available at the date being analyzed, or evaluation may be distorted by training-data and benchmark leakage (Glasserman and Lin, 2024; Didisheim et al., 2025; Z. Gao et al., 2025; Benhenda, 2026; Balloccu et al., 2024; Dong et al., 2024).

6.1 Benchmark Design

In this section, I describe how I give AI analysts tasks. Each task begins with a data packet constructed from SCE microdata. The packets come in three tranches. The clean high-inflation tranche contains genuine human observations from an episode in which inflation expectations are unanchored; because these observations come from a known high-inflation episode, I treat the tranche as unanchored by construction, and this clean-data truth is the ground truth against which I score the analyst's judgment. The contaminated high-inflation tranche starts from the same underlying data but replaces part of the distribution with AI answers, following the contamination exercises in Section 5. The normal-period placebo contains a lower-inflation distribution. The packets also vary in the information shown to the AI analyst. In some calls, the survey date is hidden. In others, the true date is shown. In a third condition, the packet is labeled with a false calm-period date. In a fourth condition, the analyst also receives later survey waves. These date and context conditions are included because an analyst may rely not only on the numbers in the packet but also on the surrounding historical frame. The later-context condition is especially useful as a positive control: an analyst shown subsequent waves should have more information for forecasting, but that information may also create look-ahead concerns.

The AI analyst performs four tasks. In the cleaning task, it receives respondent-level microdata and has discretion to exclude observations before summarizing and interpreting the packet. This task tests whether AI creates compression by cleaning away genuine tail

observations. In the unguided memo task, the analyst interprets the packet without cleaning discretion, so the effect of compressed input can be studied holding the data fixed. In the forecasting task, the analyst predicts next month's distributional statistics and is compared with simple statistical benchmarks. In the guided benchmark, the analyst receives computed distributional statistics and, in most calls, explicit instructions to examine the upper tail. The four tasks serve different purposes in the benchmark. The cleaning task asks whether compression can be created by the analyst. The unguided memo asks whether compressed input changes interpretation when the data themselves are fixed. The forecasting task asks whether the same weakness appears when the analyst must predict distributional statistics. The guided benchmark is included as a diagnostic. It gives the analyst the distributional statistics a researcher would normally compute and, in most calls, explicitly directs attention to the upper tail. This removes the simplest explanation for a mild assessment. If the AI analyst still downgrades a genuinely unanchored packet, the problem is not simply that the tail was hidden or overlooked. It is that the tail was recognized but given too little weight in the final judgment.

I score the AI analyst's judgment using two related outcomes. The broader outcome is *under-reaction*. It equals one when the clean-data truth of a packet is unanchored but the AI analyst assigns any less severe classification. This includes strict false reassurance, but it also includes intermediate downgrades. I use this broader outcome because a partial downgrade is already consequential: a researcher, central banker, or investor would read the episode as less severe than it is in the underlying data. The narrower outcome is *strict false reassurance*. It equals one when the AI analyst classifies an unanchored episode as anchored. I report it separately because it is the most severe mistake, but it is not the only economically relevant one.

The guided benchmark is a four-cell comparison. Along one dimension, the analyst sees either a clean high-inflation packet or a contaminated version in which some target-wave observations have been replaced with AI answers. Along the other dimension, the analyst sees either the packet as of the target date or the same packet plus later survey waves. Comparing clean and contaminated packets asks whether compressed input softens the AI analyst's assessment. Comparing point-in-time and later-context packets asks whether additional historical context changes the assessment. The remaining comparison asks whether the two features reinforce each other: whether contamination is especially damaging when the analyst also sees information from the future. Because the same analyst model is evaluated across cells, the exercise varies the information shown to the model rather than the model itself or its training cutoff (Z. Gao et al., 2025).

The cleaning exercise has 16 observations per tranche, and the remaining tasks reuse

a finite set of model-packet combinations. I therefore report cell means, within-task comparisons, and simple rates rather than treating the benchmark as a source of population estimates. The results should be read as evidence of mechanisms rather than population estimates of how all AI analysts behave.

6.2 Data Cleaning Discretion

This section reports findings from the cleaning task. The AI analyst receives respondent-level SCE microdata, decides whether to remove observations, and then summarizes the cleaned packet. Table 16 reports five outcomes: the exclusion rate, the share of excluded observations drawn from above the raw ninetieth percentile, and the changes in the ninetieth percentile, the interdecile spread, and the share of respondents expecting inflation above 10%. A negative change means that cleaning reduced the upper region of the distribution relative to the raw packet the analyst received.

The clean high-inflation tranche receives the most aggressive cleaning: 8.9% of observations are excluded, compared with 5.7% in the contaminated high-inflation tranche and 5.2% in the normal-period placebo. The excluded observations are also disproportionately drawn from the top of the distribution. In the clean high-inflation tranche, 39% of exclusions come from above the raw ninetieth percentile. The corresponding shares are 54% in the contaminated high-inflation tranche and 55% in the normal-period placebo. In the clean high-inflation tranche, cleaning lowers the ninetieth percentile by 3.15 points, the interdecile spread by 5.93 points, and the share above 10% by 18.66 percentage points. The last change is especially large: the share of respondents with inflation expectations above 10% falls from 24.7% before cleaning to 6.0% afterward. The contaminated high-inflation tranche also loses upper-tail mass, but less on the interdecile spread. The ninetieth percentile falls by 1.52 points, the interdecile spread by 3.02 points, and the share above 10% by 16.70 points. In the normal-period placebo, the corresponding changes are -2.72, -3.24, and -9.91 points. The tranche that contains the high-inflation signal is also the tranche in which cleaning most reduces the interdecile spread and the share of respondents with inflation expectations above 10%.

In 9 of the 16 clean high-inflation runs, the AI analyst's own exclusions move its assessment away from unanchored. These downgrades are not always strict false reassurance: some packets move from unanchored to an intermediate classification rather than all the way to anchored. The measurement problem is nevertheless already present. The analyst receives clean data from a genuinely unanchored episode, disproportionately removes observations in the upper tail, and then evaluates a cleaned distribution in which the warning signal has

been weakened.

6.3 Unguided Task

The unguided task asks whether compressed inputs change the analyst's interpretation when the analyst cannot alter the data. This is a different risk from the cleaning task. In the cleaning task, the analyst can create compression by excluding observations. Here, the packet is fixed. The question is whether an AI analyst, when asked to interpret the packet as given, assigns a less severe assessment when the same high-inflation episode has already been compressed by AI-style contamination. This distinction matters for survey users because many applications do not involve raw-data cleaning. A researcher, central banker, or investor may simply receive a prepared distribution and ask an analyst to summarize whether expectations look anchored.

I implement this task by giving the AI analyst high-inflation packets and asking for an unguided anchoring assessment. The AI analyst is not given cleaning discretion and is not instructed to delete observations. The packets vary along two dimensions. First, the input is either clean or contaminated. The clean packet contains genuine human observations from an unanchored high-inflation episode. The contaminated packet starts from the same episode but replaces part of the distribution with AI answers, following the contamination exercises above. Second, the historical frame varies across four date and context conditions: the survey date is masked, the true date is shown, a false calm-period date is shown, or later survey waves are supplied. These conditions ask whether the analyst relies only on the numbers in the packet or also on the surrounding historical frame.

Table 17 reports the resulting classifications. Each row is a date or context condition. The left block reports clean high-inflation packets, and the right block reports contaminated high-inflation packets. Within each block, the columns decompose the analyst's classification into three mutually exclusive outcomes. Unanchored is the correct classification because the benchmark design labels the high-inflation tranche as unanchored in every row. Intermediate is a downgrade: the analyst recognizes some concern but describes the episode as less severe than unanchored. Anchored is strict false reassurance. The final column reports the change in under-reaction from clean to contaminated packets within the same date or context condition.

The table shows that contamination mainly reduces correct unanchored classifications and increases intermediate downgrades. With the date masked, the analyst classifies 56% of clean high-inflation packets as unanchored, 44% as intermediate, and none as anchored. On contaminated packets, the corresponding shares are 6%, 88%, and 6%. Thus, masking the date produces the sharpest clean-contaminated contrast: contamination raises under-

reaction by 50 percentage points. The same pattern appears in the other information conditions, although the magnitudes are smaller. When the true date is shown, the correct unanchored classification falls from 38% on clean packets to 19% on contaminated packets, and under-reaction rises by 19 percentage points. Under the false calm-period label, it falls from 19% to 6%, and under-reaction rises by 13 points. With later context, it falls from 38% to 25%, and under-reaction again rises by 13 points.

The errors are usually not strict false reassurance. The anchored classification ranges from zero to 19%. The more common mistake is an intermediate downgrade. For example, in the contaminated masked-date condition, 88% of calls are classified as intermediate, while only 6% are classified as anchored. This distinction is important. The AI analyst rarely says that a genuinely unanchored episode is plainly anchored, but it often describes the episode as less severe than the clean-data truth warrants. For a user of the analysis, that is still consequential. An intermediate classification would make the episode appear less alarming even though the underlying clean packet comes from an unanchored high-inflation distribution.

