Artificial Consensus: AI and Survey Measurement*

Thomas Cauthorn

First draft: July 2026
This draft: September 2026

Abstract

Central banks, governments, firms, and investors use surveys to measure hard-to-observe economic conditions. Surveys inform economic policy by revealing beliefs that shape household spending and wage demands. I show that AI use during surveys can lead to understating household inflation expectations. I find evidence of AI use among 23.3% of 1,294 respondents in an online survey. Both respondents who use AI and who do not have plausible median inflation expectations of 3% but the tenth-to-ninetieth-percentile range of their answers span 3 and 25 percentage points, respectively. In simulations using the New York Fed's Survey of Consumer Expectations, replacing 30% of household answers with AI responses weakens warnings of elevated inflation expectations because AI lowers reported expectations and reduces disagreement. Policymakers could mistake artificial consensus for convergence in beliefs and fail to act in times of growing inflation expectations.

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 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 this link without necessarily making the response look suspicious. AI can complete a survey directly or help a human respondent formulate an answer. The survey may still appear to contain many independent observations even when some answers are from the same algorithmic source.

I study whether this difficult-to-observe change in answer origin matters in the setting of household inflation expectations. These expectations are associated with consumption and saving and may influence wage bargaining and aggregate demand (Weber et al., 2022; D'Acunto et al., 2022). The New York Fed's Survey of Consumer Expectations (SCE) helps researchers and policymakers assess whether expectations remain anchored by revealing how beliefs differ across households. If AI-generated answers cluster around the same plausible value, the survey may show less disagreement even when households' underlying beliefs remain dispersed. What appears to be a convergence of beliefs may instead be artificial consensus.

The distinction between genuine and artificial consensus matters for policy. If policymakers interpret a median low inflation expectation and shrinking disagreement as evidence of anchoring, they may understate the severity or persistence of elevated inflation expectations if a portion of the expectations is AI-generated. The issue is especially timely because the Federal Reserve is reconsidering the quality of the data it uses. At the June 17, 2026 FOMC press conference, Chair Kevin Warsh criticized current survey methods, focusing on whether surveys reach enough people, ask relevant questions, and arrive in time (Warsh, 2026). Generative AI raises a more basic question about the answers themselves: even when a survey reaches the right people, asks the right questions, and arrives on time, do its recorded responses still reflect independent household beliefs?

To examine this question, I combine three sources of evidence. First, browser-side paradata[1] from an online survey containing SCE questions provide evidence about how AI is used to answer surveys. Second, surveying fifteen AI models provides me with the distribution of answers respondents receive when they consult AI. Third, public SCE microdata then allow me to test how inserting those AI-generated answers changes the statistics used

[1]Paradata are data about the process of answering a survey, captured as a by-product of administering it (Couper, 1998; Kreuter, 2013). In this study the paradata are browser-recorded events such as copying question text, leaving the survey page, and returning to the survey page to paste new text into an answer field. Section 2.3 provides more detail on my methods.

to assess anchoring. My analysis asks whether AI-shaped responses enter surveys, whether they survive conventional quality screens while changing the economic information contained in responses, and whether disagreement is compressed by AI responses.

Respondents commonly consult AI chatbots during my survey, while only a small minority of respondents use in-browser assistants or tools that enter answers autonomously. To make this finding possible, I develop a browser-side measure of answer origin that reconstructs how a response was produced rather than judging whether the answer itself looks unusual. The survey records, inter alia, when a respondent copies a question and leaves the survey window, whether hidden text added only to the copied question later reappears in the submitted answer, how the answer enters the response field, and whether the surrounding session contains ordinary human navigation and keystrokes. Among 1,294 respondents, 23.3% leave evidence of AI involvement, although only 4.3% disclose it. 21.5% of the survey's respondents consult a chatbot to help form their answers, while 1.2% of the respondents provide answers via autonomous AI and 0.6% use an in-browser assistant.[2] These respondents are not careless participants: they outperform the control group on financial-literacy, numeracy, and cognitive-reflection screens, so conventional quality filters would not screen them out. The main problem for survey measurement is therefore respondents clicking through a survey while asking AI for help with certain questions.

AI involvement on the inflation expectation questions is associated with artificial consensus: the median inflation expectation is plausible but expectation heterogeneity shrinks. Respondents who use AI to answer the inflation questions and those who do not both report a median five-year inflation expectation of 3%. However, the tenth-to-ninetieth-percentile range is 3 percentage points for AI users and 25 percentage points for non-AI users. The main difference between AI and non-AI responses lies not in the median forecast but in the cross-sectional information around it.

I then turn to model benchmarking to investigate how AI sources can generate this pattern. Across fifteen consumer AI models, one-year inflation expectations have a standard deviation of 0.4 percentage points, compared to 12.7 in the pre-AI SCE microdata. Furthermore, 81.8% of model answers equal 3.0, 3.5, or 4.0%, with none exceeding 5%. The models therefore supply a plausible inflation expectation, but not the heterogeneous household distribution contained in the SCE.

This compression can change how policymakers perceive the data. In a stress test that replaces 30% of SCE responses from a high-inflation period with model answers, center-based rules continue to flag elevated inflation but portray a materially milder episode, while rules based on dispersion and the upper tail flag the same underlying episode in only 2–48%

[2]The Claude Chrome Extension is a good example of an in-browser assistant.

of simulations. At 50% replacement, the typical simulated wave passes every monitoring rule that a normal inflation month would pass. These simulations show how AI-shaped responses can make expectations appear closer to the anchor without making the survey answers appear implausible.

This paper contributes to three 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. However, distinguishing respondents who use AI to help answer survey questions from autonomous AI respondents is a different problem. Wang et al. (2025) use prompt-injection traps intended to provoke a model into revealing itself. While I do use prompt injections as one source, I extend the literature by also gathering paradata. I use a multi-layered detection method that can be used for almost any survey question because it infers from the response process how each answer was produced, rather than relying only on the answer itself. My method can distinguish both autonomous and semi-autonomous AI from human respondents using AI to help form their answers.

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) find that these subjective expectations predict economically important decisions while Greenwood and Shleifer (2014); Giglio et al. (2021) find that expectations elicited from investors help explain portfolios and prices. 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. Wu et al. (2025) use date-restricted prompting to replicate and retrospectively extend a human expectations survey, and Anesti et al. (2026) test whether a language model can form inflation perceptions mimicking respondents to the Bank of England's household survey. Santurkar et al. (2023) find that models reproduce plausible averages but struggle with human heterogeneity and are unrepresentative of large population segments. Furthermore, their survey responses can be sensitive to prompt wording, answer ordering, and labeling (Gao et al., 2025; Dominguez-Olmedo et al., 2024), and they fail to reproduce documented human response biases (Tjuatja et al., 2024). Closest to my inflation-expectations application, Zarifhonarvar (2026) finds that model architecture materially affects generated expectations, with proprietary mod-

els generally exhibiting less disagreement than open-source alternatives. These papers ask whether LLMs can substitute for surveys; I ask what happens when LLMs enter a survey unexpectedly without any demographic conditioning.

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 the persistence of forecast errors after structural shocks reveals information rigidities. Andre et al. (2022) trace part of this heterogeneity to the different subjective models through which households reason about macroeconomic shocks. More recent work treats changes in higher-order moments and the upper tail as an early indication that inflation expectations are becoming unanchored in household surveys (Reis, 2021) and in market-based and professional-forecaster distributions (Natoli and Sigalotti, 2018; 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 information-carrying dispersion and upper-tail statistics 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.

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, and Section 6 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 develops an answer-origin measure that classifies respondents by the strongest completion mechanism observed.

2.1 The Challenge

Imagine a researcher reading a survey to learn about respondents' expectations regarding economic conditions. A respondent forecasts inflation at 4% and explains that grocery and housing costs still feel as though they are rising. Both reasons are plausible. Neither, on its own, reveals whether the respondent answered unaided, consulted a chatbot, or was

replaced by autonomous AI. Conventional survey-quality checks ask whether respondents paid attention, followed instructions, and supplied coherent answers.[3]

AI creates two nascent respondent-level challenges. The first is substitution. Autonomous AI can complete the survey without a person in the loop, or an agentic tool can place one or more answers inside a session that a person otherwise navigates. Current autonomous systems can produce plausible responses and pass most conventional attention and quality checks (Westwood, 2025; Affonso, 2026). The second challenge is assistance. A person completes the survey but delegates part of one or more answers to an outside AI tool. This is harder to detect because the account, browser, mouse movements, and typing genuinely belong to a human respondent.

I separate these channels using four mutually exclusive groups: autonomous AI, in-browser assistant, AI chatbot, and no detected AI use. Software may place text directly into a response field; an assistant may operate inside the browser; or a respondent may route a question outside the survey and then return to answer it. For AI chatbot responses, I can observe the answer origin but cannot necessarily determine how much of the final answer came from the respondent. A simple framework shows why even partial answer replacement from AI can move reported answers toward a common model signal.

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 while obscur-

[3] I use such measures as benchmarks for conventional respondent quality rather than as measures of answer origin. I include an instructed-response attention check (Berinsky et al., 2014; Oppenheimer et al., 2009); financial-literacy, numeracy, and cognitive-reflection items (Lusardi and Mitchell, 2014; Lipkus et al., 2001; Frederick, 2005); and a set of visual-cognitive tasks based on Affonso (2026). Recent work suggests that visual-cognitive tasks may distinguish some autonomous agents from humans (Affonso, 2026), but such items need not detect a human respondent who consults AI only for selected questions.

ing meaningful heterogeneity. Note that Equation (1) has the same form as the updating equations used in randomized information-provision experiments, where posterior expectations combine a prior with a supplied signal (Cavallo et al., 2017; Coibion et al., 2022); such treatments can sharply reduce cross-sectional disagreement (Armantier et al., 2016; Coibion et al., 2022). In those studies, the researcher controls and observes the signal, randomizes its provision, and can test whether movement toward it reflects learning rather than a temporary pull toward the reported number (Haaland et al., 2023). Here, consultation is chosen by the respondent and the AI signal is unobserved, so α_i is a descriptive retention weight rather than an experimentally identified learning rate. A tighter cross-section therefore need not reflect convergence in underlying beliefs; it can be produced by shared algorithmic input.

2.3 Measuring Answer Origin

I construct an answer-origin measure from observed traces of the survey completion process rather than from whether the submitted response itself looks unusual.[4] In the next section, I present how these data can be grouped to detect the four types of respondents outlined in the previous section.

2.3.1 Survey Setup

The survey is hosted by SoSci Survey, where the questionnaire and answers are stored. Amazon Web Services records events generated during the completion process and links them to the respondent by survey page, question, answer field, and time. The instrument records:

- clipboard events, including copies matched to the survey's question text and pastes by target field, text length, and time;
- keystroke timing, key identities, trust, autorepeat;
- focus, blur, and visibility changes;
- the input source underlying each field's final value;
- extension and iframe signatures, window geometry, standard automation probes;
- browser, device, rendering, and network characteristics.

I explain the method and event-processing rules in detail in an internet appendix.

I use evidence about how a respondent interacted with the survey and how answers entered the answer field, not whether the answer is economically plausible or statistically unusual. For a small set of questions I also use prompt injection. I review each of these

[4] An internet appendix will provide the complete catalog, including intermediate components, context-only measures, and false-positive guards.

answers by hand to determine whether the prompt injection was successful. It is important to note what I cannot observe. I do not record the respondent's screen or activity inside another application.[5] Furthermore, my method cannot observe assistance obtained on a second device.

2.3.2 The Four Respondent Groups

This section assigns respondents to groups starting with autonomous AI and ending with a group of respondents for whom I detect no evidence of AI use.

Autonomous AI. First, I look at how an answer enters the answer field. Each answer field leaves an input trail. When a respondent types, pastes, or dictates an answer, the browser records those actions. I compare that history with the submitted value and look for four signs that software delivered the answer instead. A *direct field write* is a completed field with no matching keystroke, paste, or dictation event. *Synthetic input* consists of input events that the browser does not label as trusted physical input. *Metronomic typing* is a long sequence of browser-trusted keystrokes delivered with almost no pauses and at a pace outside a normal human typing distribution. *Machine-cadence typing* is sustained typing below the floor observed in large-scale human typing data.[6] Together, this evidence points to software, rather than a respondent, placing the text in the answer field.

Second, I look at how the respondent moved through the survey. Software filling selected fields while the respondent otherwise navigates normally, moves the mouse, and types ordinary, key-identified keystrokes is consistent with a human-run session using an agentic tool. Software entering answers throughout the survey, little ordinary human interaction, and corroborating browser or network evidence are instead consistent with autonomous completion. The environmental record includes browser-reported automation control; graphics-rendering characteristics; and whether the connection comes from a residential network or a commercial hosting provider. These features can corroborate the delivery and interaction evidence, but they never classify a respondent on their own.[7]

[5]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.

[6]The metronomic rule requires at least 100 trusted keydowns, fewer than 5% autorepeat events, at least ten distinct key identities, no inter-key gap above 500 ms, and a 90th-percentile inter-key interval at or below 35 ms. The broader cadence rule requires, after removing autorepeat, a median inter-key interval below 60 ms over at least thirty clean keystrokes and on at least two typed pages, below the roughly 60 ms lower bound of mean inter-key intervals observed in the large-scale typing data of Dhakal et al. (2018). I am careful to exclude plausible benign explanations involving autofill and input-method editors; a back–forward-cache restore flag was also checked and is zero on every respondent page.

[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

In-browser assistant. A respondent enters this group when the browser carries a named AI-extension signature or when a docked side panel with assistant-compatible dimensions is corroborated by an independent consultation or delivery trace. An inbound paste without a preceding copy can be informative when it occurs alongside a corroborated side panel. A side-panel assistant can read the page directly, so the question need not first leave the survey through the clipboard. In that case, the agentic tool would either enter the answer autonomously or provide the respondent with an answer, which could then be pasted into the answer field. I can observe both.

AI chatbot. Copying is not indicative of AI use in and of itself. What matters is what leaves the survey. On, for example, the compound-interest item, copying only "200" could simply be a shortcut to a calculator. Copying the full question together with the amount, interest rate, and time horizon is different. This copy takes the entire problem out of the survey in a form that can be submitted directly to a chatbot. An unfamiliar word or short phrase may be copied for a dictionary or translation lookup. I use the clipboard record to exclude benign cases. For every copy event, I compare the selected text with the question shown on the page and classify it as a complete or nearly complete question, a number-only selection, a short word or phrase, or the respondent's own draft answer. Number-only selections, short fragments, and copies of the respondent's own answer do not count as qualifying question copies.

I also use clipboard canaries (prompt injections) to provide a more specific signal of AI having been used to form an answer. I add hidden text to the copied version of three survey questions. For two questions, the copied text contains a base64-encoded instruction directing an instruction-following model to include a designated marker phrase in its response. The encoded string resembles an internal survey tag, reducing the likelihood it would be caught if a respondent inspects the copied text after pasting it into a chatbot. For the third question, the copied text contains a short planted phrase designed to ask rather than instruct AI to mention something in the response. It is important to note that, while these canaries can be useful, many models are trained to specifically ignore them for safety reasons, so they do not always work. Based on the data I can observe, there are several question-routing types I classify as AI chatbot consultation.

