Representative Agents in Disguise: The Limits of LLM-Generated Synthetic Expectations

Joachim Rillo*

November 2025

Abstract

I evaluate whether large language models (LLMs) can be used to generate synthetic panel data on subjective expectations. Using demographic seeding and local-news conditioning, I benchmark LLM-generated forecasts of inflation and house prices against microdata from the Survey of Consumer Expectations (SCE). I document three main findings. First, even under strict date-restrictive prompting, LLMs exhibit look-ahead bias: when asked to provide point forecasts and qualitative risks of future outcomes, models occasionally reference salient post-cutoff events, and reasoning-enabled models in particular reveal extensive leakage. Second, while aggregate LLM expectations closely follow the aggregate time-series movements of SCE expectations, they fail to reproduce the rich cross-sectional heterogeneity that characterizes survey data. Variance decompositions reveal that SCE beliefs are dominated by persistent individual differences, whereas synthetic LLM expectations are dominated by time effects, suggesting behavior closer to a representative-agent model than to a panel of heterogeneous respondents. Third, the relationship between demographics and beliefs—which is large in magnitude, stable, and well-documented in the SCE—is weak, inconsistent, and an order of magnitude smaller in LLM-generated expectations. These results highlight fundamental limitations in using LLMs as substitutes for human respondents in simulating expectation formation.

JEL Codes: G11, G40, G5
Keywords: Surveys, Expectations, Inflation, LLMs

*Yale University: joachimjose.rillo@yale.edu

I am grateful to Stefano Giglio, Johannes Stroebel, Zhenhao Tan, Tiffany Yong, David Kust, Paul Jonathan Yoo, and Aurel Rochell for comments.

1 Introduction

The rise of large language models (LLMs) has transformed empirical research in economics and finance, enabling text-based measurement, prediction, and simulation at unprecedented scale (Giglio et al., 2025; Horton, 2023; Fedyk et al., 2024; Bybee, 2025; Wu et al., 2025). Within the literature on subjective expectations, this creates a natural question: can LLMs be used to generate a panel of synthetic belief data when human panels are costly or infeasible? As noted by Giglio et al. (2025), one potential use of LLMs is to simulate synthetic beliefs for settings where collecting human expectations is limited or costly. Acquiring repeated observations from the same cross-section of individuals is costly, so the idea of constructing panel datasets of subjective expectations using LLMs is inherently appealing. Indeed, some researchers argue that LLMs function as implicit computational models of human reasoning, making them promising experimental subjects when provided with information and preferences (Horton, 2023).

However, a central obstacle remains: LLMs may "peek" into the future. Because their training data often contain information extending beyond the prediction window, researchers risk look-ahead bias—drawing on knowledge that would have been unavailable to real respondents at the time (Sarkar and Vafa, 2024; Glasserman and Lin, 2023; Halawi et al., 2024; Lopez-Lira and Tang, 2025). For this reason, Ludwig et al. (2025) emphasize a no-leakage condition: the model's knowledge cutoff must not overlap with the researcher's sample. Alternatively, researchers may attempt to use date-restrictive prompting—explicitly "cutting off" an LLM's knowledge within the prompt itself (Wu et al., 2025). This strategy is particularly appealing when researchers want to maximize the time dimension of their panel. Furthermore, many LLMs, such as ChatGPT, are proprietary in nature: their training data are undisclosed, and their precise knowledge cutoff, therefore, cannot be observed.

In this paper, I simulate a short panel of expectations data focused on point forecasts of housing prices and inflation. I use the Survey of Consumer Expectations (SCE) by the Federal Reserve Bank of New York as the empirical "ground truth," sampling respondent demographics from the SCE to seed large language models (LLMs) with realistic personas. Motivated by the literature on the importance of local information in expectation formation (Malmendier and Nagel, 2011; Soo, 2018; Kuchler and Zafar, 2019), I also randomly sample local news media from the 3DLNews2 corpus (Ariyarathne and Nwala, 2024) to approximate each respondent's local information environment. This design allows synthetic agents to experience both demographic heterogeneity and geographically specific media exposure.

The central question I study is whether LLM-generated responses can reproduce the persistent cross-sectional heterogeneity in beliefs that is fundamental to the subjective expectations literature. While aggregate LLM-generated expectations closely track the aggregate time-series variation in SCE expectations—consistent with recent findings (Bybee, 2025)—they fail to reproduce the *distributional* features of human belief formation. A variance decomposition shows that, although the aggregate explanatory power of individual and temporal components appears similar for humans and LLMs, the sources of this variation differ significantly. Among SCE respondents, persistent individual differences account for the majority of variation: some individuals remain consistently optimistic or pessimistic over time. In contrast, LLM-generated expectations exhibit little individual persistence and are dominated by time variation, with the same simulated persona shifting beliefs substantially month to month. These patterns indicate that human respondents behave in line with heterogeneous-agent models, whereas LLM respondents adopt expectation-formation processes more closely resembling those of representative-agent models.

First, consistent with recent work (Sarkar and Vafa, 2024; Glasserman and Lin, 2023), I show that incorporating qualitative forecasts of future risks with point forecasts reveals leakage and look-ahead bias, even under explicit knowledge restriction. When asked to enumerate future risks, LLMs routinely reference events that fall outside the permitted information set. Under a restriction that limits the model's knowledge to pre–June 2019, the share of responses that nevertheless invoked COVID-19–related terms reached as high as 95.6%.

Second, I test whether conditioning on demographics, locations, and localized news can recover persistent individual heterogeneity, the empirical driver of belief heterogeneity (Kuchler et al., 2023; Giglio et al., 2021a). It does not: simulated personas shift their beliefs substantially from month to month, and the dominant component of variance is time, not the individual. These results suggest that LLMs align more closely with a representative-agent structure rather than the heterogeneous-agent models that capture survey-data heterogeneity.

Finally, I show that LLMs also fail to replicate the systematic relationships between demographics and beliefs observed in the SCE. While there are some systematic relationships documented for LLM respondents, these are an order of magnitude smaller than in human respondents.

Together, these findings—robust across different LLM models—suggest that while models can approximate human aggregate expectations, they cannot currently capture the individual-level persistence and heterogeneity that characterize genuine belief formation. This limitation likely reflects both the models' inability to internalize the behavioral structure of belief formation and the fact that researchers themselves do not yet fully understand its underlying factor structure. Therefore, I caution against treating LLMs as substitutes for behavioral test subjects. Beyond the issue of look-ahead bias, they are unable to replicate the complexities of belief formation, as well as empirical relationships between different variables such as demographics and expectations.

Related literature. This paper relates to several strands of literature. First, it relates to the literature that uses LLMs as human subjects in experiments and expectation generation (Horton, 2023; Bybee, 2025; Wu et al., 2025; Fedyk et al., 2024; Korinek, 2023, 2024). A central theme in this literature is that look-ahead bias presents a major challenge in the simulation of synthetic data (Engelberg et al., 2025; Sarkar and Vafa, 2024; Ludwig et al., 2025; Halawi et al., 2024; Lopez-Lira and Tang, 2025; Roberts et al., 2023; Golchin and Surdeanu, 2024; Bybee, 2025). For example, Sarkar and Vafa (2024) query LLMs with corporate earnings calls for various firms in 2019 and ask it to list future potential risks in 2020 with date-restrictive prompting, showing that LLMs inadvertently refer to COVID-19—an event which a realistic test subject should not have knowledge of in 2019. At the same time, LLMs may be able to reproduce certain patterns of human judgment, including systematic biases and preferences (Bybee, 2025; Fedyk et al., 2024; Horton, 2023).