The date and context conditions also suggest a framing effect. The false calm-period label weakens the assessment even for clean packets: the share classified as unanchored falls to 19%, compared with 56% when the date is masked and 38% when the true date or later context is shown. This pattern suggests that the analyst does not rely only on the distributional facts in the packet. Historical framing can soften the interpretation of the same high-inflation data.

In summary, the unguided task shows that compression matters even without cleaning discretion. The AI analyst does not need to remove observations to under-react. When compressed input is already present, the AI analyst is less likely to classify a genuinely unanchored packet as unanchored and more likely to assign an intermediate assessment. The main error is therefore not strict false reassurance, but attenuation: the episode is recognized as concerning, yet interpreted as less severe than the clean data imply.

6.4 Forecasting Inflation

The forecasting task asks whether the same center-tail asymmetry appears when the analyst must make a quantitative prediction rather than write an anchoring assessment. This is a useful check because forecast errors are less dependent on wording or classification choices. If the AI analyst is mainly weak in the upper tail, that weakness should appear when it forecasts next month's distributional statistics, especially the upper percentiles, the interdecile spread, and the shares above high-inflation thresholds.

I implement the task by asking the AI analyst to forecast the next month's distributional statistics of inflation expectations from the information available in the packet. I compare its forecasts with three simple benchmarks computed from the same information set: a random walk, a three-month moving average, and an AR(1). Table 18 reports mean absolute errors. Each row is a distributional statistic, and the columns compare the AI analyst with the three benchmarks. Figure 3 presents the same comparison visually. Because the statistics are measured on different scales, the table should be read within rows rather than across rows.

The AI analyst performs reasonably well for central and middle-of-the-distribution statistics. For the mean, its error is 0.51, close to the random walk's 0.49. For the median, it performs somewhat worse than the random walk, with an error of 0.53 compared with 0.43. For the interquartile range, however, it does better than the random walk, with an error of 0.23 compared with 0.34, and is close to the three-month moving average and AR(1). These rows show that the analyst can produce useful forecasts for some parts of the distribution. The forecasting problem is not a general inability to summarize the data.

The upper-tail rows are less favorable. The AI analyst's error for the ninetieth percentile is 2.39, compared with 2.11 for the random walk and 2.24 for the three-month moving average. The gap is larger for the ninety-fifth percentile: the AI analyst's error is 3.78, compared with 2.89 for both the random walk and the three-month moving average and 3.02 for the AR(1). The interdecile spread shows the same pattern, with an AI analyst error of 2.44 compared with 2.19, 2.05, and 2.04 for the three benchmarks. The threshold-share rows are more mixed. For the share above 5%, the AI analyst's error is 3.75, much larger than the random walk's 2.14 but close to the moving-average and AR(1) errors. For the share above 10%, the AI analyst's error is 2.17, close to all three benchmarks.

In summary, the forecasting task reinforces the measurement interpretation. The analyst can forecast some central statistics about as well as simple benchmarks, so the problem is not generic forecasting failure. The weakness is concentrated in the parts of the distribution that are most relevant for disagreement and unanchoring. For survey users, this means that AI-assisted analysis may preserve plausible central forecasts while being less reliable for the tail-risk statistics that carry the warning signal.

6.5 Why Does the AI Analyst Fail?

This section asks why the AI analyst fails. One possibility is recognition: the AI analyst may not notice that the upper tail is important. Another possibility is weighting: the analyst may notice the tail but still not give it enough importance when classifying anchoring. Table 19

distinguishes these possibilities by giving the analyst computed distributional statistics and, in most calls, explicit instructions to examine the upper region.

Of 342 attempted model-packet cells, 330 are scorable. In the calls that explicitly direct attention to the upper region, tail recognition happens in 99.6% of the calls. Strict false reassurance is also rare, occurring in only 4.7% of responses whose ground truth is unanchored. The analyst can identify the relevant feature when prompted. The remaining failure comes from the input and from how the tail is weighted in the final judgment. In the point-in-time condition, under-reaction is 0.431 for clean packets. It rises to 0.667 when $\lambda = 0.10$, to 0.804 when $\lambda = 0.30$, and to 1.000 when $\lambda = 0.50$. Future-exposed packets show the same qualitative pattern, though later context helps in some cells: under-reaction is 0.242 for clean packets, 0.722 at $\lambda = 0.10$, 0.510 at $\lambda = 0.30$, and 1.000 at $\lambda = 0.50$. Thus, even with explicit tail instructions, compressed inputs make the analyst less likely to classify the episode as unanchored.

The later-context cells do not show a clear compounding pattern. Relative to point-in-time packets, later context lowers under-reaction for clean packets and for packets with $\lambda = 0.30$, raises it for packets with $\lambda = 0.10$, and leaves it unchanged when $\lambda = 0.50$. In this small benchmark, later context therefore does not appear to amplify the contamination problem. This does not show that look-ahead is harmless in other historical or financial applications. It shows only that the output-stage distortion documented here is not driven by the look-ahead channel the benchmark was designed to test.

Taken together, the AI analyst tasks support a narrow conclusion. The models can compute and discuss central summaries, and they can recognize the upper tail when explicitly asked to look for it. The stronger risks are specific to measurement. First, compressed inputs lead the AI analyst to under-react. Second, ordinary cleaning discretion can remove upper-tail observations that reflect subjective high-inflation beliefs rather than obvious data errors, before interpretation begins. In this benchmark, discretionary cleaning is the clearest output-stage distortion; look-ahead is not.

7 Conclusion

Surveys are valuable because they aggregate heterogeneous private beliefs. Generative AI weakens the link between the sampled person and the recorded answer at three points: it can replace the respondent, assist the respondent, and analyze the resulting data. In an independently fielded SCE-style survey, 31.4% of 1,294 respondents leave evidence of AI involvement while 4.8% disclose it, and nearly all of the evidence points to AI-assisted humans rather than bots. Across fifteen consumer models, AI answers reproduce a plausible

central forecast while containing almost none of the disagreement or tail mass that households exhibit. Replacing even a modest share of human answers leaves the center-based alarms a reader checks first still firing, though at materially milder reported levels, whereas the dispersion and upper-tail statistics that the anchoring literature treats as early warnings quietly lose their power. An AI analyst given ordinary cleaning discretion produces the same compression from the output side, removing observed upper-tail SCE responses and downgrading its own assessment of unanchoring in the majority of the benchmark's high-inflation samples. At every stage, AI makes observations more alike without making them implausible.

My results mean that the origin of a response to a survey question is no longer a given even when bots are not detected in the survey. The current modernization debate asks whether surveys reach enough people, ask the right questions, and arrive in time (Warsh, 2026). This paper adds a prior question: whose information does a recorded answer contain? My detection approach runs on standard browser instrumentation, leaves the respondent's experience unchanged, and separates machine delivery from human-mediated assistance at the level of the individual question. By using my detection method, survey users could record and publish rates of detected AI involvement alongside the other variables they release.

For researchers and policymakers who consume expectations data, the results carry a specific caution. Measured disagreement can now fall for two reasons: because beliefs converge, or because a growing share of answers passes through the same few models. The two readings have opposite policy implications, and the statistics they most affect, the cross-sectional spread and the upper tail, are exactly the ones used to detect unanchoring. Any decline in measured disagreement in a web-based survey after the arrival of consumer AI cannot be properly differentiated until answer origin is measured.

References

Affonso, Felipe M. (2026). Brief commentary: A framework for detecting AI agents in online research. *Journal of Consumer Research*, advance access, article ucag006. doi:10.1093/jcr/ucag006.

Andre, Peter, Carlo Pizzinelli, Christopher Roth, and Johannes Wohlfart (2022). Subjective models of the macroeconomy: Evidence from experts and representative samples. *Review of Economic Studies* 89(6), 2958–2991. doi:10.1093/restud/rdac008.

Ang, Andrew, Geert Bekaert, and Min Wei (2007). Do macro variables, asset markets, or surveys forecast inflation better? *Journal of Monetary Economics* 54(4), 1163–1212. doi:10.1016/j.jmoneco.2006.04.006.

Argyle, Lisa P., Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis* 31(3), 337–351. doi:10.1017/pan.2023.2.

Armantier, Olivier, Giorgio Topa, Wilbert van der Klaauw, and Basit Zafar (2017). An overview of the Survey of Consumer Expectations. *Economic Policy Review* 23(2), 51–72.

Balloccu, Simone, Patrícia Schmidtová, Mateusz Lango, and Ondřej Dušek (2024). Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 67–93. doi:10.18653/v1/2024.eacl-long.5.

Benhenda, Mostapha (2026). Look-Ahead-Bench: A standardized benchmark of look-ahead bias in point-in-time LLMs for finance. arXiv:2601.13770 (accessed July 8, 2026). doi:10.48550/arXiv.2601.13770.

