Generating Inflation Expectations with Large Language Models

Ali Zarifhonarvar*

This version: August 10, 2025 Latest version

Abstract

This paper studies the formation of inflation expectations using generative AI in survey experiments, examining diverse agents created with both proprietary and open-source large language models (LLMs). It shows that model architecture significantly impacts expectation formation, with proprietary models generally exhibiting less disagreement in their responses than open-source alternatives. Some LLMs predict higher inflation than actual rates, aligning with patterns observed in the Survey of Consumer Expectations. Information treatments, particularly forward guidance on inflation, influence LLMs' inflation expectations, though with varying magnitudes across model types. Customizing prompts with demographic personas induces heterogeneous responses that mirror human survey behaviors, with some biases similar to those documented in household surveys. The paper also demonstrates how central banks could leverage these models as communication policy tools to test messaging strategies before implementation.

Keywords: Inflation Expectation, Large Language Models, Information Provision, Survey Experiment, Forward Guidance

JEL: C90, E27, E31, E58, E71, D84

*Indiana University Bloomington - alizarif@iu.edu

I would like to thank Daniela Puzzello for her invaluable comments, help, and support throughout this project. I am also grateful for the comments and feedback from Volodymyr Lugovskyy, Ala Avoyan, Sophia Kazinnik, Yuriy Gorodnichenko, Joon Park, Hassan Afrouzi, Rupal Kamdar, Robin Horton, Benjamin Manning, Kay-Yut Chen, Leland Bybee, Gregor Schubert, Elena Asparouhova, Ran Shorrer, Rodolfo Maino, Jennifer Mangano, Costas Lambros, Talha Cakir, Eric Swanson, Anton Korinek and all the participants at the NBER Digital Economics and AI Meeting, the ICD Departmental Seminar Series at IMF, North American Meeting of the Economic Science Association, Midwest Macroeconomics Meetings, Internal Seminars at Indiana University, Experimental Economics Workshop at Purdue University, and the NBER Behavioral Macroeconomics Research Boot Camp. I am also grateful to the legendary Vernon Smith for his insightful comments. All mistakes are my own.

1 Introduction

The increasing adoption of large language models (LLMs) marks a significant shift in economic research methodology. While LLMs have demonstrated potential in assisting economists across various tasks (Charness et al., 2025; Korinek, 2023; Chang et al., 2024), their capacity to simulate and enhance our understanding of expectation formation is still emerging. This paper addresses three main questions: Can LLMs replicate key patterns observed in household survey responses regarding inflation expectations? How do these models react to different types of economic information? Can they serve as reliable tools for augmenting traditional survey methods and informing policy design?

This paper provides a comprehensive analysis of these questions, examining both the strengths and limitations of using LLMs to simulate household inflation expectations surveys. The study yields several important insights about the potential of LLMs in economic research. First, these models can successfully emulate some of the household survey patterns, including exhibiting systematic biases like higher than realized inflation expectations. Second, they demonstrate sensitivity in processing new information, particularly forward guidance on inflation, suggesting applications in testing policy communication strategies. Third, their responses show sensitivity to demographic characteristics that mirrors real-world heterogeneity, indicating potential for studying diverse population groups.

This research also explores how Artificial Intelligence (AI) agents can serve as valuable communication policy tools, allowing policymakers to predict the effects of announcements (for example, simplified vs. complex language or short-run vs. long-run messages) on inflation expectations. By deploying these agents in structured surveys and comparing their responses to human data, we can gain insights into both the models' capabilities and the process of belief formation. The ongoing debate about LLMs' capabilities highlights the importance of systematically understanding their behavioral patterns and cognitive processes (Echterhoff et al., 2024). Such understanding extends beyond identifying practical applications; it offers valuable insights into human

cognition and decision-making processes, because these models are based on training datasets of human knowledge and designed to emulate aspects of human language processing and reasoning (Zhu et al., 2024).

Previous research has explored AI agents primarily in game-theoretic scenarios (Horton, 2023; Brookins and DeBacker, 2023; Guo, 2023; Raman et al., 2024; Immorlica et al., 2024) and in forecasting of macroeconomic variables (Hansen et al., 2025; Bybee, 2025; Faria-e Castro and Leibovici, 2024). My paper extends this line of work by employing AI agents as participants in a survey experiment focused on household and consumer inflation expectations, rather than those of professional forecasters or firms. This focus is particularly valuable as household expectations play a crucial role in determining consumer behavior and aggregate economic outcomes, while gathering large-scale household response data is often costly and time-consuming. Unlike previous studies, this research enables AI agents simulating households to update their beliefs under an information provision experiment, demonstrating how they process real-world economic information. This approach is especially relevant given that household inflation expectations significantly influence central bank policy design and implementation (Coibion et al., 2022), affect consumption decisions, and play a crucial role in financial markets (Bernanke and Kuttner, 2005).

In central banking, communication strategies are critical for managing market expectations (Blinder et al., 2008; Eusepi and Preston, 2010). As AI becomes increasingly integrated into daily life through platforms like ChatGPT, understanding how these systems form and influence economic expectations becomes important. Their interpretation of central bank communications could impact markets (Hansen and Kazinnik, 2023), making it essential to study how LLMs process information about monetary policy.

The rapid adoption of Generative AI (GAI) across economic sectors adds urgency to this research. Bick et al. (2024) find that nearly 40 percent of U.S. adults used generative AI by August 2024, with 28 percent using it at work. Aldasoro et al. (2024) report almost half of U.S. households use GAI tools, though usage varies significantly

across demographic groups. As AI-assisted decision-making becomes more prevalent, understanding these patterns and their implications becomes increasingly important (Korinek, 2023).

To achieve these objectives, I introduce an experimental design using both proprietary (GPT-4.1, GPT-4o, GPT-4o-mini, Claude 3.7 Sonnet, and Claude 3.5 Haiku) and open-source LLMs (Llama 3-70B and DeepSeek V3). I explore how defining personas impacts survey results, simulating the Survey of Consumer Expectations from the Federal Reserve Bank of New York across different demographic groups.

An important finding of this research is that model architecture, parameter size, and knowledge cutoff dates significantly impact inflation expectation formation. Different model families (OpenAI, Anthropic, and open-source alternatives) exhibit distinct expectation patterns, with proprietary models generally displaying less variance compared to open-source alternatives. Additionally, there is a notable impact from model size (frontier models versus compact models), with different patterns observed across major providers. Notably, models with greater exposure to high-inflation periods in their training data tend to forecast higher inflation rates, paralleling how humans' inflation expectations are shaped by their lived economic experiences.

The results suggest that LLMs could serve as valuable tools for augmenting traditional survey methods and generating testable hypotheses about expectation formation. While AI agents' responses show qualitative similarities with human survey responses, the quantitative differences highlight the unique nature of AI-driven expectation formation. This distinction is particularly relevant for policymakers, as it suggests that AI agents may process monetary policy communications differently than traditional economic agents, offering new insights into the effectiveness of various communication strategies.

This paper contributes to the rapidly growing literature on the economics of AI while providing practical insights into AI-assisted economic forecasting and its potential impacts on policy-making. The study makes three key methodological contributions: establishing a framework for testing information treatments, developing a

novel approach to implementing demographic personas in LLMs, and demonstrating how these tools can enhance our understanding of expectation heterogeneity across population groups. By exploring the intersection of generative AI and household inflation expectations, this study advances our understanding of how economic beliefs are formed and updated in response to new information at the household level, particularly in contexts where conducting comprehensive surveys might be impractical or costly.

2 Literature Review

This paper contributes to two main strands of the literature: the experimental study of AI agents and LLMs in economics, and the formation of inflation expectations. In addition to connecting these two domains as others have done (Faria-e Castro and Leibovici, 2024; Bybee, 2025; Karger et al., 2025; Hansen et al., 2025), this paper analyzes the decision-making processes of LLMs by assessing the impact of information treatments on AI agents in an experimental setup and by implementing detailed demographic personas to capture expectation heterogeneity. The ultimate goal is to explore whether these models can be used alongside data from human respondents for economic modeling and policymaking, with the primary application for central banks being the use of LLMs as tools to help evaluate communication strategies.

2.1 Economics and LLMs

The introduction of LLMs has significantly transformed economics research, demonstrating strong capabilities in simulating complex economic scenarios (Akata et al., 2023; Heydari and Lorè, 2023; Charness et al., 2025; Korinek, 2023). This transformation has opened new avenues for understanding economic behavior through computational agents that can mimic human decision-making processes.

A growing body of research examines LLMs as simulated economic agents, building on foundational insights about artificial economic behavior. Horton (2023) draws

parallels to *'homo economicus'* by equipping LLMs with specific preferences, information, and endowments, showing LLMs emulating human behaviors. Building on this framework, Brookins and DeBacker (2023) finds that GPT agents often prioritize fairness over optimal outcomes in games like the dictator game and prisoner's dilemma. These findings suggest possibilities for fully automated social science (Manning et al., 2024; Batista and Ross, 2024; Tranchero et al., 2024), where AI agents could potentially replace human subjects in certain experimental contexts.

The application of LLMs has naturally extended into macroeconomic modeling and forecasting. Li et al. (2024) develops frameworks using LLMs as agents for macroeconomic simulations, demonstrating their potential for complex economic modeling. In forecasting applications, Faria-e Castro and Leibovici (2024) explores LLMs' inflation forecasting capabilities, showing that Google AI's PaLM outperforms traditional methods compared to the Survey of Professional Forecasters, though requiring manually limiting the AI's knowledge. Complementing this work, Bybee (2025) generates economic expectations by applying LLMs to historical news data, achieving close alignment with existing survey measures.

The evaluation of LLM forecasting performance has revealed both capabilities and limitations. Karger et al. (2025) finds that LLMs perform similarly to public forecasts but lag behind expert forecasters, highlighting the importance of expertise in economic prediction. Similarly, Hansen et al. (2025) demonstrates that AI agents exhibit patterns similar to human professional forecasters using synthetic personas, while Kazinnik and Brynjolfsson (2025) examine how central banks can strategically integrate AI to enhance operations, providing institutional context for AI adoption that complements our focus on communication strategy evaluation.

Recent advances in understanding LLMs' economic reasoning have revealed nuanced behavioral patterns. Ross et al. (2024) find that models exhibit economic behavior that is neither entirely human-like nor fully rational, suggesting a unique form of artificial economic reasoning. This finding is supported by Henning et al. (2025), who show LLMs price assets more rationally than humans, while Allard et al. (2024)

demonstrate modest improvements in inflation nowcasting through AI-enhanced analysis.

However, fundamental challenges remain in using LLMs for economic research. Lopez-Lira et al. (2025) raise critical concerns that LLMs have memorized significant economic data from training, creating challenges for forecasting research. Their findings suggest that only post-training cutoff data can reliably test genuine forecasting capabilities versus memorization skills, highlighting the need for careful experimental design.

This paper advances beyond existing approaches by incorporating persona-based prompts that modulate which aspects of the model's latent knowledge are accessed. This methodology enables more sophisticated information processing compared to non-personalized prompting. Through systematic comparison of multiple models, I demonstrate how these factors influence LLM outputs in survey experiments.

2.2 Formation of Inflation Expectation

The study of inflation expectations provides an ideal testing ground for LLM capabilities due to expectation formation's complex nature, its extensive consequences for economic decisions, and its central importance for monetary policy (Coibion et al., 2020a; Weber et al., 2022; Coibion et al., 2020b; Pfajfar and Žakelj, 2018). Understanding how these expectations form and evolve has profound implications for both economic theory and policy implementation.

The behavioral foundations of inflation expectation formation reveal systematic patterns that may be replicated in artificial agents. Candia et al. (2020) show that inflation expectations can drive consumer spending and firm price-setting behavior even without actual inflation changes, highlighting the role of subjective beliefs in economic outcomes. This finding demonstrates how perception can become reality in economic systems, making the study of artificial perception particularly relevant. Building on this insight, D'Acunto et al. (2021) demonstrate that frequently purchased grocery prices disproportionately influence households' expectations, leading to bi-

ases and excessive sensitivity to transitory fluctuations.

A consistent pattern emerges from household survey data regarding systematic biases in inflation expectations. Weber et al. (2022) show households' inflation expectations are systematically higher than actual inflation, with stronger bias for low-income and less-educated households. This systematic upward bias stems from several well-documented mechanisms that may be reflected in LLM training data. Bruine de Bruin et al. (2010) show households are disproportionately influenced by observed price increases rather than price stability, explaining why LLMs trained on human-generated text exhibit similar upward biases in their inflation forecasts.

The role of salience and personal experience in shaping expectations provides additional insight into potential LLM behavior. Salient price changes significantly impact expectations (D'Acunto et al., 2021; Cavallo et al., 2014), suggesting LLMs' upward bias may reflect training data containing more references to price increases than stability. Personal experience heterogeneity creates substantial expectation differences (Di Giacomo and Angelico, 2019; Ehrmann et al., 2017), patterns LLMs may reproduce through training data that captures diverse economic experiences. The persistence of these biases (Afrouzi and Veldkamp, 2019; Armantier et al., 2016) indicates deep-rooted information processing patterns that may be embedded in artificial systems trained on human-generated content.

Central bank communication strategies represent a crucial link between policy intentions and public expectations. Coibion et al. (2023) find households respond more to short-term interest rate information than long-term policy goals, indicating central banks need tailored communication strategies. This finding has particular relevance for testing communication effectiveness using artificial agents. Furthermore, studies show public inflation perceptions often diverge from central bank indicators (Stantcheva, 2024; Binetti et al., 2024; Afrouzi et al., 2023, 2024), with personal experiences and socioeconomic status shaping perceptions (Jiang et al., 2024).

This study contributes to understanding how AI-driven models can enhance inflation expectation analysis while potentially serving as testing grounds for policy com-

munication. By integrating LLMs with survey experiments, this paper shows how these models process information and form expectations, offering insights into expectation formation dynamics and contributing to more effective policy tools. The intersection of artificial intelligence and expectation formation thus represents a promising frontier for both understanding human economic behavior and developing novel policy instruments.

3 Experimental Design

In this experiment, I conduct surveys using a diverse set of large language models[1] via the Expected Parrot framework.[2][3] The experimental design incorporates both proprietary models (GPT-4.1, GPT-4o, GPT-4o mini, Claude 3.7 Sonnet, Claude 3.5 Haiku) and open-source models (Llama 3.3 70B, DeepSeek-V3), with varying architecture and parameter sizes. The main goal is to assess the influence of different demographic profiles and information treatments on inflation expectation formation of AI agents.[4]

Recent studies have increasingly highlighted the significance of personas in simulating human-like behavior with large language models (Horton, 2023; Chen et al., 2024; Hu and Collier, 2024). Many studies have employed persona prompting to create AI agents with specific personality traits, backgrounds, and characteristics, resulting in more realistic and nuanced behaviors. While concerns exist about whether inflation expectations or other survey responses generated by LLMs accurately reflect real-world phenomena, recent studies by Fedyk et al. (2024) and Kazinnik (2023) suggest that LLMs can effectively mimic complex human behaviors in financial contexts. These findings support the view that, with appropriate setup, LLMs can provide valu-

[1] Full details on the implementation and replication are provided in the Supplementary Appendix, Section SA-1.1.

[2] Expected Parrot provides an infrastructure that facilitates AI-powered research by providing access to many models through a single endpoint to run the experiment.

[3] In the pilot runs I employed various GPT-4 versions through OpenAI's Assistants API, with results detailed in the Supplementary Appendix.

[4] Additional exercises on different prompt strategies, temperature settings, and Retrieval-Augmented Generation (RAG) to see the impact of knowledge domain are detailed in the Supplementary Appendix.

able insights into human decision-making processes.

The experiment simulates the Survey of Consumer Expectations, utilizing micro-data comprising 7,580 observations[5]. The use of synthetic personas is increasingly recognized as a valuable experimental tool with AI models, offering a way to simulate diverse individual characteristics (Liu et al., 2024; Kwok et al., 2024). Hansen et al. (2025) does a similar exercise but using the characteristics of professional forecasters who are participating in the Survey of Professional Forecasters. [6]

I employ the EDSL framework to create AI agents with specific demographic characteristics, such as age, gender, marital status, education, income and then ran the survey on them as participants. This approach allows for a systematic analysis of how demographic characteristics influence inflation expectations while enabling the exploration of how central banks might test communication strategies across diverse population segments.[7] The experimental design, illustrated in Figure 1, followed a three-stage process: eliciting prior beliefs, information provisions, and asking updated expectations or posterior belief.

After eliciting prior beliefs, AI agents were randomly assigned to one of ten groups - one control and nine treatment groups. The experimental flow began with a system prompt[8] that established each AI agent's demographic persona (e.g., "You are a 45-year-old male who is married with a college degree and income category of $50,000-$100,000 who lives in the state of California"). Pre-treatment questions then elicited their initial expectations through probability distributions for both short and long-term expectations. Next, treatment groups received specific information—for example, "The current average rate for fixed-rate 30-year mortgage is 6.64% per year" for

[5]This is the total number of unique participants after 2020 in the survey panel. It maintains the same demographic composition as the SCE, with the shares of each category consistent with the original survey (see Armantier et al. (2016)).

[6]Additional exercises conducted without incorporating personas are detailed in the Supplementary Appendix.

[7]Additional experiments examining political affiliations (Republican/Democrat) and locations (Texas/California) showed systematic variations in inflation expectations that closely mirror patterns observed in real-world survey data, with Republican-identified agents consistently generating higher inflation predictions when a Democratic president is in office (similar to findings in Binetti et al. (2024) and Kamdar and Ray (2022)). Full results are presented in the Supplementary Appendix (Section SA-4.1).

[8]For full details on prompts refer to Appendix (Section SA-1.5).

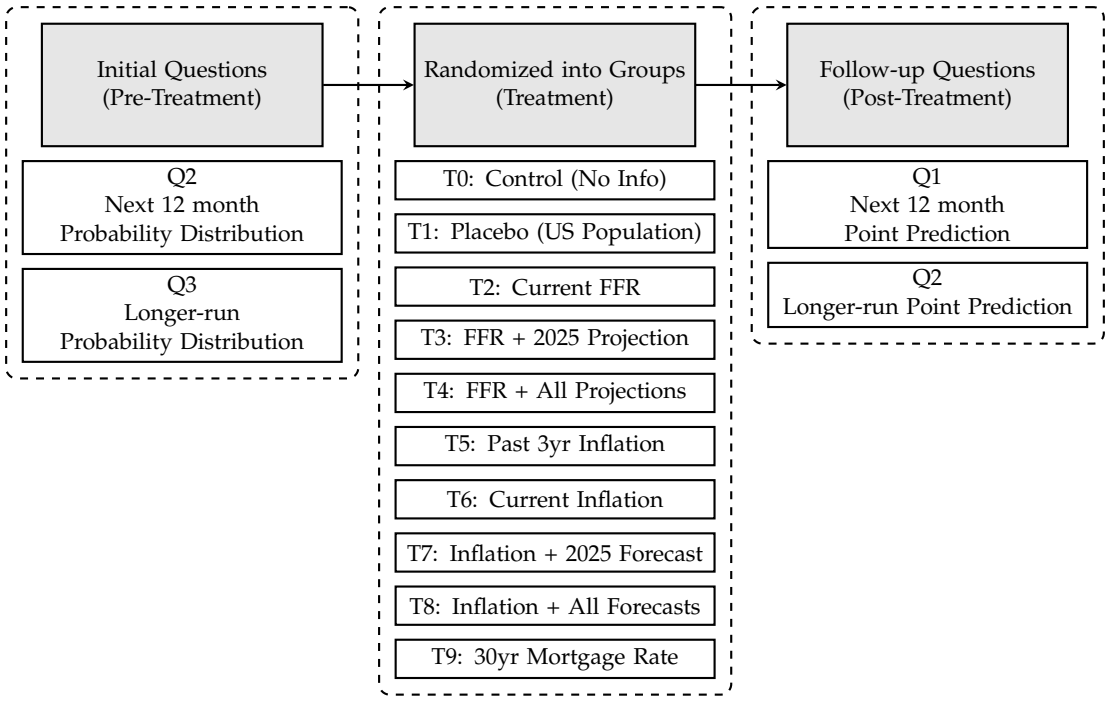

Figure 1: An Overview of the Survey Experiment Design

Note: This figure illustrates the three-stage design of the survey experiment. Participants first answer pre-treatment questions about inflation expectations, including probability distributions for different time horizons. They are then randomly assigned to one of ten treatment groups receiving different economic information. After the treatment, participants provide follow-up point predictions.

the mortgage rate treatment group, while the control group received no additional information. Finally, post-treatment questions asked for point predictions about future inflation rates. I utilized both point predictions and probability distribution questions because they capture different aspects of expectation formation. Point predictions provide a clear measure of the most likely outcome but often fail to capture uncertainty, while probability distributions allow respondents to express uncertainty across a range of outcomes (Manski, 2018; Haaland et al., 2023; Boctor et al., 2024). This approach mirrors best practices in economic surveys and enables more nuanced comparison between AI agent responses and human survey data.

4 Results

4.1 Preliminary Observations

In this section I examine how different LLMs form inflation expectations and how these expectations evolve after information treatments. [9] Table 1 presents the mean and standard deviation of inflation expectations across seven LLMs, showing both distribution forecasts (before treatment) and point forecasts (after treatment)[10]

Table 1: Inflation Expectations Across Models

Model	Distribution Forecasts *(Before Treatment)*				Point Forecasts *(After Treatment)*			
	Short-Term		Long-Term		Short-Term		Long-Term	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Panel A: OpenAI								
GPT-4.1	2.01	2.29	2.16	2.72	2.33	0.30	2.30	0.24
GPT-4o	2.29	4.39	1.80	4.41	2.89	0.40	2.71	0.34
GPT-4o-mini	4.01	4.49	3.26	4.38	3.08	0.58	2.84	0.60
Panel B: Anthropic								
Claude 3.7 Sonnet	3.48	3.82	2.65	3.50	3.65	0.55	2.92	0.37
Claude 3.5 Haiku	2.77	3.08	1.69	3.25	3.12	0.36	2.52	0.21
Panel C: Open Source								
Llama3-70B	2.67	3.92	1.85	3.86	3.18	0.55	2.77	0.51
DeepSeek-V3	2.61	3.95	2.53	3.49	3.82	37.33	3.24	23.32

Notes: This table reports average inflation expectations grouped by model provider. Short-term refers to 1-year expectations and long-term to 3-year expectations. Distribution Forecasts are calculated using the midpoint formula based on probability bins reported by the models. For comparison, the median 1-year ahead and 3-year ahead inflation expectations from the Survey of Consumer Expectations administered in April 2025 were 3.63 and 3.17 percentage points, respectively.