The first type of AI chatbot consultation evidence I introduce is what I call a question round trip: the respondent copies a complete or nearly complete question, leaves the survey page, and returns with an answer that is pasted into the response field. However, certain respondents might follow a slightly different path than my question round trip assumes. If

connection can reflect a VPN or corporate network.

a respondent is using split screens instead of multiple tabs in one screen, switching from the survey to a chatbot does not necessarily produce an away-and-return event. Respondents could also copy a question, paste it into a chatbot, and then type the answer instead of pasting the answer into the survey. Respondents could also type a question into a chatbot and then paste the answer back into the survey. I therefore also count the following as evidence of AI chatbot consultation: no text copy followed by leaving the survey page with a paste after coming back; a question copy and an away event that is not followed by a paste once returning to the survey page; a copy and paste without an away event; a paste with no observed away-and-return event.[8] The final type of evidence of AI chatbot consultation is what I call a canary marker: the answer to a question contains the designated prompt-injection marker or planted phrase from one of the previously mentioned three questions. This can either happen in conjunction with one of the previous types of evidence or in isolation.

No AI usage. Respondents who do not fall into any of the three previous groups enter the no AI usage group. During the debrief, I ask respondents if they used an AI tool during the survey. The control group used in later comparisons is the no AI usage group less the 27 respondents who self-report AI use in the debrief; I do not count these respondents as AI users unless my detection method already counted them.

2.3.3 Respondent Classification

I begin by gathering evidence at the respondent-by-question level. The raw browser events generate a large set of paradata layers, but not every layer is independent. Several layers may be components of the same physical act, such as a copy, the subsequent away episode, and the answer being entered after the respondent returns. Layers that share the same attribution root are therefore collapsed. I then assign each respondent to one of the four groups based on what I observe in the session. Respondents are only assigned to one group. Table 1 summarizes the detection method for the four groups. I go into much greater detail in a forthcoming internet appendix but spare the reader unnecessary detail here.

2.3.4 Validation

To validate the method, I ran controlled sessions in which I knew how the survey was being completed. Eleven sessions using conventional automation scripts all triggered the standard automation probes, confirming that the instrument detects signs of scripted control. I next

[8] A qualifying inbound paste contains at least eight characters. Copying never counts on its own, and pastes of the respondent's platform identifier or of the survey's own question wording are excluded.

ran four sessions with in-browser assistants to test the machine-delivery rules. When the agents wrote values directly into answer fields, the browser recorded no keystrokes and flagged the input as untrusted; when they typed through the browser protocol, inter-key intervals were 0–1 ms over several hundred keystrokes. In both cases the expected delivery signatures fired. These are the same signatures used to identify machine-delivered answers in the field. The standard automation probes, by contrast, came back clean on all 176 pages the agents visited. The sessions looked like an ordinary human browsing with few exceptions. The comparison shows why the method focuses on answer delivery rather than on a generic bot fingerprint: software-specific delivery is observable, while an agent that restricts itself to ordinary on-screen actions may leave little to no browser-side evidence.

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 distribution of respondents belonging to the four answer-origin groups. I next examine conventional quality screens and respondent characteristics, and conclude with question-specific evidence on inflation expectations.

3.1 The Surveys

The SCE (Armantier et al., 2017) is a useful setting because it measures not only a central forecast but also how beliefs vary across households. A stable median inflation expectation can coexist with a growing group of respondents who expect high inflation to persist, while disagreement and upper-tail mass reflect differences in information, experience, and exposure to prices.[9] Market measures are complementary to the SCE, but nominal yields, inflation-indexed bonds, and inflation swaps combine expectations with risk and liquidity premia (Gürkaynak et al., 2010; Chernov and Mueller, 2012), surveys have historically forecast inflation better than market-based measures (Ang et al., 2007), and none of these measures can depict the household cross-section of beliefs.

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 completions.[10] The survey contains the SCE core module, additional expec-

[9] See, for example, Mankiw et al. (2003); Coibion and Gorodnichenko (2012); Malmendier and Nagel (2016); D'Acunto et al. (2021); Reis (2021).

[10] The survey was fielded on June 3–4, 2026, using SoSci Survey (Leiner, 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. The complete survey instrument will be made available in the future. Disclaimer: Some survey questions were

tations and demographic items, an instructed-response item, 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 instructed-response item 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 evidence of AI involvement. Alongside the submitted answers, I record the completion-process paradata described in Section 2.3.

3.2 Results

3.2.1 The Prolific and SCE Samples

Table 2 compares the 1,294 Prolific respondents with 23,886 SCE respondents from 2013 to 2025. 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, with 40.7% under age 40 against 28.7% in the SCE; substantially more educated, with 70.9% in the college category against 31.9%; more likely to be employed, at 77.0% against 62.2%; less likely to own a home, at 54.5% against 68.0%; and more concentrated in the middle and upper portions of the income distribution. The Prolific panel 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 Detected Respondent Groups

Table 3 reports the four mutually exclusive respondent groups in bold and the assignment routes beneath each group. The bold rows sum to 100%. The indented detection layers can overlap, while the AI chatbot detection layers are disjoint after respondents assigned to the autonomous AI and in-browser assistant groups are removed.

Fifteen respondents (1.2% of the sample) are autonomous AI. Eight respondents (0.6%) use an in-browser assistant.[11] AI chatbot consultation is by far the largest group at 278

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. FRBNY did not participate in or endorse this study and disclaims responsibility for the administration of the survey and the analysis and interpretation of the resulting data.

[11] An assistant-compatible side panel appears in 24 sessions, but the 17 cases without independent cor-

respondents (21.5%). I do not detect AI usage for the remaining 993 respondents (76.7%). In total, 301 of the 1,294 respondents (23.3%) are classified as AI-involved. My detection method detects 28 of the 55 self-reported AI users, meaning 2.1% of the respondents self-reported using AI but were not detected by my method.[12] One reason these respondents go undetected is that I cannot observe a second device, such as a phone taking a picture of a question on the laptop screen. Respondents consulting AI chatbots are the dominant AI usage group, while only a small percentage use in-browser assistants or tools that enter answers autonomously.

Of the 278 respondents who consult AI chatbots, 245 perform a question round trip and three use answers with the canary marker alone; the remaining 30 perform a partial question round trip. Thirty-nine respondents return a canary marker, 33 of them on a question where the question round trip is also observed.[13]

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, comparing the 301 respondents with detected evidence of AI involvement with the 966 respondents with no detected evidence and no self-disclosure. The category rows report average performance within each block, while the indented rows present the individual items. Column (1) presents the AI group (all forms of AI use), column (2) presents the control group results, and column (3) presents the difference between the AI and control groups.

Respondents who use AI at some point in the survey answer 89.9% of the three financial-literacy items correctly, compared with 78.2% in the control group. The 11.7-percentage-point difference is statistically significant at the 1% level and is driven mainly by the compound-interest question, on which the AI group performs 25.8 points better. The AI group also performs 5.7 percentage points better on the real-return question and a marginally significant 3.7 percentage points better on the risk-diversification question. The numeracy

roboration from other paradata do not qualify as in-browser assistant sessions because benign explanations cannot be ruled out.

[12] A McNemar test rejects equality of the two channels' coverage ($\chi^2(1) = 201.7$, $p < 0.001$). A survey relying on disclosure alone would therefore miss roughly 90% of the respondents identified by the completion-process measure.

[13] All participants reported fluency in English. Of the 1,294 respondents, 108 reported a first language other than English. Copying rates are similar across the two groups: 25.9% versus 22.9% copied on at least one page, and 3.70% versus 3.63% copied on at least sixteen distinct pages; neither difference is statistically significant. This result makes it less likely that the routing rule is driven primarily by translation-related copying.

questions follow the same pattern. The AI group answers 94.3% of the four items correctly, compared with 85.6% in the control group. The disease-risk, lottery-probability, and infection-risk differences are all statistically significant at 4.8, 11.5, and 19.7 percentage points, respectively. Both groups perform well on the numeracy items.

The largest differences appear on the cognitive-reflection questions. Respondents in the AI group answer 75.2% of the three questions correctly, compared with 49.0% in the non-AI group, a 26.2-percentage-point statistically significant difference to which each item contributes roughly equally. The visual-illusion tasks are the exception. Respondents in the AI group answer 27.5% of the four items correctly, compared with 29.3% in the non-AI group. The difference is not statistically significant. The item-level estimates do not point in a common direction: the AI group performs 9.5 percentage points better on the Café-wall item but 7.4 points worse on the reverse Müller-Lyer item and 5.7 points worse on the Müller-Lyer item.

Taken together, the AI group submits substantially more correct answers on financial-literacy, numeracy, and cognitive-reflection items while showing no systematic advantage on the visual tasks. This finding should not be read as evidence that these respondents are intrinsically more knowledgeable. Respondents can use AI to answer these questions, but the point of this exercise is that a survey researcher would not have dropped the AI group from the sample based on these conventional quality screens. Knowledge and reasoning items are exactly what an outside tool answers well, and the item split is consistent with that interpretation: the largest gaps appear on questions an assistant can solve easily, while no comparable advantage appears on the perceptual tasks. Other channels can operate alongside this one. Experienced participants may recognize commonly used questions, and respondents who seek outside assistance may be positively selected on education, digital fluency, effort, or willingness to investigate a difficult item rather than guess.

Some of the detection layers vary within respondent, since not every answer was produced with chatbot help, which allows me to at least partially separate assistance from respondent characteristics. Table 5 estimates a respondent-by-item panel for the ten comprehension items. What enters the regressions is an observed page-level event: on the item's own page, the respondent copies the question text and then leaves the survey window. I read that pattern as AI chatbot consultation. The indicator appears on both sides of the table. In columns (1) to (4) it is the explanatory variable, with answer correctness as the outcome in columns (1) to (3) and time on the page in column (4). In column (5) it becomes the outcome, which lets me ask what predicts consulting in the first place. Column (1) contains only item fixed effects, so it reports the cross-sectional gap; column (2) adds respondent fixed effects and identifies the gap within person. Consulted items are 17.3 percentage points more likely to be

answered correctly across respondents. Within respondent the gap falls to 8.3 points but does not disappear. The respondent fixed effects absorb stable differences in ability, education, effort, and repeat-taker status, as well as common item difficulty. Chatbot consultation is not randomized, however, and respondents may be especially likely to consult on items they find difficult. I therefore interpret the estimate as a within-respondent association rather than a causal effect. Even so, the result shows that stable respondent characteristics do not explain the entire cross-sectional gap.

The repeat-taker channel is directly testable using the count of prior approved studies from Prolific. That count is a respondent-level constant, so the respondent fixed effects in column (2) already absorb it; columns (3) and (5) use it directly instead. Column (3) repeats the cross-sectional specification of column (1) with experience added. Experience does predict correctness: a one-standard-deviation higher count raises the probability of a correct answer by 3.2 percentage points. It does not, however, account for the consultation gap, which is 17.8 points with experience controlled against 17.3 points without it. Experienced participants are also slightly less likely to consult an AI chatbot, by 2.1 percentage points per standard deviation increase in experience, so the confound should not produce the gap. Column (4) adds a timing check: consulted pages take roughly 46% longer within respondent, which is more consistent with consultation than with immediate recognition.

Table 6 reports how detected AI usage and disclosure vary across respondent groups. I report descriptive shares rather than a regression. Detected involvement reflects both use and detectability, while in-survey measures such as comprehension scores and completion time can themselves be affected by AI use. Regression adjustment would therefore risk conditioning on post-treatment variables and conflating the propensity to use AI with the propensity to be detected. Each row reports the share of respondents with detected evidence and the share disclosing use in the debrief.

Detected involvement is more concentrated among younger respondents, college graduates, and working respondents. Among respondents below age 30, 29.5% used AI, as did 30.6% of those aged 30–44, compared with 20.4% of those aged 45–59 and 13.5% of those aged 60 and above. 28.4% of respondents with a bachelor's degree or higher used AI compared with 15.2% of respondents without a bachelor's degree. 26.5% of working respondents used AI compared with 15.2% of non-working respondents. Across household-income groups, the share rises from 18.1% below \$60,000 to 27.0% between \$60,000 and \$150,000 before dropping back down to 24.2% at \$150,000 or more. Self-reported use follows a different pattern. It is highest among the youngest respondents and among those not in paid work, while the education and employment findings are reversed.

In summary, the evidence runs counter to conventional screening logic. Respondents who

use AI are younger, more educated, and more successful on the comprehension battery; 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 AI Use and Inflation Expectations

A useful feature of the answer-origin measure is that it records not only whether a respondent leaves evidence of AI involvement, but also the questions on which that evidence appears. A respondent may consult a model on a numeracy question or an open-text item yet answer the inflation block using their own beliefs about future economic conditions. I therefore compare respondents who leave evidence of AI involvement on at least one inflation-expectations page with respondents for whom I observe no qualifying AI mechanism anywhere in the survey and who do not disclose AI use.

Table 7 presents descriptive statistics for one-, three-, and five-year inflation expectations. The analysis retains 1,088 of the 1,294 respondents in the full sample. The AI group contains 122 respondents: 107 with an AI-chatbot consultation trace on an inflation page[14] and 15 autonomous-AI sessions. The control group contains 966 respondents for whom I detect no qualifying AI-delivery or external-routing mechanism anywhere in the survey and who do not disclose AI use. I exclude 179 respondents whose evidence appears only on other survey questions and 27 who disclose AI use but leave no qualifying mechanical trace. Panel A reports one-year expectations, Panel B three-year expectations, and Panel C five-year expectations. Within each panel, the table reports results for the control group, the pooled AI group, and the AI-chatbot and autonomous-AI subgroups.

At the one-year horizon, the AI group reports a mean expectation of 6.40%, a median of 3.00%, and a standard deviation of 15.15 percentage points, compared with 8.06%, 4.00%, and 17.24 percentage points in the control group. The tenth and ninetieth percentiles are 1.00% and 16.30% among respondents in the AI group, versus 0.50% and 25.00% in the control group, yielding tenth-to-ninetieth-percentile ranges of 15.30 and 24.50 percentage points. At the three-year horizon, the corresponding means are 3.02% and 6.60%, the medians are 3.00% and 4.00%, and the standard deviations are 7.70 and 16.76 percentage points. The tenth-to-ninetieth-percentile range contracts from −3.00% to 20.00% in the control group to 2.00% to 5.00% in the AI group. At the five-year horizon, the means are 4.16% and 5.45%, both groups report a median of 3.00%, and the standard deviations are 6.55 and

[14]One respondent assigned to the in-browser-assistant group also leaves an AI-chatbot consultation trace on an inflation page. Because this is the only such case, I include the respondent in the AI chatbot subgroup rather than reporting a separate row. The respondent reports one-, three-, and five-year inflation expectations of 5%, 5%, and 3%, respectively.

17.89 percentage points for the AI and non-AI groups, respectively. The corresponding tenth-to-ninetieth-percentile ranges are 2.00% to 5.00% and −5.00% to 20.00%.

These comparisons show that AI involvement is associated primarily with compression of the distribution and only smaller differences in the mean. The group medians differ by no more than one percentage point at any horizon, whereas the tenth-to-ninetieth-percentile range narrows by 9.2 percentage points at one year, 20 points at three years, and 22 points at five years. At one year, the contraction comes mainly from the upper part of the distribution: the ninetieth percentile is 8.7 percentage points lower among respondents in the AI group, while the tenth percentile is only 0.5 points higher. At three and five years, both tails of the AI group's distribution lie inside those of the non-AI group. Relative to the non-AI group, the tenth percentile is 5 points higher at three years and 7 points higher at five years, while the ninetieth percentile is 15 points lower at both horizons. The five-year comparison is especially clear: the median remains at 3%, even as the standard deviation falls from 17.89 to 6.55 percentage points and the middle 80% of the distribution contracts from 25 to 3 points.