Berinsky, Adam J., Michele F. Margolis, and Michael W. Sances (2014). Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys. *American Journal of Political Science* 58(3), 739–753. doi:10.1111/ajps.12081.

Binder, Carola C. (2017). Measuring uncertainty based on rounding: New method and application to inflation expectations. *Journal of Monetary Economics* 90, 1–12. doi:10.1016/j.jmoneco.2017.06.001.

Bisbee, James, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson (2024). Synthetic replacements for human survey data? The perils of large language models. *Political Analysis* 32(4), 401–416. doi:10.1017/pan.2024.5.

Chernov, Mikhail and Philippe Mueller (2012). The term structure of inflation expectations. *Journal of Financial Economics* 106(2), 367–394. doi:10.1016/j.jfineco.2012.06.004.

Coibion, Olivier and Yuriy Gorodnichenko (2012). What can survey forecasts tell us about information rigidities? *Journal of Political Economy* 120(1), 116–159. doi:10.1086/665662.

Coibion, Olivier and Yuriy Gorodnichenko (2015). Information rigidity and the expectations formation process: A simple framework and new facts. *American Economic Review* 105(8), 2644–2678. doi:10.1257/aer.20110306.

D'Acunto, Francesco, Ulrike Malmendier, Juan Ospina, and Michael Weber (2021). Exposure to grocery prices and inflation expectations. *Journal of Political Economy* 129(5), 1615–1639. doi:10.1086/713192.

D'Acunto, Francesco, Daniel Hoang, and Michael Weber (2022). Managing households' expectations with unconventional policies. *Review of Financial Studies* 35(4), 1597–1642. doi:10.1093/rfs/hhab083.

D'Acunto, Francesco and Michael Weber (2024). Why survey-based subjective expectations are meaningful and important. *Annual Review of Economics* 16, 329–357. doi:10.1146/annurev-economics-091523-043659.

Didisheim, Antoine, Martina Fraschini, and Luciano Somoza (2025). AI's predictable memory in financial analysis. *Economics Letters* 256, 112602. doi:10.1016/j.econlet.2025.112602.

Dominguez-Olmedo, Ricardo, Moritz Hardt, and Celestine Mendler-Dünner (2024). Questioning the survey responses of large language models. In *Advances in Neural Information Processing Systems* 37, 45850–45878.

Dong, Yihong, Xue Jiang, Huanyu Liu, Zhi Jin, Bin Gu, Mengfei Yang, and Ge Li (2024). Generalization or memorization: Data contamination and trustworthy evaluation for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, 12039–12050. doi:10.18653/v1/2024.findings-acl.716.

Federal Reserve Bank of New York (2024). *Survey of Consumer Expectations: Core Module Public Questionnaire*, February 9, 2024. Available at `https:`

//www.newyorkfed.org/medialibrary/interactives/sce/sce/downloads/data/frbny-sce-survey-core-module-public-questionnaire.pdf (accessed July 8, 2026).

Frederick, Shane (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives* 19(4), 25–42. doi:10.1257/089533005775196732.

Gao, Yuan, Dokyun Lee, Gordon Burtch, and Sina Fazelpour (2025). Take caution in using LLMs as human surrogates. *Proceedings of the National Academy of Sciences* 122(24), e2501660122. doi:10.1073/pnas.2501660122.

Gao, Zhenyu, Wenxi Jiang, and Yutong Yan (2025). Detecting lookahead bias in LLM forecasts. arXiv:2512.23847, version 2, revised June 12, 2026 (accessed July 8, 2026). doi:10.48550/arXiv.2512.23847.

Giglio, Stefano, Matteo Maggiori, Johannes Stroebel, and Stephen Utkus (2021). Five facts about beliefs and portfolios. *American Economic Review* 111(5), 1481–1522. doi:10.1257/aer.20200243.

Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences* 120(30), e2305016120. doi:10.1073/pnas.2305016120.

Glasserman, Paul and Caden Lin (2024). Assessing look-ahead bias in stock return predictions generated by GPT sentiment analysis. *The Journal of Financial Data Science* 6(1), 25–42. doi:10.3905/jfds.2023.1.143.

Greenwood, Robin and Andrei Shleifer (2014). Expectations of returns and expected returns. *Review of Financial Studies* 27(3), 714–746. doi:10.1093/rfs/hht082.

Grishchenko, Olesya V. and Laura Wilcox (2024). Tale about inflation tails. Finance and Economics Discussion Series 2024-028, Board of Governors of the Federal Reserve System. doi:10.17016/FEDS.2024.028.

Gürkaynak, Refet S., Brian Sack, and Jonathan H. Wright (2010). The TIPS yield curve and inflation compensation. *American Economic Journal: Macroeconomics* 2(1), 70–92. doi:10.1257/mac.2.1.70.

Hobbs, William and Jon Green (2025). Categorizing topics versus inferring attitudes: A theory and method for analyzing open-ended survey responses. *Political Analysis* 33(3), 231–251. doi:10.1017/pan.2024.23.

Horton, John J., Apostolos Filippas, and Benjamin S. Manning (2023). Large language models as simulated economic agents: What can we learn from homo silicus? NBER Working Paper No. 31122, revised February 2026. doi:10.3386/w31122.

Leiner, D. J. (2026). SoSci Survey (Version 3.8.00) [computer software]. SoSci Survey GmbH. `https://www.soscisurvey.de`.

Lipkus, Isaac M., Greg Samsa, and Barbara K. Rimer (2001). General performance on a numeracy scale among highly educated samples. *Medical Decision Making* 21(1), 37–44. doi:10.1177/0272989X0102100105.

Ludwig, Jens, Sendhil Mullainathan, and Ashesh Rambachan (2026). Large language models: An applied econometric framework. *Annual Review of Economics* 18, in press; published online April 6, 2026. doi:10.1146/annurev-economics-120925-105620.

Lusardi, Annamaria and Olivia S. Mitchell (2014). The economic importance of financial literacy: Theory and evidence. *Journal of Economic Literature* 52(1), 5–44. doi:10.1257/jel.52.1.5.

Malmendier, Ulrike and Stefan Nagel (2016). Learning from inflation experiences. *Quarterly Journal of Economics* 131(1), 53–87. doi:10.1093/qje/qjv037.

Mankiw, N. Gregory, Ricardo Reis, and Justin Wolfers (2003). Disagreement about inflation expectations. *NBER Macroeconomics Annual* 18, 209–248. doi:10.1086/ma.18.3585256.

Manski, Charles F. (2004). Measuring expectations. *Econometrica* 72(5), 1329–1376. doi:10.1111/j.1468-0262.2004.00537.x.

Manski, Charles F. and Francesca Molinari (2010). Rounding probabilistic expectations in surveys. *Journal of Business & Economic Statistics* 28(2), 219–231. doi:10.1198/jbes.2009.08098.

Natoli, Filippo and Laura Sigalotti (2018). Tail co-movement in inflation expectations as an indicator of anchoring. *International Journal of Central Banking* 14(1), 35–71.

Oppenheimer, Daniel M., Tom Meyvis, and Nicolas Davidenko (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology* 45(4), 867–872. doi:10.1016/j.jesp.2009.03.009.

Peer, Eyal, David Rothschild, Andrew Gordon, Zak Evernden, and Ekaterina Damer (2022). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods* 54(4), 1643–1662. doi:10.3758/s13428-021-01694-3.

Reis, Ricardo (2021). Losing the inflation anchor. *Brookings Papers on Economic Activity* 2021(Fall), 307–379. doi:10.1353/eca.2022.0004.

Santurkar, Shibani, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto (2023). Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, PMLR 202, 29971–30004.

Stantcheva, Stefanie (2023). How to run surveys: A guide to creating your own identifying variation and revealing the invisible. *Annual Review of Economics* 15, 205–234. doi:10.1146/annurev-economics-091622-010157.

Tjuatja, Lindia, Valerie Chen, Tongshuang Wu, Ameet Talwalkar, and Graham Neubig (2024). Do LLMs exhibit human-like response biases? A case study in survey design. *Transactions of the Association for Computational Linguistics* 12, 1011–1026. doi:10.1162/tacl_a_00685.

Wang, Chaofan, Samuel Kernan Freire, Mo Zhang, Jing Wei, Jorge Goncalves, Vassilis Kostakos, Alessandro Bozzon, and Evangelos Niforatos (2025). Safeguarding crowdsourcing surveys from ChatGPT through prompt injection. *Proceedings of the ACM on Human-Computer Interaction* 9(7), Article CSCW322, 1–29. doi:10.1145/3757503.

Warsh, Kevin (2026). Transcript of Chairman Warsh's press conference, June 17, 2026. Board of Governors of the Federal Reserve System. Available at `https://www.federalreserve.gov/mediacenter/files/FOMCpresconf20260617.pdf` (accessed July 10, 2026).