Several notable patterns emerge from this analysis that have important implications for both economic research and policy. First, all models exhibit a relatively smaller range of expectations (1.69% to 4.01%) compared to human studies for pre-treatment distribution forecasts. After treatment, point forecasts show increased consistency, with short-term expectations ranging from 2.33% to 3.82% (excluding outliers in DeepSeek-V3) and long-term expectations between 2.30% and 3.24%. This consoli-

[9] Detailed results for all models tested in this study, including GPT-4.1, GPT-4o, GPT-4o-mini, Claude 3.5 Haiku, Claude 3.7 Sonnet, DeepSeek-V3, and Llama3-70B, are available in the Supplementary Appendix, Section SA-6. The results of the reasoning model (o3-mini) are presented in Section SA-5.

[10] Robustness checks with reversed question framing confirm that treatment effects remain consistent across different survey formats. See Appendix Section SA-3.4.

dation of expectations following information provision suggests that treatments effectively anchor the models' predictions, similar to patterns observed in human subjects (Coibion et al., 2020b). This finding has direct implications for central bank communication, suggesting that clear policy statements could have substantial anchoring effects on expectations when we are dealing with AI agents.

Second, across all models, short-term inflation expectations consistently exceed long-term expectations, mirroring a pattern observed in human survey data, where near-term inflation forecasts typically surpass longer-horizon projections (Weber et al., 2022). This finding suggests that LLMs capture the natural tendency of household economic expectations to revert toward perceived long-run equilibrium values over time. For policymakers, this implies that LLMs could serve as useful tools for understanding how forward guidance at different horizons may influence both short- and long-term expectations.

Third, proprietary models from OpenAI and Anthropic demonstrate more stable predictions with substantially lower standard deviations in their point forecasts compared to pre-treatment distribution forecasts. In contrast, DeepSeek-V3 shows extremely high standard deviations in its point forecasts (37.33 for short-term and 23.32 for long-term), indicating potential instability in its expectation formation process. This model heterogeneity highlights the importance of model selection when using LLMs for policy analysis.

Table 2 shows some patterns in how demographic characteristics influence AI agents' inflation expectations. Across models (except some cases of open-source models), lower-income and less-educated personas consistently report higher inflation expectations, mirroring well-documented socioeconomic disparities in human surveys (Bryan and Venkatu, 2001; D'Acunto et al., 2022). Age effects show a U-shaped pattern in most models, with middle-aged personas (35-44) typically reporting lower expectations than both younger and older groups, similar to findings in Bruine de Bruin et al. (2010). Gender differences are modest but consistent, with female personas generally reporting slightly higher inflation expectations, echoing gender gaps documented

Table 2: Demographic Heterogeneity in Inflation Expectations Across LLMs

	Short Run Distribution Forecast				Long Run Distribution Forecast			
	GPT-4.1	Sonnet-3.7	Llama3-70B	DeepSeek-V3	GPT-4.1	Sonnet-3.7	Llama3-70B	DeepSeek-V3
Income								
Under 50k	2.11 (2.57)	3.90 (4.06)	2.58 (4.22)	2.66 (4.09)	2.34 (3.01)	3.09 (3.67)	1.86 (4.09)	2.55 (3.60)
50k to 100k	2.01 (2.26)	3.36 (3.75)	2.65 (3.80)	2.50 (3.89)	2.15 (2.69)	2.53 (3.45)	1.87 (3.76)	2.44 (3.47)
Over 100k	1.91 (2.06)	3.23 (3.65)	2.76 (3.74)	2.68 (3.90)	2.02 (2.47)	2.38 (3.35)	1.84 (3.73)	2.59 (3.40)
Education								
High School	2.17 (2.61)	4.01 (4.12)	2.63 (4.22)	2.72 (4.08)	2.41 (3.06)	3.19 (3.69)	1.91 (4.10)	2.59 (3.60)
Some College	2.09 (2.43)	3.80 (4.00)	2.64 (4.03)	2.60 (4.01)	2.28 (2.87)	2.96 (3.61)	1.86 (3.94)	2.51 (3.54)
College	1.92 (2.13)	3.18 (3.62)	2.69 (3.78)	2.59 (3.90)	2.04 (2.55)	2.36 (3.36)	1.84 (3.76)	2.52 (3.43)
Age Group								
18-24	2.02 (2.28)	3.72 (4.03)	2.63 (4.01)	2.66 (4.06)	2.14 (2.68)	2.94 (3.64)	1.91 (3.84)	2.50 (3.56)
25-34	1.93 (2.13)	3.42 (3.79)	2.66 (3.91)	2.59 (3.92)	2.05 (2.55)	2.60 (3.47)	1.80 (3.88)	2.53 (3.46)
35-44	1.97 (2.16)	3.39 (3.75)	2.71 (3.88)	2.62 (3.95)	2.09 (2.59)	2.56 (3.45)	1.86 (3.81)	2.52 (3.47)
45-54	2.00 (2.26)	3.46 (3.81)	2.70 (3.90)	2.70 (3.95)	2.15 (2.68)	2.62 (3.50)	1.89 (3.84)	2.60 (3.49)
55-64	2.06 (2.39)	3.54 (3.85)	2.67 (3.92)	2.57 (3.95)	2.24 (2.83)	2.70 (3.53)	1.89 (3.85)	2.48 (3.49)
65+	2.08 (2.52)	3.56 (3.88)	2.59 (3.97)	2.57 (3.99)	2.30 (2.95)	2.74 (3.54)	1.83 (3.91)	2.50 (3.53)
Gender								
Female	2.04 (2.36)	3.56 (3.89)	2.64 (3.98)	2.61 (3.97)	2.21 (2.79)	2.73 (3.54)	1.84 (3.93)	2.52 (3.51)
Male	1.97 (2.22)	3.39 (3.75)	2.70 (3.85)	2.61 (3.94)	2.11 (2.65)	2.56 (3.45)	1.87 (3.78)	2.53 (3.46)
Marital Status								
Single	2.03 (2.36)	3.64 (3.93)	2.63 (4.11)	2.60 (4.00)	2.21 (2.80)	2.81 (3.60)	1.86 (4.00)	2.49 (3.54)
Married	1.99 (2.25)	3.38 (3.75)	2.69 (3.80)	2.62 (3.93)	2.14 (2.67)	2.55 (3.44)	1.85 (3.77)	2.55 (3.45)

Note: This table reports mean inflation expectations (with standard deviations in parentheses) across demographic groups and model types. Short-run refers to 1-year forecasts and long-run to 3-year forecasts. Values represent percentage points. Across models, lower-income and less-educated personas consistently report higher inflation expectations, similar to patterns observed in human surveys.

by D'Acunto et al. (2021). These demographic patterns suggest that LLMs implicitly incorporate sociological factors that influence human expectation formation, likely reflecting these relationships in their training data.

As a robustness check, I conducted additional experiments varying the temperature parameter of the model[11]. While lower temperature settings produced results nearly identical to the default temperature, higher temperature settings led to significantly more variable outcomes. For instance, GPT-4o at temperature 1.5 $\in$ [0,2] generated point estimates with a mean of 121.57 and standard deviation of 4819.63, demonstrating the critical importance of controlling hyperparameter in economic forecasting applications of LLMs.

Figure 2 shows the relationship between means and standard deviations of inflation expectations before and after information treatments. This visualization reveals

[11]The detailed results are presented in the Supplementary Appendix, Section SA-3.1. Temperature is the most important hyperparameter of each model and range from $\in [0, 2]$ for GPT and DeepSeek models (default 1.0) and $\in [0, 1]$ for Claude and Llama models (default 0.5). I conducted experiments with various temperature settings, but default values were used for main results as they best represent how these models are typically deployed in real-world applications.

a pattern of convergence: while pre-treatment expectations (circles) exhibit high variance across models, post-treatment expectations (squares) show reduced variance.

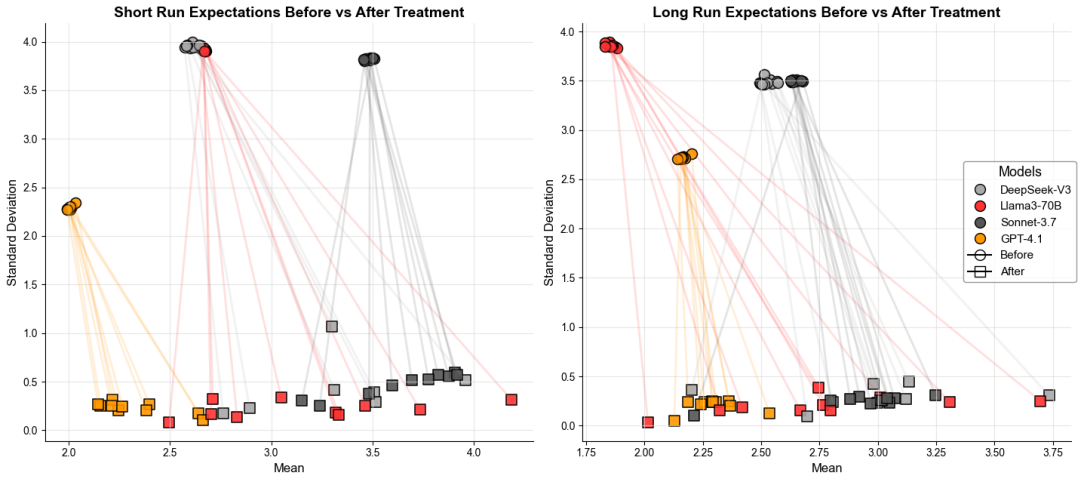

Figure 2: Standard Deviation vs. Mean of Expectations Before and After Treatment

Notes: This plot shows inflation expectations converging after information treatments, with pre-treatment values (circles) showing high variance and post-treatment values (squares) displaying significantly reduced standard deviations across all LLM models for both short and long-run forecasts.

These findings suggest information treatments have a homogenizing effect on LLMs' inflation expectations, with models converging toward similar predictions after exposure to economic data. This convergence, combined with the consistent pattern of higher short-term than long-term expectations, indicates that LLMs process economic information in ways that parallel human expectation formation, while exhibiting model-specific variations in initial beliefs and information sensitivity. From a policy perspective, this suggests central banks could test communication effectiveness using multiple LLM architectures before public release, with convergence across models indicating more robust messaging strategies.

4.2 Information Provision Treatments

This section examines how different information treatments influence inflation expectations across model architectures. Rather than focusing on prediction accuracy, the analysis explores how AI agents integrate new information into their expectation formation.

As shown in Table 3, a consistent pattern emerges across nearly all models: eco-

nomic information treatments trigger substantive revisions in inflation expectations. This suggests AI agents, like human subjects in studies like Coibion et al. (2023) and Coibion et al. (2020b), integrate new information into their belief formation processes. LLMs show exceptional sensitivity to forward guidance, particularly inflation projections, which have notably stronger impacts than interest rate information. When provided with future inflation forecasts, they demonstrate remarkable adherence to these projections in subsequent predictions, highlighting how these models prioritize explicit numerical information about inflation over other types of information.

Table 3: Changes in Inflation Expectations and Treatment Effects Across LLMs

Treatment	GPT-4.1				Claude 3.7 Sonnet			
	Short-Term		Long-Term		Short-Term		Long-Term	
	Δ SR	ATE SR	Δ LR	ATE LR	Δ SR	ATE SR	Δ LR	ATE LR
Panel A: Proprietary Models								
Control Group Change: Δ SR = 0.15, Δ LR = 0.09					*Control Group Change: Δ SR = 0.40, Δ LR = 0.39*			
Population Growth	0.23	0.08	0.07	-0.02	0.13	-0.27	0.40	0.01
Current FFR	0.18	0.03	0.09	-0.00	0.34	-0.06	0.27	-0.13
FFR + 1Y Proj	0.21	0.05	0.13	0.04	0.28	-0.11	0.23	-0.16
FFR + 1Y and Long run	0.26	0.11	0.20	0.11	0.41	0.01	0.34	-0.05
Past 3Y Inflation	0.41	0.25	0.13	0.04	0.45	0.05	0.61	0.22
Past 1Y Inflation	0.38	0.23	0.21	0.12	-0.00	-0.40	0.38	-0.01
Infl + 1Y Proj	0.64	0.49	0.39	0.30	-0.22	-0.62	0.16	-0.23
Infl + 1Y and Long run	0.65	0.50	-0.03	-0.12	-0.35	-0.75	-0.46	-0.85
Mortgage Rate	0.15	-0.00	0.05	-0.04	0.24	-0.16	0.41	0.02
Treatment	DeepSeek-V3				Llama3-70B			
	Short-Term		Long-Term		Short-Term		Long-Term	
	Δ SR	ATE SR	Δ LR	ATE LR	Δ SR	ATE SR	Δ LR	ATE LR
Panel B: Open Source Models								
Control Group Change: Δ SR = 0.87, Δ LR = 0.59					*Control Group Change: Δ SR = 0.78, Δ LR = 1.16*			
Population Growth	0.68	-0.19	0.44	-0.14	0.04	-0.75	0.87	-0.29
Current FFR	0.90	0.03	0.48	-0.10	0.64	-0.15	0.88	-0.28
FFR + 1Y Proj	1.00	0.14	0.45	-0.14	0.65	-0.13	0.81	-0.34
FFR + 1Y and Long run	0.71	-0.16	3.15	2.56	0.39	-0.40	0.56	-0.59
Past 3Y Inflation	1.35	0.48	1.22	0.63	1.06	0.27	1.44	0.28
Past 1Y Inflation	0.31	-0.56	0.31	-0.28	-0.17	-0.96	0.46	-0.70
Infl + 1Y Proj	0.19	-0.68	0.20	-0.38	0.18	-0.61	0.96	-0.19
Infl + 1Y and Long run	5.09	4.22	-0.30	-0.89	0.03	-0.76	0.15	-1.00
Mortgage Rate	1.00	0.14	0.55	-0.04	1.51	0.73	1.86	0.70

Notes: This table reports both raw changes in expectations (Δ SR for short-term, Δ LR for long-term) and average treatment effects (ATEs) relative to the control group. ATEs are calculated as the difference between the treatment group's change and the control group's change. The details on each treatment is provided in the Supplementary Appendix Section SA-7.

Importantly, these information treatments fall outside most models' training cutoffs, making them essentially out-of-sample tests that reveal how LLMs generalize their knowledge to new economic contexts. Rather than introducing entirely new in-

formation, the treatments likely act as signals prompting models to reweigh existing knowledge and their internal parameters, similar to how human subjects recalibrate their beliefs when presented with new economic data.

Several key patterns emerge: First, control group changes vary substantially across models, with open-source models displaying larger baseline shifts even without treatments. This variation is particularly notable when comparing their responses to different question formats—distribution forecasts (pre-treatment) versus point estimates (post-treatment)—suggesting that question framing (i.e., prompting) significantly influences how different architectures express uncertainty about future inflation. Second, treatment effects show significant heterogeneity, for example, treatments combining inflation data with future projections generate opposite effects in different models like DeepSeek-V3 and Claude Sonnet. Third, proprietary models generally display more consistent treatment effects than open-source alternatives, with GPT-4.1 showing the most stable pattern. Finally, inflation-related treatments elicit stronger responses than interest rate information among AI agents, in contrast to human studies where clearly framed and relevant interest rate signals, such as mortgage rates often have a larger impact on expectations and decisions (Coibion et al., 2023).

These results underscore the importance of choosing the right model architecture, as different LLMs respond to the same information in distinct ways. As AI tools become more widely used in financial markets and policy settings, it becomes increasingly important to understand how they interpret economic messages. The strong reaction of LLMs to inflation-related forward guidance suggests that, in markets where AI plays a large role, communications about future inflation could have a bigger impact, potentially strengthening the effects of monetary policy through the expectations channel.

4.2.1 Empirical Model

To empirically assess the impact of information treatments on AI agents' inflation expectations, I follow the specification used by Coibion et al. (2018, 2023), which cap-

tures how new information influences expectation formation. The empirical model is formulated as follows:

$$E_j\pi^{\text{post}} = \alpha + \theta E_j\pi^{\text{pre}} + \sum_{k=2}^{10} \beta_k \text{Treatment}_j^{(k)} + \sum_{k=2}^{10} \gamma_k \left(\text{Treatment}_j^{(k)} E_j\pi^{\text{pre}}\right) + \boldsymbol{W}_j \boldsymbol{\Psi} + \epsilon_j, \quad (1)$$

where $E_j\pi^{\text{post}}$ and $E_j\pi^{\text{pre}}$ represent the AI agent j's inflation expectations after and before receiving the treatment, respectively. $\text{Treatment}_j^{(k)}$ is a dummy variable indicating whether AI agent j was subjected to treatment k (with $k = 2, \ldots, 10$), where treatment 1 serves as the control group. The coefficients β_k and γ_k measure the level and interaction effects of the treatments on the expectations, respectively, relative to the control group. The vector $\boldsymbol{W}_j$ includes control variables relevant to each agent, and $\boldsymbol{\Psi}$ denotes the parameters associated with these controls. In line with Bayesian updating, agents revise their expectations by combining their prior beliefs with new information, as highlighted by Baley and Veldkamp (2023) and Coibion et al. (2018). The updated (posterior) expectation can be expressed as:

$$E_j\pi^{\text{post}} = (1 - \kappa_k) E_j\pi^{\text{pre}} + \kappa_k S^{(k)}, \quad (2)$$

where $\kappa_k \in [0, 1]$ represents the weight placed on the new information from treatment k, and $S^{(k)}$ is the signal provided by treatment k. By rearranging Equation (1) for agents in the control group ($k = 1$) and those who receive treatment ($k > 1$), and assuming $\boldsymbol{W}_j = 0$ for simplicity, we have: For the control group ($k = 1$):

$$E_j\pi^{\text{post}} = \alpha + \theta E_j\pi^{\text{pre}} + \epsilon_j. \quad (3)$$

For treatment groups ($k > 1$):

$$E_j\pi^{\text{post}} = (\alpha + \beta_k) + (\theta + \gamma_k) E_j\pi^{\text{pre}} + \epsilon_j. \quad (4)$$

Comparing these with the Bayesian updating formula, we can interpret: For the control group:

$$\theta = (1 - \kappa_1), \tag{5}$$

$$\alpha = \kappa_1 S^{(1)}. \tag{6}$$

For treatment groups:

$$\theta + \gamma_k = (1 - \kappa_k), \tag{7}$$

$$\alpha + \beta_k = \kappa_k S^{(k)}. \tag{8}$$

The coefficient γ_k thus captures how the weight on the prior expectation changes in response to treatment k relative to the control group. A negative γ_k implies that agents in treatment group k place less weight on their prior beliefs and more weight on the new information compared to the control group. This is consistent with Bayesian updating, where agents adjust their expectations more significantly when the new information diverges from their prior beliefs.

Specifically, a negative γ_k indicates that for agents in treatment group k, the higher the prior expectation $E_j\pi^{\text{pre}}$, the larger the adjustment in the posterior expectation $E_j\pi^{\text{post}}$ in response to the treatment, compared to the control group. This reflects the notion that agents with more extreme prior beliefs revise their expectations more upon receiving informative signals (Armantier et al., 2016).

Therefore, in the context of Equation (2), negative γ_k coefficients suggest that the information treatments effectively lead agents to revise their expectations away from their priors, with more informative treatments resulting in larger absolute values of γ_k. This aligns with Bayesian updating principles and indicates that the treatments are impactful in shaping agents' inflation expectations (Coibion et al., 2018, 2023). The β_k coefficients capture the level effect of the treatments, representing the average shift in expectations in response to the information provided in each treatment group relative to the control group.

This empirical model helps us understand how AI agents integrate new information with their pre-existing beliefs, quantifying the adjustments made in their inflation expectations in response to different types of information treatments compared to a control group.

4.2.2 Analysis

To evaluate the effects of the information treatments, I estimate Equation (2) separately across different model architectures. Table 4 presents the regression results for both short-run (1 year ahead) and long-run (3 years ahead) inflation expectations across three major model types: GPT-4.1, Claude 3.7 Sonnet, and Llama-3. The coefficients θ_{SR} and θ_{LR} represent the weight placed on prior expectations for the control group. The significant and positive values of these coefficients across all models indicate substantial persistence in expectations, consistent with adaptive expectations models, though with notable variation in magnitude.

The consistently negative and significant γ coefficients suggest that the information treatments lead AI agents to adjust their expectations away from their priors, with more informative treatments having a larger impact. [12] This pattern is consistent with a form of Bayesian updating, where new information causes agents to revise their prior beliefs, similar to findings by Armantier et al. (2016); Huber et al. (2023). For instance, in the short-run expectations for Claude 3.7, Treatment 7 (Inflation + 1Y Projection) has the largest negative γ coefficient, indicating that providing forward-looking inflation information significantly influences expectation formation. This finding aligns with some studies on human subjects which show that forward guidance can influence

[12] In this framework, κ_k is directly comparable to information-rigidity parameters in canonical expectation-formation models. From Equations (1)–(4), for the control group I recover $\kappa_1 = 1 - \theta$, and for treatment k I recover $\kappa_k = 1 - (\theta + \gamma_k)$, where θ is the coefficient on the prior and γ_k is the interaction with the prior for treatment k. In the sticky-information model of Mankiw and Reis (2002), κ maps one-for-one to the fraction of agents updating each period, with quarterly update probabilities typically around 0.25, implying gradual information diffusion. Empirical estimates for professional forecasters by Coibion and Gorodnichenko (2012, 2015) place κ between 0.10 and 0.40, consistent with substantial but incomplete updating. By contrast, the LLM-based estimates are substantially higher—ranging from 0.56 (GPT-4.1, SR) to 0.98 (Llama-3, LR), with Claude 3.7 at 0.60 (SR) and 0.73 (LR), GPT-4.1 at 0.56 (SR) and 0.69 (LR), and Llama-3 at 0.98 for both SR and LR—indicating that these models incorporate new information far more completely.

inflation expectations (Coibion et al., 2022). However, it's important to note that the literature has mixed results on the effectiveness of forward guidance, with its impact often depending on the type of information provided and how it is communicated (D'Acunto et al., 2022).

The strong responsiveness of LLMs to forward guidance mirrors established patterns in human expectation formation, though with some notable differences. Cole (2021) demonstrates that learning dynamics significantly influence the effectiveness of forward guidance in human subjects, with clarity and consistency of communication being crucial factors. Similarly, Campbell et al. (2012) show that forward guidance can effectively shape expectations when it provides clear signals about future policy paths. The AI agents' particularly strong response to forward guidance, especially in Treatments 7 and 8, aligns with Hallett and Acocella (2018)'s finding that explicit future-oriented communications can anchor expectations. However, as de Haan and Sturm (2019) note, while humans often exhibit varying degrees of attention and interpretation to forward guidance, Table 4 indicates that LLMs show a more uniform and immediate response pattern to text-based communications, which is expected given their fundamental nature as language models trained on text data.