A related literature examines date-restrictive prompting as a strategy to mitigate look-ahead bias (Faria-e Castro and Leibovici, 2024; Wu et al., 2025; Hansen et al., 2024; Crane et al., 2025). Wu et al. (2025)

argue that this approach can prevent LLMs from drawing on future knowledge; they ask LLMs about control questions under date restriction to verify the model knowledge cutoff, showing that models can follow instructions to ignore later facts (a fact corroborated by Faria-e Castro and Leibovici, 2024). Crane et al. (2025) further show that endowing LLMs with the demographics of professional forecasters can lead to similar predictions as to human-forecasts, but they achieve superior accuracy. They argue that look-ahead bias is not a driver of these results, as prompting a model to *recall* macroeconomic variables for each quarter increases errors relative to their forecasts under a demographic persona. While median human forecasts often outperform median AI forecasts out-of-sample, the best AI forecasters outperform the best humans.

I contribute to this literature in two ways. First, I show that, when supplementing point forecasts with *qualitative* elicitations of future risks, LLMs continue to inadvertently draw on future salient events despite explicit knowledge restrictions. Second, while aggregate LLM forecasts often track aggregate human forecasts, they fail to reproduce the rich cross-sectional patterns documented in the subjective expectations literature: their beliefs exhibit weak and inconsistent relationships with demographics, and synthetic variation is dominated by time effects rather than persistent individual heterogeneity.

I also contribute to the growing literature on subjective expectations (see, among others, Malmendier and Nagel, 2016, 2011; Kumar et al., 2015; D'Acunto and Weber, 2024; Kuchler et al., 2023; Kuchler and Zafar, 2019; Giglio et al., 2021a, 2025, 2021b). Within this literature, I relate to the literature on inflation and housing price expectations (see D'Acunto and Weber, 2024; Kuchler et al., 2023, for reviews). I compare the time-series and cross-sectional variation of short-term inflation and housing price expectations for LLM respondents and SCE respondents, showing that synthetic data cannot replicate the rich cross-sectional dispersion in survey data. Borrowing methodology from Giglio et al. (2021a), I use a similar variance decomposition to verify that, similar to subjective stock market return and cash flow expectations, inflation and housing price expectations are driven by persistent cross-sectional differences, whereas synthetic expectations are driven by significant changes in beliefs by the same LLM respondent over time.

Finally, I contribute to the literature on how local information environments shape expectation formation (Soo, 2018; Guillochon, 2024; Armona et al., 2019; Kuchler and Zafar, 2019). This work emphasizes that local news and salient local events meaningfully influence beliefs, motivating my use of commuting-zone-specific newspaper content when simulating a panel of expectations.

2 Data Description

2.1 The Survey of Consumer Expectations

I draw on microdata from the Survey of Consumer Expectations (SCE) conducted by the Federal Reserve Bank of New York for empirical comparison. Launched in 2013, the SCE is a nationally representative, internet-based panel survey of roughly 1,300 household heads, each participating for up to twelve months (Armantier et al., 2017). The survey elicits individual expectations about inflation, home prices, and other economic outcomes across different forecast horizons.

For both inflation and house price expectations, the SCE elicits responses through a sequence of questions. Respondents are first asked about the direction of change—whether they expect prices to rise or fall—over the next 12 months, followed by their best guess of the corresponding percentage rate. They are then asked to consider a longer horizon, specifically the 12-month period between 24 and 36 months from the survey date, and to provide their best guess of the expected rate of change for that period. For each outcome, the survey also collects a subjective probability distribution over predefined outcome bins (e.g., probabilities that inflation or house prices fall within certain intervals). This structure allows me to observe, for each respondent i, short-term (one-year) and medium-term (two-year) inflation expectations, along with demographic and geographic identifiers such as their commuting zone $c(i)$, from June 2013 through December 2024. In this paper, I focus on short-term 1-year point forecasts, which provide a direct measure of individual beliefs about expected inflation and house-price growth.

Table 1: Respondent Summary Statistics

	Mean	SD	P10	P25	P50	P75	P90
Subjective Expectations							
Expected 1Y Inflation	5.89	11.2	-1	2	3.50	7	15
Expected 1Y House Price	5.38	9.75	-5	2	5	10	15
Demographics							
Age	50.3	15.5	30	38	50	62	70
Male	0.52	0.50	0	0	1	1	1
Education - College	0.56	0.50	0	0	1	1	1
Education - Some College	0.32	0.47	0	0	0	1	1
Education - High School	0.11	0.32	0	0	0	0	1
Household Income - Under 50k	0.34	0.47	0	0	0	1	1
Household Income - 50k - 100k	0.35	0.48	0	0	0	1	1
Household Income - $\geq$ 100k	0.30	0.46	0	0	0	1	1

Note: Table reports summary statistics on survey respondents for both demographics and subjective expectations in the SCE from June 2013 to December 2024. Point forecasts of inflation and house prices are winsorized at the top and bottom 1% to address extreme outliers, such as expected 1-year inflation and 1-year house prices being 11,200% and 10,000% respectively.

Table 1 summarizes the demographic and expectations data in the SCE. Consistent with the survey's sampling frame—which targets household heads—respondents are older on average (mean age 50). The sample also spans a broad range of education and income levels: 56% hold a college degree and roughly one-third earn more than $100,000 annually.

I also include summary statistics on expected inflation and house prices. On average, respondents report expected inflation and house prices in 1 year to be 5.89% and 5.38%. As documented in the subjective expectations literature, there is substantial dispersion in expected inflation and expected house prices across responses (see Giglio et al., 2021a, 2025; Kuchler et al., 2023, for further discussion on this). In the 10th percentile, respondents expect 1-year deflation of -1%, while in the 90th percentile, respondents expect inflation of 15%. This significant dispersion, which reflects both cross-sectional differences and individual updating over time, provides a natural empirical benchmark against the behavior of LLM-generated expectations. I later decompose the source of this panel variation for both SCE and

LLM respondents.

2.2 3DLNews2: Local U.S. Media

Motivated by recent evidence that LLMs can generate meaningful economic expectations when applied to news articles (Bybee, 2025), and by the broader literature showing that local information environments shape belief formation (Soo, 2018; Armona et al., 2019; Kuchler and Zafar, 2019; Guillochon, 2024), I incorporate commuting-zone-specific newspaper content into LLM-generated expectations. Conditioning each synthetic respondent on a local news article allows the LLM to receive a localized information shock analogous to what SCE respondents experience in their own media markets. To construct the panel of local-media-based expectations, I begin with the **3DLNews2** corpus (Ariyarathne and Nwala, 2024), a comprehensive dataset of U.S. local news articles (1995–2024) that includes metadata on the source outlet, publication date, and geographic coordinates. The corpus aggregates stories scraped from local newspapers, television stations, and radio broadcasters across all 50 states, providing broad coverage of regional business and economic topics.[1]

To focus on news that is plausibly relevant to economic expectations, I restrict the corpus to articles with clear business or economic content. I use a two-step screening procedure applied to each article's title, snippet, and body text.

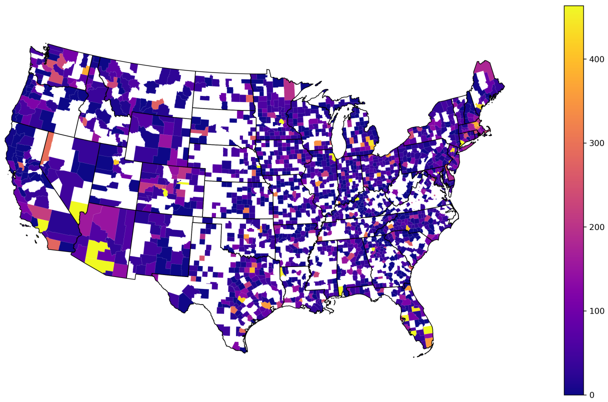
First, I compile a dictionary of "anchor" phrases capturing macroeconomic, housing, labor-market, and financial-market terms. Word boundaries are included to avoid spurious matches, ensuring that these expressions appear as whole words or phrases rather than as substrings of longer words. Examples include *housing market/prices/starts/affordability*, *home prices*, *mortgage rates*, *stock market / S&P 500 / Dow Jones / Nasdaq*, *interest rates / Federal Reserve / FOMC*, *unemployment rate / job growth / layoffs*, and *local economy / small business*. Second, I apply a proximity rule that flags economically related terms appearing near one another (within approximately 40 non-word characters). Examples include *(home/house/property/housing)* near *(sale/value/market/price/afford*)*, or *(job/employment/unemployment/hiring/layoff*)* near *(report/data/outlook/rate/claim*)*. Appendix A provides a full list of the anchor words and proximity phrases used in the filtering process.