Weber, Michael, Francesco D'Acunto, Yuriy Gorodnichenko, and Olivier Coibion (2022). The subjective inflation expectations of households and firms: Measurement, determinants, and implications. *Journal of Economic Perspectives* 36(3), 157–184. doi:10.1257/jep.36.3.157.

Westwood, Sean J. (2025). The potential existential threat of large language models to online survey research. *Proceedings of the National Academy of Sciences* 122(47), e2518075122. doi:10.1073/pnas.2518075122.

Zarifhonarvar, Ali (2026). Generating inflation expectations with large language models. *Journal of Monetary Economics* 157, 103859. doi:10.1016/j.jmoneco.2025.103859.

Table 1: Detection Signals and Their Inferential Content

Signal	Recorded event	Inference and use
Question routing	A complete or nearly complete question is copied and the survey subsequently loses focus before the respondent advances	*Supporting.* The question left the survey; neither its destination nor its effect on the submitted answer is observed.
Canary round trip	Hidden material attached only to copied text reappears in text entered into the survey	*Mechanism-specific.* An instruction-following model processed the copied question.
Machine entry	Synthetic keyboard events or a completed field with no matching keystroke, paste, or dictation record	*Mechanism-specific for delivery.* Software entered the answer; the origin of the underlying content remains separate.
Session operation and environment	Human interaction elsewhere in the session, or pervasive machine entry combined with automation, rendering, device, or network evidence	*Classification and corroboration.* Distinguishes software used inside a human-run session from autonomous completion.
Supplementary evidence	Model-like open-text patterns or respondent disclosure	*Supporting.* May reveal assistance that leaves no browser-side trace, but is vulnerable to alternative explanations or reporting error.

Notes: The table describes the inferential content of each signal, not a claim that any single signal identifies complete model authorship. Supporting signals either require corroboration or are assigned to the lower-confidence group. Environmental anomalies do not determine a classification on their own.

Table 2: The Prolific and SCE Samples

	Prolific (1)	SCE (2)
Gender		
Female	51.1	51.4
Male	48.9	48.6
Age		
Under 40	40.7	28.7
40 to 60	37.6	38.5
Over 60	21.7	32.8
Household income		
Under 50k	28.1	44.5
50k–100k	37.1	29.8
Over 100k	34.8	25.7
Education		
High school or less	13.4	36.2
Some college	15.7	32.0
College (incl. assoc.)	70.9	31.9
Race / ethnicity		
White	62.3	73.4
Hispanic	12.7	9.9
Black	12.1	8.9
Asian	5.6	2.5
Two or more races	5.2	3.5
Some other race	2.1	1.8
Marital status		
Married / partnered	57.7	62.6
Not married / partnered	42.3	37.4
Employment		
Working (full- or part-time)	77.0	62.2
Not working	23.0	37.8
Census region		
Northeast	19.3	17.3
Midwest	17.3	21.7
South	42.0	38.6
West	21.3	22.3
Homeownership		
Own	54.5	68.0
Rent	40.2	29.8
Other	5.3	2.2
Observations	1,294	23,886

Notes. Prolific data were collected on June 3 and 4, 2026. NY Fed SCE: survey-weighted respondent shares computed from the public SCE microdata (2013–2025), one record per respondent. Employment is collapsed to working (full- or part-time) versus not working; race/ethnicity orders categories by Prolific share. All SCE cells are computed from the microdata.

Table 3: Evidence of AI Involvement

Panel A. Detection-layer roster	Count	%
Question-text copy	319	24.7
Canary text present[a]	77	6.0
Direct field write	5	0.4
Synthetic keystrokes	13	1.0
AI-prose register	40	3.1
Machine fingerprint	8	0.6
Datacenter network origin	5	0.4
Self-report debrief	62	4.8
of which already detected	31	2.4
Auxiliary paradata (possibly-tier only)		
Window focus-loss	49	3.8
Alternative input	13	1.0
Panel B. Respondent classification		
Any positive evidence	406	31.4
AI-assisted human (multi-signal)	282	21.8
Possibly AI-assisted human (one-signal)	102	7.9
Agentic tool	17	1.3
Autonomous bot	5	0.4
Self-reported AI use[b]	62	4.8
No detected evidence	888	68.6
Total	1,294	100.0

Notes. $N = 1{,}294$. *Panel A* rows record individual detection layers; they overlap and are not additive. Question-text copy records a prose selection of at least twenty non-space characters; number-only copies are excluded. The auxiliary-paradata rows are single, uncorroborated signals with plausible non-AI explanations; they place a respondent only in the possibly-AI-assisted tier, never as strong evidence. *Panel B* partitions the sample into mutually exclusive tiers. "AI-assisted human (multi-signal)" requires a mechanism-specific trace or corroboration across mechanically distinct channels; "possibly AI-assisted human" contains promoted single-signal cases; "agentic tool" denotes machine delivery inside a human-navigated session; "autonomous bot" denotes completion without a human-interaction signature. "No detected evidence" is a residual category, not verified unaided completion. The reported shares are detected-evidence shares, not population estimates of true AI use. [a] The 77 canary round trips are nested within the question-text-copy channel and do not add a separate layer. [b] Self-report overlaps the classification and is not additive; all 62 self-reporters are in the 406-person positive-evidence set.

Table 4: Attention, Comprehension, and Perceptual Questions

	AI Involvement	None Detected	AI − None
	(1)	(2)	(3)
Lusardi–Mitchell Big Three (3 items)	87.4	77.9	9.5***
Compound interest	76.8	55.6	21.2***
Inflation / real return	91.9	86.8	5.0**
Risk diversification	93.3	91.2	2.1
Numeracy (4 items)	91.9	85.5	6.4***
Percentage discount	97.5	99.0	-1.4*
Lottery probability	90.4	82.2	8.2***
Disease risk	91.4	88.5	2.9
Infection risk	88.4	72.4	16.0***
Cognitive reflection (3 items)	69.5	48.3	21.3***
Bat and ball	58.1	38.1	20.1***
Widget machines	75.1	53.0	22.1***
Lily pads	75.4	53.7	21.7***
Visual illusions (4 items)	27.3	29.4	-2.1
Checker-shadow	19.5	23.1	-3.6
Müller-Lyer	17.7	22.6	-4.9*
Café wall (parallel)	53.4	45.8	7.6**
Reverse Müller-Lyer	18.7	26.0	-7.3***
Observations	406	888	

Notes. N = 1,294. Italic rows report the percent of a category's items answered correctly (equivalently, the mean score divided by the number of items); indented rows report the percent answering each item correctly. All differences are in percentage points (AI involvement minus none detected). Categories: the Lusardi–Mitchell financial-literacy items (compound interest, inflation, diversification; 0–3), the four numeracy items (0–4), the three Frederick (2005) cognitive-reflection items (0–3), and the classic visual illusions, scored as the illusion-resistant response (0–4). The compound-interest item is a reworded numeric variant of the canonical Lusardi–Mitchell question (here, $200 at a ten-percent annual rate compounding at each anniversary, asked as an open numeric response), deliberately modified so that respondents already familiar with the widely used Big Three battery would be less likely to recognize it; the inflation and diversification items use the canonical wording. "AI involvement" is any positive evidence of AI involvement (the four answer-origin tiers above no detected evidence, 406 of 1,294); "none detected" is the complement (888). All 62 respondents who self-report AI use fall in the AI-involvement group. Category-row stars are from a two-sample Welch t-test on the index score; item-row stars are from a two-proportion test. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 5: Correlates of Evidence of AI Involvement

	Dependent variable: Evidence of AI involvement				
	Baseline	Comprehension	Visual illusions	Duration	All three
	(1)	(2)	(3)	(4)	(5)
Age (per SD)	-0.075^{***}	-0.070^{***}	-0.076^{***}	-0.080^{***}	-0.074^{***}
	(0.013)	(0.012)	(0.013)	(0.012)	(0.012)
Female	-0.047^{*}	-0.022	-0.047^{*}	-0.059^{**}	-0.035
	(0.025)	(0.025)	(0.025)	(0.025)	(0.024)
College degree	0.100^{***}	0.055^{**}	0.100^{***}	0.092^{***}	0.050^{*}
	(0.027)	(0.028)	(0.027)	(0.027)	(0.027)
Income (per SD)	0.029^{**}	0.021	0.029^{**}	0.035^{***}	0.026^{**}
	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)
White	-0.050^{*}	-0.055^{**}	-0.049^{*}	-0.024	-0.031
	(0.027)	(0.026)	(0.027)	(0.026)	(0.026)
Working	0.069^{***}	0.063^{**}	0.067^{**}	0.063^{**}	0.055^{**}
	(0.026)	(0.026)	(0.026)	(0.026)	(0.026)
Comprehension (per SD)		0.106^{***}			0.098^{***}
		(0.014)			(0.014)
Visual-illusion (per SD)			-0.016		-0.024^{**}
			(0.013)		(0.012)
Survey duration (per SD)				0.075^{***}	0.061^{***}
				(0.012)	(0.012)
Observations	1,294	1,294	1,294	1,294	1,294
McFadden pseudo-R^2	0.052	0.092	0.053	0.074	0.109