Comparing the results across models reveals significant differences in treatment impacts. For instance, Treatment 8 (Inflation + 1Y and Long run) generates a positive level effect (β_8) for Claude 3.7 and GPT-4.1, but a negative effect for Llama-3. Similarly, the γ coefficients show substantial variation, with Claude 3.7 exhibiting much larger interaction effects than Llama-3 for most treatments. These cross-model differences underscore the importance of model selection in economic analysis using LLMs, as different architectures appear to process and integrate economic information in fundamentally different ways.

Figure 3 visualizes how different AI models and human survey participants respond to a range of economic information treatments. It reveals clear differences in how various architectures process information, with proprietary models (Claude 3.7, GPT-4.1) generally exhibiting more consistent patterns than open-source alternatives.

Table 4: Treatment Effects on Short-Run and Long-Run Inflation Expectations

	$E[\pi]$ **1 Year Ahead (SR)**			$E[\pi]$ **3 Years Ahead (LR)**		
	Claude 3.7	**GPT-4.1**	**Llama-3**	**Claude 3.7**	**GPT-4.1**	**Llama-3**
Intercept (Control Group)	1.65***	1.03***	3.27***	2.31***	1.36***	2.85***
θ_{SR} (Control Group)	0.65***	0.56***	0.07***	–	–	–
θ_{LR} (Control Group)	–	–	–	0.29***	0.41***	0.09***
Relative to Control:						
Population Growth	0.94***	0.38***	-0.66***	0.13***	0.14***	-0.17***
Current FFR	0.15**	-0.10*	-0.25**	-0.29***	0.03	-0.16***
FFR + 1Y Proj	0.66***	0.12**	-0.09	-0.18***	0.15***	-0.22***
FFR + 1Y and Long run	0.30***	0.26***	0.10	-0.02	0.19***	-0.39***
3Y Inflation	0.43***	0.24***	0.54***	0.06	0.13**	0.37***
1Y Inflation	1.00***	0.59***	-0.83***	0.00	0.34***	-0.55***
Infl + 1Y Proj	1.34***	1.12***	-0.47***	-0.29***	0.85***	-0.03
Infl + 1Y and Long run	0.86***	1.39***	-0.62***	-0.19***	0.69***	-0.82***
Mortgage Rate	0.47***	0.14**	0.71***	-0.06	0.01	0.81***
γ_1	-0.35***	-0.15***	-0.03	-0.05***	-0.07***	-0.05*
γ_2	-0.07***	0.07**	0.04	0.06***	-0.01	-0.05**
γ_3	-0.22***	-0.03	-0.01	0.00	-0.05*	-0.07***
γ_4	-0.08***	-0.08***	-0.19***	-0.02	-0.04	-0.11***
γ_5	-0.11***	0.00	-0.10***	0.05**	-0.04*	-0.04*
γ_6	-0.40***	-0.18***	-0.05**	-0.02	-0.11***	-0.08***
γ_7	-0.57***	-0.32***	-0.06**	0.01	-0.26***	-0.10***
γ_8	-0.46***	-0.44***	-0.06**	-0.25***	-0.38***	-0.09***
γ_9	-0.19***	-0.07***	0.01	0.02	-0.03	-0.07***
R-squared	0.77	0.63	0.85	0.82	0.52	0.83
Observations	7574	7567	7574	7569	7565	7574

Note: *** p<0.01, ** p<0.05, * p<0.1. θ_{SR} and θ_{LR} represent coefficients for pre-treatment expectations in the short-run and long-run models respectively. Treatment effects are relative to the control group.

Notably, AI models show strong and varied reactions to forward guidance treatments (T7 and T8), while diverging significantly from human survey responses, particularly for mortgage rate information (T9), which elicited larger effects in human data. These contrasts highlight the importance of model selection when using LLMs to simulate economic expectations and suggest that different architectures may internalize and represent economic signals in systematically different ways.

The observed differences between AI and human responses highlight important considerations for using LLMs in economic research. AI agents show more consistent and predictable reactions to information treatments, potentially due to their lack of personal biases or experiences. However, this consistency may overlook the nuanced and sometimes irrational ways humans process economic information (Coibion et al., 2022; Weber et al., 2022). It's crucial to remember that LLMs, as transformer-based

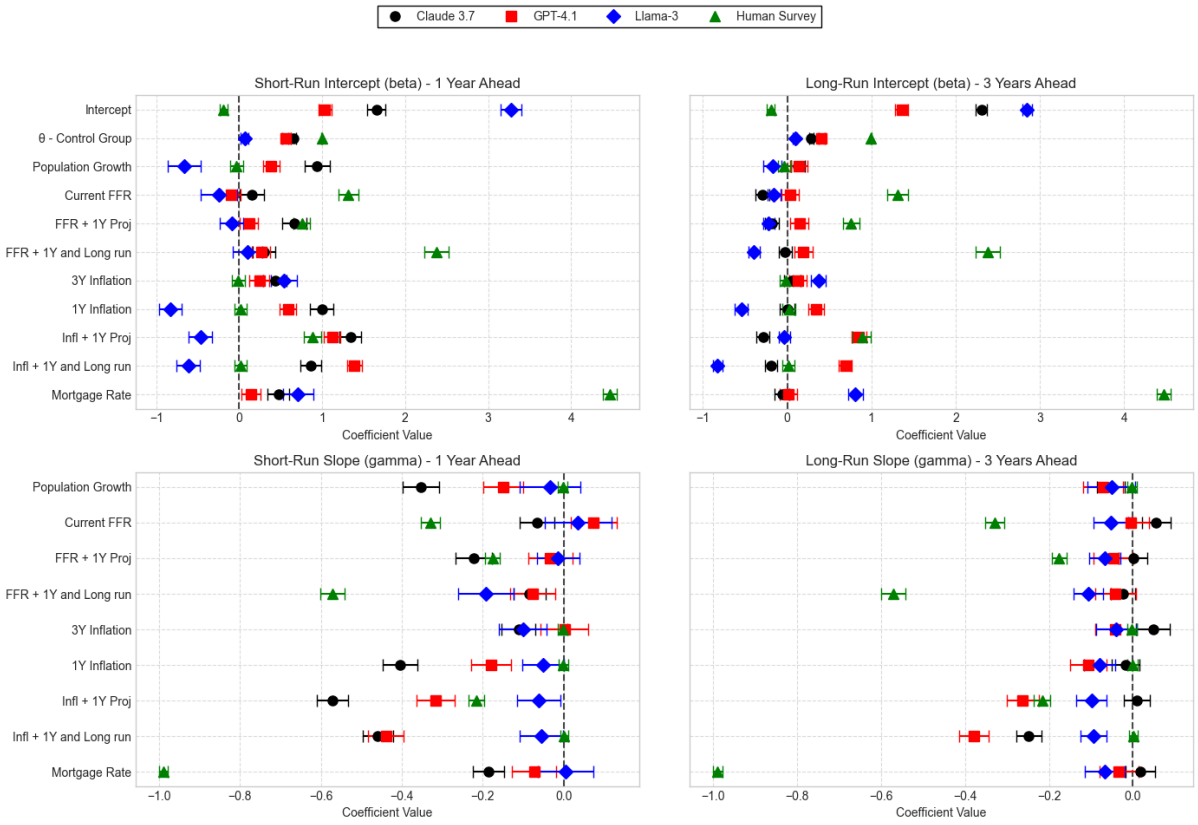

Figure 3: Comparison Between LLMs and Human Survey Data

Notes: Human survey data are drawn from Coibion et al. (2023), based on a different time period (2019) and a separate sample (Kilts–Nielsen Consumer Panel), not the Survey of Consumer Expectations. While the time and panel differ, the information treatments are designed to be equivalent across both LLMs and humans. Results are not directly comparable but highlight differences in responsiveness to the same economic signals.

models, are fundamentally designed to predict the next most likely word (token) or sequence based on patterns in their training data, rather than truly reasoning about economic concepts (Brown et al., 2020).

The integration of reinforcement learning from human feedback (RLHF) in LLM training represents a significant advancement in aligning AI behavior with human preferences and decision-making patterns (Ouyang et al., 2022). This approach could potentially bridge the gap between AI consistency and human nuance in economic reasoning, allowing LLMs to better capture the complexities of human economic decision-making (Bai et al., 2022). However, it also raises questions about whose feedback is incorporated and how this might influence the model's economic perspectives.

The primary motivation for this paper's experiments lies in the growing impor-

tance of AI-assisted decision making across various economic domains (Korinek, 2023; Chang et al., 2024). As AI systems increasingly influence financial markets, policy analysis, and individual economic choices, it becomes critical to understand their underlying mechanisms and how they interpret economic information. This understanding is essential for predicting how AI might shape market dynamics, influence policy effectiveness, and interact with human decision-makers in mixed human-AI economic environments (Fedyk et al., 2024; Zhu et al., 2024).

While LLMs' strong responsiveness to clear signals makes them highly adaptable, it may lead to overestimating the effectiveness of certain policy communications in real-world scenarios. The incorporation of RLHF could potentially mitigate this by introducing more human-like variability and context-sensitivity in responses (Christiano et al., 2017). However, these differences still underscore the need for careful calibration when using AI models to predict or simulate human economic behavior, especially as we move towards a future where AI plays a more significant role in economic decision-making processes (Korinek, 2024; Acemoglu, 2024).

4.2.3 Persona and AI Agents' Expectations

The demographic variables in the persona-based experiments (Table 5) reveal distinctive patterns across different model architectures. For Claude 3.7, inflation expectations consistently increase with age, with the oldest age groups showing the highest expectations compared to younger cohorts. Education similarly demonstrates a clear gradient effect, with less formal education associated with higher inflation expectations. Income displays a consistent negative relationship with inflation expectations across all models, reflecting the pattern that lower-income personas generally anticipate higher inflation.

Interestingly, these demographic effects vary substantially across different LLM architectures. GPT-4.1 shows much weaker and sometimes contradictory demographic patterns compared to Claude 3.7 and Llama-3. For instance, while higher education strongly correlates with lower inflation expectations in Claude 3.7 and Llama-3,

this relationship is much less pronounced in GPT-4.1. Similarly, gender effects differ markedly between models, with male personas associated with significantly lower inflation expectations in Claude 3.7 but showing the opposite relationship in Llama-3.

These cross-model differences suggest that different LLM architectures may incorporate demographic information in distinct ways when forming economic expectations. The patterns observed in Claude 3.7 and Llama-3 more closely mirror findings from human survey data (Huber et al., 2023; D'Acunto et al., 2022), where lower education and income typically correlate with higher inflation expectations. This suggests that these models may be more effectively capturing realistic demographic influences on expectation formation.

Table 5: Demographic Effects on Short-Run and Long-Run Inflation Expectations

	$E[\pi]$ **1 Year Ahead (SR)**			$E[\pi]$ **3 Years Ahead (LR)**		
	Claude 3.7	**GPT-4.1**	**Llama-3**	**Claude 3.7**	**GPT-4.1**	**Llama-3**
Demographic Variables:						
Age Group: 31–40	0.03***	0.01	0.01	0.01**	0.01	0.00
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Age Group: 41–50	0.06***	0.01	0.02**	0.02***	0.02**	0.01
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Age Group: 51–60	0.07***	0.01	0.02**	0.03***	0.01	0.00
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Age Group: 60+	0.10***	-0.02**	0.01	0.04***	0.00	0.00
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Gender: Male	-0.06***	-0.01	0.01**	-0.02***	-0.01**	0.01*
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Education: Some College	0.15***	-0.01	0.06***	0.05***	0.00	0.04***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Education: High School	0.29***	-0.03***	0.14***	0.08***	-0.02**	0.11***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Income: 50k–100k	-0.20***	0.02***	-0.07***	-0.04***	0.01***	-0.06***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Income: Over 100k	-0.22***	0.03***	-0.11***	-0.07***	0.01**	-0.10***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
R-squared	0.77	0.63	0.85	0.82	0.52	0.83
Observations	7574	7567	7574	7569	7565	7574

Note: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. The base categories are: Age Group 18–30, Gender Female, Education College, and Income Under 50k. These results are from the same regression specifications as in Table 4, controlling for treatment effects and pre-treatment expectations.

These findings have important implications for both the development of AI models in forecasting and the design of effective communication strategies for central banks

and policymakers. They suggest that AI agents, particularly those with assigned personas, can serve as valuable tools for simulating and predicting the impacts of various information dissemination strategies on public inflation expectations. However, the differences between AI and human responses highlight the need for careful calibration and interpretation when using AI models to predict human economic behavior.

4.3 Central Bank Communication Strategies

This section examines how GPT-4.1 and Llama 3-70B respond to variations in central bank communication, providing a simulation framework for testing alternative messaging strategies prior to public release. The experiment evaluates model responses to 18 systematically varied information treatments across three key dimensions: language complexity[13](simplified vs. technical), commitment framing (conditional vs. unconditional), and time horizon (short-term vs. long-term guidance)[14]. This approach extends the work of Blinder et al. (2008) and Hansen et al. (2019), who emphasize the importance of communication clarity in the effectiveness of monetary policy.

As LLMs become more integrated into economic analysis and policy environments, understanding how specific architectures interpret policy signals is increasingly important. In this context, large language models serve as a sandbox environment, allowing policymakers to test how different models like GPT-4.1 and Llama 3-70B internalize and respond to alternative communication strategies. This controlled setting reveals how variations in message framing may influence expectation formation in AI systems that increasingly assist in macroeconomic forecasting and analysis.

Results in Table 6 indicate notable variation in how different LLMs respond to central bank communication strategies, reflecting these models' sensitivity to content framing and complexity. For instance, technical language produced sharper re-

[13]In the messages used, simplified treatments average a Flesch Reading Ease score of 45.48 (higher numbers indicate easier readability on a 0-100 scale) versus 2.79 for technical treatments.

[14]Refer to Section SA-7 in the Appendix for the details on each treatment. The messages were generated by Claude 3.7 Sonnet and evaluated by GPT-4o to match the purpose and tone of each policy treatment.

visions in inflation expectations than simplified language in certain cases, particularly under hawkish scenarios, suggesting that the models interpret more formal wording as stronger policy signals. Interestingly, while GPT-4.1 shows moderate adjustments, Llama 3-70B demonstrates more substantial reactions to language complexity, with hawkish technical language reducing short-run expectations by over 1 percentage point. These patterns support the view that communication clarity and accessibility are central to effective monetary policy, consistent with insights from Bholat et al. (2019) and Haldane and McMahon (2017).

Table 6: Model Response by Communication Treatment

Treatment	GPT 4.1						Llama 3 70B					
	Short-Run			Long-Run			Short-Run			Long-Run		
	Mean	Diff	SD	Mean	Diff	SD	Mean	Diff	SD	Mean	Diff	SD
T_0 (Control)	3.34	0.00	0.23	2.81	0.00	0.29	3.37	0.00	0.46	2.52	0.00	0.12
a. Language Complexity												
T_{a1} (Neutral - Simplified)	3.30	-0.04	0.16	2.51	-0.30	0.19	2.72	-0.65	0.33	2.51	-0.01	0.10
T_{a2} (Neutral - Technical)	3.28	-0.06	0.17	2.47	-0.34	0.16	2.86	-0.51	0.36	2.53	+0.01	0.12
T_{a3} (Dovish - Simplified)	3.38	+0.04	0.24	3.16	+0.35	0.21	2.47	-0.90	0.09	2.49	-0.03	0.05
T_{a4} (Dovish - Technical)	3.36	+0.02	0.19	3.14	+0.34	0.16	2.45	-0.91	0.11	2.48	-0.04	0.06
T_{a5} (Hawkish - Simplified)	3.28	-0.06	0.19	2.49	-0.32	0.19	3.53	+0.16	0.56	2.50	-0.02	0.02
T_{a6} (Hawkish - Technical)	3.15	-0.19	0.16	2.22	-0.59	0.09	2.35	-1.02	0.31	1.93	-0.59	0.23
b. Policy Commitment Framing												
T_{b1} (Neutral - Conditional)	3.17	-0.17	0.20	2.46	-0.35	0.14	2.51	-0.85	0.10	2.50	-0.03	0.06
T_{b2} (Neutral - Unconditional)	3.22	-0.12	0.19	2.49	-0.32	0.20	2.56	-0.81	0.20	2.47	-0.05	0.08
T_{b3} (Dovish - Conditional)	3.03	-0.31	0.26	2.47	-0.34	0.19	2.21	-1.16	0.02	2.23	-0.29	0.04
T_{b4} (Dovish - Unconditional)	3.30	-0.04	0.16	2.67	-0.13	0.24	2.24	-1.13	0.11	2.32	-0.21	0.12
T_{b5} (Hawkish - Conditional)	3.17	-0.18	0.18	2.32	-0.48	0.07	3.18	-0.19	0.14	2.49	-0.03	0.05
T_{b6} (Hawkish - Unconditional)	2.96	-0.38	0.28	2.11	-0.69	0.04	3.20	-0.16	0.05	2.42	-0.10	0.13
c. Time Horizons of Guidance												
T_{c1} (Neutral - Short-term)	3.22	-0.12	0.16	2.44	-0.37	0.13	2.69	-0.68	0.33	2.51	-0.01	0.07
T_{c2} (Neutral - Long-term)	3.01	-0.33	0.25	2.19	-0.62	0.09	2.46	-0.91	0.10	2.14	-0.38	0.07
T_{c3} (Dovish - Short-term)	3.30	-0.04	0.16	2.76	-0.05	0.25	2.79	-0.58	0.42	2.50	-0.02	0.01
T_{c4} (Dovish - Long-term)	3.35	+0.01	0.20	3.15	+0.35	0.19	2.50	-0.87	0.02	2.50	-0.02	0.02
T_{c5} (Hawkish - Short-term)	3.04	-0.30	0.24	2.34	-0.47	0.09	3.48	+0.11	0.57	2.50	-0.02	0.01
T_{c6} (Hawkish - Long-term)	3.11	-0.23	0.22	2.24	-0.57	0.08	3.26	-0.11	0.22	2.44	-0.08	0.19

Note: This table reports inflation expectations generated by GPT-4.1 and Llama 3-70B across monetary policy communication treatments. "Short-Run" refers to 1-year and "Long-Run" to 3-year expectations. "Diff" indicates the difference from the control group, which received no information. Each treatment group contains the same 500 randomly selected AI agents with identical demographic characteristics to ensure comparability across treatments, totaling 19,000 observations per model. Treatments vary by language complexity, policy commitment framing, and time horizon. "Hawkish" refers to anti-inflation messaging; "Dovish" to growth-supportive messaging; "Neutral" to a balanced stance. Results on demographic heterogeneity in responses to these treatments are presented in the Supplementary Appendix Section SA-8.

Policy commitment framing also mattered: unconditional statements generally led to larger shifts in expectations than conditional ones, especially in the short run and

under hawkish guidance. This aligns with findings from Campbell et al. (2012), who show that forward guidance impacts are stronger when the communicated commitment is clear and firm. Similarly, the effects observed here support Ehrmann and Talmi (2012), who argue that clarity and consistency in communication improve expectation anchoring.

Time horizon treatments showed that long-term guidance produced more persistent effects on long-run inflation expectations compared to short-term statements. This finding is consistent with Eusepi et al. (2018), who emphasize the role of horizon framing in shaping expectation dynamics. It also resonates with Afrouzi et al. (2020), who highlight that households process economic forecasts with varying levels of attention and temporal focus.

To translate these descriptive results into a policy-relevant counterfactual, I use the estimated γ_k coefficients from Section 4.1 to simulate the impact of a 50 basis point hawkish shift in communicated inflation guidance. Focusing on the "Inflation + 1Y and Long run" treatment (T8), where the signal is an explicit numerical projection, the implied weight on the signal is given by $\kappa_k = 1 - (\theta + \gamma_k)$ for each model and horizon. Applying $\Delta S = -0.50$ pp to these κ_k values yields predicted changes in short-run (long-run) expected inflation of approximately -0.44 (-0.49) for GPT-4.1, -0.41 (-0.48) for Claude 3.7 Sonnet, and -0.50 (-0.50) for Llama 3-70B. These magnitudes indicate that a modest hawkish adjustment in communicated inflation projections could meaningfully lower aggregate expected inflation across all models, with the pass-through close to one-for-one in some cases. This exercise illustrates how the descriptive updating patterns in Table 4 can be mapped into concrete policy effects, offering central banks a quantitative sense of the stakes when using AI-based expectation models to test alternative communication strategies.

This experiment illustrates how multiple LLMs can serve as a testing environment for evaluating the design of monetary policy communication. By simulating expectation formation in a controlled setting, this framework offers central banks a novel sandbox for optimizing their messaging strategies before public release. In doing so,

it complements the broader literature on heterogeneity in the reception and impact of monetary policy communications (Blinder et al., 2008; D'Acunto et al., 2022).

5 Discussion

5.1 Heterogeneity Across Models

Running the experiment across multiple models reveals important differences in how various LLM architectures form and update inflation expectations. Three key dimensions of variation emerge: model size, provider differences, and proprietary versus open-source implementation.

First, model size significantly influences expectation formation. Larger models (e.g., GPT-4.1 and Claude 3.7 Sonnet) generally produce more consistent responses to information treatments, with lower variance in their predictions compared to smaller models (e.g., GPT-4o-mini, Claude 3.5 Haiku), as shown in Table 1. However, this relationship is not strictly linear. For example, Claude 3.5 Haiku occasionally outperforms its larger counterpart in aligning with human survey patterns, suggesting that architectural design can sometimes compensate for parameter count (Brown et al., 2020).

Second, systematic differences are observable across model providers. OpenAI models tend to produce lower baseline inflation expectations than Anthropic models. This provider effect persists across model sizes and treatments, indicating underlying differences in training data, modeling assumptions, or design choices across companies (Bommasani et al., 2021).

Third, proprietary models exhibit more consistent and moderate responses to information treatments than open-source alternatives. While Llama 3-70B performs competitively on many metrics, open-source models show greater volatility in response to identical treatments. DeepSeek-V3, in particular, can swing wildly under the same conditions. One likely reason is that commercial models get an extra layer of "polishing" with extensive human feedback that nudges them toward steadier answers, whereas most open-source models are released in a rougher, less-filtered form that

lets small prompt changes produce much bigger jumps (Kukreja et al., 2024).

These findings underscore the importance of using diverse model architectures when conducting economic simulations. Systematic differences across models suggest that researchers and policymakers should adopt multi-model approaches, especially when evaluating communication strategies or forecasting heterogeneous responses to policy announcements.

5.2 Limitations of LLM-Based Economic Simulations

While LLMs offer a flexible and low-cost laboratory for studying expectation formation, they present clear limitations when used as stand-ins for real households in economic analysis.