Figure 1 provides a visual summary of these distributions. At each horizon, it compares the same 122 respondents in the AI group with the 966 respondents without AI involvement. The boxes span the interquartile range, the horizontal lines designate the medians, and the whiskers extend from the tenth to the ninetieth percentile. The vertical axis is split above 9% for readability.

When looking at the figure, the patterns from Table 7 stand out: respondents in the AI group have much more homogeneous inflation expectations than those of the non-AI group. Around similar centers, the non-AI group has a much longer upper tail at every horizon and a visibly wider lower tail at the three- and five-year horizons. The AI group instead clusters tightly around the center. A two-sample Kolmogorov–Smirnov test rejects equality of the two distributions at the one-, three-, and five-year horizons ($p < 0.001$ in each case). Taken together, the table and figure show much less heterogeneity among respondents whose answers to inflation questions were at least partially informed by AI.

Table 7 shows that respondents in the AI group report a much narrower distribution of inflation expectations, but not where the narrowing comes from. A lower mean and a thinner upper tail are consistent with two different patterns: a common downward shift that empties the upper tail and fills the lower one, or answers moving inward from both tails toward a focal value. The two patterns can be told apart by the lower tail, which grows under a shift and shrinks under compression. Table 8 separates them by asking, region by region, whether respondents in the AI group are more or less likely than the control group to report an expectation in that region, holding respondent demographics fixed.

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 122 respondents in the AI group and zero for the 966 respondents in the control group, so every regression uses the same 1,088 respondents as Table 7. The vector X_i contains controls for age, gender, education, household income, race, and employment. Five regions and three horizons give fifteen separate regressions, one per cell of the upper block. The final row adds three more, which replace the indicator on the left-hand side with the difference $\mathbf{1}\{Y_{ih} = 3\} - \mathbf{1}\{Y_{ih} = 4\}$.

Each cell reports β_{ch} in percentage points, an HC1 standard error in parentheses, and the mean of the outcome in the control group in brackets, which is the base against which the coefficient should be read. Columns (1), (2), and (3) correspond to the one-, three-, and five-year horizons. The first row presents the probability of an inflation expectation of 0 or lower. The next two rows present probabilities of reporting inflation above 5% and above 10%. The next two rows present the probabilities of reporting exactly 3% and exactly 4%, and the final row tests whether the change at 3% differs from the change at 4%. The regions overlap (an answer above 10% is also above 5%), and the coefficients should not be added together.

I begin with the first row. At the one-year horizon, AI involvement is not associated with a significant difference from the non-AI group. However, AI involvement is associated with a 10.3-percentage-point reduction at three years, relative to a control-group mean of 18.6%, and a 20.7-percentage-point reduction at five years, relative to a mean of 26.5%. Both estimates are statistically significant at the 1% level. The next two rows show upper-tail contraction at every horizon. The probability of reporting inflation above 5% is 20.3 percentage points lower among respondents in the AI group at one year, 29.9 points lower at three years, and 21.4 points lower at five years. The corresponding means in the non-AI group are 33.2%, 32.4%, and 28.7%. Using the 10% threshold produces the same pattern. AI involvement is associated with reductions of 9.4, 17.0, and 10.7 percentage points at the one-, three-, and five-year horizons, respectively, against non-AI group means of 19.5%, 17.9%, and 16.9%. All six upper-tail estimates are statistically significant at the 1% level.

I then investigate the median values from Table 7. AI involvement is associated with a 22.9-percentage-point higher probability of reporting exactly 3% at one year, a 29.7-point higher probability at three years, and a 31.3-point higher probability at five years. These estimates are large relative to the corresponding control-group means of 15.6%, 16.1%, and

15.2%, and all are statistically significant at the 1% level. None of the horizons shows a significant difference in the probability of selecting 4%. The final row compares these two estimated changes. The increase in the probability of reporting 3% exceeds the change in the probability of reporting 4% by 18.9 percentage points at one year, 31.2 points at three years, and 34.9 points at five years. These differences are statistically significant at all three horizons. The additional concentration among respondents in the AI group occurs primarily at 3%, rather than being spread evenly across 3% and 4%.[15]

In summary, respondents in the AI group are less likely to report a high-inflation expectation at every horizon, less likely to report a nonpositive expectation at the three- and five-year horizons, and substantially more likely to report exactly 3% This finding is inconsistent with a simple common downward shift, which would move more observations into the lower tail. Instead, answers move inward from both sides toward a focal center. These differences remain after adjustment for observed demographic characteristics, but because AI involvement is not randomly assigned, the estimates should be interpreted as conditional associations.

These differences point to a measurement problem but cannot be used to screen individual respondents. The two distributions still overlap substantially, so a respondent's inflation forecasts alone cannot reveal whether the answers were produced unaided or with AI. If anything, extreme expectations are less common among respondents with detected AI involvement, but it is not possible to identify AI use from anchored or moderate inflation expectations alone. With these findings in hand, I then turn to a calibration in which the source of every answer is known.

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, all of which could potentially affect their inflation expectation answers. I therefore turn from detected respondent behavior to the distribution of answers produced by the models themselves. I ask whether a consumer AI model, when given 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

[15]Eleven of the fifteen estimates in the upper block remain significant at the 5% level after a Benjamini–Hochberg correction across the fifteen tests.

consumer AI models the same SCE inflation items and comparing their answers with public SCE microdata.

4.1 Calibration Design

I use the public SCE microdata as the human benchmark for my analysis.[16] Because the model calls were made in 2026 while the human benchmarks come from earlier SCE waves,[17] I do not interpret level differences as evidence that AI forecasts are too high or too low because the mean and median inflation expectation can vary over time. However, field respondents in the non-AI group, answering three weeks after the model calls, report a one-year standard deviation of 17.2 percentage points (Table 7). I therefore use the calibration to compare distributional shape rather than contemporaneous forecasting performance. I focus on the pre-AI period from June 2013 through November 2022 while also investigating the normal-inflation period of 2019 and a high-inflation period from 2022. I choose these three periods to further mitigate concerns that my model data from May 2026 might be slightly different from the SCE data from May 2026, which have yet to be released.

Each of the fifteen consumer AI models receives the SCE inflation items in two formats. A 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 or allows autonomous AI to take the survey. Both formats use a short system prompt instructing the model to answer as a respondent in a United States household survey, "as if you were a typical American adult thinking about your own financial situation and the economy." No demographic identity was assigned for this initial analysis. Together, the two formats produce 886 valid one-year answers and 865 three-year answers.

4.2 The Distribution of AI Inflation Expectations

Table 9 presents the main calibration. Panel A reports one-year expectations and Panel B reports three-year expectations. Within each panel, the first three columns report the public SCE microdata for the pre-AI, 2019 normal-inflation, and 2022 high-inflation periods. The last two columns report the pooled AI-model answers from the single-question and complete-

[16] 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.

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

survey calls. The rows present 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 number of observations.

I begin with the center of the one-year distribution in Panel A. The single-question and complete-survey means are 3.3% and 3.4%, compared with 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%. The 2022 SCE median is 7.0%. Thus, the model answers are close to the center of the human distribution in normal periods. The dispersion rows show a much larger difference. The one-year 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 corresponding model standard deviation is 0.4 points in both prompt formats. The SCE interquartile ranges are 5.0, 3.0, and 6.0 percentage points, compared with 0.5 points in each model column. The tenth-to-ninetieth-percentile range runs from 1% to 15% in the pre-AI SCE, from 1% to 12% in 2019, and from −2% to 20% in the high-inflation period. In both model formats, the tenth-to-ninetieth-percentile range runs only from 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 share of SCE respondents reporting exactly 3.0, 3.5, or 4.0% is 22.4% in the full pre-AI sample, 23.7% in 2019, and 9.4% in the high-inflation period. The corresponding model shares are 76.0% in the single-question calls and 87.5% in the complete-survey calls. Most model answers therefore collapse onto a small set of adjacent central values.

Panel B reports the three-year expectations. The single-question and complete-survey means are 2.9% and 2.8%, compared with SCE means of 5.3% in the full pre-AI sample, 4.7% in 2019, and 4.2% in 2022. The SCE medians are 3.0%, 3.0%, and 4.0%, while the model medians are 3.0% and 2.8%. As in Panel A, the model answers are not implausible as central forecasts even though they have a slightly lower mean. The three-year dispersion measures again differ sharply. The SCE standard deviations are 13.1, 12.1, and 16.0 percentage points across the three human benchmarks, compared with 0.3 points in both model formats. The SCE interquartile ranges are 4.0, 3.0, and 10.0 points, while the model interquartile range is 0.5 points in each column. 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 model range is 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 share of SCE respondents reporting 3.0, 3.5, or 4.0% three-year inflation expectations is 23.7% in the pre-AI period, 24.8% in 2019, and 15.0% in 2022. The corresponding model shares are 63.6% in the single-question calls

and 46.0% in the complete-survey calls.

In summary, the row-by-row comparison shows AI homogenization compared to SCE microdata across both horizons, both prompt formats, and all three human benchmark periods. The models produce answers that resemble plausible central forecasts but not draws from the household cross-section. They preserve a normal-times center while failing to reproduce the surrounding heterogeneity. The human mean exceeds the human median because some households report high inflation while few report deflation expectations. Those observations contain information about disagreement, uncertainty, and the risk of unanchoring. A model distribution that reproduces the median while eliminating those observations hides information that is an important part of the SCE. Models provide a centralized forecast rather than an independent household belief.

4.3 Across Models, Prompts, and Personas

As previously mentioned, the data in Table 9 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 choice is intentional. My 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. The main prompts therefore match the natural use case most relevant for AI-assisted respondents.

This design 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 case, the compression in Table 9 might reflect the absence of demographic conditioning. While I do not believe the average survey respondent performs demographic conditioning on a model before asking it to aid in a survey, it is possible that a model learns about a user over time and attempts to take this information into account. I address this concern in two steps. I first ask whether the result is driven by a particular model, unequal call counts, or one of the two prompt formats. 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.

Table 10 reports the one-year inflation expectation separately for each model-by-prompt group. Across the fifteen 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 tenth-to-ninetieth-percentile range is 1.0 percentage point, and 81.8% of answers equal 3.0, 3.5, or 4.0%. Giving each model equal weight produces a mean and median of 3.5%, a standard deviation of 0.4 points, and a focal-value share of 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 prompt formats leads to the same conclusion. 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 calls should generate substantially more variation. They do not. The complete-survey answers are as concentrated as the single-question answers in Table 9. Adding ordinary survey context therefore does not make the model behave like a heterogeneous cross-section of households.

The estimates remain conditional on the prompt wording. The calls presented 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. It is that, under ordinary prompts resembling quick respondent assistance, every model in the panel supplies a highly centralized answer distribution. Table 11 investigates 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. I use six models for the exercise: OpenAI's gpt-chat-latest and gpt-5.4-mini, Anthropic's claude-haiku-4.5, Google's gemini-3-flash-preview, xAI's grok-4.3, and DeepSeek's deepseek-v4-pro, accessed through OpenRouter on May 13–14, 2026. I construct 231 personas from New York Fed SCE sampling-frame margins tabulated from 177,927 SCE respondent-months. Each persona is a cell defined by age group, education, household-income group, census region, and numeracy (the numeracy category is missing for 15 of the 231 cells). Each model answers the full survey instrument once for each persona, giving an intended 1,386 full-survey calls at each inflation horizon. A few calls did not finish properly, so I have 1,383, 1,381, and 1,384 observations at the one-, three-, and five-year horizons. To be consistent with the treatment of the survey and SCE data, I drop eight one-year answers above 100%. This design uses many calls because the relevant variation is across demographic profiles and model families. If the compression in the previously observed calls was mainly due to the absence of demographic conditioning, the persona pool should lead

to a substantial widening of the inflation expectation distribution. Each persona is queried once per model and enters the pool with equal weight. The pool is balanced by design cell rather than representative of the household population, so its dispersion measures how far model answers move across demographic prompts, not how much a population of households disagrees. It should therefore not be compared with population-weighted SCE statistics.

Persona prompting widens the AI distribution, but it does not make it look like the SCE. The first block of Table 11 reports a mean one-year inflation expectation of 4.4%, a median of 3.8%, a standard deviation of 3.3 percentage points, an interquartile range of 1.0 percentage point, and a tenth-to-ninetieth-percentile range from 3.0% to 6.0%. These values are substantially wider than those from the calls in Table 9, where the one-year standard deviation is about 0.4 points. However, they remain far inside the pre-AI human distribution, where the 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, the persona-prompted distribution has a 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 9 runs from -2% to 15%. At five years, the persona pool has 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%. 45.7% of one-year answers, 42.3% of three-year answers, and 31.3% of five-year answers equal 3.0, 3.5, or 4.0%.

The remaining blocks show which characteristics generate the additional spread. Age produces almost no variation. The one-year means are 4.4%, 4.3%, and 4.4% for young, middle-aged, and older personas; 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 produces the clearest variation. 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 economically sensible, but they are not large enough to recover the household heterogeneity seen in the SCE microdata. Across the demographic splits, interquartile ranges remain between 0.5 and 1.3 percentage points, tenth percentiles cluster between 3.0% and 3.2%, and large shares of answers still fall on the same three values. Persona prompting therefore adds some demographic structure, especially by income, while leaving the distributions concentrated around a narrow set of

central expectations.

In summary, demographic conditioning can move model answers away from the distribution in 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 9: consumer AI systems can supply plausible inflation expectations, but they do not reproduce the cross-sectional heterogeneity that the SCE contains.

4.4 Density Forecasts

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

Table 12 presents the results of examining this possibility. Panel A reports the one-year horizon and Panel B the three-year horizon. Within each panel, the rows compare SCE respondents in the pre-AI, 2019 normal-inflation, and 2022 high-inflation periods used in Table 9, field respondents in the control group, and the two AI prompt formats. The largest-bin share, the probability of placing at least 90% in one bin, entropy, and implied standard deviation describe the shape of an individual density forecast. The final column reports disagreement, measured as the cross-sectional standard deviation of the density-implied means across respondents or model calls.

I begin with the shape of individual forecasts in Panel A. At the one-year horizon, the average largest-bin share is 0.58 in the pre-AI SCE, 0.60 in the normal period, and 0.56 in the high-inflation period. The corresponding values are 0.51 among field respondents, 0.40 in the single-question calls, and 0.42 in the complete-survey calls. Entropy is 1.02 in the pre-AI SCE, 0.98 in the normal period, and 1.08 in the high-inflation period, compared with 1.27 in the field and 1.54 and 1.51 in the two model formats. The implied within-forecast standard deviation is 2.79 points in the pre-AI SCE, 2.62 points in the normal period, and 3.24 points in the high-inflation period. It is 3.81 points among field respondents and 3.53

and 3.38 points in the single-question and complete-survey calls.

Panel B shows the same general pattern at the three-year horizon. The average largest-bin share is 0.57 in the pre-AI SCE, 0.59 in the normal period, and 0.53 in the high-inflation period, compared with 0.50 in the field and 0.38 and 0.47 in the two model formats. Entropy is 1.02, 0.98, and 1.13 across the three SCE periods, 1.26 in the field, and 1.61 and 1.44 in the model calls. The implied within-forecast standard deviations are 2.77, 2.61, and 3.21 points across the three SCE periods, compared with 3.67 points among field respondents and 3.75 and 3.02 points in the two model formats.

Across the SCE periods, the shape of individual densities moves with the inflation environment. Relative to the normal period, the high-inflation period has a lower largest-bin share, higher entropy, and a larger implied standard deviation at both horizons. The pre-AI values generally fall between the normal- and high-inflation-period values. Human respondents therefore express more uncertainty within their forecasts during the high-inflation period. The model densities differ from the human densities in another direction: their largest-bin shares are below, and their entropy values above, those of every SCE period and the field sample at both horizons, while their implied standard deviations remain comparable to the human values. The AI-generated density forecasts therefore appear well formed at the individual-response level. They allocate probability to several outcomes rather than collapsing to a point mass. Conventional data-quality screens would likely not flag or exclude such observations.