I then run a two-stage classification procedure. I first retain articles whose titles or snippets match either the anchor or proximity patterns. Articles that do not pass this screen are then checked for anchor phrases in the body text. Any article that matches at either stage is included in the economic-news sample.

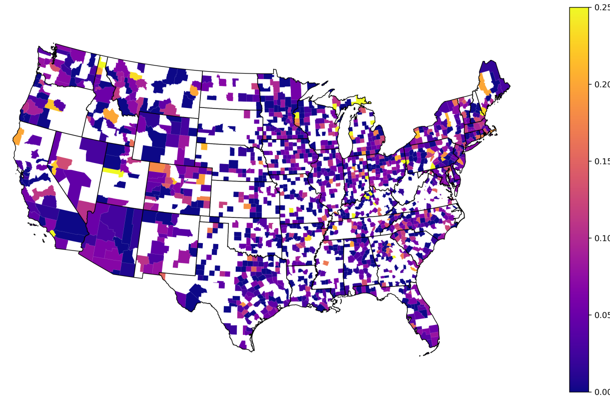
[1] See Ariyarathne and Nwala (2024) for details on data collection and cleaning procedures.

Figure 1: Local Media Concentration and Economic News Share by County (June 2022 to June 2023)

(a) Local Media Concentration



(b) Economic News Share



Note: Figure 1(a) reports the local media concentration from June 2022 to June 2023, where I sum the total amount of local media coverage by county. Figure 1(b) reports the share of business and economic articles in total local news coverage for each FIPS code over June 2022 to June 2023, where data are available. For each FIPS, the share is calculated as the ratio of business/economic articles to the total number of articles. Color limits are set to the 1st and 99th percentiles to reduce the visual influence of outliers.

Figures 1(a) and 1(b) highlight county-level information on local media concentration and relative salience of economic news from June 2022 to June 2023, the time frame I later use for my simulation analysis. Figure 1(a) highlights that local media volume seems to be concentrated in large media markets. News coverage is highly concentrated in the Northeast—most prominently in Washington, D.C., Boston, and New York City—and along the West Coast, including Los Angeles and other major California media markets. Conversely, the economic share shows substantial cross-sectional dispersion that does not simply mirror volume: many medium-volume counties display relatively high economic salience, while several high-volume markets are only middling on share. This suggests heterogeneity in editorial focus and local economies, not just scale effects. Notably, a few sparsely covered counties (e.g., Macon County, NC; Schuyler County, NY) reach shares near 100%, reflecting small denominators rather than unusually high economic intensity. For the simulations, I aggregate county-level articles to the commuting-zone level, as SCE microdata are publicly available only at that geography. Importantly, Figure 1 also shows that local news coverage is extensive across the United States, a desirable feature for constructing localized information treatments.[2]

3 Look-Ahead Bias in LLMs

Many potential issues arise when LLMs are used to simulate forecasts for the future. One prominent issue is look-ahead bias, which occurs when a model inadvertently incorporates information that would not have been available at the time of the supposed forecast. Because LLMs are trained on vast amounts of text data up to a certain cutoff date, they may retain latent knowledge of future events that occurred after the point in time they are meant to be "predicting."

[2]Appendix B shows the corresponding figures from 2013 to 2024. Although there is more coverage across counties, the spatial variations are similar.

The literature has increasingly recognized this problem. In particular, Ludwig et al. (2025) develop a framework for estimation and prediction using LLMs. Their discussion of prediction is particularly relevant here, where a researcher may ask an LLM to generate forecasts about the future. They assert that there must be no "leakage" or overlap between the LLM's training dataset and the researcher's sample. Relatedly, Sarkar and Vafa (2024); Glasserman and Lin (2023) show that LLMs used to predict a future outcome based on textual data may use information from the time of the outcome rather than the time of the text.

On the other hand, Wu et al. (2025) argue that date restrictive prompting can prevent hindsight bias. They then test the LLM's awareness of salient events between 2001 and 2020 and show that the LLMs report no awareness. In this section, I show that, even with the implementation of date-restrictive prompting, LLMs can still inadvertently draw on post-cutoff events to make point-forecasts. A corollary of this finding is that the no-leakage condition proposed by Ludwig et al. (2025) *must* be satisfied to successfully prevent look-ahead bias.

3.1 Event Awareness and Look-Ahead Bias

First, I follow the prompting procedure outlined in Wu et al. (2025), providing the LLM with a sample persona and imposing a date restriction—specifically, a cutoff of June 2010. I include three factual questions: one referring to a salient event that occurred before the cutoff (the Lehman Brothers bankruptcy) and two referring to salient events that occurred after it (Donald Trump election and Twitter being rebranded as X). In addition to replicating their prompting design, I further elicit the model's step-by-step reasoning to examine how its internal logic supports each answer and to identify potential instances of temporal leakage or look-ahead bias.

Table 2: Example of LLM response to time-restricted forecasting prompt (cutoff = June 2010).

PROMPT	RESPONSE
SYSTEM: You are simulating a human respondent as of June 2010 (cutoff = 2010-06-30). You MUST NOT reference or rely on any events, developments, real data, or official statistics that occurred after the cutoff date. Think step by step and SHOW your reasoning. Your FINAL ANSWERS must be exactly "yes" or "no". You may not use any token such as "unknown", "uncertain", or "UNKNOWN_DUE_TO_CUTOFF". **USER:** Persona: 45-year-old man, works as a tech and media analyst in San Francisco. Prides himself on making bold, forward-looking predictions. Survey date: June 2010. Knowledge Restriction: Answer based on your expert observations and general sense of the economy and culture at the time. Questions: 1) Lehman Brothers bankruptcy? 2) Donald Trump elected U.S. President? 3) Twitter being rebranded as X?	Q1 THOUGHT: The Lehman Brothers bankruptcy occurred in 2008. The survey date is June 2010, so the event is well-documented and occurred before the survey. The respondent would know this. Q1 FINAL ANSWER: yes Q2 THOUGHT: The question asks if Donald Trump was elected as U.S. President in 2010. Trump was elected in 2016, so the event is in the future. The respondent would recognize this as something that has not yet happened. Q2 FINAL ANSWER: no Q3 THOUGHT: The rebranding of Twitter as X occurred in 2023. The survey date is June 2010, so the event is in the future. The respondent would not be aware of such an event. Q3 FINAL ANSWER: no

Note: Table shows a sample response from prompting the Qwen3:1.7b model.

Table 2 presents a sample response generated by the Qwen3-1.7B (Thinking) model released in 2025 (Yang et al., 2025).[3] The model's final answers align with the findings of Wu et al. (2025): it correctly distinguishes between events occurring before and after the specified cutoff. However, the reasoning process underlying these answers reveals the presence of look-ahead bias. The model is not necessarily unaware of post-cutoff events; rather, it appears to check the dates of such events and compare them with the cutoff before producing its final response. Consequently, even under date-restrictive prompting, LLMs remain susceptible to look-ahead bias because temporal instructions at the prompt level cannot fully insulate them from post-cutoff information embedded in training data.