First, LLMs do not experience real-world constraints, such as intertemporal budget constraints, income shocks, or actual consumption. Their reported "beliefs" are grounded in statistical patterns extracted from text, not lived experience, and thus cannot fully capture how households form expectations during economic shocks or periods of uncertainty.

Second, the way LLMs process information differs fundamentally from human cognition. LLMs process information through learned attention patterns that may not align with human cognitive salience, while households typically focus on salient or frequently encountered prices (e.g., groceries, gas) and often overlook more abstract variables (D'Acunto et al., 2021). Although prompt engineering can nudge LLMs toward more human-like behavior, such adjustments are imperfect and lack full transparency. In some cases, alignment methods such as Reinforcement Learning from Human Feedback (RLHF) may introduce unintended biases rather than eliminate them.RetryClaude can make mistakes. Please double-check responses.

Third, LLMs are limited by the data they were trained on. Most frontier models incorporate information only up to a few months before deployment (or earlier in the case of open-source models). Without regular retraining or online access, they cannot dynamically update their knowledge in response to new data. In contrast, households

continuously revise their expectations in response to changing economic conditions and real-world experiences (Coibion et al., 2018).

Fourth, the statistical properties of LLM-generated forecasts may not align with those of human surveys, as shown in Figure 3. Their outputs are often overly smooth, lacking the variability or representation of tail risks that characterize real-world expectations data. Finally, LLMs can exhibit representational biases. They are primarily trained on English-language, internet-based, and often affluent or urban-centric content. As a result, they may underrepresent the perspectives of low-income, less digitally connected, or rural households (Guo et al., 2024). Even when detailed persona prompts are used, these deeper sample composition issues may persist and could bias aggregate estimates if LLMs are treated as substitutes for survey data.

For all these reasons, LLMs are best viewed as complementary tools, well suited for rapid, low-cost testing of central bank communication strategies and for generating new research hypotheses. However, they should not be considered substitutes for traditional household surveys or field-based research.

6 Concluding Remarks

This study provides new insights into the use of LLM-based AI agents for modeling inflation expectations and explores the broader potential of Generative AI and LLMs in the economics of expectations. While LLMs possess notable forecasting capabilities, the main contribution of this work is to show how these models process economic information, respond to new data, and update their beliefs in ways that often resemble patterns observed among human respondents (Bybee, 2025; Hansen et al., 2025).

Although LLMs can simulate key aspects of expectation formation, they are subject to important constraints. Their forecasts and belief-updating processes are shaped by biases in training data, fixed knowledge cutoffs, and limitations in representing real-world behavior (Bender et al., 2021). As such, LLMs offer valuable perspectives on how expectations might evolve, while also reflecting some of the same challenges seen

in traditional survey studies.

A central finding of this research is that information treatments, especially those involving forward guidance on inflation, show the strongest influence on LLMs' inflation expectations. This parallels similar effects documented in studies of human participants. Additionally, some LLMs tend to predict higher inflation than realized values, echoing the persistent upward bias observed in household surveys such as the Survey of Consumer Expectations.

The use of persona-based prompts further illustrates the potential of LLMs to capture heterogeneity in expectation formation. Assigning agents demographic characteristics such as age, gender, income, education, or marital status generates systematic variation in responses, reflecting patterns often found in real survey data. For instance, lower-income and less-educated personas consistently report higher inflation expectations across models, mirroring well-documented socioeconomic disparities in human surveys (Bryan and Venkatu, 2001; D'Acunto et al., 2022). This suggests that LLMs may serve as useful tools for exploring how different economic agents interpret the same information.

As AI tools become more integrated into economic forecasting and policy analysis, understanding how LLMs reason about expectations takes on growing importance. While AI-generated insights can help individuals and policymakers process complex data, they may also introduce biases . For instance, if a model consistently overweights certain economic indicators, users may unknowingly adopt similarly skewed interpretations, especially when prior beliefs are weak or uncertain.

Despite their promise, LLMs face several important limitations. They may struggle to simulate responses to unprecedented shocks, lack nuanced understanding of cultural or emotional factors, and, despite having persona, cannot fully replicate the complex dynamics of experience, social context, and psychology that shapes real-world economic behavior (Echterhoff et al., 2024). These limitations are particularly relevant in novel or fast-changing environments where historical patterns may be insufficient guides.

This research also demonstrates how central banks could leverage these models as communication policy tools to test messaging strategies before implementation (Blinder et al., 2008; Hansen et al., 2019). The experiment with different communication approaches—varying language complexity, commitment framing, and time horizon—show that LLMs respond differently to alternative formulations, potentially offering insights into optimizing policy communications.

In summary, this paper demonstrates both the promise and boundaries of integrating generative AI into economic research. By leveraging the capabilities of LLMs, researchers can gain new insights into expectation formation and improve the design of communication strategies and monetary policy. At the same time, LLMs are best understood as complementary tools, useful for generating hypotheses, testing ideas, and supporting rapid analysis, rather than as substitutes for traditional survey methods or structural economic models.

Supplementary Appendix

Supplementary material related to this article can be found online at ???.

Data availability

Replication Package can be found at: https://github.com/alizarif/JME_AI-and-Inflation
The data is available at: Harvard Dataverse. https://doi.org/doi:10.7910/DVN/UJUFIN

References

D. Acemoglu. The simple macroeconomics of ai. *National Bureau of Economic Research*, 2024.

H. Afrouzi and L. L. Veldkamp. Biased inflation forecasts. Unpublished manuscript, 2019.

H. Afrouzi, S. Y. Kwon, A. Landier, Y. Ma, and D. Thesmar. Overreaction and working memory. *National Bureau of Economic Research*, 2020.

H. Afrouzi, S. Y. Kwon, A. Landier, Y. Ma, and D. Thesmar. Overreaction in expectations: Evidence and theory. *The Quarterly Journal of Economics*, 138(3):1713–1764, 2023.

H. Afrouzi, A. Dietrich, K. Myrseth, R. Priftis, and R. Schoenle. Inflation preferences. *National Bureau of Economic Research*, 2024.

E. Akata, L. Schulz, J. Coda-Forno, S. J. Oh, M. Bethge, and E. Schulz. Playing repeated games with large language models. *arXiv preprint arXiv:2305.16867*, 2023.

I. Aldasoro, O. Armantier, S. Doerr, L. Gambacorta, and T. Oliviero. Survey evidence on gen ai and households: job prospects amid trust concerns. *Bank for International Settlements*, 2024.

M.-A. Allard, P. Teiletche, and A. Zinebi. Enhancing inflation nowcasting with llm: Sentiment analysis on news, 2024.

O. Armantier, S. Nelson, G. Topa, W. van der Klaauw, and B. Zafar. The price is right: Updating inflation expectations in a randomized price information experiment. *The Review of Economics and Statistics*, 98(3):503–523, 2016.

Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

I. Baley and L. Veldkamp. Bayesian learning. In *Handbook of economic expectations*, pages 717–748. Elsevier, 2023.

R. Batista and J. Ross. Words that work: Using language to generate hypotheses. 2024.

E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623. ACM, 2021.

B. S. Bernanke and K. N. Kuttner. What explains the stock market's reaction to federal reserve policy? *The Journal of finance*, 60(3):1221–1257, 2005.

D. Bholat, N. Broughton, J. Ter Meer, and E. Walczak. Enhancing central bank communications using simple and relatable information. *Journal of Monetary Economics*, 108:1–15, 2019.

A. Bick, A. Blandin, and D. J. Deming. The rapid adoption of generative ai. *National Bureau of Economic Research*, 2024.

A. Binetti, F. Nuzzi, and S. Stantcheva. People's understanding of inflation. *Journal of Monetary Economics*, 148:103652, 2024.

A. S. Blinder, M. Ehrmann, M. Fratzscher, J. De Haan, and D.-J. Jansen. Central bank communication and monetary policy: A survey of theory and evidence. *Journal of economic literature*, 46(4):910–945, 2008.

V. R. Boctor, O. Coibion, Y. Gorodnichenko, and M. Weber. On eliciting subjective probability distributions of expectations. *National Bureau of Economic Research*, 2024.

R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

P. Brookins and J. M. DeBacker. Playing games with gpt: What can we learn about a large language model from canonical strategic games? *Available at SSRN 4493398*, 2023.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

W. Bruine de Bruin, W. Vanderklaauw, J. S. Downs, B. Fischhoff, G. Topa, and O. Armantier. Expectations of inflation: The role of demographic variables, expectation formation, and financial literacy. *Journal of Consumer Affairs*, 44(2):381–402, 2010.

M. F. Bryan and G. Venkatu. The demographics of inflation opinion surveys. *Economic Commentary, Federal Reserve Bank of Cleveland*, 2001.

L. Bybee. The ghost in the machine: Generating beliefs with large language models. *arXiv preprint arXiv:2305.02823*, 2025.

J. R. Campbell, C. L. Evans, J. D. Fisher, A. Justiniano, C. W. Calomiris, and M. Woodford. Macroeconomic effects of federal reserve forward guidance. *Brookings papers on economic activity*, pages 1–80, 2012.

B. Candia, O. Coibion, and Y. Gorodnichenko. Communication and the beliefs of economic agents. *National Bureau of Economic Research*, 2020.

A. Cavallo, G. Cruces, and R. Perez-Truglia. Inflation Expectations, Learning and Supermarket Prices. *National Bureau of Economic Research*, 10 2014.

S. Chang, A. Kennedy, A. Leonard, and J. A. List. 12 best practices for leveraging generative ai in experimental research. *National Bureau of Economic Research*, 2024.

G. Charness, B. Jabarian, and J. A. List. The next generation of experimental research with llms. *Nature Human Behaviour*, Mar. 2025.

C. Chen, B. Yao, Y. Ye, D. Wang, and T. J.-J. Li. Evaluating the llm agents for simulating humanoid behavior. 2024.

P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. In *Advances in neural information processing systems*, pages 4299–4307, 2017.

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

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

O. Coibion, Y. Gorodnichenko, and S. Kumar. How do firms form their expectations? new survey evidence. *American Economic Review*, 108(9):2671–2713, 2018.

O. Coibion, Y. Gorodnichenko, S. Kumar, and M. Pedemonte. Inflation expectations as a policy tool? *Journal of International Economics*, 124:103297–103297, 2020a.

O. Coibion, Y. Gorodnichenko, and T. Ropele. Inflation expectations and firm decisions: New causal evidence. *The Quarterly Journal of Economics*, 135(1):165–219, 2020b.

O. Coibion, Y. Gorodnichenko, and M. Weber. Monetary policy communications and their effects on household inflation expectations. *Journal of Political Economy*, 130(6): 1537–1584, 2022.

O. Coibion, D. Georgarakos, Y. Gorodnichenko, and M. Weber. Forward guidance and household expectations. *Journal of the European Economic Association*, pages 2131–2171, 2023.

S. J. Cole. Learning and the effectiveness of central bank forward guidance. *Journal of Money, Credit and Banking*, 53(1):157–200, 2021.

F. D'Acunto, U. Malmendier, J. Ospina, and M. Weber. Exposure to grocery prices and inflation expectations. *Journal of Political Economy*, 129(5):1615–1639, 2021.

F. D'Acunto, D. Hoang, M. Paloviita, and M. Weber. Cognitive abilities and inflation expectations. *American Economic Review*, 112(4):1147–1191, 2022.

J. de Haan and J.-E. Sturm. *Central Bank Communication*, pages 231–262. Oxford University Press, mar 14 2019.

F. Di Giacomo and C. Angelico. Heterogeneity in Inflation Expectations and Personal Experience. *SSRN Electronic Journal*, 2019.

F. D'Acunto, U. Malmendier, and M. Weber. Gender roles produce divergent economic expectations. *Proceedings of the National Academy of Sciences*, 118(21), may 18 2021.

F. D'Acunto, D. Hoang, and M. Weber. Managing households' expectations with unconventional policies. *The Review of Financial Studies*, 35(4):1597–1642, 2022.

J. Echterhoff, Y. Liu, A. Alessa, J. McAuley, and Z. He. Cognitive bias in high-stakes decision-making with llms. *arXiv preprint arXiv:2403.00811*, 2024.

M. Ehrmann and J. Talmi. Communicating about macro-prudential supervision–a new challenge for central banks. *International Finance*, 20(2):97–115, 2012.

M. Ehrmann, D. Pfajfar, and E. Santoro. Consumer's attitudes and their inflation expectation. *International Journal of Central Banking*, 13:225, 2017.

S. Eusepi and B. Preston. Central bank communication and expectations stabilization. *American Economic Journal: Macroeconomics*, 2(3):235–271, 2010.

S. Eusepi, M. P. Giannoni, and B. Preston. The limits of forward guidance. *Journal of Economic Dynamics and Control*, 2018:1–4, 2018.

M. Faria-e Castro and F. Leibovici. Artificial intelligence and inflation forecasts. *Federal Reserve Bank of St. Louis Review*, 106(12):1–14, 2024.

A. Fedyk, A. Kakhbod, P. Li, and U. Malmendier. Chatgpt and perception biases in investments: An experimental study. *Available at SSRN 4787249*, 2024.

F. Guo. Gpt in game theory experiments. *arXiv preprint arXiv:2305.05516*, 2023.

Y. Guo, M. Guo, J. Su, Z. Yang, M. Zhu, H. Li, M. Qiu, and S. S. Liu. Bias in large language models: Origin, evaluation, and mitigation, 2024.

I. Haaland, C. Roth, and J. Wohlfart. Designing information provision experiments. *Journal of Economic Literature*, 61(1):3–40, 2023.

A. Haldane and M. McMahon. A little more conversation a little less action. *Bank of England Speech*, 2017.

A. H. Hallett and N. Acocella. Stabilization and commitment: Forward guidance in economies with rational expectations. *Macroeconomic Dynamics*, 22(1):122–134, 2018.

A. L. Hansen and S. Kazinnik. Can chatgpt decipher fedspeak? *Available at SSRN 4399406*, 2023.

A. L. Hansen, J. J. Horton, S. Kazinnik, D. Puzzello, and A. Zarifhonarvar. Simulating the survey of professional forecasters. *Available at SSRN*, 2025.

S. Hansen, M. McMahon, and A. Prat. Transparency, deliberation, and monetary policy. *The Quarterly Journal of Economics*, 133(2):801–870, 2019.

T. Henning, S. M. Ojha, R. Spoon, J. Han, and C. F. Camerer. Llm trading: Analysis of llm agent behavior in experimental asset markets, 2025.

B. Heydari and N. Lorè. Strategic behavior of large language models: Game structure vs. contextual framing. *preprint*, 2023.

J. J. Horton. Large language models as simulated economic agents: What can we learn from homo silicus? *National Bureau of Economic Research*, 2023.

T. Hu and N. Collier. Quantifying the persona effect in llm simulations. *arXiv preprint arXiv:2402.10811*, 2024.

S. J. Huber, D. Minina, and T. Schmidt. *The pass-through from inflation perceptions to inflation expectations*. Number 17/2023. Deutsche Bundesbank Discussion Paper, 2023.

N. Immorlica, B. Lucier, and A. Slivkins. Generative ai as economic agents. *ACM SIGecom Exchanges*, 22(1):93–109, 2024.

J. H. Jiang, R. Kamdar, K. Lu, and D. Puzzello. How do households respond to expected inflation? an investigation of transmission mechanisms. *Bank of Canada Staff Working Paper*, 2024.

R. Kamdar and W. Ray. Polarized expectations, polarized consumption. *Polarized Consumption (October 31, 2022)*, 2022.

E. Karger, H. Bastani, C. Yueh-Han, Z. Jacobs, D. Halawi, F. Zhang, and P. E. Tetlock. Forecastbench: A dynamic benchmark of ai forecasting capabilities, 2025.

S. Kazinnik. Bank run, interrupted: Modeling deposit withdrawals with generative ai. *SSRN Electronic Journal*, October 2023.

S. Kazinnik and E. Brynjolfsson. Ai and the fed, 2025.

A. Korinek. Generative ai for economic research: Use cases and implications for economists. *Journal of Economic Literature*, 61(4):1281–1317, 2023.

A. Korinek. Economic policy challenges for the age of ai. 2024.

S. Kukreja, T. Kumar, A. Purohit, A. Dasgupta, and D. Guha. A literature survey on open source large language models. In *Proceedings of the 2024 7th International Conference on Computers in Management and Business*, pages 133–143, 2024.

L. Kwok, M. Bravansky, and L. D. Griffin. Evaluating cultural adaptability of a large language model via simulation of synthetic personas. *arXiv preprint arXiv:2408.06929*, 2024.

N. Li, C. Gao, M. Li, Y. Li, and Q. Liao. EconAgent: Large language model-empowered agents for simulating macroeconomic activities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2024.

R. Liu, J. Wei, F. Liu, C. Si, Y. Zhang, J. Rao, S. Zheng, D. Peng, D. Yang, D. Zhou, et al. Best practices and lessons learned on synthetic data for language models. *arXiv preprint arXiv:2404.07503*, 2024.

A. Lopez-Lira, Y. Tang, and M. Zhu. The memorization problem: Can we trust llms' economic forecasts? *Available at SSRN 5217505*, 2025.

N. G. Mankiw and R. Reis. Sticky information versus sticky prices: A proposal to replace the new keynesian phillips curve. *The Quarterly Journal of Economics*, 117(4): 1295–1328, 2002. doi: 10.1162/003355302320935034.

B. S. Manning, K. Zhu, and J. J. Horton. Automated social science: Language models as scientist and subjects. *National Bureau of Economic Research*, 2024.

C. F. Manski. Survey measurement of probabilistic macroeconomic expectations: progress and promise. *NBER Macroeconomics Annual*, 32(1):411–471, 2018.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

D. Pfajfar and B. Žakelj. Inflation expectations and monetary policy design: Evidence from the laboratory. *Macroeconomic dynamics*, 22(4):1035–1075, 2018.

N. Raman, T. Lundy, S. J. Amouyal, Y. Levine, K. Leyton-Brown, and M. Tennenholtz. Steer: assessing the economic rationality of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 42026–42047, 2024.

J. Ross, Y. Kim, and A. W. Lo. Llm economicus? mapping the behavioral biases of llms via utility theory. *arXiv preprint arXiv:2408.02784*, 2024.

S. Stantcheva. Why do we dislike inflation? *National Bureau of Economic Research*, 2024.

M. Tranchero, C.-F. Brenninkmeijer, A. Murugan, and A. Nagaraj. Theorizing with large language models. *National Bureau of Economic Research*, 2024.

M. Weber, F. D'Acunto, Y. Gorodnichenko, and O. Coibion. The subjective inflation expectations of households and firms: Measurement, determinants, and implications. *Journal of Economic Perspectives*, 36(3):157–184, 2022.

J.-Q. Zhu, H. Yan, and T. L. Griffiths. Language models trained to do arithmetic predict human risky and intertemporal choice. *arXiv preprint arXiv:2405.19313*, 2024.

Supplementary Appendix

Generating Inflation Expectations with Large Language Models

Zarifhonarvar (2025)

SA-1 Technical Implementations

Running surveys or experiments on AI agents or LLMs can be conducted through multiple approaches. For proprietary models such as GPT models by OpenAI, Claude models by Anthropic, or Gemini by Google, researchers typically use native API endpoints provided directly by these companies. For open-source models such as Llama or DeepSeek, experiments can be executed either locally—requiring significant computational resources and specialized applications like LM Studio[15] or Ollama[16]—or via third-party cloud-based platforms like Together AI[17], DeepInfra[18], Groq[19], and Amazon Bedrock[20].

When we need to run an experiment across different models, it's easier to use third-party integration frameworks that provide a single endpoint, requiring only one code to run the experiment. LangChain[21] is one of the most widely adopted frameworks for creating second-layer applications on LLM models, and Expected Parrot[22] is another AI infrastructure that facilitates AI-powered research by providing access to many models. With Expected Parrot's easy-to-follow design, we can run experiments that are easily reproducible.

SA-1.1 General Considerations

The implementation process involved several key considerations and strategies:

1. **Model Selection:** I employed a diverse set of both proprietary and open-source LLMs to ensure coverage across model architectures, training methodologies, and parameter sizes. From the proprietary domain, I utilized GPT-4.1, GPT-4o, GPT-4o-mini, Claude 3.7 Sonnet, and Claude 3.5 Haiku, which represent the state-of-the-art commercial offerings. In addition to these, I used leading open-source models including Llama 3-70B and DeepSeek-V3 to provide architectural diversity and to assess whether inflation expectation patterns generalize across different model families. This approach allowed me to test whether observed behaviors were model-specific or reflected broader patterns in how LLMs process economic information.

2. **Temperature:** For all experiments in this paper, I used the default temperature

[15] https://lmstudio.ai/
[16] https://ollama.com/
[17] https://www.together.ai/
[18] https://deepinfra.com/
[19] https://groq.com/
[20] https://aws.amazon.com/bedrock/
[21] https://www.langchain.com/
[22] https://www.expectedparrot.com/

setting for each model.[23] I also conducted some experiments with temperatures lower and higher than the default (See Section SA-3.1).

3. **Prompts:** In the main experiment user prompts are generated from question details and system prompts from agent details. To ensure robust results, I experimented with prompt variations. I found that imprecise or overly broad prompts often led to the AI stating it couldn't predict the future or giving some vague responses. I designed prompts for both numerical answers and open-ended explanations (see Section SA-1.5)

4. **Knowledge Access:** It is possible to provide LLMs with a knowledge source. As another application, I implemented Retrieval-Augmented Generation (RAG) using API Assistants. This allowed models to access relevant economic data during queries. The process involved feeding PDF files from FOMC meetings and other sources to the API assistant, as illustrated in Figure A.7. OpenAI's embedding and indexing system then identified relevant information for context-aware responses from the AI agent. The results of this are provided in Section SA-2.1.

5. **Structured Responses:** There are three types of questions asked from the agents: 1- Point Estimation, 2- Density Forecast (Probability Distribution), 3- Open-ended Question. EDSL provides different question types to get the response of each type in a structured way. This approach ensured AI outputs aligned with traditional survey data formats, facilitating easier analysis and comparison with human surveys.

6. **Context Window:** The input and output sequences in my experiments were relatively short, so the context window was not a limiting factor. Current language models can handle very long contexts, often exceeding 128,000 tokens, which is much larger than needed here. In general, for experiments on AI agents, the context window is an important factor to consider.

[23] Temperature is the most important hyperparameter of each model and range from $\in [0, 2]$ for GPT and DeepSeek models (default 1.0) and $\in [0, 1]$ for Claude and Llama models (default 0.5).