I next turn to disagreement across forecasts. At the one-year horizon, disagreement is 3.87 percentage points in the pre-AI SCE, 3.45 points in the normal period, and 4.50 points in the high-inflation period. It is 3.42 points among field respondents, compared with 0.85 points in the single-question calls and 0.75 points in the complete-survey calls. At the three-year horizon, disagreement is 4.00 points in the pre-AI SCE, 3.58 points in the normal period, and 5.12 points in the high-inflation period. The corresponding values are 3.99 points in the field, 0.78 points in the single-question calls, and 0.93 points in the complete-survey calls.

The period-specific SCE rows sharpen the comparison. Human disagreement is lowest in the normal period and highest in the high-inflation period at both horizons, indicating that the cross-sectional distribution of density-implied expectations widens when inflation expectations are elevated. Yet even the normal-period disagreement values are about four times the largest corresponding model values. The field comparison makes the distinction especially clear: the implied within-forecast standard deviations are similar for field respondents and model calls, but disagreement across field respondents is roughly four times as large. The models therefore express uncertainty within an individual forecast while producing very similar density-implied means across calls. The density instrument shows the same

homogenization as the point forecasts: individual model 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 itself informative. At the one-year horizon, 17% of pre-AI SCE forecasts place at least 90% of the probability in a single bin, compared with 19% in the normal period and 16% in the high-inflation period. The corresponding share is 6% among field respondents and zero in both model formats. At the three-year horizon, the shares are 18%, 20%, and 15% across the three SCE periods, 8% in the field, and again zero in both model formats. Coarse reporting is therefore somewhat less common in the high-inflation period, when individual uncertainty is greater, but it remains present in every human sample and absent from the model forecasts. 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 does more than smooth the answer. It replaces a human measurement object with a model-generated one that is valid-looking, centralized, and behaviorally different.

5 Consequences for Survey Measurement

The preceding section shows that consumer AI models can return plausible central inflation forecasts while producing much less dispersion and upper-tail mass than observed in the SCE. I now ask how inserting an answer distribution of this shape changes the statistics that researchers, central banks, and market participants use to monitor inflation expectations. The exercise uses one-year point expectations. Let F denote the cross-sectional distribution of household answers in a high-inflation episode and F_{AI} the model-answer distribution calibrated in Section 4. When a fraction λ of recorded answers is fully replaced by model answers, the case $\alpha_i = 0$ for affected answers in Equation (1), the observed cross-section is

$$G_\lambda = (1 - \lambda)\, F + \lambda\, F_{\text{AI}}. \tag{3}$$

Equation (3) is the ε-contamination model of Huber (1964), a standard framework for robust statistics (Chen et al., 2016). Here the contaminating distribution is not a hypothetical worst case: it is the empirically realized distribution of 886 model answers. Rather than construct an estimator that withstands contamination, I ask how familiar survey statistics change when model answers enter the recorded cross-section, the kind of sensitivity analysis motivated by Lin (2026). I treat λ as a stress parameter, not an estimate of AI involvement in the official SCE.

For this analysis, I consider how an economist might monitor inflation expectations. I assume that the economist uses normal-period survey data as a benchmark for interpreting each new survey wave. I implement this comparison using monitoring rules that compare each survey statistic with a threshold calibrated on normal-period months. These rules follow the general logic of monitoring against a benchmark (Chu et al., 1996), but use empirical monthly thresholds. A monitor "flags an event" when a simulated wave's statistic exceeds its threshold. A flag therefore indicates that the economist would recognize the wave as unusual relative to the normal-period benchmark. It does not indicate detection of AI involvement.

I summarize the simulations using both reported values and alarm rates. The reported value is the numerical value of a survey statistic, such as the median inflation expectation or the cross-sectional standard deviation. For each replacement share, I report the average of these wave-specific values across simulations. The alarm rate is the fraction of simulated waves in which that statistic exceeds its monitoring threshold. Reported values therefore show how AI replacement changes the measured level, disagreement, or tail mass of inflation expectations. Alarm rates show how often the corresponding rule still identifies the high-inflation episode.

The distinction matters because a statistic can be materially distorted without losing its ability to flag the episode. For example, the median expectation can fall substantially after replacement but remain above its normal-period threshold. An economist would still recognize elevated expectations, but the reported median would understate their level. Further replacement can push the statistic below its threshold, removing the warning even though the underlying human beliefs have not changed. Examining both measures allows me to distinguish a change in what the survey reports from a failure to flag the high-inflation period.

5.1 Design and Monitoring Rules

I use the SCE during the 2022 high-inflation period as my human benchmark. During this period, expectations were elevated at the center and in the upper tail.[18] Each simulated wave first draws one of the six months from April through September 2022 with equal probability

[18] A concern is that high reported expectations reflect response errors, in which case removing them could improve rather than weaken survey measurement. Table A1 examines this concern by checking whether high expectations persist across months and are also reflected in respondents' density forecasts. In the full SCE sample, the probability of reporting an expectation above 10% the following month is 58% among respondents currently above 10%, compared with an unconditional rate of 12%. During the high-inflation episode, the corresponding probabilities are 65% and 22%. In that period, 60% of respondents with point expectations above 10% also have a density-implied mean above 10%, compared with 5% of other respondents. These patterns suggest that the upper tail contains information about what households believe, even if the exact values they report are imprecise.

and then draws 1,000 respondents along official SCE weights. I replace a fraction λ of their answers with draws from the model-answer pool. Eight answers from the SCE data are outside ± 100 percentage points and are removed before sampling. I simulate 3,000 waves at each value of λ. I report results for $\lambda \in \{0, 0.1, 0.3, 0.5\}$.

I investigate the following statistical monitors: the median, mean, standard deviation, interquartile range, tenth-to-ninetieth-percentile range (P90−P10), ninetieth and ninety-fifth percentiles, and the shares of answers above 5%, 8%, and 10%. For each monitor, the threshold is the empirical 95th percentile of its 72 realized weighted monthly values from January 2014 through December 2019. I specify that a monitor flags a wave of SCE data as a high-inflation month only when its statistic exceeds the normal-period threshold.

5.2 Loss of Survey Signals

I first ask how replacing SCE responses with AI-generated responses changes reported expectations and whether an economist would still recognize a high-inflation period. Table 13 presents the results. Columns (1)–(4) of Panel A report the average value of each statistic at replacement shares of 0, 10%, 30%, and 50%. Column (5) reports the normal-period cutoff for each statistic. These cutoffs are the 95th percentiles of the corresponding monthly statistics from January 2014 through December 2019 and remain fixed across replacement shares. For example, the median rule flags a simulated wave when its median inflation expectation exceeds 4.23%. Panel B reports how often each rule flags a wave across the simulations. Columns (1)–(4) of Panel B report alarm rates from the same simulated waves. The final two columns of Panel B report the breakdown share λ^* and its 90% confidence interval. The last row of Panel B gives the share of waves in which the monitors miss the high-inflation period.

I begin with Panel A. At 10% replacement, the average median falls from 7.8% to 7.2%, and the mean from 10.7% to 10.0%. At 30% replacement, they fall to 4.7% and 8.5%. At 50% replacement, they fall below their thresholds of 4.23% and 7.67% to 3.7% and 7.1%. The dispersion rows show the same contraction. At replacement shares of 0, 30%, and 50%, the standard deviation of inflation expectations is 17.5, 15.0, and 12.9 percentage points; the interquartile range is 8.9, 6.7, and 4.8 points; and the tenth-to-ninetieth-percentile range is 29.2, 20.4, and 13.9 points. All three average dispersion measures are below their thresholds at a replacement share of 30%. The next two rows show that much of the contraction is in the upper tail. The ninetieth percentile falls from a baseline of 27.7% to 21.2% at 30% replacement and 16.3% at 50%. The corresponding ninety-fifth percentiles are 46.3%, 36.8%, and 27.9%. The final three rows show declining upper-tail mass. The shares of

inflation expectations above 5%, 8%, and 10% fall from 64.2%, 45.3%, and 29.0% without replacement to 45.0%, 31.8%, and 20.5% at 30% replacement, and to 32.1%, 22.7%, and 14.6% at 50%. The distribution of inflation expectations is both lower and less dispersed when AI-generated answers replace SCE microdata observations in a high-inflation period.

Panel B shows how these changes affect the warning rules. Without any AI replacement, the median and the three threshold-share rules flag the high-inflation episode in every simulated wave, while the mean flags it in 98% of the simulations. The remaining rules do not always flag the episode even when the data contain only SCE respondents. The standard deviation, IQR, tenth-to-ninetieth-percentile range, ninetieth percentile, and ninety-fifth percentile flag between 60% and 92% of the simulated waves. These clean-episode alarm rates provide the baseline against which I evaluate how the warning signals change as AI replacement increases. They also point to the importance of assessing the statistics as a group rather than in isolation. At a replacement share of 10%, the median rule and the 5% and 8% threshold-share rules flag 100% of the simulated waves, the 10% threshold-share rule flags 99%, and the mean flags 92%. The dispersion and upper-quantile rates to flagging 73% of the simulated waves for the standard deviation, 46% for the IQR, 76% for the tenth-to-ninetieth-percentile range, and 81% and 85% for the two upper percentiles.

At a replacement share of 30%, the difference across monitoring rules becomes larger and certain monitoring rules almost always miss the high-inflation period. The median and mean have alarm rates of 70% and 74%. The rates are 24% for the standard deviation, 2% for the IQR, 5% for the tenth-to-ninetieth-percentile range, and 40% for the ninetieth and 48% for the ninety-fifth percentile. The threshold-share rules are more resilient, with rates of 100%, 78%, and 72% for the shares of inflation expectations above 5%, 8%, and 10%. The dispersion and upper-quantile rules lose much of their warning signal earlier than the median, mean, and threshold-share rules. The surviving signals nevertheless describe a materially milder episode, as the lower reported values in Panel A show. The joint no-alarm rate remains 0% at this replacement share.

At a replacement share of 50%, the median and all three dispersion alarm rates are 0%, while the ninetieth- and ninety-fifth-percentile rates are 1% and 3%. The threshold-share rates are 13%, 0%, and 2%, while the mean is the strongest remaining monitor at 28%. The final row presents the joint outcome: none of the ten rules fires in 70% of simulated waves, compared to a reported rate of 0% both without replacement and at the two lower replacement shares. A wave drawn from the high-inflation episode can therefore fail to trigger any of these warnings when half its recorded answers are replaced.

I next turn to the breakdown shares in columns (5)–(6) of Table 13. These estimates ask a different question from the alarm rates: how much AI replacement is required before a

statistic from the high-inflation episode falls back below its normal-period warning threshold? To answer this question, I pool the six months of the 2022 high-inflation episode into one weighted human distribution, F, and combine it with the AI distribution according to $G_\lambda = (1-\lambda)F + \lambda F_{\text{AI}}$. I then increase λ in increments of 0.005 and recompute each statistic directly on this mixture. The breakdown share, λ^*, is the first replacement share at which a rule that flags the clean pooled episode no longer flags it. For example, a breakdown share of 0.20 means that once 20% of the pooled human distribution is replaced by AI answers, that statistic falls below its fixed normal-period threshold.

This calculation differs from the alarm rates in columns (1)–(4). The alarm rates come from repeated simulated survey waves and report the fraction of those waves that still trigger a warning at a given replacement share. The breakdown share instead removes finite-sample variation and asks when the statistic of the underlying pooled mixture itself crosses the threshold. It is therefore a monitoring analogue of the breakdown point in robust statistics (Donoho and Huber, 1983), rather than the contamination share at which an estimator diverges. Column (6) reports 90% confidence intervals from 300 bootstrap replications. In each replication, I reconstruct the pooled six-month episode distribution by resampling its respondent observations and the model-answer pool while re-estimate the monitoring threshold by resampling the 72 normal-period months. I then recompute the breakdown share.

The median breakdown lies at 34% and the mean at 41.5%. The standard deviation crosses its threshold earlier, at 20%. The tenth-to-ninetieth-percentile range follows at 24.5%, the ninetieth percentile at 25.5%, and the ninety-fifth percentile at 29.5%. For the shares of answers above 5%, 8%, and 10%, the breakdown is at 46%, 38%, and 36.5%. Which warning disappears first depends on both the change in the statistic and its distance from the calibrated threshold.

Figure 2 provides a visual summary across replacement shares. The left panel plots the average median, standard deviation, tenth-to-ninetieth-percentile range, and two upper percentiles. The right panel plots alarm rates for the median, mean, share above 5%, standard deviation, tenth-to-ninetieth-percentile range, and ninetieth percentile. Reported values decline gradually, while alarm rates can fall sharply as more waves cross below a fixed threshold. The figure reinforces the distinction between a statistic that continues to signal a milder episode and a rule that no longer signals at all.

5.3 Compression versus Level Shift

A natural concern with the baseline results is that the model answers differ from the 2022 SCE not only because they are more concentrated, but also because they are centered at a lower inflation expectation. I therefore ask whether the monitoring results are driven by this difference in the center of the model forecast or by the compression of the model-answer distribution itself. Table 14 separates these two channels. Columns (1)–(2) present the alarm rates at a λ of 30% and 50% from Table 13 for ease of comparison. Columns (3)–(4) shift every model answer upward by 4.5 percentage points so that the pool's median matches the weighted median of the pooled episode. This changes the location of the model distribution but not its shape. Columns (5)–(6) instead use the persona-prompted responses of Table 11, which introduce more dispersion: their standard deviation is 3.3 percentage points and their ninetieth percentile is 6.0%, compared with 0.4 points and 4.0% in the baseline pool. The final row again reports the joint no-alarm rate.

I begin with the recentered pool. At 30% and 50% replacement, the median alarm rate is 100% and the mean rates are 98% and 97%. The dispersion rates remain low: 14% for the standard deviation, 0% for the IQR, and 4% for the tenth-to-ninetieth-percentile range at 30% replacement, with all three rounding to zero at a 50% replacement share. The ninetieth- and ninety-fifth-percentile alarm rates fall from 41% and 48% at a replacement share of 30% to 1% and 2% at a 50% replacement share. The share-above-5% rule flags the high-inflation wave in 100% of the waves, the share-above-8% falls from 100% to 87%, and the share-above-10% falls from 73% to 2%. The joint no-alarm rate is 0% at both replacement shares. Recentering therefore preserves the center-based warnings without restoring the dispersion signal. The dispersion warnings weaken even when the model pool has the same median as the human benchmark.

I next turn to the persona pool in columns (5)–(6). At 30% replacement, the median alarm flags the high-inflation period in 100% of the waves while the mean does so in 76%. The standard-deviation, IQR, and tenth-to-ninetieth-percentile-range rates are 23%, 3%, and 6%, while the ninetieth- and ninety-fifth-percentile rates are 42% and 53%. The threshold-share rates are 100%, 83%, and 70%. At 50% replacement, the median and mean retain rates of 65% and 50%, compared to 0% and 28% in the baseline. All three dispersion rates nevertheless round to zero, and the two upper-percentile rates fall to 1% and 6%. The shares above 5%, 8%, and 10% have rates of 78%, 1%, and 3%. The joint no-alarm rate rises from 0% at 30% replacement to 21% at 50% replacement. This is substantially below the 70% no-alarm rate under the baseline pool, meaning that at 50% replacement at least one monitor still flags the high-inflation episode in 79% of persona-pool waves, compared with only 30% under the baseline pool. In summary, both alternative pools preserve the main dispersion

result, but giving model answers a higher center or more persona-induced variation allows some warnings to survive higher levels of replacement.