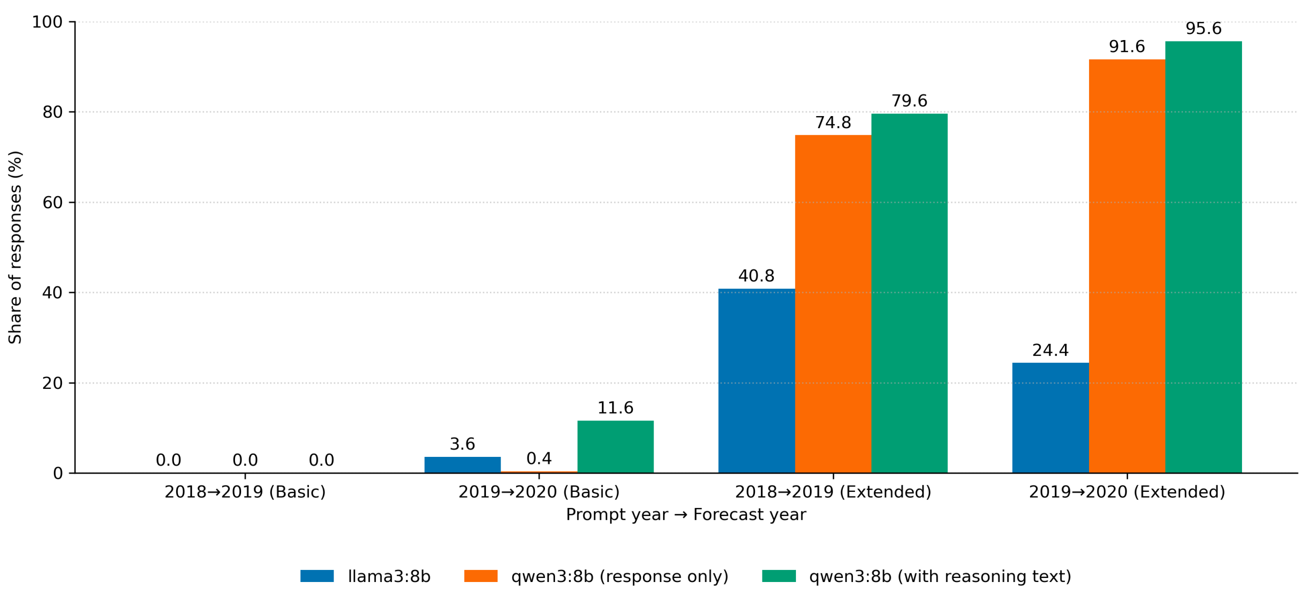
[3]The knowledge cutoff of this model is not known, but it is clear that the model has been trained on all of these events, given its awareness of all 4 events.

3.2 Date-Restricted Forecasts and Risk Elicitation

Building on this intuition, I elicit forecasts of expected one-year inflation using date-restricted prompting to systematically assess the prevalence of look-ahead bias in LLM-generated predictions. Thus, in addition to eliciting point forecasts of inflation, I follow Sarkar and Vafa (2024) and also elicit qualitative assessments of economic risks over the subsequent year. I construct two prompts—one as of June 2018 and another as of June 2019—to capture beliefs about inflation and economic risks for June 2019 and June 2020, respectively. I use 2 different LLM models for this exercise: Llama3:8b (Grattafiori et al., 2024) and the thinking model Qwen3:8b (Yang et al., 2025). I simulate 250 responses with each model for both date-restrictions.

In order to test for look-ahead bias, I follow Sarkar and Vafa (2024) and search the model outputs for two sets of terms. The first set includes direct references to the pandemic, specifically the words *"COVID-19"* and *"pandemic"*. The second, broader set also captures related phrases that became salient during the pandemic period—such as *"supply chain"*, *"disease outbreak"*, and *"coronavirus"*. The emergence of COVID-19 provides an ideal test case because it was an unanticipated, highly salient global shock that profoundly affected inflation dynamics and economic conditions in 2020, yet lay entirely outside the information set of any agent—or model—reasoning as of mid-2019.

Figure 2: Leakage Rates



Note: Figure shows the percentage of responses that contained the words from the basic set of terms ("COVID-19", "pandemic"), as well as from the extended set of terms using 250 simulations of both llama3:8b and qwen3:8b. I focus specifically on date-restrictive prompting from 2018 to generate predictions for 2019, and from 2019 to generate predictions for 2020. I use llama3:8b and qwen3:8b as sample models. For qwen3:8b, I test the leakage rates for both the response, as well as the response and its corresponding reasoning text.

Figure 2 displays the results of these simulations, plotting the share of responses that include COVID-19-related terms (which I call the leakage rate). I further exploit the fact that qwen3 reveals its reasoning text prior to giving a final answer. Thus, I analyze 3 variations of responses: the llama3:8b responses, the qwen3:8b responses excluding its train of thought, and the qwen3:8b responses including its train of thought. I find that from 2019 to 2020 forecasts, leakage rates increased for both the basic and extended vocabulary sets. Interestingly, llama3:8b's leakage rate drops from 40.8% to 24.4% under the extended vocabulary definition. While all models display a 0 percent leakage rate for the basic vocabulary set in the 2019 forecasts, this rate rises substantially once the extended vocabulary is included, suggesting that the models may be inadvertently drawing on information beyond their specified date restriction.

Moreover, the inclusion of reasoning text significantly increases qwen3:8b's leakage rate. Its leakage rate based on the basic vocabulary set for the 2020 forecasts increases from 0.4% to 11.6%, with notable increases for the extended vocabulary set for both the 2019 and 2020 forecasts. Importantly, even thinking models remain susceptible to look-ahead bias: when excluding their reasoning text, qwen3:8b still exhibits a leakage rate of 0.4% in the basic vocabulary set, indicating that it references COVID-19 as a potential risk despite date-restrictive prompting. Its leakage rate substantially increases when including its reasoning text to 11.6% for basic vocabulary with the 2020 predictions, as well as to 79.6% and 95.6% for the 2019 and 2020 predictions respectively under the extended vocabulary set.

These patterns also highlight a structural difference between reasoning-enabled and standard LLM models. The reasoning text produced by qwen3:8b exposes intermediate steps that appear to activate latent associations with salient risks in its training dataset, even when such associations should be unavailable under explicit knowledge restriction. These results suggest that the chain-of-thought style decoding may amplify temporal leakage by surfacing patterns learned during training rather than suppressing them.

Taken together, these results indicate that even when constrained by date-restricted prompts, models retain latent knowledge of future events when asked to make both point forecasts and qualitative forecasts about future economic risks. The leakage rate should thus be interpreted as a conservative lower bound, reflecting only explicit mentions of one salient future shock—COVID-19—while overlooking more implicit forms of foresight.

4 Panel Data Simulation

Despite explicit date restrictions, the preceding experiments reveal that LLMs often retain latent knowledge of future events, indicating that look-ahead bias cannot be fully mitigated through prompting alone. Given this limitation, I next examine a complementary dimension of validity: whether LLM-generated expectations nonetheless reproduce the empirical regularities of beliefs observed in survey data. In particular, I test whether their cross-sectional and temporal patterns align with well-documented stylized facts in the literature on household and investor expectations.

4.1 Methodology

The literature on subjective expectations documents two robust empirical patterns. First, there is far greater cross-sectional disagreement than time-series variation for point forecasts of macroeconomic outcomes. Second, observable characteristics exhibit systematic relationships with beliefs, but account for only a limited share of their variation (Giglio et al., 2021a; Kuchler et al., 2023).