SA-1.2 Step by Step Walkthrough

The full code[24] and dataset[25] are publicly available in an online repository. This section provides a step-by-step guide for implementing and running the inflation expectations experiment across multiple LLMs. The setup relies on the Expected Parrot framework and the EDSL package in Python to design reproducible AI-based survey experiments.

SA-1.2.1 Conceptual Framework using EDSL

The implementation relies on EDSL's[26] concept of *Questions* being answered by *Agents* in different *Scenarios (Information Treatments)* using large language *Models* to generate responses that are returned as formatted *Results*.

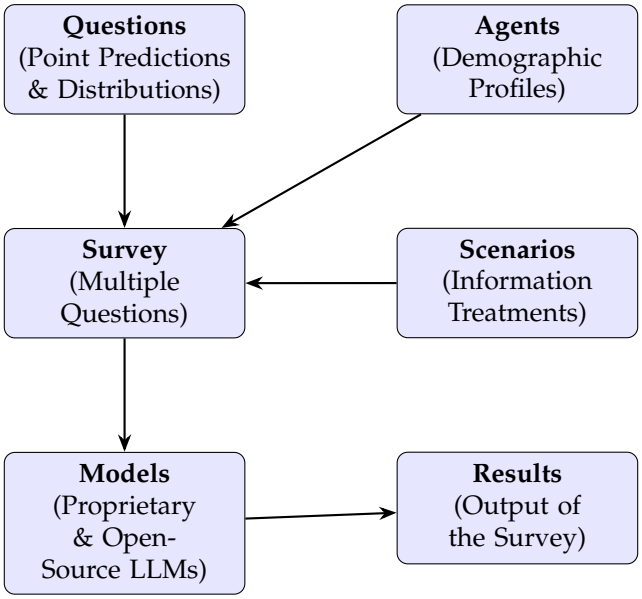

Figure A.1: EDSL Experimental Framework for an Inflation Expectation Survey

SA-1.2.2 API Key

EDSL provides a single endpoint for API access, eliminating the need to have API keys for all models. The safest approach is to create an `.env` file and store your key in it. Alternatively, we can import:

```
EXPECTED_PARROT_API_KEY = 'api-key'
```

[24] https://github.com/alizarif/JME_AI-and-Inflation
[25] Available on Harvard Dataverse: https://doi.org/doi:10.7910/DVN/UJUFIN
[26] Documentation: https://docs.expectedparrot.com/

SA-1.2.3 Data Preparation and Setup

First, I import the required modules from the EDSL library, which provides the building blocks for my survey experiment:

```
from edsl import QuestionFreeText, QuestionNumerical, QuestionDict
from edsl import QuestionList, Survey, Scenario, ScenarioList
from edsl import Agent, AgentList, Model, ModelList, FileStore
```

I access demographic data for the experiment from a CSV file stored on the Expected Parrot platform:

```
fs = FileStore('persona.csv')

if refresh := True:
    fs.push(
        description = 'Evidence on Inflation Expectations Formation Using Large Language Models: AI Personas',
        alias = 'inflation-expectation-survey-ai-personas',
        visibility = 'public' # it can be 'public' or 'unlisted'
    )
else:
    fs.patch('https://www.expectedparrot.com/content/alizarif/inflation-expectation-survey-ai-personas', value = fs)

```

```
fs = FileStore.pull('https://www.expectedparrot.com/content/alizarif/inflation-expectation-survey-ai-personas')
```

Then we can create Scenarios from our data and then inspect it. The experiment simulates the Survey of Consumer Expectations, utilizing microdata comprising 7,580 observations[27].

```
import pandas as pd

temp_file_path = fs.to_tempfile()
df = pd.read_csv(temp_file_path)

# Convert the DataFrame into a ScenarioList
demographics_scenarios = ScenarioList([
    Scenario(row.to_dict()) for _, row in df.iterrows()
demographics_scenarios[0] # inspect the first row
])
```

SA-1.2.4 Demographic Data Transformation

The demographic data includes attributes like age, gender, marital status, state, education, and income for different simulated respondents. I add text representations of dummy variables.

[27] This is the total number of unique participants after 2020 in the survey panel. It maintains the same demographic composition as the SCE, with the shares of each category consistent with the original survey

SA-1.2.5 Agent Creation with Demographic Traits

I convert the demographic scenarios into AI agents with specific persona traits, leveraging EDSL's agent capabilities to simulate AI respondents:

```
# Convert demographic scenarios to agents
agents = demographics_scenarios.to_agent_list()
########
# Define how the demographic traits should be presented to the LLMs
def add_traits_presentation(agentlist):
    new_agents = None

    for a in agentlist:
        a = Agent(
            traits = a.traits,
            traits_presentation_template = """
            You are a {{ age }} year old {{ gender_text }} who is {{ marital_text }}
            with an education level of {{ education }} degree and income category of {{ income }}
            who lives in the state of {{ state_name }}.
            """
        )

        if new_agents is None:
            new_agents = [a]
        else:
            new_agents.append(a)

    return AgentList(new_agents)
########
agents = add_traits_presentation(agents)
```

The experiment includes 10 groups: one control group and nine treatment groups (T_1 through T_9). These groups are balanced to maintain identical demographic composition across all experimental conditions, ensuring that demographic variables (age, gender, income, education, etc.) are evenly distributed. For implementation purposes, I randomly assigned each created agent to one of these balanced groups (labeled agents0 through agents9), maintaining demographic equivalence while randomizing individual assignment.

```
import random

# Create 10 random agent groups
num_groups = 10
total_agents = len(agents)
base_agents_per_group = total_agents // num_groups
remaining_agents = total_agents % num_groups

agent_groups = {}

indices = list(range(total_agents))
random.shuffle(indices)

extra_agent_groups = random.sample(range(num_groups), remaining_agents)

current_idx = 0
```

```python

for i in range(num_groups):
    group_size = base_agents_per_group + (1 if i in extra_agent_groups else 0)

    group_indices = indices[current_idx:current_idx + group_size]
    current_idx += group_size

    agent_groups[f"agents{i}"] = AgentList([agents[idx] for idx in group_indices])

print(f"Created {num_groups} agent groups with even distribution")
for group_name, group in agent_groups.items():
    print(f"{group_name}: {len(group)} agents")


# we can create variables in the global namespace for each agent group
for group_name, group in agent_groups.items():
    globals()[group_name] = group

agents0 # we check each group
agents0[0] # or we can check one particular agent
```

ScenarioList scenarios: 10; keys: ['Education', 'Marital', 'Income', 'userid', 'Age', 'Gender', 'STATE'];

	userid	Age	Gender	Marital	STATE	Education	Income
0	70111962	49	2	2	VA	College	Over 100k
1	70111963	69	2	1	NJ	College	Over 100k
2	70111970	60	1	2	FL	College	50k to 100k
3	70111982	78	2	1	NV	Some College	Under 50k
4	70111984	74	2	1	PA	Some College	50k to 100k
5	70111995	64	2	1	FL	High School	Under 50k
6	70112005	33	1	2	MI	Some College	50k to 100k
7	70112013	29	2	1	NJ	College	Over 100k
8	70112015	70	2	1	MD	College	Over 100k
9	70112019	46	2	1	TX	College	Over 100k

Figure A.2: An Example of an Agent Traits

SA-1.2.6 Scenario Component

I define 10 different treatment groups (one for control and nine for information treatments) as scenarios, utilizing EDSL's scenario functionality to be able to ask the questions with different economic contexts:

```python
scenarios0 = ScenarioList([
    Scenario({
        "treatment": "T_0",
        "description": "Control with no information",
        "info": ""
    })
])

scenarios1 = ScenarioList([
```

```
    Scenario({
        "treatment": "T_1",
        "description": "Placebo group",
        "info": "Population of the U.S. grew by 1% between 2022 and 2024."
    })
])

scenarios2 = ScenarioList([
    Scenario({
        "treatment": "T_2",
        "description": "Current rate, FFR",
        "info": "The interest rate set by the Federal Reserve, known as the Federal Funds
    Rate, is currently at 4.25%-4.5% range."
    })
])


##############################################
##############################################
##############################################
############Rest of the Scenarios############
##############################################
##############################################
##############################################

scenarios9 = ScenarioList([
    Scenario({
        "treatment": "T_9",
        "description": "Current fixed-rate 30-year mortgage",
        "info": "The current average rate for fixed-rate 30-year mortgage is 6.64% per year."
    })
])
```

The data for these treatments are as follows: Population Growth statistics (T_1) from the U.S. Census Bureau; Federal Funds Rate (T_2) from FRED; Current Inflation data (T_3-T_4) from BLS CPI reports; Long-Run Inflation projections (T_5-T_7) from the Federal Reserve's Summary of Economic Projections; and Mortgage Rate information (T_8-T_9) from FRED. These represent the most accurate and timely economic indicators available for the U.S. economy.

SA-1.2.7 Question Component

EDSL supports multiple question formats including checkboxes, free text, linear scales, multiple choice, numerical questions, etc. For this study, I utilized two specific EDSL question types: **QuestionDict** for density forecasts (or probability distributions) and **QuestionNumerical** for point estimates of expected inflation.

```
# Dictionary-based questions for short-run inflation probability distributions
q1_dict = QuestionDict(
    question_name = "Q1_S_Before_dict",
    question_text = """
    Please estimate the probability (as a percentage) for each of the following
    inflation/deflation scenarios over the next 12 months.
    Each probability must be between 0% and 100%.
```

```
    You may use up to 2 decimal points (e.g., 7.25%).
    The sum of all probabilities must equal exactly 100%.
    Return only a list of the numbers (i.e., 50 instead of '50%').

    Inflation of 12% or more: _____%
    Inflation between 8% and 12%: _____%
    Inflation between 4% and 8%: _____%
    Inflation between 2% and 4%: _____%
    Inflation between 0% and 2%: _____%
    Deflation between 0% and 2%: _____%
    Deflation between 2% and 4%: _____%
    Deflation between 4% and 8%: _____%
    Deflation between 8% and 12%: _____%
    Deflation of 12% or more: _____%
    """,
    answer_keys = ["inflation_forecast"],
    value_types = [list],
    value_descriptions = ["List of probability percentages for different inflation/deflation
    scenarios, must sum to 100"]
)

# Numerical questions for point predictions after treatment
q3 = QuestionNumerical(
    question_name = "Q1_S_After",
    question_text = """
    Consider the following current event: {{ info }}
    What do you expect the rate of inflation to be over the next 12 months? Please give your
    best guess.
    """,
    # min_value = 1
    # max_value = 100
)
```

The structure of these questions closely follows the Survey of Consumer Expectations conducted by the New York Federal Reserve, ensuring comparability with human survey responses. In EDSL, the user prompt for the each question above also includes an instruction for the model to provide a comment about its answer: *"After the answer, you can put a comment explaining why you chose that option on the next line."* which can be can be useful for understanding the context of a response or reasoning behind each answer.

SA-1.2.8 Survey Creation and Model Selection

Then we can compile all questions into a full survey and select the models we want to run the survey for.

```
survey = Survey([q1_list, q2_list, q3, q4])
survey = survey.set_full_memory_mode()
# to have the full memory of the survey path for each agent

# Selecting the model to run the survey
models = ModelList([
    Model("gpt-4o", service_name = "openai", temperature = 1),
    #Model("gpt-4o-mini", service_name = "openai", temperature = 1),
```

```
    #Model("gemini-2.0-pro-exp", service_name = "google", temperature = 1),
    #Model("gemini-2.0-flash-lite", service_name = "google", temperature = 1),
    #Model("mistral-large-2411", service_name = "mistral"),
    #Model("meta-llama/Llama-3.3-70B-Instruct", service_name = "deep_infra"),
    #Model("meta-llama/Meta-Llama-3-8B-Instruct", service_name = "deep_infra"),
    #Model("claude-3-7-sonnet-20250219", service_name = "anthropic"),
    #Model("claude-3-5-haiku-20241022", service_name = "anthropic"),
    #Model("deepseek-ai/DeepSeek-V3", service_name = "deep_infra", temperature = 1),
])
```

EDSL updates the list of the models frequency and we can use any model we want for each experiment[28].

SA-1.2.9 Running the Experiment

Finally, we can run the experiment. Each agent group will be placed in their corresponding survey with their treatment group. We can see the progress as shown in figure A.3 and .A.4.

```
# First I create lists of all scenarios (treatments) and agents
all_scenarios = [scenarios0, scenarios1, scenarios2, scenarios3, scenarios4,
                 scenarios5, scenarios6, scenarios7, scenarios8, scenarios9]
all_agents = [agents0, agents1, agents2, agents3, agents4,
              agents5, agents6, agents7, agents8, agents9]

# Loop through each scenario list and its corresponding agent list
for i, (scenario_list, agent_list) in enumerate(zip(all_scenarios, all_agents)):
    # Run the survey with matching scenario and agent lists
    results = survey.by(scenario_list).by(agent_list).by(models).run()

    # Convert to pandas and save to CSV
    results.to_pandas().to_csv(f"results{i}.csv")

    # Print progress
    print(f"Completed survey for scenario {i}")
```

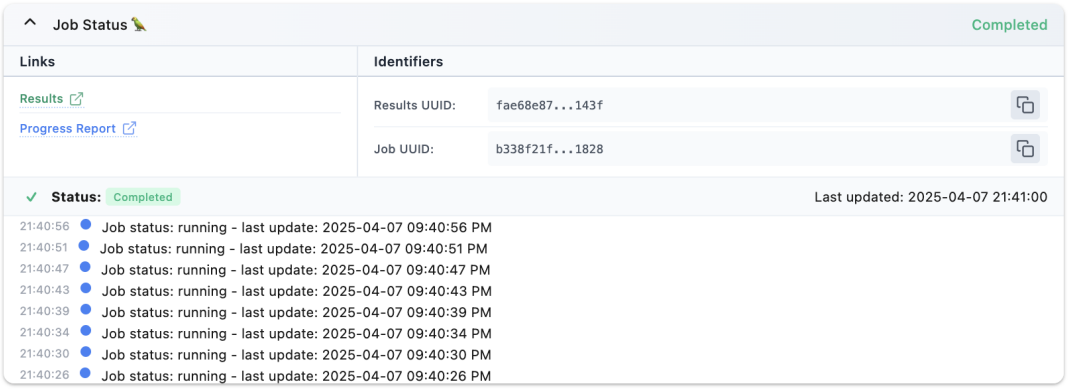

Figure A.3: Progress of the Survey

[28] https://www.expectedparrot.com/models

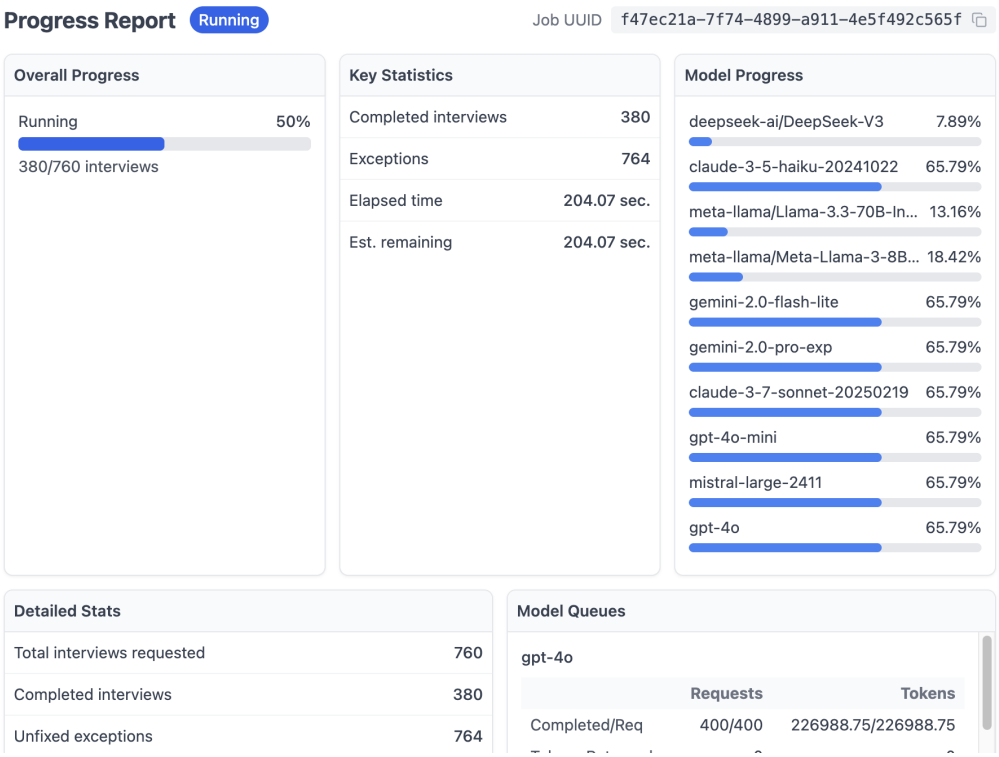

Figure A.4: Detailed Progress Report

EDSL has many other features, including a way to estimate the cost of the survey, which can be found in the package documentation.

SA-1.3 Pilots and Preliminary Tests

Before conducting the main experiments across multiple model architectures using EDSL as described in Section SA-1.2, I ran a series of pilot studies and preliminary tests using various GPT-4 variants.

The pilot experiments were conducted using Open AI's Assistants API, which allows for the creation of specialized AI agents with different capabilities. The implementation involved creating AI assistants with and without file search capabilities. This is just an example of how we can create an AI assistant powered by Open AI's model[29]:

```
# Example of creating an Assistant with Retrieval capabilities
assistant_with_retrieval = client.beta.assistants.create(
    name="Assistant M", # Powered by FOMC Minutes
    instructions="Answer questions about monetary policy and inflation expectations based on
    FOMC documents.",
    model="gpt-4-turbo",
    tools=[{"type": "retrieval"}],
    file_ids=[fomc_minutes_file_id]
)

# Example of creating an Assistant without Retrieval
assistant_no_retrieval = client.beta.assistants.create(
    name="Assistant N", # No Retrieval
    instructions="Answer questions about monetary policy and inflation expectations.",
    model="gpt-4-turbo",
    tools=[]
)
```

The complete implementation code is available in the GitHub repository[30].

[29] Another approach is to create the assistant directly using the platform: https://platform.openai.com/

[30] https://github.com/alizarif/JME_AI-and-Inflation

SA-1.4 Knowledge Cutoffs

SA-1.4.1 Main Experiment with Multiple Models

For the main experiment conducted in April 2025, I employed a diverse range of both proprietary and open-source LLMs to ensure comprehensive coverage across different model architectures, training methodologies, and parameter sizes. Table A.1 provides details on the models used in this experiment.

Model	Provider	Type	Knowledge Cutoff
GPT-4.1	OpenAI	Proprietary	June 2024
GPT-4o (3 Temperature Setting)	OpenAI	Proprietary	Apr 2024
GPT-4o-mini	OpenAI	Proprietary (Compact)	October 2023
Claude 3.7 Sonnet	Anthropic	Proprietary	Oct 2024
Claude 3.5 Haiku	Anthropic	Proprietary (Compact)	July 2024
Llama 3 (70B)	Meta	Open-source	Dec 2023
DeepSeek V3	DeepSeek	Open-source	July 2024

Table A.1: Models Used in the Main Experiment (April 2025)

This balanced approach allowed me to test whether observed inflation expectation behaviors were model-specific or represented more fundamental patterns in how LLMs process economic information. The experimental design specifically accounted for different context window capabilities, knowledge cutoff dates, and parameter scales across both proprietary and open-source models.

SA-1.4.2 Preliminary Experiments with GPT-4 Variants

Prior to the main experiment, I conducted several tests using various versions of GPT-4 models to develop and refine the methodology. These earlier experiments, summarized in Table A.2, served as pilot studies, and some robustness checks.

ID	Model	Experiment	Date	Knowledge Cutoff
1	GPT 4 Turbo	Knowledge Source	Apr 14, 2024	2023-01-12
2	GPT 4 Turbo	Pilot	Apr 15, 2024	2023-01-12
3	GPT 4 Turbo	Model Selection	Jul 15, 2024	2023-01-12
4	GPT 4o	Model Selection	May 15, 2024	2024-05-01
5	GPT 4o	Basic Persona	May 22, 2024	2024-05-01
6	GPT 4omini	Main Survey Experiments	Sep 18-22, 2024	2024-07-18
7	o1 preview	Reasoning	Sep 14, 2024	2023-01-10
8	GPT 4o	Prediction of Survey Results	Dec 22, 2024	2024-05-01

Table A.2: Preliminary Experiments with GPT-4 Variants (2024)

SA-1.5 Prompts

SA-1.5.1 For Main Experiment

Sample User Prompt

Please estimate the probability (as a percentage) for each of the following inflation/deflation scenarios over the next 12 months.

Each probability must be between 0% and 100%.
You may use up to 2 decimal points (e.g., 7.25%).
The sum of all probabilities must equal exactly 100%.
Return only a list of the numbers (i.e., 50 instead of '50%').

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

Please respond with a dictionary using the following keys: inflation_forecast.

Do not include "python" for create a code block. Just return the dictionary.
Here are descriptions of the values to provide:
- "inflation_forecast": "List of probability percentages for different inflation/deflation scenarios, must sum to 100"
The values should be formatted in the following types: "inflation_forecast": "list"
If you do not have a value for a given key, use "null".
After the answer, you can put a comment explaining your response on the next line.

System Prompt

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

SA-1.5.2 For pilot runs

I experimented with other prompts as well. These prompts were used for the pilot runs and only for OpenAI models. They included instructions for numerical percentage estimates and specific formats like single numbers or lists of percentages summing to 100%. By providing clear examples and using simple, neutral language, I ensured accurate instruction following. This approach helped with the collection of consistent and comparable data across different AI agents and treatments.

System Prompt

The GPT is designed to answer users' questions about the expectations for future inflation. It responds with numerical percentage estimates, reflecting its opinions and anticipations about inflation while acknowledging the inherent uncertainty in such forecasts. It presents answers in two formats: single percentage points for point prediction and a list of percentages for probability distribution questions. The GPT ensures responses are solely numerical and not any written statements with the alphabets and formatted accordingly: for point prediction, it uses [___ %,] and for probability distribution questions, it uses [___ %,___ %,___ %,___ %,___ %,___ %,___ %,___ %,___ %,___ %,].

RAG Prompt (Only for Retrieval Mode)

Use your general understanding of the document including the sentiments of the policy and all the information around it to answer. These are questions about inflation expectations and the perception of inflation, not inflation prediction. Do not answer nothing.