Notes: Estimates are average marginal effects from logit models for the probability of any positive evidence (406 of 1,294 completers); heteroskedasticity-robust delta-method standard errors are in parentheses. Age, household income, survey duration, and the two cognitive composites are standardized (per SD). "Working" equals one for full- or part-time employment; the omitted race category is non-White. The comprehension score is the number correct on the ten cognitive items of Table 4 (the Lusardi–Mitchell Big Three, four numeracy items, and three cognitive-reflection items); the visual-illusion score is the number of illusion-resistant responses on the four illusion items. Columns (2)–(4) add one measure to the common demographic baseline; column (5) adds all three jointly. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

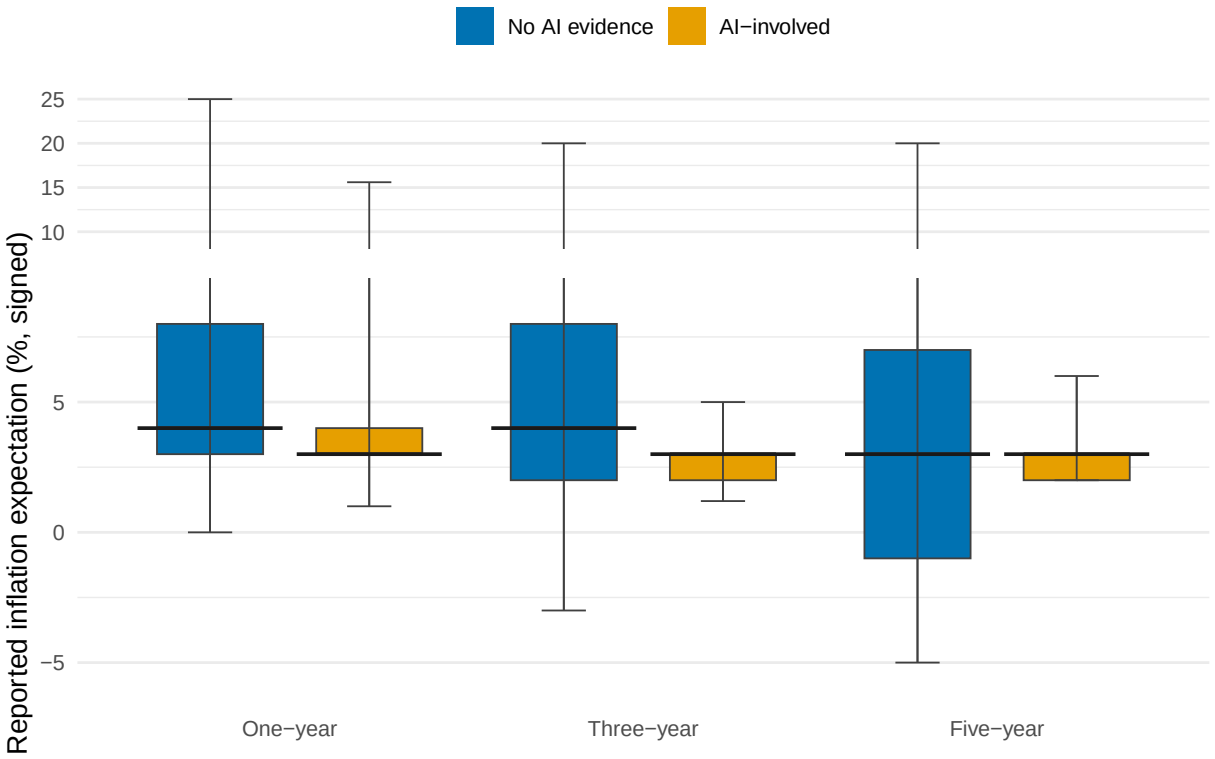


Figure 1: Inflation expectations by evidence of AI involvement on the inflation question, by horizon

Notes: Signed inflation point expectations at the one-, three-, and five-year horizons. Each pair compares the 133 *AI-involved* respondents with the 888 respondents with no detected evidence of AI involvement. Boxes span the interquartile range, the central line is the median, and the whiskers span the tenth and ninetieth percentiles; the vertical axis is split above 10 for readability. A two-sample Kolmogorov–Smirnov test on the full distributions rejects equality of the group distributions at every horizon ($p < 0.001$ for the one-, three-, and five-year point expectations).

Table 6: Inflation point expectations by inflation-question routing evidence and respondent classification

	Mean	Median	SD	P10	P90	N
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: One-year-ahead inflation expectation (%)						
No AI evidence	7.96	4.00	16.79	0.00	25.00	888
AI-involved	7.08	3.00	16.63	1.00	15.60	133
AI-assisted	7.42	3.00	17.67	2.00	20.00	111
Agentic tool	6.18	3.00	11.16	2.20	6.20	17
Autonomous bot	2.40	3.00	3.05	−0.80	5.20	5
Panel B: Three-year-ahead inflation expectation (%)						
No AI evidence	6.55	4.00	16.92	−3.00	20.00	888
AI-involved	3.20	3.00	8.15	1.20	5.00	133
AI-assisted	3.84	3.00	6.51	2.00	5.00	111
Agentic tool	−0.82	3.00	15.31	0.80	4.40	17
Autonomous bot	2.60	2.00	2.30	0.40	5.00	5
Panel C: Five-year-ahead inflation expectation (%)						
No AI evidence	5.64	3.00	18.31	−5.00	20.00	888
AI-involved	4.53	3.00	7.90	2.00	6.00	133
AI-assisted	4.37	3.00	7.58	2.00	6.00	111
Agentic tool	4.94	3.00	10.18	2.00	4.80	17
Autonomous bot	6.80	5.00	7.53	2.00	14.00	5

Notes: The table reports inflation expectations for 1,021 respondents: 133 in the AI-involved group and 888 with no detected AI evidence. The indented rows divide the AI-involved group into 111 AI-assisted respondents, 17 agentic-tool sessions, and 5 autonomous bots. Respondents with AI evidence only on other survey questions ($N = 273$) are excluded. P10 and P90 denote the 10th and 90th percentiles.

Table 7: Evidence of AI involvement on the inflation question and distributional outcomes

	One-year	Three-year	Five-year
	(1)	(2)	(3)
$P(y \leq 0)$	-1.0	-9.3^{***}	-20.1^{***}
	(2.5)	(2.7)	(2.6)
	[10.1]	[19.1]	[27.0]
$P(y > 5\%)$	-19.9^{***}	-28.9^{***}	-19.3^{***}
	(3.7)	(2.8)	(3.4)
	[33.7]	[32.4]	[29.3]
$P(y > 10\%)$	-8.4^{***}	-15.7^{***}	-10.8^{***}
	(3.1)	(2.4)	(2.8)
	[19.1]	[18.1]	[17.5]
$P(y = 3)$	20.9^{***}	25.2^{***}	26.6^{***}
	(4.5)	(4.6)	(4.5)
	[16.1]	[16.0]	[14.8]
$P(y = 4)$	4.0	-1.1	-3.6^{*}
	(4.1)	(3.0)	(2.1)
	[21.7]	[11.9]	[7.4]
$P(y = 3) - P(y = 4)$	16.9^{**}	26.3^{***}	30.2^{***}
	(7.3)	(6.1)	(5.3)
Demographic controls	Yes	Yes	Yes
Observations	1,021	1,021	1,021

Notes: Each cell in the upper block is the coefficient on an AI-involved indicator from a linear probability model with controls for age, sex, education, household income, race, and employment; coefficients and HC1 standard errors are reported in percentage points, and the bracketed third line gives the mean of the outcome in the no-evidence group (also in percentage points). The estimating sample contains 1,021 respondents: 133 classified as AI-involved (copied the inflation-question text on an inflation page, or completed the session through an agentic tool or autonomous bot) and 888 with no detected AI evidence; a further 273 respondents with AI evidence only elsewhere in the survey are excluded from this comparison. The outcomes are overlapping events rather than a partition of the support, so the table characterizes selected regions of the distribution rather than a complete accounting of where probability mass moves. The first outcome is a nonpositive signed inflation report, $y \leq 0$, and therefore includes both negative and zero values; it is not a pure deflation indicator. The final row reports the difference $\beta_{y=3} - \beta_{y=4}$, estimated as the coefficient on AI involvement in a single linear probability model for $\mathbf{1}[y = 3] - \mathbf{1}[y = 4]$, and tests whether the excess heaping is specific to 3 percent rather than common to both round values. Multiple-testing adjustments and the associational interpretation of these estimates are discussed in Section 3. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 8: New York Fed Survey of Consumer Expectations versus AI