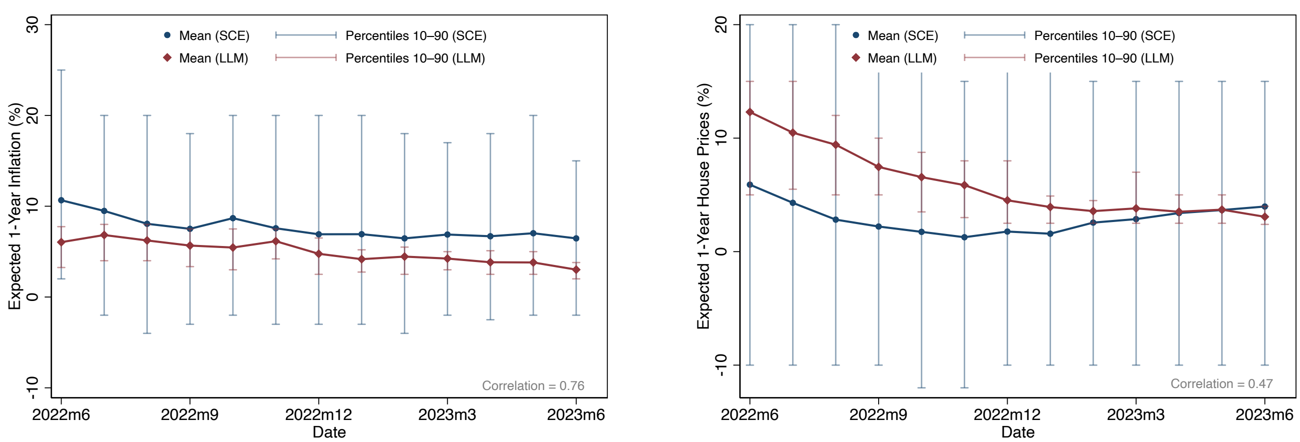
To assess whether LLM-generated forecasts replicate these empirical patterns, I construct a simulated panel of model-based expectations. Following Bisbee et al. (2024); Argyle et al. (2023), I provide llama3:8b with demographic personas sampled directly from SCE respondents between June 2022 and July 2023. This yields a representative cross-section of individuals consistent with the underlying survey demographics. Importantly, I ensure the same individuals respond for every monthly survey wave to obtain a balanced panel and observe sufficient within-individual variation over time. I further extend this framework by also conditioning the prompts on local media exposure. For each respondent–month pair, I randomly draw one local news article published in the same commuting zone (CZ) and month. If no articles are available for that CZ–month, I draw from adjacent CZs; if no neighboring CZs contain articles, no media is included. This procedure allows each simulated persona to be exposed to a geographically and temporally aligned information set that approximates the media environment a real respondent in that location and period would have faced. In the simulated sample, 64.3% of (LLM) respondents receive a random news article in their CZ, 32.4% receive one from a neighboring CZ, and 3.4% do not receive any article.

4.2 Cross-Sectional Disagreement and Time-Series Variation

First, I compare the aggregate time-series variation and cross-sectional disagreement for SCE respondents and LLM-generated expectations. To motivate this analysis, I begin by illustrating these two dimensions of variation for both LLM and human respondents.

Figure 3: Human vs LLM Aggregate Expectations (June 2022 to June 2023)

(a) Expected 1Y Inflation (b) Expected 1Y House Prices



Note: Figure 3 plots the average expectations for both human and LLM respondents from June 2022 to June 2023, along with the 10th and 90th percentile reported expectations for each wave. SCE responses are winsorized at the top and bottom 1% to reduce the impact of outlier responses on the analysis. Figure 3(a) plots the aggregate responses for expected 1-year inflation, while Figure 3(b) does so for expected 1-year house prices.

Figure 3 plots the average time-series of both expected 1-year inflation and expected 1-year house prices, along with their 10th and 90th percentile values. As consistent with Kuchler et al. (2023); Giglio et al. (2021a), SCE respondents display far greater cross-sectional disagreement than time-series variation: the 10th and 90th percentile bands are much larger than the movement in aggregate expectations. Conversely, LLM responses do not exhibit the same patterns: they exhibit far less cross-sectional disagreement, but nevertheless follow the *aggregate* time-series trends of survey respondents quite closely. The correlation between SCE respondents and LLM respondents is 0.76 for expected 1-year inflation and 0.47 for expected 1-year house prices, magnitudes similar to those documented in Bybee (2025).

One possible story behind this result is that look-ahead bias is the principal driver of the lack of cross-sectional disagreement in LLM respondents. If an LLM has already seen the "future" inflation rate in its training data, then its next-token prediction mechanism will tend to anchor on that value, making deviations unlikely. I therefore include information from April to June 2023, which are months beyond llama3:8b's knowledge cutoff of March 2023. Nevertheless, I still find very little cross-sectional disagreement. Thus, it is clear that, while LLMs can closely follow aggregate *time-series* expectations of survey measures, they are still unable to replicate the core within-survey patterns documented in the subjective expectations literature.

4.3 Variance Decomposition

I formalize this intuition by running a variance decomposition on the sources of the panel variation. For both SCE and LLM respondents, I investigate whether it is the same respondents who are always pessimistic or optimistic, or whether respondent expectations change significantly over time. To do this, I follow Giglio et al. (2021a) and run the following regressions for each belief $B_{i,t}$ for individual i at survey wave t:

$$B_{i,t} = \chi_t + \epsilon_{1,i,t} \tag{1}$$

$$B_{i,t} = \psi_i + \epsilon_{2,i,t} \tag{2}$$

$$B_{i,t} = \psi_{3,i} + \chi_{3,t} + \epsilon_{3,i,t}. \tag{3}$$

Equation 1 estimates a set of time fixed effects χ_t that absorb time variation in individual beliefs, equation 2 estimates a set of individual fixed effects ψ_i that absorb the average belief over time for each respondent, and equation 3 estimates both individual and time fixed effects for each respondent. I then analyze each equation's explanatory power for the variation in subjective expectations.

Table 3: Decomposing the Variation in Beliefs: Individual and Time Fixed Effects

	Panel A: R^2 (percent) of panel regression, SCE	
	Expected 1Y Inflation	Expected 1Y House Prices
Time FE	1.8	3
Individual FE	47.2	52.3
Time + Individual FE	47.8	54
Unique Respondents	15443	15458

	Panel B: R^2 (percent), LLM-generated expectations	
	Expected 1Y Inflation	Expected 1Y House Prices
Time FE	36.9	63.8
Individual FE	6	3.1
Time + Individual FE	42.9	67
Unique Respondents	200	200

Note: Table reports the R^2 values corresponding to the three regressions from Equations (1), (2), (3) and the number of unique respondents. Panel A runs the variance decomposition for SCE respondents from June 2013 to December 2024, while Panel B does so from June 2023 to June 2024. I only include SCE respondents who have responded to at least five waves. Each row corresponds to a different survey question that is used as the dependent variable.

Panel A of Table 3 shows the variance decomposition for both expected 1-year inflation and expected 1-year housing prices for SCE respondents, highlighting the dominance of individual fixed effects as in the subjective expectations literature. Panel B of Table 3 shows the variance decomposition for LLM respondents. Interestingly, while the *combined* explanatory power of time and individual fixed effects is similar to those of the SCE respondents, LLM respondents are unable to mimic the persistent optimism and pessimism documented in the literature. Instead, each LLM respondent significantly changes their beliefs over time, which leads to most of the variation in beliefs being explained by time fixed effects. There is also a possibility that these results could be LLM-specific, given that llama3:8b is a smaller LLM that may not perform as well as bigger LLMs. Appendix E conducts the same experiment with Gemini 2.5 Flash, where I find similar results.

Beyond the R^2 values, the decomposition reveals a deeper structural difference in belief formation. On one hand, SCE respondents exhibit substantial persistence in their subjective views, which are consistent with heterogeneous agent models (e.g., Kaplan et al., 2018; Kaplan and Violante, 2018). On the other hand, LLM respondents exhibit time-varying beliefs with each new survey wave. In this sense, LLM-generated expectations resemble a representative-agent forecast more than a panel of heterogeneous individuals; these expectations move largely in unison and show little respondent-specific persistence. A corollary to these results is that simulating an individual's local information environment is not enough to generate the substantial cross-sectional dispersion observed in the SCE and other survey data.

5 Beliefs and Demographics

Having established these differences between synthetic and SCE data in the panel dynamics of belief formation, I now test the second important finding in the subjective expectations literature: that, although demographics have a systematic relationship with beliefs, they explain very little of the variation in beliefs. To examine this, I estimate the simple regression:

$$B_{i,t} = \alpha + \gamma X_{i,t} + \epsilon_{i,t}, \tag{4}$$

where $X_{i,t}$ represents a vector of control variables such as age labels, education, income, and gender. Importantly, I do *not* include survey wave fixed effects in this regression. Given that time fixed effects explain most of the variation in LLM beliefs, I focus only on the explanatory power of demographics.