Main Prompt (For example, for T_3 with Current rate, FFR)

Initial: "Q2_I In this question, you will be asked about the probability (PERCENT CHANCE) of something happening. The percent chance must be a number between 0 and 100 and the sum of your answers must add up to 100. What do you think is the percent chance that, over the next 12 months... [RANGE OF EACH OPTION BELOW is 0-100 and each option can be 2 DECIMAL POINTS but the most important thing is that the total should be 100%] ... Give your answer as a list like this: [____ %, ____ %, ____ %, ____ %, ____ %, ____ %, ____ %, ____ %, ____ %, ____ %] "

Additional Context(Information Provision Step): "The interest rate set by the Federal Reserve, known as the Federal Funds Rate, is currently at 5.25%."

Follow-up: "What do you expect the rate of inflation to be over the next 12 months? Please give your best guess. Over the next 12 months, I expect the rate of inflation to be ____ %."

SA-1.5.3 Prompt of the LLMs as Predictor of Survey Results

You are tasked with prediction of inflation expectations based on survey data for different demographic groups.

First, you will be provided with survey data from the Survey of Consumer Expectations. The data in json format for each entry of the microdata is in your file searches.

Your task is to analyze this data and make predictions for the following demographic groups:

- **By Gender:**
  - Female
  - Male
- **By Age Group:**
  - Under 40
  - 40 to 60
  - Over 60
- **By Education Group:**
  - High School
  - Some College
  - College
- **By Income Group:**
  - Under 50k
  - 50k to 100k
  - Over 100k

For each demographic group, you need to predict:

1. Short-run inflation expectations (12 months from now)
2. Medium-run inflation expectations: (average of 12 months between 24 months from the survey date and 36 months from the survey date)
3. Long-run inflation expectations: (average of 12 months between 48 months from the survey date and 60 months from the survey date)

SA-1.6 Consistency in Responses

A common concern in experiments involving AI and language models is the consistency of their responses. In a preliminary test, I used two AI assistants: one without retrieval capabilities (N) and another with four years of FOMC minutes data (M2). Figure A.5 illustrates the distribution of responses across these two waves, showing a strong consistency between them.

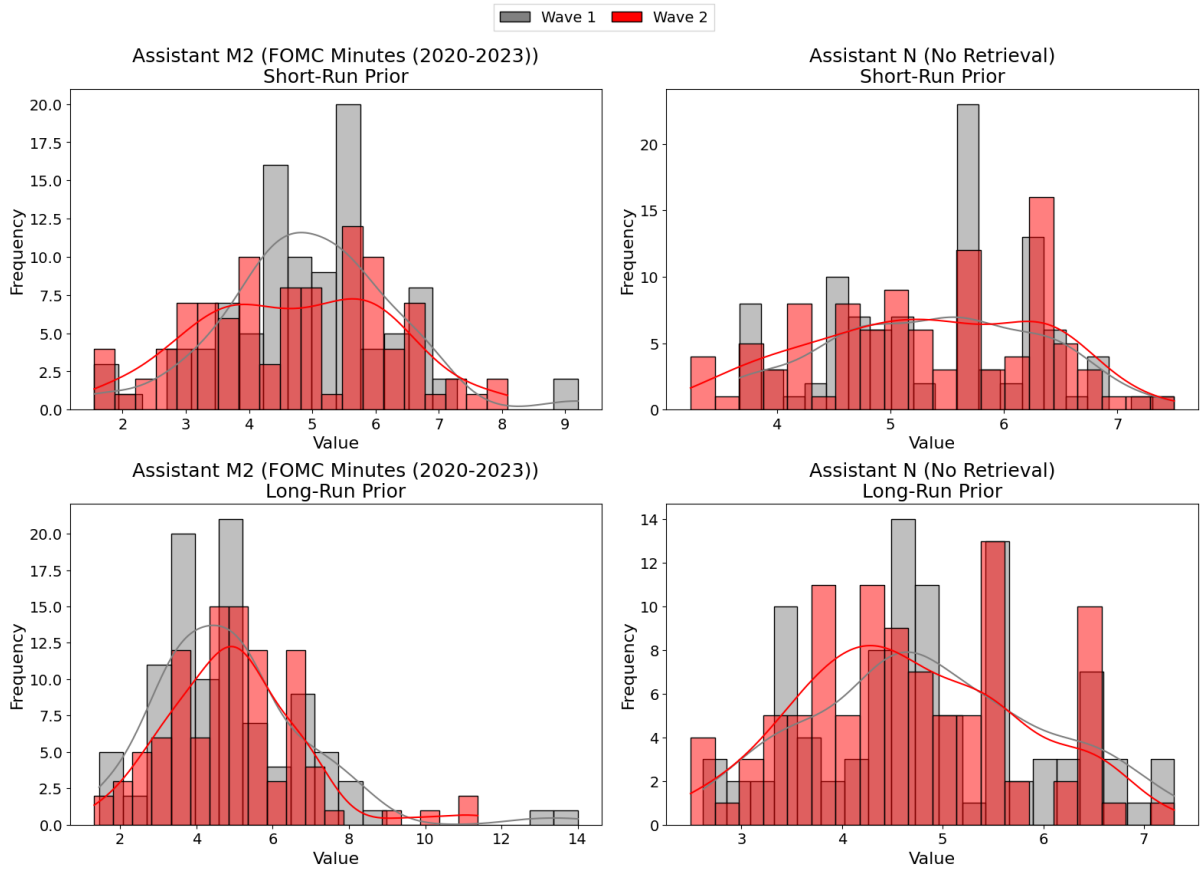

Figure A.5: Consistency of Responses in two Waves

Note: t-test shows that there is no statistically significant variation between the two runs.

The strong consistency observed between the two waves has significant implications for the reliability of LLMs in economic experiments. This consistency suggests that LLMs can produce stable and reproducible results across multiple runs, a crucial factor for reproducibility. Also, this consistency demonstrates that LLM responses are time-invariant, suggesting that their economic reasoning remains stable over short periods. However, as we observe in Section SA-1.7, while LLMs exhibit temporal consistency, they are model-variant, with different architectures or versions producing varying results. This reliability within models enhances the potential of LLMs as tools for economic research, allowing for more confident extrapolation of findings and enabling comparative studies across different economic scenarios.

SA-1.7 Model Selection

The selection of large language models is a critical factor that can significantly impact the results and conclusions drawn from experiments involving these models. This is particularly important in exploring the formation of inflation expectations using LLMs. The architectural differences between model families can lead to systematic variations in how economic information is processed and interpreted, making it essential to test across multiple models rather than relying on a single model family.

In another preliminary test (Figure A.6), I compared different models of OpenAI GPTs. These preliminary findings demonstrating model-specific variations led this study to use a comprehensive follow-up experiment in April 2025 that included multiple model architectures from different providers, including both proprietary models (GPT-4.1, GPT-4o, Claude 3.7 Sonnet, Claude 3.5 Haiku) and open-source alternatives (Llama 3-70B, DeepSeek-V3)

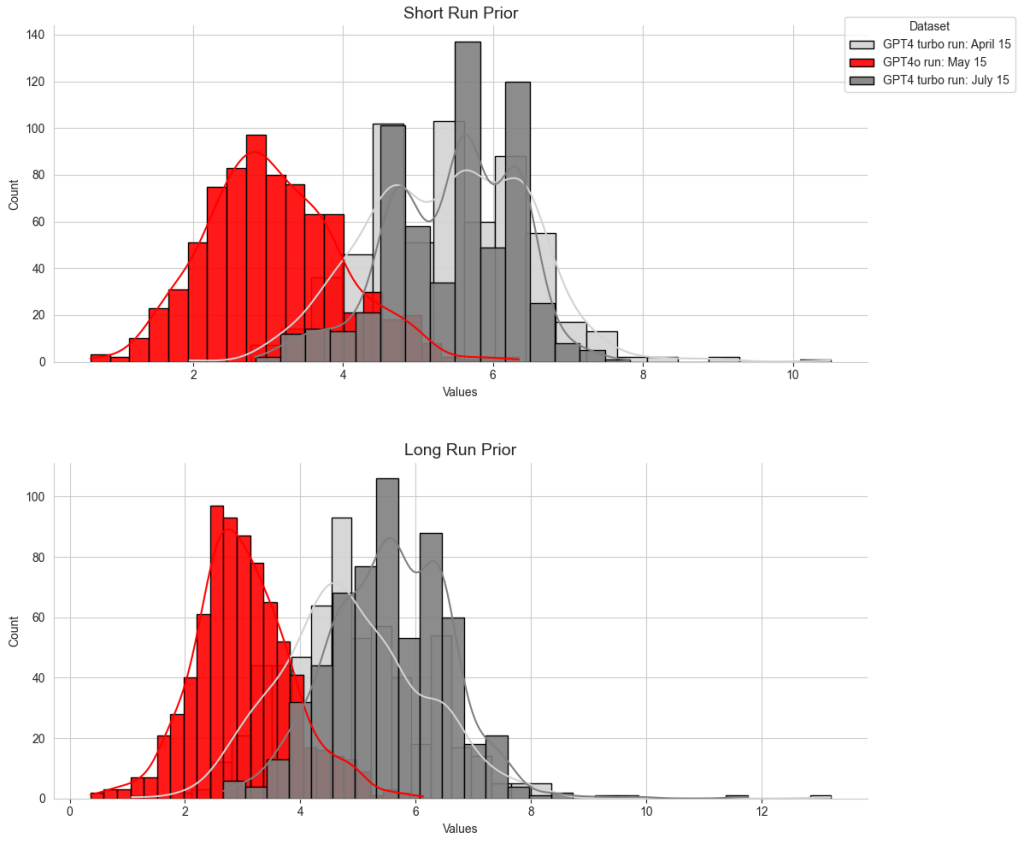

Figure A.6: Short-Run and Long-Run Inflation Expectations for Different Models

Note: The figure compares responses to short-run (SR) and long-run (LR) inflation expectation questions for different GPT-4 models. Results are presented for three runs: GPT-4 turbo on April 15, GPT-4o on May 15, and GPT-4 turbo on July 15. The GPT-4o model has lower inflation expectations. However, the overall shape of the distributions remains similar across models and runs, suggesting consistent response patterns despite updates and architectural variations.

SA-2 Additional Applications

SA-2.1 Knowledge Domain (RAG)

SA-2.1.1 Design

The survey on LLMs works similarly to an actual survey with humans, as shown in Figure A.7. For this experiment, I utilize Retrieval-Augmented Generation (RAG). RAG allows the model to access relevant external knowledge during runtime, enhancing its responses with new source of information. This dynamic knowledge integration enables the model to generate more accurate and contextually grounded answers. Unlike fine-tuning, which adjusts a model's internal parameters to improve performance on a specific task, RAG actively pulls in external data to better answer questions, providing a significant advantage in keeping the model's responses up-to-date and reliable. The RAG process can be shown as follows (Gao et al., 2023):

$$p(x|y) = \sum_{z \in \text{Retrieve}(y)} p(x|y,z) \cdot p(z|y) \tag{9}$$

where y is the input, x is the output, and z is the retrieved information from the external knowledge source. This expression shows how RAG combines the likelihood of generating x given the input y and the retrieved information z, weighted by the probability of retrieving z given y. As shown in Figure A.7, RAG operates with two prompts instead of one. The first prompt instructs the LLM on how to retrieve information, while the second prompt is the main question.

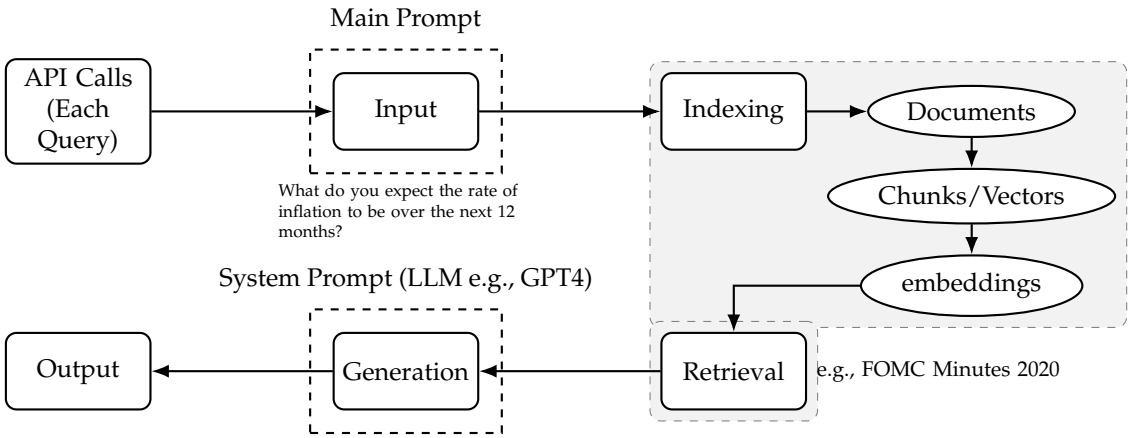

Figure A.7: An illustration of how RAG works

Note: This diagram shows a survey experiment for AI agents, which is very similar to the structure of a traditional survey. The system prompt serves a similar role to survey instructions, the main prompts are the questions, and the RAG prompt resembles the thought process that, while typically implicit in human surveys, is explicitly requested from the AI to utilize the provided context in generating responses.

As shown in Table A.3, I categorized the knowledge domains into six distinct types: (1) no retrieval (N), which does not provide any additional knowledge beyond the pre-trained GPT-4 data; (2) placebo (W), using general information from Wikipedia texts; (3) economic reports of the president (E), providing a broad view of the US economic status; (4) FOMC minutes from 2018 (M0) and (5) FOMC minutes from 2023 (M1), to compare the difference between old and recent monetary policy status; and (6) FOMC minutes from 2020 to 2023 (M2), offering a wider range of data on monetary policy. This structure allows me to analyze the impact of varying types of economic knowledge on agents' expectations.

Table A.3: Knowledge Domains in Survey Experiment

Code	Description
N	No Retrieval
W	Placebo (Wikipedia Texts)
E	Economic Reports of the President
M0	FOMC Minutes (only 2018)
M1	FOMC Minutes (only 2023)
M2	FOMC Minutes (2020-2023)

Note: This experiment evaluates the impact of different information retrieval contexts. It contrasts no retrieval with retrieval, non-economic with economic texts, and texts on monetary policy versus broader economic content. It also examines the influence of the historical context window.

SA-2.1.2 Results

This experiment involved 600 subjects across six types of assistants, where I ask three main questions to assess AI agents' inflation expectations without any information provision. Each assistant type accessed one of the previously described knowledge sources. The questions covered: (1) a point prediction of the perceived inflation rate for the past 12 months, (2) a density forecast (probability distribution) of expected inflation for the upcoming 12 months, and (3) a density forecast for the period between two and three years into the future.

Rather than focusing primarily on forecasting accuracy, these preliminary observations serve as validation checks for understanding how AI agents process economic information and form expectations. Table A.4 presents the summary statistics. For question 1, I report the point predictions. For questions 2 and 3, I used the midpoint formula to calculate the mean. While many studies account for both with and without deflation responses in the calculations, the amount that was assigned to deflation by AI agents in my survey is negligible.[31]

[31] In the Survey of Consumer Expectations, the median is usually reported instead of the mean due

Table A.4: Summary Statistics for (No Information Treatment)

Assistant (Obs.)	Past Inflation				1 Year Ahead				3 Years Ahead			
	min	max	mean	std	min	max	mean	std	min	max	mean	std
N (100)	2	3.5	2.43	0.26	3.24	7.5	5.31	1	2.49	7.3	4.68	1.08
W (100)	0	11.58	2.72	2.4	0.94	14	5.96	2.27	1.7	14	5.17	1.92
E (100)	0	7.1	3.01	1.29	-1	14	4.93	2.2	-2.25	11	4.84	1.98
M0 (100)	1.7	3.6	2.07	0.21	1.08	14	4.36	1.9	1.2	10.12	4.54	1.8
M1 (100)	2	8	3.85	1.1	1	14	4.67	2.37	1.09	14	4.55	2.23
M2 (100)	1.7	6.85	3.86	0.86	1.55	8	4.72	1.5	1.32	11.66	4.97	1.8
All (600)	0	11.58	2.99	1.40	-1	14	4.99	2	-2.25	14	4.79	1.84

Note: Assistant N demonstrates the most concentrated predictions with the smallest standard deviations, while Assistant W shows the highest variability. Mean predictions across all observations remain relatively consistent.

Figure A.8 shows the distribution of expectations reported by AI agents, compared to the actual observed US inflation of 3.48% at the time of the experiment. The gray bars show the distribution of responses from the Survey of Consumer Expectations with human subjects. The red bars reflect short-run expectations of AI agents, with a mean of 4.99%, and the gray bars show long-run expectations of AI agents, averaging 4.80%. This indicates that while the distributions are similar, the AI agents' expectations are more concentrated around the mean, suggesting less variance in their inflation expectations compared to human respondents.

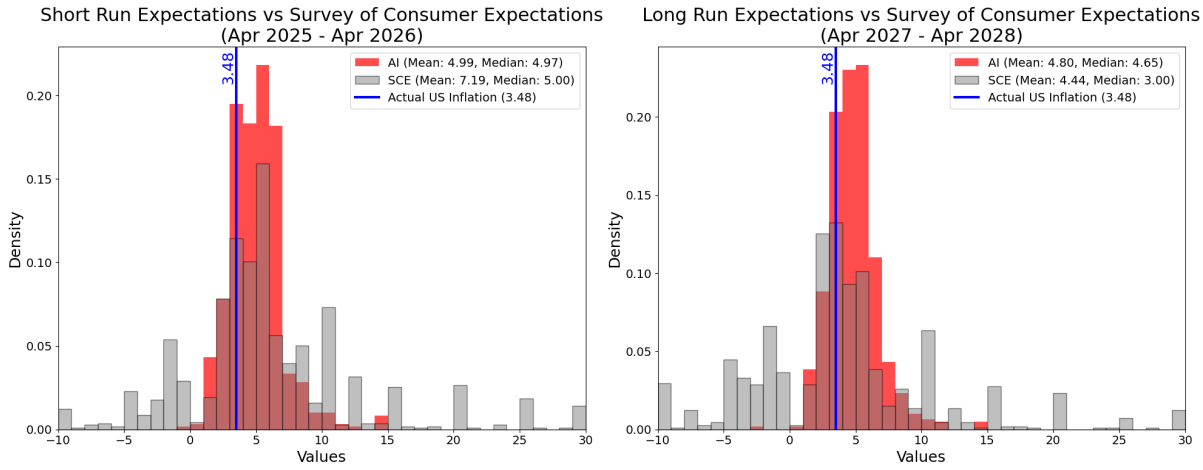

Figure A.8: Distribution of responses for human and AI respondents

Note: Data from SCE is trimmed from -10 to 30 as there are many outliers outside this range.

Figure A.9 shows the inflation expectations from different assistants, categorized into past inflation (black), short-term inflation (gray), and long-term inflation (red).

to the presence of outliers. In this experiment, the absence of significant outliers can be attributed to two factors: firstly, AI agents may follow the instructions more strictly, and secondly, their underlying pre-trained models and data may have a more robust knowledge base about the notion and possibility of deflation.

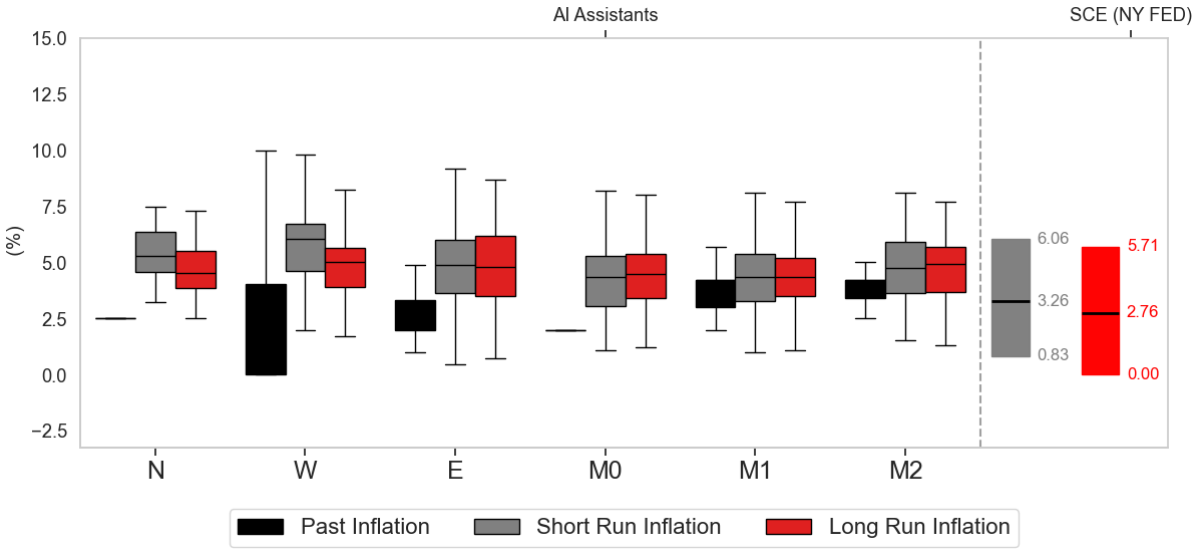

Figure A.9: Inflation Expectations in Different Horizon by Assistant

Note: The right side displays inflation expectations from the SCE, showing the 25th, median, and 75th percentiles for comparison. Assistant M2 shows greater comparability, potentially due to broader data access, while Assistant N, without additional retrieval capability, exhibits less variation.

The perceptions differ across assistants. For instance, Assistant M2, who had access to the last four years of minutes, showed more consistent predictions. Interestingly, Assistant N, which had no additional knowledge, also provided precise answers. This suggests that while additional information can enhance prediction accuracy and reliability, it may also introduce greater variability in responses, acting as a double-edged sword.

We can compare the inflation expectations from AI agents with other well-known economic surveys such as the Survey of Professional Forecasters (SPF). Table A.5 shows that professional forecasters (SPF) predict lower inflation rates (3.8% for 1-year and 2.5% for long-term) compared to both AI agents (around 5% for 1-year and 4.8% for long-term) and survey of consumer expectations (SCE).

Table A.5: Comparison with SCE and SPF

1 Year Ahead	**N**	**W**	**E**	**M0**	**M1**	**M2**	**All**	**SCE**	**SPF**
Mean	5.34	5.99	5.07	4.41	4.67	4.72	5.03	5.38	3.8
Median	5.3	6.06	5.02	4.43	4.38	4.76	5	3.26	3.8

2-3 Years Ahead	**N**	**W**	**E**	**M0**	**M1**	**M2**	**All**	**SCE**	**SPF***
Mean	4.84	5.2	4.92	4.57	4.58	4.99	4.85	2.80	2.5
Median	4.62	5.05	4.75	4.5	4.4	5	4.7	2.76	2.6

Note: Survey of Professional Forecasters (SPF) is from the Second Quarter 2024 report and Survey of Consumer Expectations (SCE) is from April 2024.