5.4 Assistance Rather Than Replacement

A second concern with the baseline results is that they assume every AI-involved answer is fully replaced by a model answer. The field evidence in Section 3, however, shows that the dominant mechanism is human respondents consulting AI rather than autonomous AI replacing them. Respondents might modify the AI's answer or adopt only the parts they agree with. I therefore ask whether the loss of survey signals remains when an AI-involved answer still contains some of the respondent's own beliefs. I examine this question by holding the affected-answer share fixed at $\lambda = 0.3$ and varying how much of the original human answer is retained. For each affected respondent, I record $\text{round}(\alpha \cdot \text{human} + (1 - \alpha) \cdot \text{AI}, 1)$, where α is the weight placed on the respondent's original answer. Table 15 presents the results of this analysis. Column (1), with $\alpha = 0$, is the pure replacement from column (3) of Panel B in Table 13. Columns (2)–(4) increase the human-answer weight to 0.3, 0.5, and 0.7. The rows report alarm rates for the same ten statistics, followed by the share of waves in which none of these statistics flags the high-inflation episode.

The median alarm rate rises from 70% under pure replacement to 100% for answers that retain at least 30% of the respondents' beliefs. The mean rates rise from 74% to 83%, 89%, and 93% as the retained share of the own answer rises from 0 to 0.7. The dispersion measures are slower to react to higher respondent belief weights. The standard-deviation rates are 24%, 25%, 36%, and 57%; the IQR rates are 2%, 6%, 18%, and 37%; and the tenth-to-ninetieth-percentile-range rates are 5%, 33%, 70%, and 81%. Even when the human answer receives a weight of 70%, each dispersion rule remains below its clean-episode alarm rate in Table 13. The ninetieth-percentile rates are 40%, 56%, 78%, and 84%, and the ninety-fifth-percentile rates are 48%, 48%, 65%, and 75%. The threshold-share monitors are quicker to respond to increased respondent weight in the answer generation process than the other dispersion measures. The share-above-5% flags the high-inflation waves in 100% of the simulations while the share-above-8% moves from 78% in the baseline to 99% at a weight of 30%, and 100% for weights of 50% or higher. The share-above-10% rates are slightly slower to respond but move from 89% in column (2) to 100% in column (4). The joint no-alarm rate remains 0% in all four columns. In summary, retaining part of the respondent's own answer restores the center and threshold-share warnings quickly but restores the disagreement signal only partially: at a weight of 70%, each dispersion rule still fires less often than on the clean episode.

5.5 Robustness

I next take up several robustness checks to investigate whether my results depend on the monitoring rules, the sampling of model and human answers, or how affected answers are formed and selected. Appendix Tables A2–A10 report these checks.

5.5.1 Monitoring Rules and Episode Months

I begin by asking whether the main results depend on how the monitoring thresholds are constructed. Because the alarm rates compare each simulated wave with a normal-period cutoff, poorly calibrated thresholds could make a monitor appear too sensitive or too insensitive even before AI replacement is introduced. I therefore compare my monthly calibration with a pooled-resampling alternative and test how well the thresholds behave in held-out normal-period months. Table A2 reports these comparisons together with the calibration-period and high-inflation-period monthly means. The comparison suggests that the monthly calibration is conservative rather than mechanically generating alarms: allowing for realized month-to-month variation raises the cutoff for key statistics relative to pooled resampling. The held-out false-alarm rates are noisy because each split contains only 36 months and several statistics take repeated values, but they do not point to a common direction of miscalibration; I therefore do not rely on the exact cutoff alone and confirm the main pattern using alternative thresholds below.

I next ask how much uncertainty comes from estimating the monitoring thresholds with only 72 normal-period months. Table A3 addresses this concern by redrawing the calibration months in 300 bootstrap replicates, re-estimating the thresholds, and evaluating the same simulated waves against each new set of cutoffs. The resulting intervals can be wide when a statistic lies close to its threshold. For example, the median alarm rate at $\lambda = 0.3$ has a 90% interval of $[0.02, 0.71]$, while the share-above-5% rate at $\lambda = 0.5$ has an interval of $[0.04, 0.42]$. Threshold uncertainty therefore makes the exact alarm rate difficult to pin down for monitors close to their crossing point. It does not change the underlying reported statistic, but it does affect when that statistic is classified as unusual. I therefore interpret individual alarm rates near a threshold cautiously and next examine whether the main pattern survives when I do not rely on a single cutoff.

Table A4 evaluates each statistic across all possible thresholds using the area under the receiver operating characteristic curve (AUC), where higher values indicate better separation between high-inflation and normal-period waves in the direction of elevated expectations. At $\lambda = 0.3$, the mean, median, and share above 5% have AUC values of 0.967, 0.957, and 1.000, compared with 0.752 for the standard deviation and 0.720 for the tenth-to-

ninetieth-percentile range. At $\lambda = 0.5$, the standard-deviation and interdecile-spread AUCs fall further, to 0.331 and 0.186. The IQR is an important qualification: its AUC is 0.914 at $\lambda = 0.3$ even though its fixed-threshold alarm rate is only 0.02, showing that the particular cutoff understates how much information that statistic still contains. Taken together, the AUC results show that some exact alarm rates are sensitive to the threshold choice, but the deterioration of the dispersion-based warning signal as replacement increases is not driven only by the 95th-percentile cutoff. AUC values below 0.5 indicate that the statistic separates the two periods in the opposite direction from the upward monitoring rule, not that it contains no information.

I next ask whether pooling the six high-inflation months masks important differences across individual months. The breakdown shares in Table 13 treat April through September 2022 as one pooled episode distribution, but the strength of the inflation signal differs from month to month. Table A5 therefore repeats the breakdown calculation separately by episode month and reports the bootstrap classifications underlying the pooled estimates. The IQR illustrates why this check matters. It does not signal on the clean pooled distribution, but it does signal in five of the six individual episode months, with month-specific breakdown shares between 0.010 and 0.265. The exact replacement share at which a monitor fails therefore varies across months. I interpret the pooled breakdown share as a summary of the high-inflation episode rather than as an exact failure point that should apply to every month.

Finally, I ask whether excluding higher-order moments from the original analysis ignores potentially useful warnings. Table A6 examines skewness and kurtosis using the same monitoring framework. Neither statistic flags the clean high-inflation episode: their alarm rates are both 0.00 before any AI replacement. The reason is visible in Table A2: average kurtosis is lower during the high-inflation episode than during the normal period, while skewness changes little. Under heavy replacement, the upward alarms become more likely to fire because the increasingly concentrated distribution creates a different shape, not because the rules recover the original high-inflation signal. I therefore retain skewness and kurtosis as distributional diagnostics but do not include them among the operational monitors.

In summary, these robustness checks support the same overall pattern. Threshold estimates are noisy, individual episode months differ, and some statistics such as the IQR retain useful separation under alternative cutoffs, so I do not interpret any single alarm rate or pooled breakdown share as a universal failure point. However, across the monthly calibration, threshold bootstrap, alternative-threshold analysis, and month-specific breakdowns, increasing AI replacement consistently weakens the dispersion-based warning signal. Skewness and kurtosis do not provide an alternative upward warning in this episode because

neither statistic flags the clean 2022 high-inflation period.

5.5.2 Model Answers and Sampling Weights

I next ask whether the main results depend on particular features of the model-answer pool rather than on its low dispersion. The baseline simulations draw from the empirical distribution of 886 model answers, which is highly concentrated. I therefore replace the empirical model pool with a Gaussian distribution matched to its mean and standard deviation. Table A7 reports this comparison. The resulting alarm rates are similar to those obtained with the empirical pool, indicating that the main compression result does not depend on the precise shape of the observed model-answer distribution. What matters for this exercise is primarily that the replacement distribution is centered near conventional expectations and has much less dispersion than the household distribution.

I also examine the role of SCE sampling weights and present the results in Table A7. The baseline draws human answers using the official respondent weights and calibrates the monitoring thresholds using weighted monthly statistics. As a diagnostic, I instead draw human answers without the survey weights while leaving the weighted monitoring thresholds unchanged. This changes the clean-episode benchmark because the simulated statistics and the thresholds are no longer constructed under the same weighting scheme. I therefore do not interpret this exercise as a fully unweighted replication. Instead, it shows why the weighting convention used to construct a survey statistic should also be used to calibrate the threshold against which that statistic is evaluated.

A separate concern is that the baseline model pool combines answers from fifteen models, whereas actual AI use may be concentrated in a smaller number of providers. Table A8 therefore changes how model answers enter a simulated wave. I compare the baseline call-weighted pool with equal weighting across models and with a specification in which a single model supplies all AI answers within a wave. Average alarm rates remain close to the baseline across these sampling schemes. The identity of the model can nevertheless matter for individual monitoring rules: at $\lambda = 0.3$, for example, the median alarm rate ranges from 0.61 to 1.00 across the single-model exercises. The main average pattern is therefore not driven by unequal call counts or by mixing models within a wave, although the exact effect of contamination can vary depending on which model respondents use.

In summary, the results do not seem to be driven by the exact shape of the empirical model-answer pool, by unequal numbers of calls across models, or by mixing answers from different models within a simulated wave. The identity of the model can matter for individual rules, but the broader result survives these alternatives: replacing heterogeneous household answers with a much more concentrated model distribution weakens the survey's warning

signals.

5.5.3 Answer Formation and Selection

I finally ask whether the results depend on how AI assistance changes an individual answer. The partial-assistance exercise in Table 15 combines the human and model answers mechanically and then rounds the result to one decimal place. One concern is that this procedure may create unusually concentrated answers because it contains no additional respondent-specific variation. Table A9 addresses this concern by adding idiosyncratic noise to the blended answer before rounding. Most alarm rates change little as the noise standard deviation increases from 0 to 0.3 and 0.6 percentage points. The main exception is the median under pure replacement, whose alarm rate rises from 0.70 without added noise to 0.77 and 0.95 with noise. The broad pattern is therefore not driven by the absence of small idiosyncratic variation, although the median can be sensitive when many model answers lie close to its monitoring threshold.

The baseline simulations also assume that the respondents whose answers are affected are selected at random. This is deliberately neutral, but AI consultation may not be random in practice. Respondents with unusually high or otherwise extreme beliefs could be more likely to seek outside assistance, in which case replacing the same number of answers would remove more information from the tails. Table A10 examines this possibility by replacing either the largest answers or the answers farthest from the weighted monthly median. The effects are substantially larger than under random replacement. Under the largest-answer rule, replacing only 5% of answers reduces the standard-deviation alarm rate to 0.00 and the ninety-fifth-percentile alarm rate to 0.01. The field data do not reveal respondents' beliefs before consultation, so these targeting rules should not be interpreted as estimates of actual selection into AI use. They instead show that random replacement is not the most damaging case: if consultation is concentrated among respondents carrying unusual beliefs, relatively little AI involvement can remove a disproportionate amount of the survey's disagreement and upper-tail signal.

Taken together, the robustness checks reinforce the main mechanism while clarifying its limits. The results are not specific to the exact empirical model-answer distribution, to unequal numbers of calls across models, or to the absence of small idiosyncratic noise. The magnitude of the effect does depend on which model supplies the answer, how much of the respondent's own information remains, and which respondents seek assistance. These qualifications are economically important, but they preserve the central result: introducing a concentrated common signal can reduce measured disagreement and weaken the distributional warnings in the SCE even when the resulting individual answers remain plausible.

6 Conclusion

Surveys are valuable because they aggregate heterogeneous private beliefs. Generative AI weakens the link between the sampled person and the recorded answer in two ways: it can replace the respondent or assist a human respondent in forming an answer. I develop a browser-side measure of answer origin that distinguishes autonomous completion from respondents consulting AI chatbots without classifying respondents based on whether their submitted answers look unusual. In an independently fielded SCE-style survey, 23.3% of 1,294 respondents leave evidence of AI involvement, while only 4.3% disclose it. Nearly all of this evidence points to human respondents consulting AI rather than software replacing them. At the five-year horizon, the AI and control groups both report a median inflation expectation of 3%, but their tenth-to-ninetieth-percentile ranges are 3 and 25 percentage points, respectively. The center remains plausible while the surrounding heterogeneity contracts.

Because the field comparison is observational, I calibrate the model-answer distribution directly. Across fifteen AI models, AI answers cluster around a plausible normal-period center while reproducing only a small fraction of the disagreement and tail mass observed among households. The same distinction appears in the density questions: individual model forecasts are well formed and express uncertainty across several outcomes, but their density-implied means vary little across model calls. I then use the calibrated model distribution to stress-test statistics of one-year point expectations in the SCE during the 2022 high-inflation episode. Dispersion and upper-percentile monitors fail before center-based monitors, and only the tail-share rules remain reliable once 30% of answers are replaced. AI makes recorded answers more alike without making them individually implausible.

The results imply that answer origin can no longer be taken for granted, even when a human appears to complete the survey and autonomous AI is not detected. Debates about survey modernization ask whether surveys reach enough people, ask the right questions, and arrive in time. This paper adds a prior question: whose information does a recorded answer contain? My approach uses standard browser instrumentation, leaves the respondent's experience unchanged, and distinguishes detected machine delivery from detected human-mediated assistance at the level of the individual survey page. Survey administrators could use this approach to record and publish rates of detected AI involvement alongside the other survey-quality measures they release.

For researchers and policymakers who use expectations data, the results carry a specific caution. Measured disagreement can fall because household beliefs genuinely converge or because a growing share of recorded answers passes through the same few models. Conditional

on a stable central expectation, the first explanation may indicate stronger anchoring; the second indicates that the survey has lost independent information even though its individual answers remain plausible. The same concern applies beyond inflation expectations whenever disagreement and tail beliefs are themselves economically meaningful objects of measurement. Without information about answer origin, a decline in measured disagreement cannot distinguish genuine convergence from artificial consensus.

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.

Anesti, Nikoleta, Edward Hill, and Andreas Joseph (2026). Inflation attitudes of large language models. Bank of England Staff Working Paper No. 1190, June 2026. Also arXiv:2512.14306.

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, Scott Nelson, Giorgio Topa, Wilbert van der Klaauw, and Basit Zafar (2016). The price is right: Updating inflation expectations in a randomized price information experiment. *Review of Economics and Statistics* 98(3), 503–523. doi:10.1162/REST_a_00499.

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.

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.

Cavallo, Alberto, Guillermo Cruces, and Ricardo Perez-Truglia (2017). Inflation expectations, learning, and supermarket prices: Evidence from survey experiments. *American Economic Journal: Macroeconomics* 9(3), 1–35. doi:10.1257/mac.20150147.

Chen, Mengjie, Chao Gao, and Zhao Ren (2016). A general decision theory for Huber's ϵ-contamination model. *Electronic Journal of Statistics* 10(2), 3752–3774. doi:10.1214/16-EJS1216.

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.

Chu, Chia-Shang James, Maxwell Stinchcombe, and Halbert White (1996). Monitoring structural change. *Econometrica* 64(5), 1045–1065. doi:10.2307/2171955.

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, Yuriy Gorodnichenko, and Michael Weber (2022). Monetary policy communications and their effects on household inflation expectations. *Journal of Political Economy* 130(6), 1537–1584. doi:10.1086/718982.

Couper, Mick P. (1998). Measuring survey quality in a CASIC environment. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 41–49.

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.