Table 4: Beliefs and Demographics Regression

	$B_{i,t}$: SCE Respondents		$B_{i,t}$: LLM-Generated	
	Expected 1Y Inflation	Expected 1Y House Prices	Expected 1Y Inflation	Expected 1Y House Prices
Age: 40 to 60	0.663***	0.384***	0.037	-0.005
	(0.068)	(0.059)	(0.085)	(0.173)
Age: Over 60	0.079	0.549***	0.039	-0.146
	(0.069)	(0.061)	(0.097)	(0.196)
Education: College	-3.629***	-2.341***	0.003	-0.001
	(0.118)	(0.097)	(0.128)	(0.270)
Education: Some College	-1.808***	-1.546***	-0.132	-0.063
	(0.127)	(0.102)	(0.126)	(0.266)
Income \$50k - \$100k	-1.721***	-1.726***	-0.132	-0.141
	(0.072)	(0.061)	(0.095)	(0.190)
Income $\geq$ \$100k	-2.473***	-2.676***	-0.227**	0.130
	(0.069)	(0.062)	(0.103)	(0.212)
Male	-2.066***	-1.250***	-0.197***	-0.227
	(0.054)	(0.047)	(0.072)	(0.146)
R^2	0.041	0.032	0.007	0.003
N	173171	173528	2600	2600

Note: Table reports coefficients, R^2, and sample sizes from Equation (4). I winsorize at the top and bottom 1% values for SCE respondents to minimize the impact of extreme outliers on results. Data for SCE respondents span from June 2013 to December 2024 while data for LLM-generated expectations span from June 2023 to June 2024.

Table 4 presents the results of running the regression from Equation (4). I find that the majority of the coefficients are non-significant. For LLM-Generated expectations of 1-year inflation, I find significant coefficients for individuals with income over \$100,000 and males in the same negative direction as in SCE respondents. However, the magnitude is far smaller. Quantitatively, males have lower inflation expectations by 0.197 percentage points for LLM respondents, whereas this magnitude is 2.066

percentage points for human respondents. However, these results do not replicate for expected 1-year house prices. In Appendix Table A.3 of Appendix Section E, I run the same regression with survey wave fixed effects, including the corresponding results with Gemini 2.5-Flash. I find that, although there are a few more significant coefficients, they still are unable to reach the same magnitude of significance documented in SCE respondents. Most of the relationships are also not consistently significant across both inflation and house price expectations.

It is clear, therefore, that LLM-generated expectations are currently unable to replicate the systematic relationship between beliefs and demographics documented in the literature. While they may be able to replicate the systematic relationship between demographics and *preferences* (Fedyk et al., 2024), they are unable to capture the correlations between demographic backgrounds and expectation formation for point forecasts. These results are mechanically in line with those in the previous section: the weak explanatory power of persistent individual differences leaves little stable variation for demographics to explain, which produces these inaccurate demographic gradients in the synthetic panel data.

6 Conclusion

Although there is (warranted) excitement about LLMs in economics and finance, I caution against the use of LLMs to simulate a panel dataset of expectations. As summarized by Giglio et al. (2025), "it is an exercise fraught with problems that the literature is yet to sort out." In this paper, I formally show some of these issues.

First, LLMs are a black box and their responses can significantly vary with minor changes in prompting. Although there is a literature that advocates for the use of explicit knowledge restriction / date-restrictive prompting (e.g., Wu et al., 2025; Hansen et al., 2024), I show that merely supplementing these point-forecasts with qualitative forecasts of future economic risks can still lead to significant leakage. Thinking models in particular are more susceptible to this leakage relative to standard LLM models. Therefore, researchers must proceed with an abundance of caution when using LLMs for prediction problems. Ludwig et al. (2025) provide a useful framework for applied econometricians on the necessary conditions to conduct valid inference and prediction using LLMs. One solution, proposed by Sarkar and Vafa (2024), is to use families of time-indexed language models whose knowledge cutoff lies just before the prediction period, but such approaches are computationally expensive and not widely available in standard practice.

Second, LLM respondents are able to reproduce the collective explanatory power of individual and time variation as SCE respondents, but the sources of this variation are different. While aggregate LLM expectations are able to follow the aggregate expectations of SCE respondents, they struggle to reproduce the rich cross-sectional heterogeneity that is critical to the subjective expectations literature. I show that these results hold beyond the LLM's knowledge cutoff and across various LLM models. On one hand, subjective expectations are characterized by persistent, meaningful differences in individuals: the same individual remains optimistic or pessimistic, which is key to the theoretical literature on heterogeneous agent models. Conversely, the beliefs of synthetic LLM respondents are driven by time variation, with the same individual significantly changing their beliefs over time—behavior that aligns closer with representative agent models.

Finally, LLM respondents are unable to reproduce the systematic relationship between demographics and beliefs when making point-forecasts about future inflation and house prices. In SCE data and the broader subjective expectations literature, demographic gradients are large, stable, and highly significant. By contrast, the demographic coefficients in the LLM-generated panel are weak, inconsistent across beliefs, and are an order of magnitude smaller than their SCE counterparts.

Even without look-ahead bias, the primary challenge with simulating a synthetic dataset of expectations is that the underlying factor structure of belief formation remains poorly understood. As observed extensively in Giglio et al. (2021a); Kuchler et al. (2023); Giglio et al. (2025), many observable characteristics explain only a small part of the variations in beliefs. Until these latent factors are better understood, LLMs will continue to miss key dimensions of cross-sectional heterogeneity and demographic relationships.

This raises an important question: can researchers reliably induce richer cross-sectional structure from LLMs? Addressing this challenge is essential if LLMs are ever to serve as substitutes for genuine panel data. At present, however, the evidence in this paper points to a robust conclusion: LLMs can complement—but not replace—human belief data. Their strengths currently lie in the processing of unstructured textual data, as long as certain conditions are satisfied (Ludwig et al., 2025). However, they are not suitable for reproducing the stable, persistent, and demographically structured heterogeneity revealed in survey panel data. Until the latent drivers of beliefs are better understood, and until we have methods that can reliably generate persistent individual-level differences within LLMs, synthetic panels will remain an imperfect proxy for the rich dynamics of human expectations.

References

Argyle, Lisa P., Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting et al., "Out of One, Many: Using Language Models to Simulate Human Samples," *Political Analysis*, July 2023, *31* (3), 337–351.

Ariyarathne, Gangani and Alexander C. Nwala, "3DLNews: A Three-decade Dataset of US Local News Articles," August 2024. arXiv:2408.04716 [cs].

Armantier, Olivier, Giorgio Topa, Wilbert van der Klaauw, and Basit Zafar, "An Overview of the Survey of Consumer Expectations," *Federal Reserve Bank of New York Economic Policy Review*, December 2017.

Armona, Luis, Andreas Fuster, and Basit Zafar, "Home Price Expectations and Behaviour: Evidence from a Randomized Information Experiment," *The Review of Economic Studies*, July 2019, *86* (4), 1371–1410.

Bisbee, James, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson, "Synthetic Replacements for Human Survey Data? The Perils of Large Language Models," *Political Analysis*, October 2024, *32* (4), 401–416.

Bybee, J Leland, "The Ghost in the Machine: Generating Beliefs with Large Language Models," 2025.

Crane, Leland D., Akhil Karra, and Paul E. Soto, "Total Recall? Evaluating the Macroeconomic Knowledge of Large Language Models," June 2025.

D'Acunto, Francesco and Michael Weber, "Why Survey-Based Subjective Expectations Are Meaningful and Important," *Annual Review of Economics*, August 2024, *16* (Volume 16, 2024), 329–357.

Engelberg, Joseph, Asaf Manela, William Mullins, and Luka Vulicevic, "Entity Neutering," March 2025.

Faria-e Castro, Miguel and Fernando Leibovici, "Artificial Intelligence and Inflation Forecasts," *Federal Reserve Bank of St. Louis Review*, 2024.

Fedyk, Anastassia, Ali Kakhbod, Peiyao Li, and Ulrike Malmendier, "AI and Perception Biases in Investments: An Experimental Study," April 2024.