	Public SCE			AI Models	
	Pre-AI 2013–22	Normal 2019	High 2022	Single question	Complete survey
Panel A: One-year-ahead inflation point estimate					
Mean	5.9	4.7	9.4	3.3	3.4
Median	3.0	3.0	7.0	3.2	3.5
Std. deviation	12.7	11.8	15.2	0.4	0.4
Interquartile range	5.0	3.0	6.0	0.5	0.5
10th percentile	1.0	1.0	−2.0	3.0	3.0
90th percentile	15.0	12.0	20.0	4.0	4.0
% at 3/3.5/4	22.4	23.7	9.4	76.0	87.5
Observations	147,311	15,765	7,548	437	449
Panel B: Three-year-ahead inflation point estimate					
Mean	5.3	4.7	4.2	2.9	2.8
Median	3.0	3.0	4.0	3.0	2.8
Std. deviation	13.1	12.1	16.0	0.3	0.3
Interquartile range	4.0	3.0	10.0	0.5	0.5
10th percentile	−2.0	−1.0	−7.0	2.5	2.5
90th percentile	15.0	12.0	15.0	3.5	3.0
% at 3/3.5/4	23.7	24.8	15.0	63.6	46.0
Observations	147,417	15,791	7,558	426	439

Notes. Inflation point estimates in percentage points. *Public SCE* columns: New York Fed Survey of Consumer Expectations public microdata: one-year inflation expectations are in Panel A, and three-year in Panel B. Pre-AI is the period from June 2013 to November 2022; the normal inflation expectation period is January 2019 to December 2019; and the high inflation expectation period is April 2022 to September 2022. *AI* columns: 15 large language models answering the verbatim SCE item, collected on 11–13 May 2026; the "Single question" call poses the item in isolation, the "Complete survey" call embeds it in the full survey instrument. The focal-value share "% at 3/3.5/4" is the share of a column's respondents reporting a point estimate landing exactly on 3.0, 3.5, or 4.0%. Humans report these three round values 9–25% of the time; the AI columns concentrate 46–88% of all answers on just these three values. Observation counts differ across panels because the one- and three-year point questions have different item non-response (SCE columns) and different valid-answer counts by horizon (AI columns).

Table 9: AI response distributional statistics by model and prompt format (one-year horizon)

Model	Format	N	Mean	Median	SD	P90−P10	% >5	% at 3/3.5/4
claude-haiku-4.5	A	40	2.7	2.8	0.2	0.3	0.0	0.0
claude-haiku-4.5	B	40	2.9	2.8	0.2	0.4	0.0	0.0
claude-opus-4.7	A	10	4.0	4.0	0.0	0.0	0.0	100.0
claude-opus-4.7	B	10	4.0	4.0	0.0	0.0	0.0	100.0
claude-sonnet-4.6	A	10	4.0	4.0	0.0	0.0	0.0	100.0
claude-sonnet-4.6	B	10	4.0	4.0	0.0	0.0	0.0	100.0
deepseek-r1-0528	A	10	4.1	4.0	0.8	2.0	0.0	60.0
deepseek-r1-0528	B	10	3.3	3.4	0.3	0.5	0.0	80.0
deepseek-v4-pro	A	40	3.5	3.5	0.3	0.6	0.0	92.5
deepseek-v4-pro	B	40	3.7	3.5	0.4	0.8	0.0	87.5
gemini-3-flash-preview	A	40	3.5	3.5	0.0	0.0	0.0	100.0
gemini-3-flash-preview	B	40	3.5	3.5	0.2	0.0	0.0	100.0
gemini-3.1-pro-preview	A	10	3.6	3.5	0.3	0.6	0.0	100.0
gemini-3.1-pro-preview	B	9	3.8	4.0	0.4	1.0	0.0	100.0
gpt-5.4-mini	A	40	3.0	3.0	0.2	0.1	0.0	100.0
gpt-5.4-mini	B	40	3.1	3.0	0.3	0.0	0.0	100.0
gpt-5.5	A	10	3.2	3.1	0.2	0.5	0.0	80.0
gpt-5.5	B	10	3.4	3.5	0.2	0.3	0.0	90.0
gpt-chat-latest	A	40	3.2	3.2	0.0	0.0	0.0	0.0
gpt-chat-latest	B	40	3.5	3.5	0.0	0.0	0.0	97.5
grok-4.3	A	40	3.5	3.5	0.4	1.0	0.0	77.5
grok-4.3	B	40	3.4	3.5	0.5	1.0	0.0	92.5
kimi-k2.6	A	27	3.8	4.0	0.4	0.7	0.0	96.3
kimi-k2.6	B	40	3.7	3.8	0.3	0.5	0.0	100.0
llama-4-maverick	A	40	3.0	3.0	0.0	0.0	0.0	100.0
llama-4-maverick	B	40	3.4	3.5	0.2	0.0	0.0	100.0
mistral-large-2512	A	40	3.5	3.5	0.0	0.0	0.0	100.0
mistral-large-2512	B	40	3.5	3.5	0.1	0.0	0.0	92.5
qwen3.6-plus	A	40	3.2	3.0	0.2	0.5	0.0	85.0
qwen3.6-plus	B	40	3.2	3.0	0.3	0.5	0.0	97.5
One-year pool under three weighting schemes:								
Pooled (call-weighted)	—	886	3.4	3.5	0.4	1.0	0.0	81.8
Equal-weight by model	—	886	3.5	3.5	0.4	0.5	0.0	84.3
Equal-weight by model×prompt	—	886	3.5	3.5	0.5	0.4	0.0	84.3

One-year inflation point estimate by model and prompt format. Format A is the single-question call, which asks the SCE item as a single stand-alone question; format B is the complete-survey call, which answers it within a full survey-completion conversation. Unweighted within each model-format cell. % at 3/3.5/4 = share of trials landing exactly on 3.0, 3.5, or 4.0%. The N column shows the call-count imbalance (40 per format for high-traffic models, 9–10 for the rest) that motivates the equal-weight rows. The three summary rows give the pooled, equal-weight-by-model, and equal-weight-by-model×prompt one-year distributional statistics; in the equal-weight rows the N column repeats the underlying call count for reference and Median/P90−P10/% are the equal-weight average of the per-group statistic. AI responses were collected via OpenRouter, 11–13 May 2026; no winsorization is applied because no AI answer falls outside the SCE range $[-5, 50]$. Two cells have $N < 40$: kimi-k2.6 format A ($N = 27$; 13 calls returned empty after exhausting the output-token budget on internal reasoning) and gemini-3.1-pro format B ($N = 9$; one trial left incomplete after a provider outage). These were not re-run and do not affect the concentration result.

Table 10: Persona-prompted AI inflation expectations

	By horizon (persona pool)			By age			By household income			By education	
	One-year	Three-year	Five-year	Young	Middle	Older	Low	Middle	High	College	Non-college
Mean	4.4	3.1	2.9	4.4	4.3	4.4	5.2	4.1	3.9	4.1	4.5
Median	3.8	3.0	2.7	3.8	3.8	3.5	4.0	3.5	3.5	3.5	3.8
Std. deviation	3.3	1.1	2.0	3.3	2.9	3.6	4.7	2.7	1.4	2.9	3.4
Interquartile range	1.0	0.4	0.5	1.0	1.0	1.3	1.2	0.5	0.8	0.8	1.0
10th percentile	3.0	2.4	2.3	3.0	3.0	3.0	3.2	3.0	3.0	3.0	3.2
90th percentile	6.0	4.0	4.0	6.0	5.6	6.2	8.0	5.0	5.0	5.0	6.0
% at 3/3.5/4	45.7	42.3	31.3	43.8	44.2	49.4	33.1	50.9	52.8	46.1	45.5
Observations	1,375	1,381	1,384	443	495	437	447	468	460	466	909

Notes. Persona-prompted (Mode D) AI inflation point estimates in percentage points (signed; deflation negative), the companion to Table 8. Six large language models each answer the full SCE instrument once for each of 231 demographic personas built from New York Fed SCE sampling-frame margins (see text). The first panel reports the persona pool at each horizon; the age, income, and education panels decompose the one-year pool (age: under 40 / 40–60 / over 60; income: under \$50k / \$50–100k / over \$100k; education: college versus high school or some college). No winsorization; responses with $|\text{value}| > 100$ percentage points are dropped, as in Table 8. "% at 3/3.5/4" is the share landing exactly on 3.0, 3.5, or 4.0%. Persona dispersion (one-year SD 3.3) sits between the near-degenerate canonical AI calls (SD ≈ 0.4) and the pre-AI human distribution (SD 12.7); the clearest demographic gradient is income (mean 5.2% for low income to 3.9% for high). Calls collected via OpenRouter, 11–13 May 2026.