* 5 Year Forecast.

The observed differences in numerical estimates between AI agents and various human groups (experts and households) reflect the distinct nature of AI expectation formation. While AI agents demonstrate systematic upward bias similar to household surveys, their quantitative estimates often fall between expert forecasts and household expectations. This positioning suggests that AI agents combine aspects of both expert analysis and general public perception, likely due to their training on diverse datasets encompassing both professional economic analysis and broader public discourse. This hybrid characteristic of AI expectations becomes particularly relevant when considering the future role of AI in economic decision-making, as it implies that AI agents may process and respond to monetary policy communications in ways that don't perfectly align with any single existing group of economic agents.

SA-2.2 LLMs as Predictor of Survey Responses

Building on the previous findings, we can also explore whether LLMs can effectively predict inflation expectations across different demographic groups. This extension addresses a crucial challenge which is predicting heterogeneous beliefs with limited survey data. By leveraging LLMs' ability to process historical survey responses and demographic characteristics, we can develop a framework for generating out-of-sample predictions of inflation expectations. In a similar study, Brynjolfsson et al. (2025) show that LLMs can augment and accurately predict human survey responses across various domains using Panel Study on Income Dynamics (PSID).

The approach involves providing the survey microdata to the model on survey data up to time t. I then task the model with predicting inflation expectations for various demographic groups at $t+1$.[32] This methodology allows us to generate predictions for granular demographic subgroups even when survey data is sparse, while incorporating both historical patterns and recent economic developments. The model can provide predictions across multiple time horizons, from short-run to long-run expectations. This example demonstrates just one potential application of LLMs in survey research; the framework could be expanded to include additional demographic dimensions and more complex expectations.

Table A.6 presents the model's (GPT-4o) predictions for inflation expectations across different demographic groups and time horizons. The results reveal several notable patterns consistent with historical observations: females tend to have slightly higher inflation expectations, possibly due to perceived higher exposure to household cost fluctuations. Older age groups, particularly those over 60, expect higher inflation, potentially due to their fixed-income status and sensitivity to healthcare expenses. Education and income levels appear inversely correlated with inflation expectations, with

[32]The full prompt is provided in SA-1.5

Demographic Group	Short-run (12 months)	Medium-term (3 years)	Long-run (5 years)
Gender			
Female	4.5%	5.5%	5.0%
Male	4.0%	5.0%	4.5%
Age Group			
Under 40	5.0%	6.0%	5.5%
40 to 60	4.0%	5.0%	4.5%
Over 60	6.0%	7.0%	6.5%
Education Level			
High School	5.0%	6.0%	5.5%
Some College	4.5%	5.5%	5.0%
College	4.0%	5.0%	4.5%
Annual Income			
Under $50k	5.5%	6.5%	6.0%
$50k–$100k	4.5%	5.5%	5.0%
Over $100k	4.0%	5.0%	4.5%

Table A.6: Predicted Inflation Expectations by Demographic Group

Note: Values represent the mean expected inflation rate generated by GPT-4o for each demographic group across different time horizons. Experiment conducted in December 2024.

less educated and lower-income groups generally expecting higher inflation. This may reflect varying levels of access to economic information and the more immediate impact of price increases on daily budgets.

These results suggest that LLMs could serve as valuable tools for anticipating how different population segments might respond to changing economic conditions. Such predictions could help policymakers better target their communications and assess the effectiveness of monetary policy across diverse demographic groups. However, these capabilities should be viewed as complementary to, rather than replacements for, traditional survey methods.

SA-3 LLM Robustness Checks

SA-3.1 Impact of Temperature

Large language models rely on a set of hyperparameters that control their behavior during text generation. Among these, the temperature parameter plays a central role by modulating the degree of randomness in the model's output distribution. A lower temperature (e.g., 0.5) makes the model more deterministic and conservative, favoring high-probability tokens and generating more stable, repetitive responses. In contrast, a higher temperature (e.g., 1.5) injects greater randomness into the sampling process, allowing the model to explore more diverse outputs at the cost of coherence and numerical stability. Other hyperparameters—such as top-k, top-p sampling, or frequency penalties—also influence generation, but are either held constant or shown to have less pronounced effects in the inflation expectation setting analyzed here.

The results in Table A.7 demonstrate how sensitive GPT-4o is to this single hyperparameter. At moderate temperatures (0.5 and 1.0), the model yields stable and plausible expectations both before and after receiving information treatments. However, increasing the temperature to 1.5 destabilizes the model's behavior.

Table A.7: Summary Statistics for GPT-4o Models with Different Temperature Settings

Model	Short-term (1-year) Expectations			Long-term (3-year) Expectations			N
	Before	After	Change	Before	After	Change	
GPT-4o (t=0.5)	2.46	2.93	0.47	1.74	2.66	0.92	7,580
	(4.40)	(0.43)	(0.31)	(4.45)	(0.29)	(0.42)	
GPT-4o (t=1.0)	2.29	2.89	0.60	1.80	2.71	0.91	7,580
	(4.39)	(0.40)	(0.24)	(4.41)	(0.34)	(0.28)	
GPT-4o (t=1.5)	2.29	121.57	119.28	1.94	165.11	163.17	7,512
	(4.38)	(4819.63)	(4817.68)	(4.43)	(9196.67)	(9194.73)	

Notes: This table reports summary statistics for GPT-4o models with different temperature settings. Mean values are shown with standard deviations in parentheses. "Before" refers to expectations prior to receiving information treatments, "After" refers to expectations after treatments, and "Change" is the difference between After and Before values. N is the number of observations. The data shows that while GPT-4o models with temperatures 0.5 and 1.0 produce stable and reasonable expectations, the temperature 1.5 setting generates highly unstable outputs with extreme values and variance, suggesting numerical instability at higher randomness settings.

Figure A.10 visualizes how GPT-4o models with varying temperature (0.5, 1.0, and 1.5) allocate probability mass across inflation and deflation bins in response to a density forecast question. It highlights how increasing temperature affects the dispersion and skewness of inflation expectations. Table A.8 presents the estimated effects of different information treatments on inflation expectations for each temperature setting. Panel A reports the raw changes (post-treatment minus pre-treatment), while Panel B shows the average treatment effects relative to the control group (T0), offering insight

into the interaction between model randomness and information treatments.

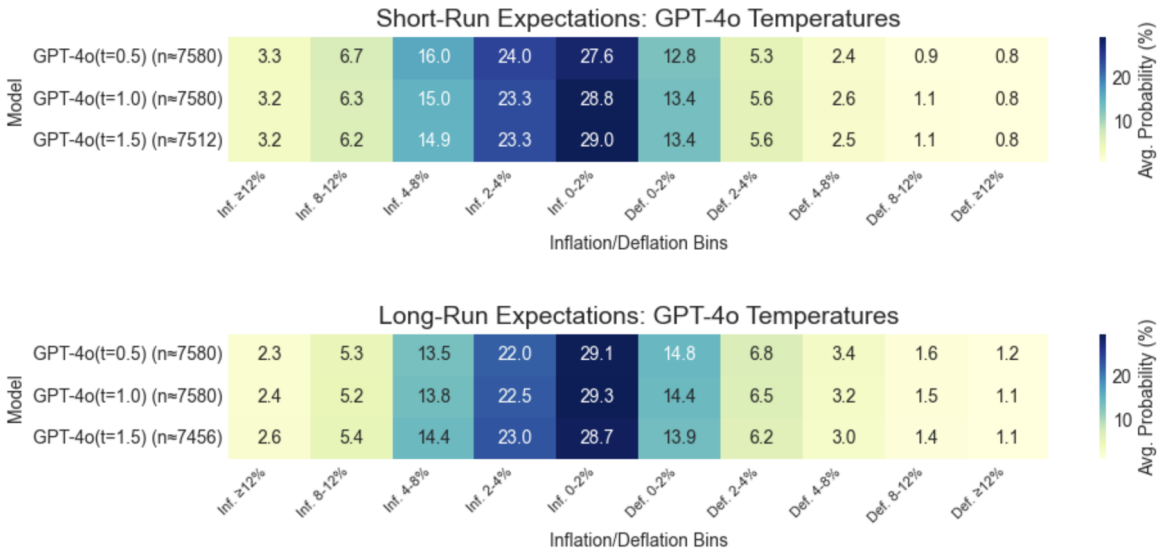

Figure A.10: Density Forecast

Table A.8: Effects of Information Treatments: GPT-4o with Different Temperature

	Short-term (1-year) Expectations			Long-term (3-year) Expectations		
Treatment	GPT-4o (t=0.5)	GPT-4o (t=1.0)	GPT-4o (t=1.5)	GPT-4o (t=0.5)	GPT-4o (t=1.0)	GPT-4o (t=1.5)
Panel A: Changes in Inflation Expectations (After − Before)						
T0 (Control)	0.12	0.23	238.31	0.90	0.85	10.04
T1 (Placebo)	0.28	0.48	7.13	0.84	0.86	7.41
T2 (Current FFR)	0.53	0.63	3.95	0.90	0.86	973.57
T3 (FFR + Next Year)	0.73	0.76	7.45	0.85	0.85	431.92
T4 (FFR + Full Projections)	0.78	0.79	255.52	0.93	0.86	11.70
T5 (Current + Past Inflation)	0.92	1.07	280.00	1.35	1.44	8.05
T6 (Current Inflation)	0.21	0.39	5.55	0.96	0.94	15.63
T7 (Inflation + Next Year)	0.33	0.49	365.18	0.94	0.88	151.05
T8 (Inflation Full Projections)	0.13	0.39	2.27	0.53	0.49	6.98
T9 (Mortgage Rate)	0.59	0.69	25.38	1.01	1.02	11.85
Panel B: Average Treatment Effects (relative to T0 control)						
T1 (Placebo)	0.16	0.26	-231.18	-0.06	0.02	-2.63
T2 (Current FFR)	0.41	0.40	-234.36	0.00	0.02	963.53
T3 (FFR + Next Year)	0.61	0.53	-230.86	-0.05	0.00	421.88
T4 (FFR + Full Projections)	0.66	0.57	17.21	0.03	0.02	1.66
T5 (Current + Past Inflation)	0.80	0.85	41.69	0.45	0.59	-1.99
T6 (Current Inflation)	0.09	0.16	-232.76	0.06	0.09	5.59
T7 (Inflation + Next Year)	0.21	0.27	126.87	0.04	0.03	141.01
T8 (Inflation Full Projections)	0.01	0.16	-236.04	-0.37	-0.36	-3.06
T9 (Mortgage Rate)	0.47	0.46	-212.93	0.11	0.17	1.81

Notes: This table reports the effects of different information treatments on inflation expectations using three variants of the GPT-4o model with different temperature settings (0.5, 1.0, and 1.5). Panel A shows the raw changes in inflation expectations (after treatment minus before treatment). Panel B shows the Average Treatment Effects (ATEs) relative to the control group (T0). The temperature parameter controls the randomness in the model's outputs, with higher values producing more varied responses.

Running the same experimental design with DeepSeek-V3, an open-source alter-

native, reveals comparable temperature sensitivity patterns to GPT-4o. As shown in Table A.9, DeepSeek-V3 exhibits similar stability at lower temperatures (0.5 and 1.0) with modest expectation changes, but displays markedly different behavior at temperature 1.5. While both models show increased variance and larger effects at higher temperatures, DeepSeek-V3 maintains greater numerical stability, avoiding the extreme values observed in GPT-4o. The treatment-specific effects in Table A.10 further demonstrate that temperature impacts remain substantial across different model architectures, though the magnitude and specific patterns vary between proprietary and open-source implementations.

Table A.9: Summary Statistics for DeepSeek-V3 Models with Different Temperature

Model	Short-term (1-year) Expectations			Long-term (3-year) Expectations			N
	Before	After	Change	Before	After	Change	
DeepSeek-V3 (t=0.5)	2.61	4.13	1.52	2.47	3.05	0.58	758
	(4.74)	(15.86)	(16.55)	(4.64)	(0.54)	(4.67)	
DeepSeek-V3 (t=1.0)	2.64	3.42	0.78	2.13	3.09	0.96	758
	(3.94)	(1.16)	(4.11)	(3.82)	(1.17)	(4.00)	
DeepSeek-V3 (t=1.5)	2.61	22.36	19.75	1.09	37.46	36.37	758
	(4.85)	(95.84)	(95.96)	(5.37)	(179.69)	(179.77)	

Notes: This table reports summary statistics for DeepSeek-V3 models with different temperature settings. Mean values are shown with standard deviations in parentheses.

Table A.10: Treatment Effects: DeepSeek-V3 with Different Temperature

	Short-term (1-year) Expectations			Long-term (3-year) Expectations		
Treatment	DS-V3 (t=0.5)	DS-V3 (t=1.0)	DS-V3 (t=1.5)	DS-V3 (t=0.5)	DS-V3 (t=1.0)	DS-V3 (t=1.5)
Panel A: Changes in Inflation Expectations (After − Before)						
T0 (Control)	1.08	0.99	11.02	0.64	0.83	44.78
T1 (Placebo)	0.49	0.54	20.68	0.46	0.99	19.35
T2 (Current FFR)	0.87	0.90	19.19	0.51	1.11	35.79
T3 (FFR + Next Year)	0.76	0.83	10.16	0.78	1.05	58.42
T4 (FFR + Full Projections)	0.81	0.94	22.12	0.43	1.25	113.15
T5 (Current + Past Inflation)	7.22	1.80	8.52	1.59	1.98	23.05
T6 (Current Inflation)	0.57	0.22	27.02	0.32	0.78	20.72
T7 (Inflation + Next Year)	1.63	0.15	21.67	0.11	0.48	8.60
T8 (Inflation Full Projections)	0.20	0.17	24.16	-0.05	-0.05	16.55
T9 (Mortgage Rate)	1.35	1.20	32.79	1.09	1.15	14.38
Panel B: Average Treatment Effects (relative to T0 control)						
T1 (Placebo)	-0.59	-0.45	9.66	-0.18	0.16	-25.43
T2 (Current FFR)	-0.21	-0.09	8.17	-0.13	0.28	-8.99
T3 (FFR + Next Year)	-0.32	-0.16	-0.86	0.14	0.22	13.64
T4 (FFR + Full Projections)	-0.27	-0.05	11.10	-0.21	0.42	68.37
T5 (Current + Past Inflation)	6.14	0.81	-2.50	0.95	1.15	-21.73
T6 (Current Inflation)	-0.51	-0.77	16.00	-0.32	-0.05	-24.06
T7 (Inflation + Next Year)	0.55	-0.84	10.65	-0.53	-0.35	-36.18
T8 (Inflation Full Projections)	-0.88	-0.82	13.14	-0.69	-0.88	-28.23
T9 (Mortgage Rate)	0.27	0.21	21.77	0.45	0.32	-30.40

Notes: This table reports the effects of different information treatments on inflation expectations using the DeepSeek-V3 model with different temperature settings (0.5, 1.0, and 1.5). Panel A shows the raw changes in inflation expectations (after treatment minus before treatment). Panel B shows the Average Treatment Effects (ATEs) relative to the control group (T0).

SA-3.2 Chain-of-Thought Quality Analysis

To provide quantitative measures of reasoning quality, I implement two metrics that evaluate LLMs' chain-of-thought processes in inflation expectation formation.

Self-Consistency for Numerical Responses: This measures how reliably each model generates similar responses when identical personas face the same economic scenario. For each treatment group k, I calculate:

$$\text{Self-Consistency}_k = \frac{1}{1 + CV_k} = \frac{1}{1 + \frac{\sigma_k}{\mu_k}} \tag{10}$$

where $_k$ is the Coefficient of Variation, σ_k and μ_k are the standard deviation and mean of responses within treatment k. Higher scores indicate greater consistency, with 1 means perfect agreement.

Token-Level Depth for Reasoning Comments: This quantifies the complexity of models' explanatory comments after providing numerical responses. For each explanation, I calculate:

$$\text{Token Count} = \text{number of words in explanation} \tag{11}$$

$$\text{Unique Word Ratio} = \frac{\text{unique words}}{\text{total words}} \tag{12}$$

Table A.11 presents the results across all models. Proprietary models show higher self-consistency than open-source models, except for the extremely poor performance of GPT-4o at temperature 1.5 (0.095), highlighting the importance of parameter control.

Table A.11: Chain-of-Thought Quality Analysis Across LLMs

	Self-Consistency Scores		**Token-Level Depth**	
Model	**Numerical Responses**	**Probability Distributions**	**Avg Tokens**	**Uniqueness Ratio**
Panel A: Proprietary Models				
GPT-4.1	0.914	0.847	66.6	0.794
GPT-4o(t=0.5)	0.920	0.747	43.0	0.820
GPT-4o(t=1.0)	0.908	0.720	45.6	0.835
GPT-4o(t=1.5)	0.095	0.716	51.3	0.929
GPT-4o-mini	0.894	0.818	49.9	0.826
Sonnet-3.7	0.905	0.830	57.3	0.817
Haiku-3.5	0.959	0.791	64.0	0.797
Panel B: Open Source Models				
Llama3-70B	0.935	0.820	76.0	0.687
DeepSeek-V3	0.762	0.764	61.9	0.766

Notes: Numerical responses refer to point forecasts after treatment; probability distributions refer to pre-treatment expectation distributions.

These results show that model architecture significantly affects reasoning quality in economic contexts.

SA-3.3 Hallucination Detection and Out-of-Sample Verification

To address concerns about response reliability and experimental validity, I systematically detect hallucinations and verify the out-of-sample nature of the experimental design. This ensures findings represent genuine expectation formation rather than memorized patterns. I define hallucinations as responses that have extreme cases: inflation expectations exceeding 50% or below -20%, negative probabilities in distribution forecasts, non-numeric responses when numeric values are expected.

Table A.12 presents hallucination rates across all tested models, revealing heterogeneity in response reliability, with rates ranging from 0.00% for most stable models to 5.50% for GPT-4o at temperature 1.5.

Table A.12: Hallucination Rates and Out-of-Sample Verification

	Hallucination Components				Overall	Knowledge
Model	**Extreme Inflation**	**Non-Numeric**	**Invalid Dist.**	**Negative Prob.**	**Rate**	**Cutoff**
Panel A: Proprietary Models						
GPT-4.1	0	16	0	0	0.21%	June 2024
GPT-4o(t=0.5)	0	0	0	0	0.00%	Apr 2024
GPT-4o(t=1.0)	0	0	0	0	0.00%	Apr 2024
GPT-4o(t=1.5)	281	135	0	1	5.50%	Apr 2024
GPT-4o-mini	0	0	0	0	0.00%	Oct 2023
Sonnet-3.7	0	5	0	0	0.07%	Oct 2024
Haiku-3.5	0	0	0	0	0.00%	July 2024
Panel B: Open Source Models						
Llama3-70B	0	0	0	0	0.00%	Dec 2023
DeepSeek-V3	5	62	2	0	0.91%	July 2024

Notes: Hallucination rates calculated as percentage of total responses (N=7,580 per model). Extreme inflation includes predictions >50% or <-20%. All experimental treatments use March 2025 data, post-dating model training cutoffs by 4-16 months, confirming out-of-sample design validity.

Regarding look-ahead bias concerns, it is important to note that this study focuses on expectation formation rather than forecasting accuracy. The experimental design examines how LLMs process and integrate new economic information to form beliefs about future inflation, not their ability to predict actual future outcomes. In this context, the temporal relationship between training data and experimental treatments serves to ensure that models cannot rely on memorized patterns when forming expectations. All experimental treatments use March 2025 data sources, which post-date every model's training cutoff by 4-16 months, confirming genuine out-of-sample expectation formation rather than data retrieval.

SA-3.4 Framing of the Survey Questions

A key methodological concern involves potential prompt artifacts in the observed convergence of forecast variance after information treatments. The substantial reduction in standard deviations from distribution forecasts (pre-treatment) to point forecasts (post-treatment) could reflect question format changes rather than genuine information processing effects. To address this concern, I conducted a robustness check that reverses the original experimental design: participants now provide point estimates before treatment and probability distributions after treatment.

Table A.13: Treatment Effects: GPT-4o (Reversed Question Framing)

	Short-Term (1 year ahead)				Long-Term (3 year ahead)			
Treatment	Before	After	After−Before	Δ vs T_0	Before	After	After−Before	Δ vs T_0
T_0	3.49 (0.34)	2.23 (3.73)	−1.26 (4.07)	–	2.85 (0.29)	1.96 (3.53)	−0.89 (4.04)	–
T_1	3.44 (0.28)	1.87 (3.82)	−1.57 (4.10)	−0.31	2.80 (0.25)	1.57 (3.65)	−1.23 (4.07)	−0.34
T_2	3.40 (0.22)	2.34 (3.68)	−1.06 (4.06)	0.20	2.81 (0.27)	1.90 (3.49)	−0.91 (3.98)	−0.02
T_3	3.39 (0.29)	2.02 (3.50)	−1.37 (3.84)	−0.11	2.76 (0.26)	1.75 (3.22)	−1.01 (3.56)	−0.12
T_4	3.43 (0.27)	2.09 (3.45)	−1.34 (3.72)	−0.08	2.81 (0.28)	1.84 (3.07)	−0.97 (3.35)	−0.08
T_5	3.37 (0.20)	2.92 (4.58)	−0.45 (4.35)	0.81	2.81 (0.23)	2.36 (4.15)	−0.45 (4.38)	0.44
T_6	3.42 (0.22)	1.39 (3.98)	−2.03 (4.20)	−0.77	2.78 (0.25)	1.22 (3.80)	−1.56 (4.06)	−0.67
T_7	3.44 (0.22)	1.46 (3.20)	−1.98 (3.43)	−0.72	2.76 (0.25)	1.35 (2.99)	−1.41 (3.24)	−0.52
T_8	3.45 (0.25)	1.48 (2.69)	−1.97 (2.94)	−0.71	2.79 (0.23)	1.52 (2.55)	−1.27 (2.78)	−0.38
T_9	3.44 (0.28)	1.82 (3.90)	−1.62 (4.18)	−0.36	2.80 (0.25)	1.58 (3.70)	−1.22 (4.03)	−0.33
Total	**3.43 (0.26)**	**1.96 (3.71)**	−1.47 **(3.97)**	**–**	**2.80 (0.26)**	**1.71 (3.46)**	−1.09 **(3.84)**	**–**

Notes: Standard deviations in parentheses. Before = point estimates (pre-treatment), After = probability distribution midpoints (post-treatment). T_0 is the control group. Δ vs T_0 represents treatment effects relative to control.

The reversed framing experiment reveals that question format significantly influences variance patterns, but not in the direction that would invalidate the main findings. Contrary to the original design where variance decreased after treatment, the reversed framing shows variance *increasing* from point estimates to probability distributions. This confirms that format changes drive variance shifts.