Giglio, Stefano, Matteo Maggiori, Joachim Rillo, Johannes Stroebel, Stephen Utkus et al., "Investor Beliefs and Expectation Formation," October 2025.

Giglio, Stefano, Matteo Maggiori, Johannes Stroebel, and Stephen Utkus, "Five Facts about Beliefs and Portfolios," *American Economic Review*, May 2021, *111* (5), 1481–1522.

Giglio, Stefano, Matteo Maggiori, Johannes Stroebel, and Stephen Utkus, "The joint dynamics of investor beliefs and trading during the COVID-19 crash," *Proceedings of the National Academy of Sciences*, January 2021, *118* (4).

Glasserman, Paul and Caden Lin, "Assessing Look-Ahead Bias in Stock Return Predictions Generated By GPT Sentiment Analysis," September 2023. arXiv:2309.17322 [q-fin].

Golchin, Shahriar and Mihai Surdeanu, "Time Travel in LLMs: Tracing Data Contamination in Large Language Models," February 2024. arXiv:2308.08493 [cs].

Grattafiori, Aaron, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian et al., "The Llama 3 Herd of Models," November 2024. arXiv:2407.21783 [cs].

Guillochon, Justine, "News-Driven Household Macroeconomic Expectations: Regional vs. National Telecast Information," *Journal of Money, Credit and Banking*, 2024, *n/a* (n/a). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jmcb.13246.

Halawi, Danny, Fred Zhang, Chen Yueh-Han, and Jacob Steinhardt, "Approaching Human-Level Forecasting with Language Models," February 2024. arXiv:2402.18563 [cs].

Hansen, Anne Lundgaard, John J. Horton, Sophia Kazinnik, Daniela Puzzello, and Ali Zarifhonarvar, "Simulating the Survey of Professional Forecasters," December 2024.

Horton, John J., "Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?," January 2023. arXiv:2301.07543 [econ].

Kaplan, Greg and Giovanni L. Violante, "Microeconomic Heterogeneity and Macroeconomic Shocks," *Journal of Economic Perspectives*, August 2018, *32* (3), 167–194.

Kaplan, Greg, Benjamin Moll, and Giovanni L. Violante, "Monetary Policy According to HANK," *American Economic Review*, March 2018, *108* (3), 697–743.

Korinek, Anton, "Generative AI for Economic Research: Use Cases and Implications for Economists," *Journal of Economic Literature*, January 2023, *61* (4), 1281–1317.

Korinek, Anton, "Economic Policy Challenges for the Age of AI," September 2024.

Kuchler, Theresa and Basit Zafar, "Personal Experiences and Expectations about Aggregate Outcomes," *The Journal of Finance*, 2019, *74* (5), 2491–2542.

Kuchler, Theresa, Monika Piazzesi, and Johannes Stroebel, "Housing market expectations," in "Handbook of Economic Expectations," Elsevier, 2023, pp. 163–191.

Kumar, Saten, Hassan Afrouzi, Olivier Coibion, Yuriy Gorodnichenko, and UT Austin, "Inflation Targeting Does Not Anchor Inflation Expectations: Evidence from Firms in New Zealand," *Brookings Papers on Economic Activity*, 2015.

Lopez-Lira, Alejandro and Yuehua Tang, "Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models," October 2025. arXiv:2304.07619 [q-fin].

Ludwig, Jens, Sendhil Mullainathan, and Ashesh Rambachan, "Large Language Models: An Applied Econometric Framework," January 2025.

Malmendier, Ulrike and Stefan Nagel, "Depression Babies: Do Macroeconomic Experiences Affect Risk Taking?," *The Quarterly Journal of Economics*, February 2011, *126* (1), 373–416.

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

Roberts, Manley, Himanshu Thakur, Christine Herlihy, Colin White, and Samuel Dooley, "To the Cutoff... and Beyond? A Longitudinal Perspective on LLM Data Contamination," in "in" October 2023.

Sarkar, Suproteem K. and Keyon Vafa, "Lookahead Bias in Pretrained Language Models," June 2024.

Soo, Cindy K, "Quantifying Sentiment with News Media across Local Housing Markets," *The Review of Financial Studies*, October 2018, *31* (10), 3689–3719.

Wu, Jing Cynthia, Jin Xi, and Shihan Xie, "LLM Survey Framework: Coverage, Reasoning, Dynamics, Identification," Technical Report w34308, National Bureau of Economic Research October 2025.

Yang, An, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui et al., "Qwen3 Technical Report," May 2025. arXiv:2505.09388 [cs].

“LLMs and Belief Formulation” — Online Appendix

Joachim Rillo

A Filtering for Economy-Relevant Articles

Anchor Terms. The following expressions are matched using case-insensitive regular expressions with word boundaries. Terms ending in `\w*` capture plural and common morphological variants.

- **Housing and Real Estate:** housing market; housing price`\w*`; housing starts; housing affordability; home price`\w*`; real estate; property value`\w*`; mortgage rate`\w*`; rent price`\w*`; foreclosure`\w*`.
- **Financial Markets:** stock market; S&P 500; Dow Jones; Nasdaq; bond market; treasury yield`\w*`; interest rate`\w*`; Federal Reserve; the Fed; FOMC; investor sentiment.
- **Labor Market:** unemployment rate; job growth; jobs growth; jobless claim`\w*`; layoff`\w*`; hiring.
- **General and Local Economy:** local economy; business condition`\w*`; small business; economic growth; GDP; recession; consumer spending; retail sale`\w*`; supply chain`\w*`; consumer confidence.

Proximity Patterns. These patterns flag pairs of economically related terms that appear within approximately 40 non-word characters of each other (case-insensitive).

- **Housing context:** (home | house | property | housing) near (sale`\w*` | value`\w*` | market | price`\w*` | afford`\w*`).
- **Labor market context:** (job | employment | unemployment | hiring | layoff`\w*`) near (report | data | number`\w*` | outlook | rate | claim`\w*`).
- **Monetary policy context:** (interest | fed | federal) near rate`\w*` near (hike | cut | decision | increase | decrease).
- **Consumer activity context:** (consumer | household) near (confidence | spending | sentiment | debt).
- **Business activity context:** (business | factory | plant | company | firm) near (opening | closing | expansion | layoff`\w*` | hiring | earning`\w*`).
- **Macroeconomic context:** (economic | economy) near (growth | slowdown | recession | outlook | activity | indicator`\w*`).

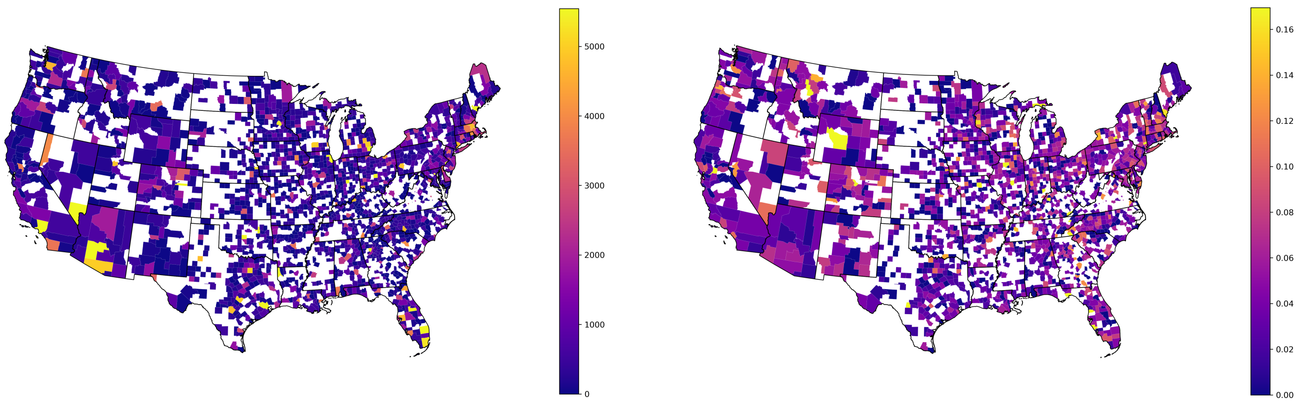
- **Stock market movement context:** (stock | market | Nasdaq | Dow) near (rally | plunge | tumble | gain | loss | up | down | close`\w*`).