Dhakal, Vivek, Anna Maria Feit, Per Ola Kristensson, and Antti Oulasvirta (2018). Observations on typing from 136 million keystrokes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. doi:10.1145/3173574.3174220.

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.

Donoho, David L. and Peter J. Huber (1983). The notion of breakdown point. In P. J. Bickel, K. A. Doksum, and J. L. Hodges Jr. (eds.), *A Festschrift for Erich L. Lehmann*, 157–184. Belmont, CA: Wadsworth.

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.

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.

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.

Haaland, Ingar, Christopher Roth, and Johannes Wohlfart (2023). Designing information provision experiments. *Journal of Economic Literature* 61(1), 3–40. doi:10.1257/jel.20211658.

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.

Huber, Peter J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics* 35(1), 73–101. doi:10.1214/aoms/1177703732.

Kreuter, Frauke (Ed.) (2013). *Improving Surveys with Paradata: Analytic Uses of Process Information*. Hoboken, NJ: Wiley. doi:10.1002/9781118596869.

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

Lin, Zhicheng (2026). Artificial-intelligence-mediated contamination in online research: Taxonomy, risk gradient, and recommendations. *Advances in Methods and Practices in Psychological Science* 9(3). doi:10.1177/25152459261454231.

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.

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.

Wu, Jing Cynthia, Jin Xi, and Shihan Xie (2025). LLM survey framework: Coverage, reasoning, dynamics, identification. NBER Working Paper 34308. doi:10.3386/w34308.

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: Answer-origin classification rules and paradata inputs

Qualifying rule	Paradata inputs	What the rule establishes
Panel A: Autonomous AI		
Direct field write	Final value; keystroke, paste, and dictation history; autofill, input-method-editor, and back–forward-cache guards	No ordinary input event explains the completed field.
Synthetic input	Trust flag, event type, and key identity	Programmatic rather than ordinary physical input.
Metronomic typing	Trusted key timing, identities, pause distribution, and autorepeat	A long trusted-key sequence has a cadence inconsistent with physical typing.
Machine-cadence typing	Inter-key timing across fields and pages, net of autorepeat	Sustained typing below the externally calibrated human floor.
Panel B: In-browser assistant		
Named AI-extension signature	Content-script identity and iframe origin	A known AI extension is present in the browser session.
Corroborated assistant-compatible side panel	Window geometry plus an independent consultation or delivery trace	An assistant-compatible panel appears with separate mechanism evidence.
Panel C: AI chatbot		
Question round trip	Qualifying question-text copy; away episode and observed return; subsequent answer entry; event ordering	The question left the survey and an answer followed the return.
Canary marker	Instrument-owned clipboard payload (prompt injection) and exact returned marker	An instruction-following model processed the payload.
Split-screen copy–paste	Qualifying question copy followed by a qualifying paste on the same page without an away event	A same-page consultation-and-delivery route consistent with split-screen use.
Question copy and away episode, no delivery	Qualifying question copy plus focus, blur, or visibility evidence that the respondent left the survey	The question left the survey; no subsequent delivery event is observed.
Away episode followed by paste, no observed copy	Focus, blur, or visibility history followed by a qualifying paste into a survey-specific answer field	Text composed or obtained off-page is delivered on return; export is unobserved.
Qualifying paste only	Qualifying paste into a survey-specific answer field with no observed copy or away episode	Outside text entered the field while the preceding route remains unobserved.
Panel D: No AI usage		
No positive predicate satisfied	No qualifying rule above; weak or contextual traces may remain	No qualifying completion mechanism was observed.

Notes: This table lists the rules that assign a respondent to a group, the paradata each rule reads, and what a satisfied rule establishes. Context-only measures never classify on their own. Exact thresholds, guards, and validation fixtures are documented in the replication package.

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: This table presents the demographic composition of the Prolific sample, collected on June 3–4, 2026, and of the New York Fed Survey of Consumer Expectations (SCE). SCE shares are survey-weighted and computed from the public microdata for 2013–2025 with one record per respondent. Employment is collapsed to working (full- or part-time) versus not working.

Table 3: Detected respondent groups

Group or assignment route	Respondents	Percent
Autonomous AI	**15**	**1.2**
Direct field write	8	0.6
Synthetic input	3	0.2
Metronomic typing	3	0.2
Machine-cadence typing	5	0.4
In-browser assistant	**8**	**0.6**
Named AI-extension signature	1	0.1
Corroborated assistant-compatible side panel	7	0.5
AI chatbot	**278**	**21.5**
Question round trip	245	18.9
Canary marker only	3	0.2
Split-screen copy–paste	5	0.4
Question copy and away episode, no delivery	8	0.6
Away episode followed by paste, no observed copy	14	1.1
Qualifying paste only	3	0.2
Self-reported AI use, held out of classification	55	4.3
Of which also detected	28	2.2
Detected involvement	**301**	**23.3**
No AI usage	**993**	**76.7**
Total	**1,294**	**100.0**

Notes: This table presents the number of respondents in each detected group and the assignment route within each group.

Table 4: Attention, comprehension, and perceptual questions

	AI involvement	Control group	AI − control
	(1)	(2)	(3)
Lusardi–Mitchell Big Three (3 items)	89.9	78.2	11.7***
Compound interest	82.1	56.2	25.8***
Inflation / real return	93.0	87.4	5.7***
Risk diversification	94.7	91.0	3.7*
Numeracy (4 items)	94.3	85.6	8.7***
Percentage discount	97.7	99.0	-1.3
Lottery probability	93.7	82.2	11.5***
Disease risk	93.0	88.2	4.8**
Infection risk	92.7	73.0	19.7***
Cognitive reflection (3 items)	75.2	49.0	26.2***
Bat and ball	64.1	38.6	25.5***
Widget machines	80.7	53.8	26.9***
Lily pads	80.7	54.5	26.3***
Visual illusions (4 items)	27.5	29.3	-1.8
Checker-shadow	19.6	23.1	-3.5
Müller-Lyer	16.6	22.4	-5.7**
Café wall (parallel)	55.5	46.0	9.5***
Reverse Müller-Lyer	18.3	25.7	-7.4**
Observations	301	966	

Notes: This table presents the percent of attention, comprehension, and perceptual questions answered correctly by respondents with AI involvement (autonomous AI, in-browser assistant, or AI chatbot) and by the control group; the 27 respondents who disclose AI use in the debrief without a qualifying mechanical trace are excluded from both columns. Italic rows report the percent of a category's items answered correctly and indented rows the percent answering each item correctly; column (3) is the difference in percentage points. The financial-literacy items are the Lusardi–Mitchell questions on compound interest, inflation, and diversification, with the compound-interest item reworded as an open numeric response; the cognitive-reflection items are from Frederick (2005). $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 5: AI chatbot consultation and comprehension within respondent

	Correct (0/1)			log seconds	Consulted (0/1)
	(1)	(2)	(3)	(4)	(5)
AI chatbot consultation	0.173***	0.083***	0.178***	0.382***	
	(0.010)	(0.030)	(0.011)	(0.073)	
Prior approved studies (SD)			0.032***		−0.021**
			(0.006)		(0.009)
Item fixed effects	Yes	Yes	Yes	–	–
Page fixed effects	–	–	–	Yes	Yes
Respondent fixed effects	–	Yes	–	Yes	–
Observations	12,940	12,940	12,940	3,882	3,882

Notes: This table presents estimates from a respondent×item panel over the ten scored comprehension items (three cognitive-reflection, four numeracy, and three Lusardi–Mitchell financial-literacy items), which are on three survey pages; all 1,294 respondents are included. AI chatbot consultation is the explanatory variable in columns (1) to (4) and the outcome in column (5); column (2) is identified by the 93 respondents who consulted on some but not all three item pages. Column (4) uses the log of the survey engine's page dwell time on the respondent×page panel. Prior approved studies is the log count of a respondent's previously approved Prolific studies in standard deviation units; it is constant within respondent and enters only the specifications without respondent fixed effects. Standard errors clustered by respondent are in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 6: AI involvement by respondent group

	N	Detected (%)	Self-reported (%)
	(1)	(2)	(3)
All respondents	1,294	23.3	4.3
Gender			
Female	661	21.9	5.1
Male	633	24.6	3.3
Age			
18–29	271	29.5	6.6
30–44	330	30.6	3.9
45–59	383	20.4	2.9
60+	310	13.5	4.2
Education			
Bachelor's degree or higher	789	28.4	3.7
No bachelor's degree	505	15.2	5.1
Household income			
Under $60,000	485	18.1	4.7
$60,000 to $149,999	627	27.0	4.3
$150,000 or more	182	24.2	2.7
Employment			
Working	762	26.5	3.9
Not working	394	15.2	6.1

Notes: This table presents, for each respondent group, the share of respondents with detected evidence of AI involvement (autonomous AI, in-browser assistant, or AI chatbot) and the share disclosing AI use in the post-survey debrief, which is held out of the classification. Age, gender, and employment come from the Prolific profile; education and household income come from the demographics module. Respondents with missing group information appear only in the all-respondents row.

Table 7: Inflation expectations by AI involvement

	Mean	Median	SD	P10	P90	N
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: One-year-ahead inflation expectation (%)						
Control group	8.06	4.00	17.24	0.50	25.00	966
AI group	6.40	3.00	15.15	1.00	16.30	122
AI chatbot	6.93	3.00	15.00	1.60	14.00	107
Autonomous AI	2.60	4.00	16.24	−0.80	13.00	15
Panel B: Three-year-ahead inflation expectation (%)						
Control group	6.60	4.00	16.76	−3.00	20.00	966
AI group	3.02	3.00	7.70	2.00	5.00	122
AI chatbot	2.58	3.00	7.35	2.00	4.00	107
Autonomous AI	6.13	4.00	9.55	2.00	6.80	15
Panel C: Five-year-ahead inflation expectation (%)						
Control group	5.45	3.00	17.89	−5.00	20.00	966
AI group	4.16	3.00	6.55	2.00	5.00	122
AI chatbot	4.15	3.00	6.20	2.00	4.40	107
Autonomous AI	4.20	3.00	8.91	−1.20	5.00	15

Notes: This table presents inflation point expectations for the 122 respondents in the AI group and the 966 control-group respondents. The indented rows divide the AI group into 107 respondents with an AI chatbot trace on an inflation page (106 from the AI chatbot group of Table 3 and one in-browser assistant respondent) and 15 autonomous AI sessions. I exclude 179 respondents with AI evidence only on other survey questions and 27 who disclose AI use in the debrief without me detecting them. P10 and P90 denote the 10th and 90th percentiles.

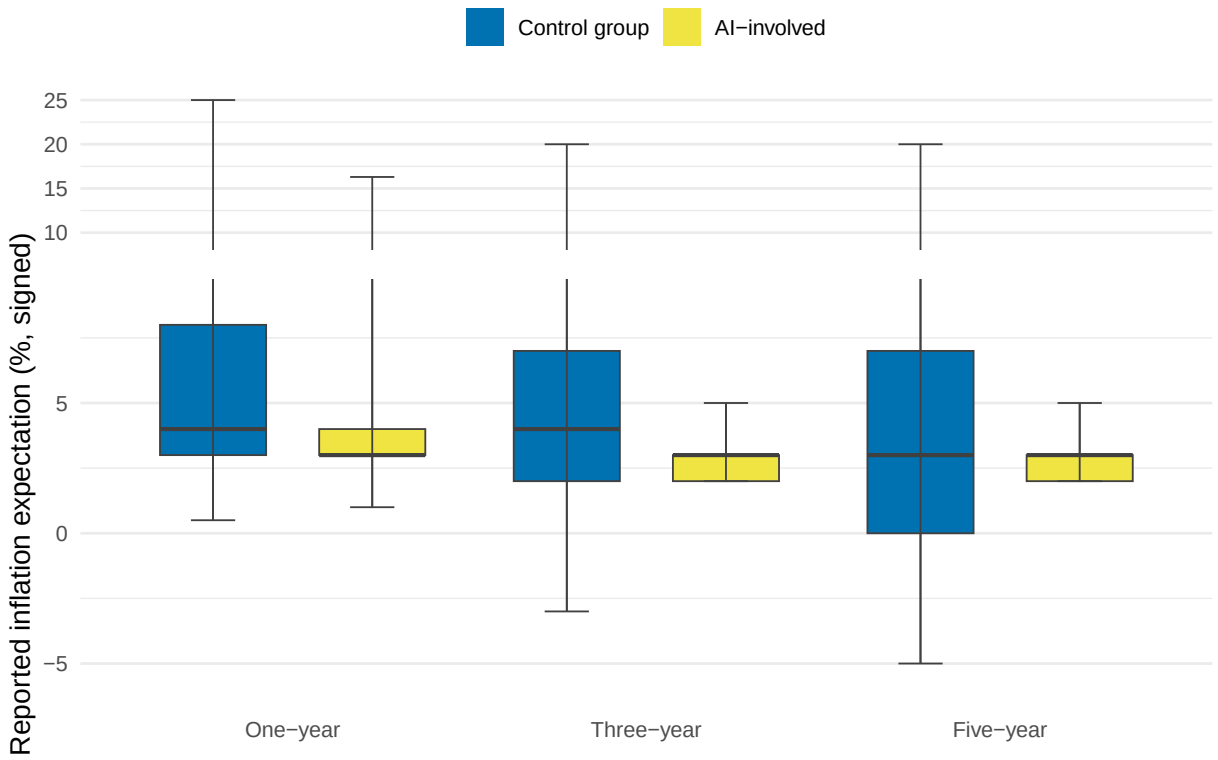


Figure 1: Inflation expectations by evidence of AI involvement on the inflation question, by horizon

Notes: This figure presents inflation point expectations at the one-, three-, and five-year horizons for the 122 respondents in the AI group and the 966 control-group respondents of Table 7. Boxes span the interquartile range, the central line is the median, and the whiskers span the 10th and 90th percentiles; the vertical axis is split above 9%. A two-sample Kolmogorov–Smirnov test rejects equality of the two distributions at every horizon ($p < 0.001$).

Table 8: AI involvement and the distribution of inflation expectations

	One-year	Three-year	Five-year
	(1)	(2)	(3)
$P(y \leq 0)$	-0.2	-10.3^{***}	-20.7^{***}
	(2.6)	(2.5)	(2.6)
	[10.0]	[18.6]	[26.5]
$P(y > 5\%)$	-20.3^{***}	-29.9^{***}	-21.4^{***}
	(3.7)	(2.8)	(3.2)
	[33.2]	[32.4]	[28.7]
$P(y > 10\%)$	-9.4^{***}	-17.0^{***}	-10.7^{***}
	(3.2)	(2.2)	(2.8)
	[19.5]	[17.9]	[16.9]
$P(y = 3)$	22.9^{***}	29.7^{***}	31.3^{***}
	(4.7)	(4.7)	(4.7)
	[15.6]	[16.1]	[15.2]
$P(y = 4)$	4.0	-1.5	-3.6
	(4.3)	(3.0)	(2.3)
	[22.0]	[12.0]	[8.2]
$P(y = 3) - P(y = 4)$	18.9^{**}	31.2^{***}	34.9^{***}
	(7.6)	(6.3)	(5.6)
Demographic controls	Yes	Yes	Yes
Observations	1,088	1,088	1,088

Notes: This table presents the coefficient on an AI-group indicator from separate linear probability models for each outcome and horizon, with controls for age, gender, education, household income, race, and employment. Coefficients and standard errors in parentheses are in percentage points; the bracketed value is the control-group mean of the outcome. The sample contains the 1,088 respondents of Table 7. The first outcome, $y \leq 0$, includes zero. The final row is the coefficient on AI involvement in a linear probability model for $\mathbf{1}[y = 3] - \mathbf{1}[y = 4]$. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 9: 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: This table presents inflation point estimates in percentage points from the New York Fed Survey of Consumer Expectations (SCE) public microdata and from 15 AI models. The pre-AI period is June 2013 to November 2022, the normal period January 2019 to December 2019, and the high-inflation period April 2022 to September 2022; SCE statistics are unweighted. The AI columns answer the SCE item text as a stand-alone question (single question) or embedded in the full survey instrument (complete survey), collected on May 11–13, 2026. "% at 3/3.5/4" is the share of point estimates equal to exactly 3.0, 3.5, or 4.0%.