Table 11: AI answers and the survey's uncertainty instrument

	N (1)	Largest bin (2)	$\geq$0.90 one bin (3)	Entropy (4)	Implied SD (5)	Disagreement (6)
Panel A: one-year horizon						
SCE respondents, 2013–2021	133,788	0.58	0.18	1.01	2.72	3.70
Field respondents, no evidence (2026)	888	0.50	0.06	1.28	3.88	3.39
AI, single-question call	382	0.40	0.00	1.54	3.53	0.85
AI, complete-survey call	361	0.42	0.00	1.51	3.38	0.75
Panel B: three-year horizon						
SCE respondents, 2013–2021	133,644	0.58	0.18	1.01	2.70	3.84
Field respondents, no evidence (2026)	888	0.50	0.08	1.27	3.72	3.97
AI, single-question call	348	0.38	0.00	1.61	3.75	0.78
AI, complete-survey call	379	0.47	0.00	1.44	3.02	0.93

Notes: The table summarizes responses to the SCE's ten-bin probabilistic inflation-expectations question, which elicits uncertainty within each respondent's forecast. For each response, the reported bin probabilities are normalized to sum to one; responses are retained when the raw probabilities sum to between 50 and 150. Unless otherwise indicated, the reported statistics are averages across valid responses. *Largest bin* is the probability assigned to the highest-probability bin. ≥ 0.90 *in one bin* is the share of responses assigning at least 90% probability to a single bin. *Entropy* is Shannon entropy, $-\sum_b p_b \ln p_b$, with a maximum of $\ln(10) \approx 2.30$. *Implied SD* is the within-response standard deviation computed using the SCE bin midpoints $(14, 10, 6, 3, 1, -1, -3, -6, -10, -14)$. *Disagreement* is the cross-sectional standard deviation of the corresponding midpoint-implied means across respondents or model calls. Thus, *implied SD* measures uncertainty within an individual forecast, whereas *disagreement* measures dispersion across forecasts. The latter is the density-question counterpart to point-forecast disagreement, but is computed from density-implied means rather than reported point forecasts. SCE statistics are unweighted and use public microdata from June 2013 through December 2021. Field statistics use the 888 respondents with no detected AI evidence described in Section 3. AI statistics use all valid density forecasts from the fifteen-model panel in Section 4, separately for the two canonical prompt formats.

Table 12: Replacement effects on monitored distributional statistics

Statistic	Statistic value				Monitor power		
	Clean	$\lambda = 0.1$	$\lambda = 0.3$	$\lambda = 0.5$	$\lambda = 0.1$	$\lambda = 0.3$	$\lambda = 0.5$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Median	8.0	7.1	5.0	3.5	1.00	1.00	1.00
Mean	10.7	10.0	8.5	7.0	1.00	1.00	0.89
SD	17.5	16.8	15.0	12.9	0.96	0.37	0.00
IQR	8.6	8.0	6.8	5.0	1.00	1.00	0.00
P90−P10	30.3	27.5	19.5	13.4	0.99	0.21	0.00
P90	28.1	25.9	21.0	16.1	0.98	0.26	0.00
P95	47.9	45.1	36.1	28.0	1.00	0.78	0.05
Share $> 5\%$	64.2	57.7	44.9	32.1	1.00	1.00	0.99
Share $> 8\%$	45.4	40.8	31.8	22.7	1.00	1.00	0.04
Share $> 10\%$	29.1	26.2	20.4	14.6	1.00	1.00	0.04
Skewness	1.2	1.3	1.7	2.2	0.01	0.17	0.58
Kurtosis	10.1	11.1	13.9	19.1	0.98	0.34	0.00

Notes: The clean benchmark is the 2022 Q2–Q3 high-inflation episode. Each cell is the mean over 3,000 monthly bootstrap samples of $N = 1,000$, using official SCE respondent weights. A random fraction λ of answers is replaced with draws from the fifteen-model pool, single-question and complete-survey calls equally weighted. The empirical model pool is centered near 3–4%, so this exercise combines a level shift with lower dispersion. Monitor power is the probability that a monitoring rule crosses its normal-period threshold. Thresholds are calibrated on 2014–2019 months under a disjoint calibrate-and-read split. Retained power does not mean the statistic level is unchanged: a rule can continue to fire while the statistic it reports is materially distorted.

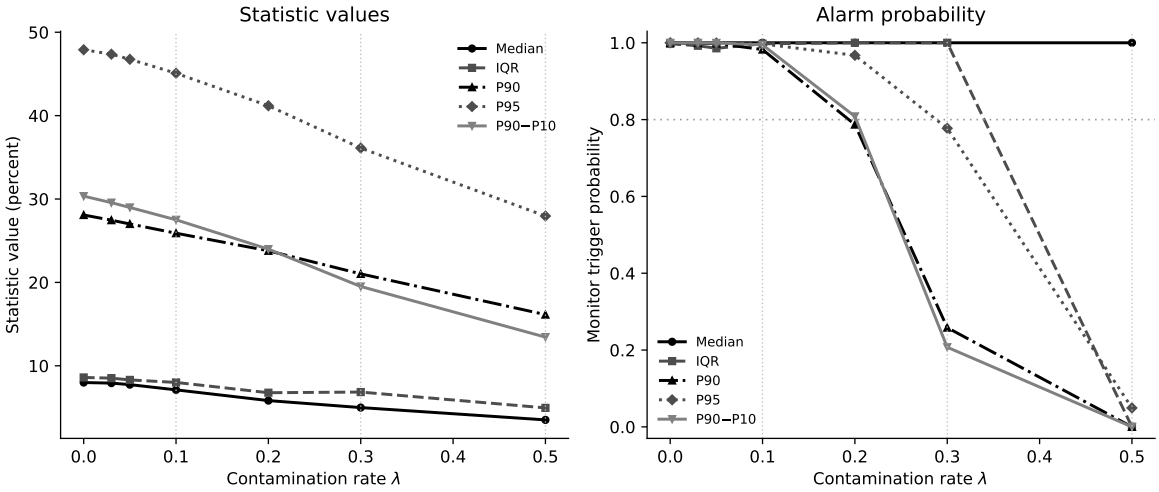


Figure 2: Tail erasure across the contamination calibration grid

Notes: Left panel: mean statistic levels (the median, interquartile range, P90, P95, and P90–P10 spread) of the one-year inflation distribution as a function of the AI contamination share λ over the 2022 Q2–Q3 high-inflation episode (3,000 weighted monthly bootstrap samples of $N = 1{,}000$, official SCE respondent weights). Right panel: the probability that each monitor's alarm fires; thresholds are calibrated on 2014–2019 normal-period months under a disjoint calibrate-and-read split, and the horizontal dashed line marks the 0.80 power-retention threshold. Vertical dashed markers locate the three calibration levels $\lambda = 0.1$, 0.3, and 0.5.

Table 13: Median-preserving recentering

Statistic	Statistic value				Monitor power		
	Clean	$\lambda = 0.1$	$\lambda = 0.3$	$\lambda = 0.5$	$\lambda = 0.1$	$\lambda = 0.3$	$\lambda = 0.5$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Median	8.0	8.0	8.0	8.0	1.00	1.00	1.00
Mean	10.7	10.4	9.9	9.3	1.00	1.00	1.00
SD	17.6	16.7	14.7	12.5	0.94	0.23	0.00
IQR	8.6	6.5	4.6	1.3	0.69	0.00	0.00
P90−P10	30.4	27.5	19.6	12.9	1.00	0.22	0.00
P90	28.1	26.0	21.1	16.2	0.98	0.27	0.00
P95	48.0	45.1	36.4	28.0	1.00	0.80	0.05
Share >5%	64.2	67.8	74.9	82.1	1.00	1.00	1.00
Share >8%	45.4	42.7	37.1	31.7	1.00	1.00	1.00
Share >10%	29.1	26.2	20.4	14.6	1.00	1.00	0.03

Notes: Identical design to Table 12, except that the fifteen-model answer pool is shifted by a constant (+4.5 percentage points) so that its median equals the weighted median of the clean 2022 Q2–Q3 episode; the median of the contaminated distribution is therefore fixed by construction. Because the recentered draws lie above the 5% cutoff, the reported share above 5% is distorted upward rather than downward, so that rule's survival is mechanical. Each cell is the mean over 3,000 weighted monthly bootstrap samples of $N = 1{,}000$; thresholds are calibrated on 2014–2019 normal-period months under a disjoint calibrate-and-read split. Small differences from Table 12 in the clean column (for example, an SD of 17.6 versus 17.5) reflect independent simulation draws.

Table 14: Effort-reduction blend across the retained human-belief weight

α	P90−P10 @.30 (1)	P95 @.30 (2)	Mon. P90−P10 (3)	Mon. P95 (4)	Mon. Median (5)	Mon. >5% (6)
0.0	19.5	36.2	0.20	0.78	1.00	1.00
0.3	22.5	36.2	0.59	0.82	1.00	1.00
0.5	25.7	38.9	0.96	0.96	1.00	1.00
0.7	27.7	42.0	1.00	1.00	1.00	1.00

Notes. Effort-reduction blend: the recorded answer is $\alpha \cdot \text{human} + (1-\alpha) \cdot \text{AI} + \mathcal{N}(0, 0.3)$; α=0 is pure replacement and higher α retains more of the human answer. Columns evaluated at λ=0.30. 'Mon.' is the monitor power.