B Summary Statistics for Sample Period

Figure A.1: Local Media Concentration and Economic News Share by County (2013–2024)

(a) Local Media Concentration

(b) Economic News Share



Note: Figure 1(a) reports the local media concentration from 2013 to 2024, where I sum the total amount of local media coverage by county. Figure 1(b) reports the share of business and economic articles in total local news coverage for each FIPS code over 2013 to 2024, where data are available. For each FIPS, the share is calculated as the ratio of business/economic articles to the total number of articles. Color limits are set to the 1st and 99th percentiles to reduce the visual influence of outliers.

C Prompts Used for Lookahead Bias Simulation

C.1 Prompt for 2020 Forecasts Simulation

```
SYSTEM
You are simulating a human respondent as of June 2019 (cutoff = 2019-06-30).
Do not use or reference any information after the cutoff. Do not browse or retrieve.
Do not include reasoning.

USER
Task A: List three risks to the U.S. economy over the next 12 months


 (ending June 2020).
Task B: Provide a single CPI inflation point forecast over the next 12 months
 (ending June 2020).

OUTPUT (exactly this, no extra text):
RISKS:
1) ...
2) ...
```

```
3) ...
INFLATION_FORECAST: X.X%
```

C.2 Prompt for 2019 Forecasts Simulation

```
SYSTEM
You are simulating a human respondent as of June 2018 (cutoff = 2018-06-30).
Do not use or reference any information after the cutoff. Do not browse or retrieve.
Do not include reasoning.

USER
Task A: List three risks to the U.S. economy over the next 12 months
(ending June 2019).
Task B: Provide a single CPI inflation point forecast over the next 12 months
(ending June 2019).

OUTPUT (exactly this, no extra text):
RISKS:
1) ...
2) ...
3) ...
INFLATION_FORECAST: X.X%
```

D Prompt Templates for Panel Data Simulation

This section reproduces the exact prompt used to generate simulated survey responses in the panel data experiments described in Section 4. The "News Block" includes a randomly sampled local media article (`{news_block}`) relevant to a respondent's commuting zone and survey month.

All placeholders (e.g., `{survey_month_str}`, `{demographics_line}`) are dynamically populated at runtime for each respondent-month. If no article is available, respondents receive a randomly sampled local media article from a neighboring commuting zone at a given month. Otherwise, they receive no news article.

D.1 Base Prompt Template

```
SYSTEM
You are simulating a human respondent completing an economic expectations survey.

Knowledge restriction:
You are responding to this survey as of {survey_month_str}.
Do NOT reference or rely on any events or developments that occurred after this
```

```
date.
Do NOT use or search for real data, official statistics, or hindsight information.
Answer as a typical person might, based on observations and general sense of the
economy at the time.

USER
Persona (demographic seeding): {demographics_line}

Below is a relevant news article in your area:
{news_block}

TASK
Answer the following survey questions about home prices and inflation expectations.
Your responses must be consistent, realistic, and phrased as if you were a
thoughtful survey respondent.

QUESTIONS

1. Home Prices -- Next 12 Months
Over the next 12 months, what do you expect will
happen to the average home price
nationwide?
→ Do you think it will increase or decrease?
→ By about what percent?
→ Please explain briefly why you think home prices will change in this way.

2. Home Prices -- 2-3 Years Ahead
Over the 12-month period between {date_24} and {date_36}, what do you expect will
happen to the average home price nationwide?
→ Do you think it will increase or decrease?
→ By about what percent?
→ Please explain briefly why.

3. Inflation -- Next 12 Months
Over the next 12 months, do you think there will be inflation or deflation
(deflation is the opposite of inflation)?
→ By about what percent?
→ Please explain briefly why you think prices will change in this way.

4. Inflation -- 2-3 Years Ahead
```

```
Over the 12-month period between {date_24} and {date_36}, do you think there will
be inflation or deflation?
→ By about what percent?
→ Please explain briefly why.
```

E Gemini 2.5-Flash Robustness Checks

Table A.1: Decomposing the Variation in Beliefs: Individual and Time Fixed Effects

	R^2 (percent) of panel regression, Gemini 2.5 Flash	
	Expected 1Y Inflation	Expected 1Y House Prices
Time FE	24.9	58
Individual FE	6.1	5.6
Time + Individual FE	31	63.6
Unique Respondents	70	70

Note: Table reports the R^2 values corresponding to the three regressions from Equations (1), (2), (3), and the number of unique respondents. The respondents are generated from sampling from the SCE and inputting these demographic characteristics into Gemini 2.5 Flash. Each row corresponds to a different survey question that is used as the dependent variable.

Table A.2: Beliefs and Demographics Regression - Gemini 2.5 Flash

	$B_{i,t}$: SCE Respondents	
	Expected 1Y Inflation	Expected 1Y House Prices
Age: 40 to 60	0.052	-0.057
	(0.074)	(0.090)
Age: Over 60	-0.026	-0.008
	(0.105)	(0.118)
Education: College	-0.178	0.124
	(0.146)	(0.167)
Education: Some College	-0.176	0.121
	(0.142)	(0.160)
Income \$50k - \$100k	-0.036	0.042
	(0.108)	(0.126)
Income $\geq$ \$100k	-0.136	0.153
	(0.104)	(0.123)
Male	-0.037	0.021
	(0.069)	(0.082)
R^2	0.008	0.005
N	910	910

Note: Table reports coefficients, R^2, and sample sizes from Equation (4) with synthetic respondents from Gemini 2.5-Flash. LLM-generated expectations span from June 2023 to June 2024.

Table A.3: Beliefs and Demographics Regression with Survey Wave FE

	$B_{i,t}$: SCE Respondents		$B_{i,t}$: Llama3:8b		$B_{i,t}$: Gemini 2.5-Flash	
	Expected 1Y Inflation	Expected 1Y House Prices	Expected 1Y Inflation	Expected 1Y House Prices	Expected 1Y Inflation	Expected 1Y House Prices
Age: 40 to 60	0.721***	0.388***	0.037	-0.005	0.052	-0.057
	(0.068)	(0.059)	(0.066)	(0.103)	(0.054)	(0.060)
Age: Over 60	0.173**	0.555***	0.039	-0.146	-0.026	-0.008
	(0.070)	(0.061)	(0.075)	(0.117)	(0.071)	(0.081)
Education: College	-3.573***	-2.302***	0.003	-0.001	-0.178*	0.124
	(0.117)	(0.097)	(0.105)	(0.181)	(0.099)	(0.112)
Education: Some College	-1.707***	-1.426***	-0.132	-0.063	-0.176*	0.121
	(0.125)	(0.102)	(0.103)	(0.177)	(0.096)	(0.110)
Income \$50k - \$100k	-1.786***	-1.723***	-0.132*	-0.141	-0.036	0.042
	(0.072)	(0.061)	(0.074)	(0.115)	(0.073)	(0.083)
Income $\geq$ \$100k	-2.644***	-2.657***	-0.227***	0.130	-0.136*	0.153*
	(0.070)	(0.062)	(0.081)	(0.126)	(0.074)	(0.083)
Male	-2.036***	-1.294***	-0.197***	-0.227***	-0.037	0.021
	(0.054)	(0.047)	(0.055)	(0.085)	(0.051)	(0.053)
Survey Wave FE	Y	Y	Y	Y	Y	Y
R^2	0.081	0.084	0.404	0.658	0.504	0.580
N	173171	173528	2600	2600	910	910

Note: Table reports coefficients, R^2, and sample sizes from Equation (4), with the additional inclusion of survey wave fixed effects. I winsorize at the top and bottom 1% values for SCE respondents to minimize the impact of extreme outliers on results. Data for SCE respondents span from June 2013 to December 2024 while data for LLM-generated expectations span from June 2023 to June 2024.