Table 10: AI answers by model and prompt format

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	1.0	0.0	84.3
Equal-weight by model×prompt	–	886	3.5	3.5	0.4	1.0	0.0	84.3

Notes: This table presents one-year inflation point estimates by model and prompt format. Format A is the single-question call and format B the complete-survey call. Statistics are unweighted within each model-format cell; "% at 3/3.5/4" is the share of answers equal to exactly 3.0, 3.5, or 4.0%. The three summary rows give the pooled, equal-weight-by-model, and equal-weight-by-model×format statistics. The equal-weight rows weight each response by the inverse of its model's (or cell's) call count, so that every model (or cell) carries the same total weight, and compute all statistics on that weighted distribution; the N column repeats the call count. Two cells have fewer valid calls than the others: kimi-k2.6 format A ($N = 27$, where 13 calls returned empty) and gemini-3.1-pro-preview format B ($N = 9$).

Table 11: 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: This table presents persona-prompted AI inflation point estimates in percentage points. Six AI models each answer the full SCE instrument once for each of 231 demographic personas built to be representitive of the SCE sample. The first panel reports the persona pool at each horizon; the remaining panels split the one-year pool by age (under 40, 40–60, over 60), household income (under \$50k, \$50–100k, over \$100k), and education (college versus high school or some college). Responses with |value| $>$ 100 are dropped. "% at 3/3.5/4" is the share equal to exactly 3.0, 3.5, or 4.0%. Calls ran on May 13–14, 2026.

Table 12: 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, pre-AI (2013–22)	146,989	0.58	0.17	1.02	2.79	3.87
SCE, normal (2019)	15,752	0.60	0.19	0.98	2.62	3.45
SCE, high-inflation (2022)	7,562	0.56	0.16	1.08	3.24	4.50
Field, control group (2026)	966	0.51	0.06	1.27	3.81	3.42
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, pre-AI (2013–22)	147,082	0.57	0.18	1.02	2.77	4.00
SCE, normal (2019)	15,770	0.59	0.20	0.98	2.61	3.58
SCE, high-inflation (2022)	7,555	0.53	0.15	1.13	3.21	5.12
Field, control group (2026)	966	0.50	0.08	1.26	3.67	3.99
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: This table presents summary statistics of responses to the SCE ten-bin probabilistic inflation question. Bin probabilities are normalized to sum to one; responses whose raw probabilities sum to outside 50–150 are dropped. Columns (2)–(5) are averages across responses of the probability in the largest bin, an indicator for at least 90% in one bin, entropy, and the standard deviation implied by the SCE bin midpoints. Column (6) is the cross-sectional standard deviation of the midpoint-implied means. SCE rows use unweighted public microdata over the periods of Table 9; the field row is the control group; the AI rows pool the fifteen models by prompt format.

Table 13: Alarm rates and breakdown shares under replacement

Panel A: Reported statistic (mean over simulated waves)

Monitor	Replacement share $\lambda = 0$ (1)	$\lambda = 0.1$ (2)	$\lambda = 0.3$ (3)	$\lambda = 0.5$ (4)	Threshold (5)
Median	7.8	7.2	4.7	3.7	4.23
Mean	10.7	10.0	8.5	7.1	7.67
SD	17.5	16.8	15.0	12.9	16.00
IQR	8.9	8.5	6.7	4.8	8.00
P90−P10	29.2	26.6	20.4	13.9	23.00
P90	27.7	26.0	21.2	16.3	20.00
P95	46.3	44.4	36.8	27.9	37.93
Share $> 5\%$	64.2	58.0	45.0	32.1	34.77
Share $> 8\%$	45.3	41.0	31.8	22.7	28.35
Share $> 10\%$	29.0	26.3	20.5	14.6	18.51

Panel B: Alarm rate and breakdown share

Monitor	Replacement share $\lambda = 0$ (1)	$\lambda = 0.1$ (2)	$\lambda = 0.3$ (3)	$\lambda = 0.5$ (4)	λ^* (5)	90% CI (6)
Median	1.00	1.00	0.70	0.00	0.340	[0.220, 0.350]
Mean	0.98	0.92	0.74	0.28	0.415	[0.365, 0.510]
SD	0.83	0.73	0.24	0.00	0.200	[0.105, 0.265]
IQR[a]	0.60	0.46	0.02	0.00	–	–
P90−P10	0.92	0.76	0.05	0.00	0.245	[0.210, 0.270]
P90	0.87	0.81	0.40	0.01	0.255	[0.085, 0.285]
P95	0.92	0.85	0.48	0.03	0.295	[0.105, 0.350]
Share $> 5\%$	1.00	1.00	1.00	0.13	0.460	[0.445, 0.480]
Share $> 8\%$	1.00	1.00	0.78	0.00	0.380	[0.350, 0.425]
Share $> 10\%$	1.00	0.99	0.72	0.02	0.365	[0.330, 0.410]
No operational monitor fires	0.00	0.00	0.00	0.70		

Notes: This table presents the effect of replacing a fraction λ of survey answers with AI-model answers on ten monitored statistics of one-year inflation expectations. Each simulated wave draws $n = 1{,}000$ respondents with probability proportional to their SCE weight from one of the six months April–September 2022 and replaces a fraction λ of the answers with count-weighted draws from the fifteen-model AI answer pool; the few answers with $|x| > 100$ are removed from both sources. Columns (1)–(4) rest on 3,000 waves per replacement share. A monitor's threshold is the empirical 95th percentile of its 72 monthly values of 2014–2019, and an alarm fires when the wave statistic exceeds it. Panel A reports the mean statistic over the waves and, in column (5), its threshold. Panel B reports the alarm rate and, in columns (5)–(6), the breakdown share λ^*, the smallest λ at which the monitor stops firing on the pool $(1 - \lambda)F_{\text{episode}} + \lambda F_{\text{AI}}$ (grid step 0.005), and its 90% interval from 300 bootstrap replicates. The last row is the share of waves in which none of the ten monitors fires.

[a] The clean pooled IQR is 7.5, below its 8.0 threshold, so no breakdown share exists; Panel A averages finite-sample waves.

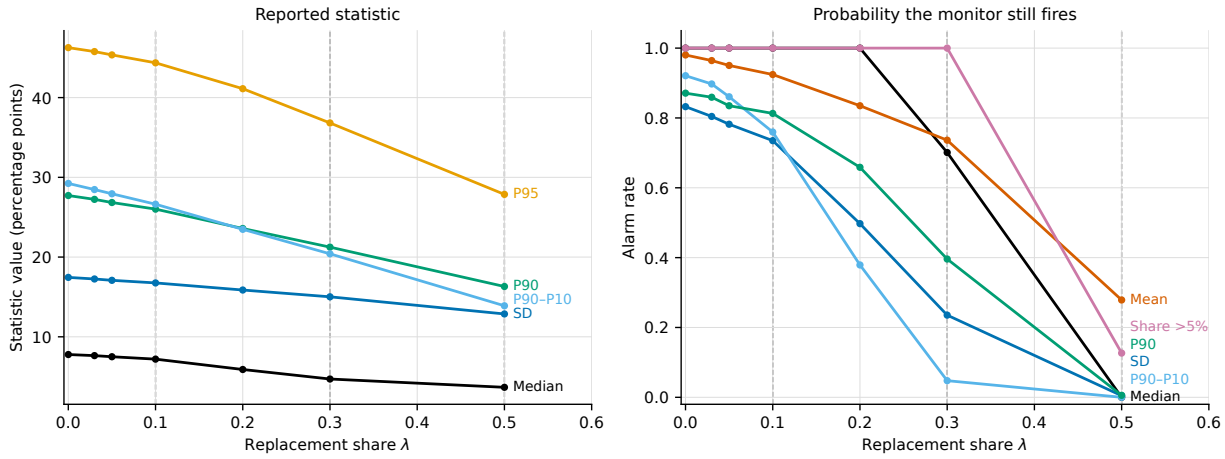


Figure 2: Tail erasure across the contamination grid

Notes: This figure presents, in the left panel, the mean reported median, standard deviation, 90th and 95th percentiles, and tenth-to-ninetieth-percentile range of one-year inflation expectations in percentage points across replacement shares λ, and, in the right panel, the probability that each of six one-sided monitoring rules still fires. Waves, thresholds, and the alarm rule follow Table 13. Vertical dashed lines mark $\lambda = 0.1$, 0.3, and 0.5.

Table 14: Alarm rates under alternative AI pools

	Baseline pool		Recentered pool		Persona pool	
Monitor	$\lambda = 0.3$	$\lambda = 0.5$	$\lambda = 0.3$	$\lambda = 0.5$	$\lambda = 0.3$	$\lambda = 0.5$
	(1)	(2)	(3)	(4)	(5)	(6)
Median	0.70	0.00	1.00	1.00	1.00	0.65
Mean	0.74	0.28	0.98	0.97	0.76	0.50
SD	0.24	0.00	0.14	0.00	0.23	0.00
IQR	0.02	0.00	0.00	0.00	0.03	0.00
P90−P10	0.05	0.00	0.04	0.00	0.06	0.00
P90	0.40	0.01	0.41	0.01	0.42	0.01
P95	0.48	0.03	0.48	0.02	0.53	0.06
Share $> 5\%$	1.00	0.13	1.00	1.00	1.00	0.78
Share $> 8\%$	0.78	0.00	1.00	0.87	0.83	0.01
Share $> 10\%$	0.72	0.02	0.73	0.02	0.70	0.03
No operational monitor fires	0.00	0.70	0.00	0.00	0.00	0.21

Notes: This table presents alarm rates at replacement shares $\lambda = 0.3$ and $\lambda = 0.5$ for three contaminant distributions. The baseline pool repeats columns (3)–(4) of Panel B in Table 13. The recentered pool shifts every AI answer up by 4.5 percentage points, so that the median of the pool equals the median of the pooled 2022 episode. The persona pool draws the contaminant from the persona-prompted model responses of Table 11 instead of the default pool. Each simulated wave draws $n = 1{,}000$ respondents with probability proportional to their SCE weight from one of the six months April–September 2022 and replaces a fraction λ of the answers, chosen at random, with independent draws from the respective contaminant pool; the few answers with $|x| > 100$ are removed from both sources. Each cell is based on 3,000 waves. A monitor's threshold is the empirical 95th percentile of its 72 monthly values of 2014–2019, and an alarm fires when the wave statistic exceeds it. The row *No operational monitor fires* is the share of waves in which none of the ten monitors listed fires.

Table 15: Alarm rates under partial assistance ($\lambda = 0.3$)

Monitor	Share of the own answer retained			
	$\alpha = 0$	$\alpha = 0.3$	$\alpha = 0.5$	$\alpha = 0.7$
	(1)	(2)	(3)	(4)
Median	0.70	1.00	1.00	1.00
Mean	0.74	0.83	0.89	0.93
SD	0.24	0.25	0.36	0.57
IQR	0.02	0.06	0.18	0.37
P90−P10	0.05	0.33	0.70	0.81
P90	0.40	0.56	0.78	0.84
P95	0.48	0.48	0.65	0.75
Share $> 5\%$	1.00	1.00	1.00	1.00
Share $> 8\%$	0.78	0.99	1.00	1.00
Share $> 10\%$	0.72	0.89	0.97	1.00
No operational monitor fires	0.00	0.00	0.00	0.00

Notes: This table presents alarm rates at replacement share $\lambda = 0.3$ when each contaminated respondent record is a blend of the respondent's own answer and the AI draw. Column (1), $\alpha = 0$, is pure replacement and repeats column (3) of Panel B in Table 13. Each simulated wave draws $n = 1{,}000$ respondents with probability proportional to their SCE weight from one of the six months April–September 2022 and replaces a fraction λ of the answers draws from the fifteen-model AI answer pool; a few answers with $|x| > 100$ are removed from both sources. Each cell is based on 3,000 waves. A monitor's threshold is the empirical 95th percentile of its 72 monthly values of 2014–2019, and an alarm fires when the wave statistic exceeds it. The row *No operational monitor fires* is the share of waves in which none of the ten monitors listed fires.

Appendix

Table A1: Persistence of upper-tail expectations in the SCE

Panel A: Probability of reporting above 10% in the following month

	Unweighted		SCE-weighted		
Consecutive-month pairs	$> 10\%$ now	All	$> 10\%$ now	All	Pairs
	(1)	(2)	(3)	(4)	(5)
All months, 2013–2025	0.58	0.12	0.60	0.16	145,493
Episode, April–August 2022	0.65	0.22	0.67	0.27	5,397

Panel B: Corroboration in the episode months, by current point expectation

	$> 10\%$	$\leq 10\%$
	(1)	(2)
Median next-month expectation	15.0	6.0
Respondent-month pairs	1,293	4,104
Share with density mean > 10	0.60	0.05
Respondent-months	1,726	5,735

Notes: This table presents the within-respondent persistence of upper-tail one-year point inflation expectations in the SCE panel. Panel A uses every pair of consecutive monthly interviews of the same respondent from June 2013 to April 2025. Columns (1) and (3) are the probability that a respondent reporting above 10% reports above 10% again in the following month, and columns (2) and (4) the unconditional probability. Weighted columns use the current-month SCE sampling weight. The episode rows use pairs whose current month is April–August 2022, so that the following month lies inside the episode. Panel B compares, within the episode months, respondents whose current point expectation is above 10% with the rest: the median of their next-month point expectation in percentage points (pair sample of Panel A) and the share whose independently elicited density forecast has an implied mean above 10 (respondent-months of April–September 2022 with a density forecast). Within the top decile of each month's point expectations, the rank correlation between the point value and the density mean is -0.03 (906 respondent-months), which is why the tail monitors are defined on tail shares and quantiles rather than on individual extreme values.

Table A2: Threshold calibration

Monitor	Threshold	Monthly mean		Held-out size		Pooled resampling
		2014–2019	2022 episode	Odd→even	Even→odd	
	(1)	(2)	(3)	(4)	(5)	(6)
Median	4.23	3.20	7.67	0.056	0.056	3.00
Mean	7.67	5.82	10.74	0.056	0.111	6.54
SD	16.00	13.81	17.48	0.028	0.139	15.42
IQR	8.00	4.75	8.96	0.000	0.000	6.00
P90−P10	23.00	18.18	28.91	0.056	0.111	22.00
P90	20.00	17.73	27.58	0.028	0.028	20.00
P95	37.93	28.12	47.50	0.056	0.056	30.00
Share $> 5\%$	34.77	27.26	64.24	0.028	0.139	29.50
Share $> 8\%$	28.35	22.52	45.46	0.000	0.167	24.70
Share $> 10\%$	18.51	14.45	29.15	0.056	0.083	16.30
Skewness	2.59	1.10	1.14	0.139	0.056	2.03
Kurtosis	19.36	16.27	10.18	0.167	0.000	20.01

Notes: This table presents the calibration of the monitoring thresholds. Column (1) is the one-sided threshold: the empirical 95th percentile of the 72 realized weighted monthly statistics of January 2014–December 2019. Columns (2) and (3) are the means of the monthly statistic over the 72 calibration months and over the six high inflation months April–September 2022. Columns (4)–(5) are a held-out size check: the calibration months are split by alternating position into 36 odd and 36 even months, the threshold is estimated on one half, and the realized size is the share of the other half's months that trigger an alarm. Column (6) is the one-sided 95th percentile obtained instead by resampling $n = 1{,}000$ draws from the pooled 2014–2019 answers (1,000 simulations). Skewness and kurtosis are reported as diagnostics: neither fires on the clean episode under the one-sided rule, and the episode mean of kurtosis lies below its calibration mean.