Table 15: Selective replacement, persona, and generic-shrinkage robustness

Spec	λ (1)	Median (2)	P90−P10 (3)	>10% (4)	Mon. P90−P10 (5)	Mon. >10% (6)
Tail-selected	0.10	6.2	18.1	19.1	0.10	0.97
Tail-selected	0.30	4.0	12.2	0.2	0.00	0.00
Tail-selected	0.50	3.5	8.5	0.0	0.00	0.00
Persona pool	0.10	7.3	27.7	26.3	0.99	1.00
Persona pool	0.30	5.1	19.8	20.5	0.26	1.00
Persona pool	0.50	4.4	13.9	14.8	0.00	0.07
Generic Gaussian	0.10	7.1	27.6	26.2	1.00	1.00
Generic Gaussian	0.30	5.0	19.6	20.4	0.21	1.00
Generic Gaussian	0.50	3.7	13.6	14.5	0.00	0.04

Notes. Robustness of selective tail erasure. *Tail-selected*: the λ share of responses with the largest answers is replaced (the worst case for the tail monitors). *Persona pool*: contamination drawn from the persona-prompted AI distribution of Table 10. *Generic Gaussian*: contamination drawn from a non-AI Gaussian matched to the AI pool's mean and standard deviation.

Table 16: AI analyst cleaning discretion

		Exclusions		Upper-tail Δ (cleaned − raw)		
	Runs	Rate	P90	P90	P90−P10	>10%
		(1)	(2)	(3)	(4)	(5)
Clean high-inflation	16	0.089	0.39	−3.15	−5.93	−18.66
Contaminated high-inflation	16	0.057	0.54	−1.52	−3.02	−16.70
Normal-period placebo	16	0.052	0.55	−2.72	−3.24	−9.91

Notes. Each observation is one model-packet cleaning run; each tranche contains 16 runs (48 in total). Every run removes at least one observation, so the any-exclusion share is 1.00 in all three tranches. The exclusion rate (1) is the share of shown observations the analyst deletes, and the tail share of exclusions (2) is the fraction of deleted observations drawn from above the raw ninetieth percentile (a region that holds only one tenth of the data). Columns (3)–(5) report *erosion*: the analyst's final statistic minus the mechanical statistic of the raw packet, so a negative value means cleaning shrinks the upper region. In 9 of the 16 clean high-inflation runs (56%), cleaning downgrades the analyst's assessment of unanchoring; this is not necessarily a change all the way to "anchored."

Table 17: Unguided anchoring memo: classification of unanchored packets

	Clean high-inflation				Contaminated high-inflation				
Condition	N	Unanchored	Interm.	Anchored	N	Unanchored	Interm.	Anchored	Δ under-react
Masked date	16	0.56	0.44	0.00	16	0.06	0.88	0.06	0.50
True date shown	16	0.38	0.50	0.12	16	0.19	0.81	0.00	0.19
False calm date	16	0.19	0.62	0.19	16	0.06	0.88	0.06	0.13
Later context	16	0.38	0.50	0.12	20	0.25	0.60	0.15	0.13

Notes: The clean-data truth is unanchored in every row. "Unanchored" is the share of calls in which the analyst correctly classifies the packet as unanchored; "Interm." (intermediate) is the share downgraded to a less severe but not anchored classification; "Anchored" is strict false reassurance. Δ under-react is the contaminated-packet under-reaction rate minus the clean-packet under-reaction rate within the same date/context condition. The recognition diagnostics from the earlier version are omitted because they contain little cross-cell variation: tail recognition equals one in all but one cell, and center-only readings never occur.

Table 18: Next-month forecast errors: AI analyst versus benchmarks

Statistic	N (1)	AI (2)	Random walk (3)	3m MA (4)	AR(1) (5)
Mean	139	0.51	0.49	0.66	0.55
Median	139	0.53	0.43	0.67	0.59
IQR	139	0.23	0.34	0.20	0.21
P90	139	2.39	2.11	2.24	2.59
P95	139	3.78	2.89	2.89	3.02
P90−P10	139	2.44	2.19	2.05	2.04
Share $>5\%$	139	3.75	2.14	3.86	3.62
Share $>10\%$	139	2.17	2.01	2.26	2.33

Notes. AbsErr = |forecast − realized| for the next-month statistic, averaged across packets (N per statistic in column 1). The three benchmark forecasts, a random walk, a three-month moving average (3m MA), and an AR(1), are computed from the same information set (history through the current month). Because the statistics are on different scales, the comparison is most informative within each row.

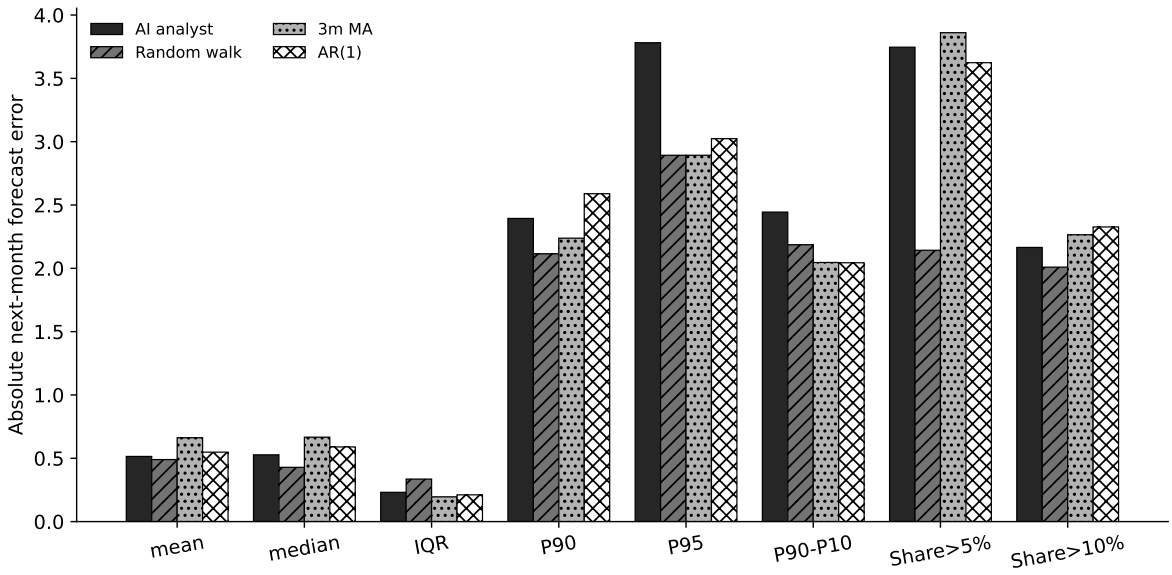


Figure 3: AI analyst forecast errors by distributional statistic: center versus tail

Notes: Absolute next-month forecast errors by distributional statistic for the AI analyst and three benchmarks (random walk, three-month moving average, AR(1)), computed from the same information set.

Table 19: AI-analyst under-reaction in the guided benchmark, by condition

Condition	All UR	Frontier UR	Open UR	Strict FR	Tail miscal.
Point-in-time, clean (λ=0)	0.431 (51)	0.278 (18)	0.515 (33)	0.000	0.431
Point-in-time, λ=0.10	0.667 (18)	0.500 (6)	0.750 (12)	0.000	0.667
Point-in-time, λ=0.30	0.804 (51)	0.833 (18)	0.788 (33)	0.098	0.765
Point-in-time, λ=0.50	1.000 (18)	1.000 (6)	1.000 (12)	0.056	1.000
Future-exposed, clean (λ=0)	0.242 (33)	0.167 (12)	0.286 (21)	0.030	0.152
Future-exposed, λ=0.10	0.722 (18)	0.667 (6)	0.750 (12)	0.000	0.722
Future-exposed, λ=0.30	0.510 (51)	0.500 (18)	0.515 (33)	0.039	0.431
Future-exposed, λ=0.50	1.000 (18)	1.000 (6)	1.000 (12)	0.167	1.000

Notes: UR = under-reaction, the share of genuinely unanchored packets that the analyst does not classify as unanchored (n in parentheses). Strict FR = the share classified "anchored"; Tail miscal. = the share in which the analyst recognizes the upper tail yet does not classify the packet as unanchored. "All" pools the six analyst models; "Frontier" contains Claude Opus and GPT-5.5; "Open" contains the four open-weight models. A random fraction λ of the target wave's answers is replaced with model draws; historical and future context remains clean. Of 342 attempted model-packet cells, 12 were unscorable after retries, leaving 330 scored responses: the 258 with an unanchored ground truth are shown here; the remaining 72 have an anchored ground truth. Cells pool the benchmark's structured and unstructured prompt variants. Tail recognition is 99.6% on the 270 calls whose instructions explicitly direct attention to the upper region (71.7% on the 60 unstructured calls); strict false reassurance occurs in 12 of the 258 unanchored-truth calls (4.7%).