Table A3: Alarm rates with threshold uncertainty

	$\lambda = 0.1$		$\lambda = 0.3$		$\lambda = 0.5$	
Monitor	Rate	90% CI	Rate	90% CI	Rate	90% CI
	(1)	(2)	(3)	(4)	(5)	(6)
Median	1.00	[0.97, 1.00]	0.70	[0.02, 0.71]	0.00	[0.00, 0.00]
Mean	0.92	[0.88, 0.99]	0.74	[0.68, 0.86]	0.28	[0.17, 0.57]
SD	0.73	[0.61, 0.80]	0.24	[0.12, 0.33]	0.00	[0.00, 0.01]
IQR	0.46	[0.46, 0.46]	0.02	[0.02, 0.02]	0.00	[0.00, 0.00]
P90−P10	0.76	[0.74, 0.95]	0.05	[0.04, 0.39]	0.00	[0.00, 0.00]
P90	0.81	[0.80, 0.81]	0.40	[0.37, 0.40]	0.01	[0.00, 0.01]
P95	0.85	[0.66, 0.87]	0.48	[0.24, 0.52]	0.03	[0.01, 0.04]
Share $> 5\%$	1.00	[1.00, 1.00]	1.00	[1.00, 1.00]	0.13	[0.04, 0.42]
Share $> 8\%$	1.00	[1.00, 1.00]	0.78	[0.72, 0.95]	0.00	[0.00, 0.03]
Share $> 10\%$	0.99	[0.97, 1.00]	0.72	[0.69, 0.82]	0.02	[0.01, 0.12]
Skewness	0.00	[0.00, 0.02]	0.02	[0.01, 0.19]	0.29	[0.20, 0.58]
Kurtosis	0.00	[0.00, 0.00]	0.01	[0.01, 0.01]	0.43	[0.42, 0.48]

Notes: This table presents the alarm rates of Table 13, Panel B, together with 90% confidence intervals that reflect threshold-estimation uncertainty. Each of 300 bootstrap replications redraws the 72 calibration months with replacement, re-estimates every threshold, and re-evaluates the same 3,000 simulated waves against it; the interval is the 5th–95th percentile of the resulting alarm rates. Intervals are wide where a statistic crosses its threshold over a narrow range of replacement shares and degenerate where the threshold sits on an atom of the integer-valued answer distribution.

Table A4: ROC area under the curve by monitor

Monitor	$\lambda = 0.1$	$\lambda = 0.3$	$\lambda = 0.5$
	(1)	(2)	(3)
Median	0.999	0.957	0.868
Mean	0.992	0.967	0.848
SD	0.934	0.752	0.331
IQR	0.962	0.914	0.556
P90−P10	0.979	0.720	0.186
P90	0.944	0.774	0.428
P95	0.965	0.856	0.555
Share $> 5\%$	1.000	1.000	0.853
Share $> 8\%$	1.000	0.979	0.570
Share $> 10\%$	0.998	0.950	0.554
Skewness	0.580	0.709	0.840
Kurtosis	0.046	0.217	0.772

Notes: This table presents the area under the ROC curve for each monitor and replacement share λ. The curve sweeps the one-sided alarm threshold over the 72 monthly statistics of 2014–2019; the true-positive rate is the share of the 3,000 contaminated waves of Table 13 exceeding the threshold, and the false-positive rate is the share of the 72 calibration months exceeding it.

Table A5: Breakdown-share diagnostics

		Share of bootstrap replicates			Episode months	
Monitor	λ^*	Crossed	Never fires	Survives	Range of λ^*	Months firing
	(1)	(2)	(3)	(4)	(5)	(6)
Median	0.340	1.00	0.00	0.00	0.270–0.385	6 of 6
Mean	0.415	1.00	0.00	0.00	0.110–0.540	6 of 6
SD	0.200	1.00	0.00	0.00	0.135–0.330	5 of 6
IQR	–	0.31	0.69	0.00	0.010–0.265	5 of 6
P90−P10	0.245	1.00	0.00	0.00	0.105–0.215	5 of 6
P90	0.255	1.00	0.00	0.00	0.190–0.410	5 of 6
P95	0.295	1.00	0.00	0.00	0.020–0.405	6 of 6
Share $> 5\%$	0.460	1.00	0.00	0.00	0.395–0.500	6 of 6
Share $> 8\%$	0.380	1.00	0.00	0.00	0.275–0.430	6 of 6
Share $> 10\%$	0.365	1.00	0.00	0.00	0.180–0.470	6 of 6
Skewness	–	0.00	1.00	0.00	–	0 of 6
Kurtosis	–	0.00	1.00	0.00	–	0 of 6

Notes: This table presents diagnostics for the breakdown shares of Table 13, Panel B. Column (1) repeats λ^* on the pooled weighted episode. Columns (2)–(4) classify the 300 bootstrap replications that redraw episode respondents, AI-pool answers, and the 72 calibration months: the share in which the monitor fires at $\lambda = 0$ and stops firing on the grid, the share in which it does not fire on the clean replicate, and the share in which it still fires at the end of the grid ($\lambda = 0.6$). The confidence intervals in Table 13 use the replicates of column (2) only. Columns (5)–(6) repeat the exercise month by month within the episode: the range of the six monthly breakdown shares and the number of months in which the monitor fires on the clean month. A dash marks a monitor with no breakdown share because it does not fire on the clean distribution.

Table A6: Skewness and kurtosis under replacement

		Statistic value				Alarm rate			
Statistic	Threshold	$\lambda = 0$	$\lambda = 0.1$	$\lambda = 0.3$	$\lambda = 0.5$	$\lambda = 0$	$\lambda = 0.1$	$\lambda = 0.3$	$\lambda = 0.5$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Skewness	2.59	1.15	1.29	1.66	2.19	0.00	0.00	0.02	0.29
Kurtosis	19.36	10.14	11.06	13.86	19.08	0.00	0.00	0.01	0.43

Notes: This table presents the two shape statistics excluded from the operational comparison in Table 13. Each simulated wave draws $n = 1{,}000$ respondents with probability proportional to their SCE weight from one of the six months April–September 2022 and replaces a fraction λ of the answers, chosen at random, with draws from the fifteen-model AI answer pool; the few answers with $|x| > 100$ are removed from both samples. Each cell is based on 3,000 waves. A monitor's threshold is the empirical 95th percentile of its 72 monthly values of 2014–2019, and an alarm fires when the wave statistic exceeds it. Columns (2)–(5) are the mean of the statistic over the simulated waves and columns (6)–(9) the share of waves in which the one-sided upward rule fires. Neither statistic fires on the clean episode: the episode months lower kurtosis relative to the calibration months and leave skewness almost unchanged (Table A2, columns (2)–(3)). The alarms that appear at $\lambda = 0.5$ respond to the central spike that heavy replacement creates, not to the high-inflation episode.

Table A7: Contaminant shape and weighting variants

	Gaussian-matched pool			Unweighted draws		
Monitor	$\lambda = 0$	$\lambda = 0.3$	$\lambda = 0.5$	$\lambda = 0$	$\lambda = 0.3$	$\lambda = 0.5$
	(1)	(2)	(3)	(4)	(5)	(6)
Median	1.00	0.71	0.00	1.00	0.68	0.00
Mean	0.98	0.73	0.29	0.86	0.56	0.02
SD	0.85	0.23	0.00	0.25	0.00	0.00
IQR	0.61	0.03	0.00	0.00	0.00	0.00
P90−P10	0.93	0.05	0.00	0.36	0.00	0.00
P90	0.88	0.40	0.01	0.47	0.01	0.00
P95	0.92	0.48	0.02	0.48	0.02	0.00
Share $> 5\%$	1.00	1.00	0.13	1.00	1.00	0.02
Share $> 8\%$	1.00	0.78	0.00	1.00	0.53	0.00
Share $> 10\%$	1.00	0.71	0.02	0.90	0.18	0.00
Skewness	0.00	0.02	0.29	0.01	0.30	0.63
Kurtosis	0.00	0.01	0.44	0.00	0.41	0.95
No operational monitor fires	0.00	0.00	0.69	0.00	0.00	0.97

Notes: This table presents alarm rates for two variants of the replacement exercise of Table 13. *Gaussian-matched* replaces the empirical AI pool with normal draws matching its count-weighted mean and standard deviation; over the full replacement grid its alarm rates differ from the baseline by at most 0.05 in any operational cell. *Unweighted draws* samples episode respondents with equal probability instead of probability proportional to weight while keeping the weighted thresholds of Table A2; it is a diagnostic of mismatched weighting, not a recalibrated unweighted design. At $\lambda = 0$ the unweighted episode statistic lies below its weighted threshold for SD, IQR, P90−P10 and P95, which is why clean-episode firing drops for those monitors. Each simulated wave draws $n = 1{,}000$ respondents (with probability proportional to their SCE weight in the Gaussian-matched columns and with equal probability in the unweighted columns) from one of the six months April–September 2022 and replaces a fraction λ of the answers with independent draws from the respective contaminant distribution; the few answers with $|x| > 100$ are removed from both samples. Each cell is based on 3,000 waves. A monitor's threshold is the empirical 95th percentile of its 72 monthly values of 2014–2019, and an alarm fires when the wave statistic exceeds it. The row *No operational monitor fires* is the share of waves in which none of the ten monitors listed fires.

Table A8: Alarm rates by contaminating model

	Equal-share mixture			Single model per wave			By model, $\lambda = 0.3$	
Monitor	$\lambda = 0.1$	$\lambda = 0.3$	$\lambda = 0.5$	$\lambda = 0.1$	$\lambda = 0.3$	$\lambda = 0.5$	Lowest	Highest
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Median	1.00	0.71	0.00	1.00	0.70	0.02	0.61	1.00
Mean	0.91	0.74	0.31	0.92	0.72	0.31	0.65	0.77
SD	0.72	0.22	0.00	0.71	0.23	0.00	0.17	0.27
IQR	0.40	0.02	0.00	0.45	0.02	0.00	0.00	0.06
P90−P10	0.75	0.05	0.00	0.75	0.05	0.00	0.02	0.06
P90	0.81	0.40	0.01	0.80	0.38	0.01	0.35	0.41
P95	0.84	0.48	0.03	0.85	0.48	0.02	0.40	0.55
Share $> 5\%$	1.00	1.00	0.13	1.00	1.00	0.12	1.00	1.00
Share $> 8\%$	1.00	0.78	0.00	1.00	0.77	0.00	0.72	0.84
Share $> 10\%$	0.98	0.72	0.02	0.98	0.70	0.02	0.64	0.77
Skewness	0.00	0.02	0.27	0.00	0.02	0.29	0.00	0.03
Kurtosis	0.00	0.01	0.44	0.00	0.01	0.43	0.00	0.04

Notes: This table presents alarm rates when the contaminant draws are clustered by model instead of pooled. In the baseline of Table 13 every replacement answer is an independent draw from the pool of fifteen models. *Equal-share mixture* draws a model uniformly for each contaminated respondent and then an answer from that model's own responses. *Single model per wave* draws one model per wave and takes every replacement answer from it. Columns (7)–(8) report, across the fifteen single-model exercises at $\lambda = 0.3$, the lowest and highest alarm rate of each monitor. Averaged over waves, the clustered processes differ from the baseline by at most 0.06 in any operational cell of the full replacement grid, and by at most 0.02 in at least 94% of them. Each simulated wave draws $n = 1{,}000$ respondents with probability proportional to their SCE weight from one of the six months April–September 2022 and replaces a fraction λ of the answers with draws from the fifteen-model AI answer pool under the stated clustering; the few answers with $|x| > 100$ are removed from both samples. Each cell is based on 3,000 waves. A monitor's threshold is the empirical 95th percentile of its 72 monthly values of 2014–2019, and an alarm fires when the wave statistic exceeds it.

Table A9: Partial assistance with idiosyncratic noise

	$\sigma = 0.3$				$\sigma = 0.6$			
Monitor	$\alpha = 0$	$\alpha = 0.3$	$\alpha = 0.5$	$\alpha = 0.7$	$\alpha = 0$	$\alpha = 0.3$	$\alpha = 0.5$	$\alpha = 0.7$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Median	0.77	1.00	1.00	1.00	0.95	1.00	1.00	1.00
Mean	0.74	0.83	0.89	0.93	0.74	0.83	0.89	0.93
SD	0.24	0.25	0.36	0.57	0.23	0.25	0.36	0.57
IQR	0.02	0.06	0.18	0.37	0.02	0.06	0.19	0.38
P90−P10	0.05	0.32	0.70	0.81	0.06	0.32	0.70	0.82
P90	0.40	0.55	0.78	0.84	0.40	0.55	0.78	0.84
P95	0.48	0.48	0.65	0.75	0.48	0.48	0.65	0.75
Share $> 5\%$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Share $> 8\%$	0.78	0.99	1.00	1.00	0.78	0.99	1.00	1.00
Share $> 10\%$	0.72	0.89	0.97	1.00	0.72	0.90	0.97	1.00
Skewness	0.02	0.01	0.00	0.00	0.02	0.01	0.00	0.00
Kurtosis	0.01	0.01	0.00	0.00	0.01	0.01	0.00	0.00

Notes: This table repeats Table 15 with an idiosyncratic noise term $\varepsilon \sim N(0, \sigma^2)$ added to the blend before rounding, so that each contaminated respondent records round($\alpha \cdot$human$+(1-\alpha)\cdot$AI$+\varepsilon$, 1) at replacement share $\lambda = 0.3$. Wave construction, thresholds, and the alarm rule follow Table 13.

Table A10: Tail-targeted replacement

	Largest k replaced			Most extreme k replaced		
Monitor	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.3$	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.3$
	(1)	(2)	(3)	(4)	(5)	(6)
Median	1.00	1.00	0.27	1.00	1.00	0.98
Mean	0.57	0.08	0.00	0.79	0.64	0.00
SD	0.00	0.00	0.00	0.00	0.00	0.00
IQR	0.13	0.00	0.00	0.19	0.01	0.00
P90−P10	0.29	0.00	0.00	0.26	0.03	0.00
P90	0.30	0.00	0.00	0.48	0.03	0.00
P95	0.01	0.00	0.00	0.09	0.00	0.00
Share $> 5\%$	1.00	1.00	0.60	1.00	1.00	1.00
Share $> 8\%$	1.00	0.93	0.00	1.00	1.00	0.20
Share $> 10\%$	0.86	0.64	0.00	0.92	0.72	0.00
Skewness	0.00	0.00	0.00	0.00	0.00	0.00
Kurtosis	0.03	0.41	0.99	0.00	0.00	0.00
No operational monitor fires	0.00	0.00	0.40	0.00	0.00	0.00

Notes: This table presents alarm rates when the replaced answers are selected by value rather than at random. Under the largest-k rule the $k = \lceil \lambda n \rceil$ largest answers in a wave are replaced with AI-pool draws; under the most-extreme-k rule the k answers farthest from the weighted month median are replaced. Wave construction, thresholds, and the alarm rule follow Table 13. The row *No operational monitor fires* is the share of waves in which none of the ten monitors listed fires.