More importantly, the systematic downward revision of expectations persists across both experimental designs. Nearly all treatments show negative "AfterBefore" values, indicating that LLMs consistently revise their inflation expectations downward when transitioning from point estimates to probability distributions. This suggests that format changes trigger recalibration effects where LLMs become more conservative when expressing uncertainty through distributions.

The treatment effects (Δ vs T_0) remain substantively meaningful even after controlling for format effects. Treatments involving explicit inflation information (T_6, T_7, T_8) continue to show the largest effects relative to control, suggesting that information content matters beyond formatting artifacts. These findings demonstrate that while question framing introduces systematic effects, the core findings about information treatment efficacy remain robust.

SA-4 Additional Results of Persona Prompting

SA-4.1 Partisan Expectations

In examining the effects of persona attributes on inflation expectations, I conducted a pilot analysis across different agent configurations: a baseline with no persona specification, geographic variations (Texas and California residents), and political identifications (Republican and Democrat). This analysis focused on testing the hypothesis that LLMs demonstrate systematic variation in beliefs across different demographic and political characteristics.

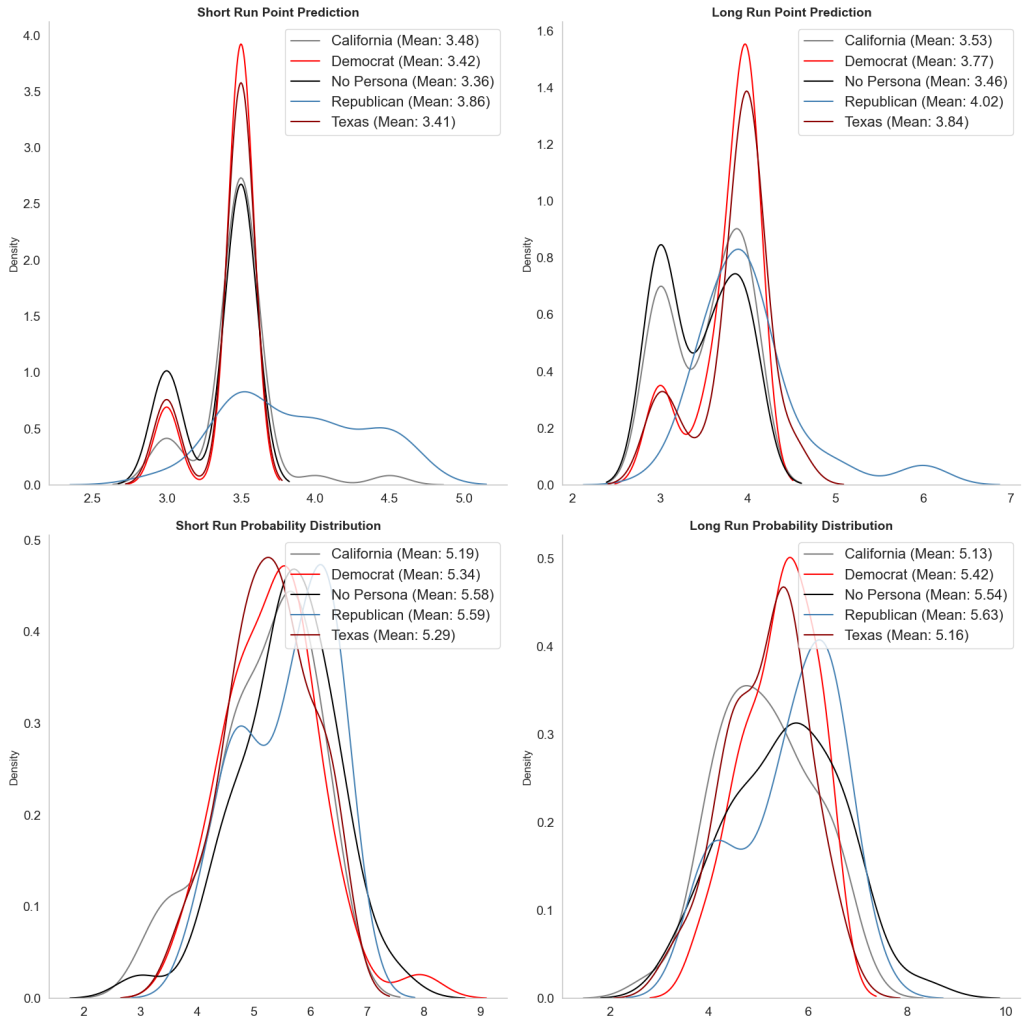

Figure A.11: Inflation expectations by persona attributes

Note: This figure shows the inflation expectation outputs of AI agents with different political persona attributes, including no persona, Republican, Democrat, living in Texas, and living in California. The top row shows the density plots for the short-run (left) and long-run (right) point prediction questions. The bottom row shows the distribution of probability distribution questions for the short-run (left) and long-run (right). The Republican agent consistently expresses higher inflation expectations in the point prediction questions compared to the other agents. For the probability distribution questions, the agents tend to assign probabilities more uniformly across the inflation bins, with less variation by political persona. The total number of observations is 200 (40 for each category).

To further validate these results, I do a comparison with data from the Survey of Consumer Expectations. The data in Table A.14 shows a big difference in inflation expectations between California (a proxy for being Democrat and residing in a "blue" state) and Texas (a proxy for being Republican and residing in a "red" state), with Texas consistently showing higher expectations for both short-term (1 year ahead) and long-term (3 years ahead) horizons. This aligns with my results, where the Texas/Republican persona generally exhibited higher inflation expectations compared to the California/Democrat persona.

Table A.14: Statistics Summary of Subjects in SCE (Point Prediction)

State	Metric	Mean	Std	Median
CA	1 Year Ahead	5.39	13.25	4.50
CA	3 Year Ahead	2.12	13.37	3.00
TX	1 Year Ahead	9.54	15.42	5.50
TX	3 Year Ahead	6.06	16.77	4.00

Note: This table presents the average of the aggregated data for Texas and California from January 2023 to September 2023 of the Survey of Consumer Expectations.

These comparisons between AI experiment results and the Survey of Consumer Expectations data provide compelling evidence that LLM agents with carefully crafted personas can generate inflation expectations that reflect real-world patterns. While there are certainly differences and limitations, the overall trends and relative positions of different groups (e.g., Republican vs. Democrat, Texas vs. California) are largely consistent between the simulated data and actual survey responses. This opens up new possibilities for using LLM-based experiments to explore economic expectations and decision-making processes.

SA-4.2 Persona vs. No Persona

Table A.15 compares inflation expectations between AI agents with personas and those without, based on a simulation of the SCE panel. The results are from a pilot study conducted in September 2024.

Table A.15: Summary Statistics of Inflation Expectations

Stat	Dataset	P_Inflation	Short-run Prior	Long-run Prior	Short-run Post.	Long-run Post.
Mean	With Persona	5.047	4.591	5.297	4.091	3.431
	No Persona	3.901	3.691	4.757	3.368	2.945
SD	With Persona	0.748	2.244	2.090	0.783	0.838
	No Persona	0.787	1.781	1.565	0.691	0.631
Min	With Persona	3.000	-1.300	-0.400	2.000	2.000
	No Persona	2.000	-0.100	-0.220	2.000	2.000
Median	With Persona	5.000	4.400	5.000	4.000	3.000
	No Persona	3.500	2.727	4.300	3.500	3.000
Max	With Persona	8.000	10.560	10.640	7.000	6.000
	No Persona	8.500	10.840	10.780	6.000	5.000
N	With Persona	6,528	6,528	6,528	6,528	6,528
	No Persona	6,402	6,528	6,528	6,402	6,402

Table A.16 shows the demographic distribution of the personas used in the study, based on microdata from the Survey of Consumer Expectations. These results are from the pilot study, which includes fewer observations (6,528) compared to the main experiment (7,580), as participants were added to the panel incrementally each month. However, the demographic composition remains identical across both samples.

Table A.16: Demographic Distribution of Persona Used from the SCE

Variable	Category	Distribution
Gender	Male	3,297 (50.5%)
	Female	3,233 (49.5%)
Marital Status	Married	4,117 (63.0%)
	Not Married	2,413 (37.0%)
Age	Under 40	2,071 (31.7%)
	40 to 60	2,602 (39.8%)
	Over 60	1,857 (28.5%)
Education	High School	786 (12.0%)
	Some College	2,068 (31.7%)
	College	3,676 (56.3%)
Income	Under $50k	2,026 (31.0%)
	$50k to $100k	2,236 (34.2%)
	Over $100k	2,268 (34.8%)
Total	(Unique Participants from 01/2020 to 09/2023)	6,528

Figures A.12 and A.13 illustrate the distribution of posterior inflation expectations for both the short-run (1 year ahead) and long-run (3 years ahead) horizons, com-

paring scenarios with and without persona attributes.[33] Each figure shows the density of posterior expectations across different treatment groups. In the short run, we observe more dispersed distributions—particularly in the persona-based condition—suggesting that different information treatments lead to more varied expectation updates. In contrast, long-run expectations exhibit greater convergence, especially in the non-persona case, indicating that AI agents tend to align their long-term inflation forecasts more closely regardless of the information received. Notably, the "Current Inflation + Longer Run" treatment (T5) consistently yields a more concentrated distribution across both time horizons and model types, suggesting that this form of forward guidance has a strong anchoring effect on AI agents' expectations.

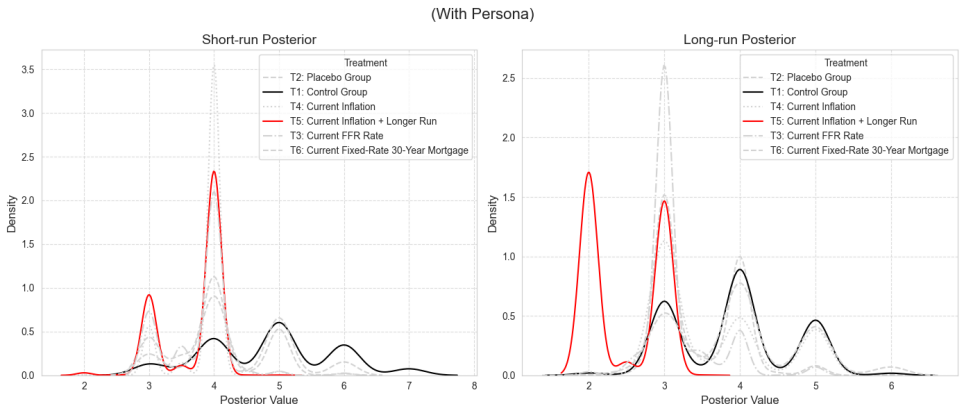

Figure A.12: Distribution of Posterior for Short-run(1 Year Ahead)

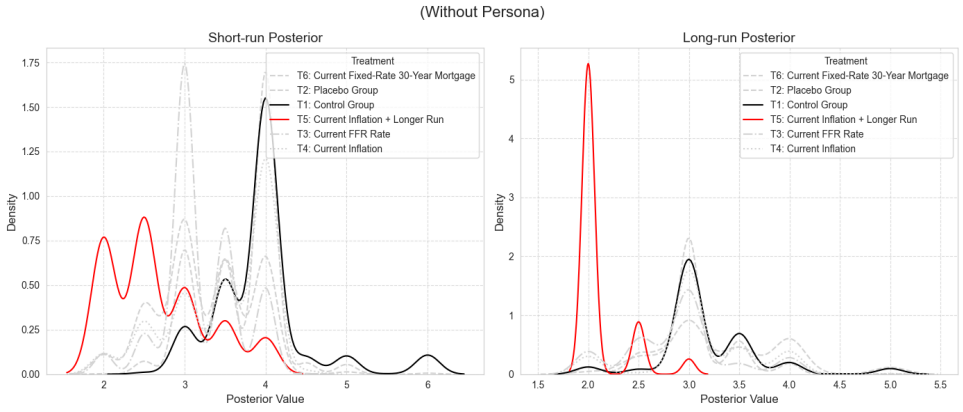

Figure A.13: Distribution of Posterior for Long-run(3 Year Ahead)

[33] In this pilot experiment, I include only 5 treatments; the main experiment incorporates 9 treatments.

SA-5 Reasoning Models

OpenAI's recent release of o models and the subsequent launch of GPT-5 in August 2025 represent a substantial development in Generative AI's ability to engage in complex reasoning processes. The o series models, as explained in OpenAI's release notes, are designed to spend more time "thinking" before they respond, achieved through a specific version of the chain-of-thought (CoT) prompting pattern. This method, introduced by Wei et al. (2023), has been refined through a large-scale reinforcement learning algorithm that teaches the model to "think" productively using its chain of thought. Unlike previous versions, these models spend more time processing problems before responding, much like a person would. Through training, they learn to refine their thinking process, try different strategies, and recognize their mistakes. Building on these advances, GPT-5 introduces a hybrid reasoning architecture that unifies the o-series reasoning capabilities with traditional GPT-style fast responses.

Table A.17: Effects of Information Treatments: o3-mini

Treatment	Short-term (1-year) Expectations	Long-term (3-year) Expectations
Panel A: Changes in Inflation Expectations (After − Before)		
T0 (Control)	0.38	0.53
T1 (Placebo)	-0.02	0.41
T2 (Current FFR)	0.34	0.36
T3 (FFR + Next Year)	0.19	0.24
T4 (FFR + Full Projections)	-0.05	0.32
T5 (Current + Past Inflation)	0.99	0.84
T6 (Current Inflation)	0.11	0.41
T7 (Inflation + Next Year)	-0.03	0.35
T8 (Inflation Full Projections)	-0.12	0.21
T9 (Mortgage Rate)	0.53	0.39
Panel B: Average Treatment Effects (relative to T0 control)		
T1 (Placebo)	-0.40	-0.12
T2 (Current FFR)	-0.04	-0.17
T3 (FFR + Next Year)	-0.19	-0.29
T4 (FFR + Full Projections)	-0.43	-0.21
T5 (Current + Past Inflation)	0.61	0.31
T6 (Current Inflation)	-0.27	-0.12
T7 (Inflation + Next Year)	-0.41	-0.18
T8 (Inflation Full Projections)	-0.50	-0.32
T9 (Mortgage Rate)	0.15	-0.14
Panel C: Summary Statistics		
Before Treatment (Total)	2.85	2.18
After Treatment (Total)	3.08	2.59
Overall Change	0.23	0.41

*Notes:*This table reports the effects of different information treatments on inflation expectations using the o3-mini model. Panel A shows the raw changes in inflation expectations (after treatment minus before treatment). Panel B shows the Average Treatment Effects (ATEs) relative to the control group (T0). Panel C shows overall summary statistics. The o3-mini model shows moderate and stable responses across treatments, with the largest positive effect observed for T5 (Current + Past Inflation) treatment. Sample size: n=758 (except T8 and T9: n=75). Standard deviations show substantial reduction after treatment (from 4.8 to 0.6-0.8), indicating convergence in responses.

To examine how these reasoning models respond to information treatments, I conducted experiments using o3-mini, on of the most recent iteration of OpenAI's reason-

ing model series. Table A.17 presents the results of these experiments across different information treatments. The model exhibits moderate sensitivity to information treatments, with Average Treatment Effects (ATEs) ranging from -0.50 to 0.61 for short-term expectations and -0.32 to 0.31 for long-term expectations.

These new models are particularly relevant to my study of inflation expectations. While previous experiments provided valuable insights into how AI models form and update inflation expectations, the o1 model (as an example of reasoning models) offers the potential to go deeper into the reasoning processes behind these expectations and address some criticisms regarding LLMs as "black boxes." However, OpenAI does not provide access to the reasoning tokens through their API, making it impossible to directly observe the model's chain-of-thought process and because of that I conducted manual experiments with o1. Through these manual experiments, I was able to capture how the model systematically works through economic information, revealing a chain of thought comprising the stages shown in Table A.18. Due to limited access to this model at the time of the pilots, I ran the experiment with five AI agents, one for each treatment group, in addition to the three pre-treatment questions, manually documenting their reasoning when explicitly prompted to explain their thinking process.

Table A.18: Different Stages Mentioned by AI Agents

Category	**Number of Topics**
Inflation Analysis	13
Probability Assessment	10
Economic Analysis	10
Data Interpretation	20
Guideline Compliance	10

The o1 model's responses exhibited a higher degree of nuance and consideration of multiple economic factors compared to previous models. For instance, when asked about short-term inflation expectations after receiving the current Federal Funds Rate information (T3), the model provided a detailed explanation with several stages of thinking:

- **Formulating Expectations:** I'm thinking about the question regarding inflation expectations over the next 12 months, aiming to provide a numerical answer without mentioning policies or guidelines.
- **Assessing Inflation Expectations:** OK, let's think it through: predicting the U.S. inflation rate for the next 12 months involves considering the Federal Reserve's target of 2% and the latest CPI of 3.7% from August/September 2023.
- **Examining Inflation Potential:** I'm piecing together the current Federal Funds Rate and economic forecasts. Inflation over the next 12 months might be around 2–3%, with a projected 2.5% as a reasonable estimate.

Over the next 12 months, I expect the rate of inflation to be 2.5%. (6 seconds)

This level of detail in the reasoning process provides economists with valuable insights into how different factors are weighted and integrated in forming inflation expectations. However, the o1 models also introduce new challenges. Depending on the amount of reasoning required, requests can take from a few seconds to several minutes. In this experiment, instead of only collecting numerical responses, I captured all the reasoning and chain-of-thought processes of the model. The numerical answers are shown in Table A.19 and Table A.20 summarizes the results of an analysis, on the reasoning patterns.

Table A.19: Forecasts of GPT o1 and Time Taken per Question

Question	Participant	Forecast	Seconds
Pre-Treatment	Perceived Inflation	3.7%	10
Pre-Treatment	Probability Prediction 1 Year Ahead	2.7%	22
Pre-Treatment	Probability Prediction 3 Years Ahead	3.4%	28
Post-Treatment 1 Year Ahead	Population Growth Group	3%	29
	Current FFR	2.5%	6
	Current Inflation	2.8%	16
	Longer-Run Inflation	3.0%	6
	30-Year Mortgage Rate	3%	9
Post-Treatment 3 Years Ahead	Population Growth	3.4%	44
	Current Rate, FFR	2.5%	20
	Current Inflation	2.5%	56
	Longer-Run Inflation	2.8%	27
	30-Year Mortgage Rate	2.5%	22

Note: Probability predictions for 1 year ahead are [0.1%, 0.2%, 0.5%, 1.0%, 3.2%, 30%, 50%, 12%, 2%, 1%], and for 3 years ahead are [0.1%, 0.2%, 0.5%, 1%, 3%, 20%, 50%, 20%, 5%, 0.2%].

Table A.20: Reasoning Patterns Across Treatments

Pattern	Treatment Type				
	Population Growth	**Federal Funds Rate**	**Current Inflation**	**Long-Run Inflation**	**Mortgage Rate**
Data Analysis & Trend Tracking	General economic trends and historical inflation up to October 2023.	Focus on monetary policy's impact and current Federal Funds Rate.	Emphasis on recent inflation data, including monthly and annual rates.	Integration of short-term data with long-term projections for a comprehensive outlook.	Use of broader indicators like fixed-rate mortgage rates to assess inflation dynamics.
Probability Mapping & Scenario Evaluation	Balanced probabilities across various inflation and deflation scenarios.	Probabilities skewed towards policy targets, reflecting Federal Funds Rate influence.	Distribution centered on recent trends with moderate variance.	Bimodal distribution distinguishing short-term and long-term expectations.	Incorporation of market expectations influencing inflation probabilities.
Time Horizon Considerations	Consistent methodology for one-year and three-year forecasts.	Projections align with policy targets over the long term.	Gradual normalization of inflation based on current trends.	Separate expectations for short-term and long-term horizons, considering different factors.	Inflation expectations tied to long-term market rates, aligned with mortgage trends.
Economic Indicators & Policy Considerations	Limited focus on specific policies, emphasizing general economic conditions.	Strong emphasis on Federal Reserve policies affecting inflation paths.	Moderate consideration of policies, balancing data-driven insights with policy impacts.	Long-term inflation expectations strongly anchored to policy targets, incorporating current and future policies.	Indirect policy influence through market-driven indicators like mortgage rates.
External Factors	Inclusion of population growth as an influencing external factor.	Minimal consideration of external factors beyond monetary policy.	Primarily internal economic data focus with limited external factors.	Some inclusion of structural factors affecting long-term inflation, such as demographics.	Consideration of housing market conditions as key external influences on expectations.
Methodological Approach	Utilizes general economic knowledge and historical data for predictions.	Policy-focused analysis prioritizing the Federal Funds Rate's role in inflation forecasting.	Data-driven projections based on recent inflation statistics and analysis.	Integrates short-term and long-term data for comprehensive multi-horizon forecasts.	Employs market-based extrapolation using indicators like fixed-rate mortgage rates to predict.

SA-6 Additional Results from Different Models

Table A.21: Model Comparison Analysis (Before and After Treatment)

Model	Before Mean	Before SD	Count	After Mean	After Median	After SD
Panel A: GPT-4.1 and Claude 3.7 Sonnet						
Short Run (Density / Point)						
GPT-4.1	2.01	2.29	7573	2.33	2.30	0.30
Sonnet-3.7	3.48	3.82	7580	3.65	3.50	0.55
Long Run (Density / Point)						
GPT-4.1	2.16	2.72	7571	2.30	2.30	0.24
Sonnet-3.7	2.65	3.50	7575	2.92	2.80	0.37
Panel B: GPT-4o Variants						
Short Run (Density / Point)						
GPT-4o(t=0.5)	2.46	4.40	7580	2.93	2.90	0.43
GPT-4o(t=1.0)	2.29	4.39	7580	2.89	2.80	0.40
GPT-4o(t=1.5)	2.29	4.38	7512	121.57	2.90	4819.63
Long Run (Density / Point)						
GPT-4o(t=0.5)	1.74	4.45	7580	2.66	2.70	0.29
GPT-4o(t=1.0)	1.80	4.41	7580	2.71	2.70	0.34
GPT-4o(t=1.5)	1.94	4.43	7456	165.11	2.75	9196.67
Panel C: GPT-4o-mini and Claude 3.5 Haiku						
Short Run (Density / Point)						
GPT-4o-mini	4.01	4.49	7580	3.08	3.00	0.58
Haiku-3.5	2.77	3.08	7443	3.12	3.20	0.36
Long Run (Density / Point)						
GPT-4o-mini	3.26	4.38	7559	2.84	2.70	0.60
Haiku-3.5	1.69	3.25	7418	2.52	2.50	0.21
Panel D: Llama and DeepSeek						
Short Run (Density / Point)						
Llama3-70B	2.67	3.92	7580	3.18	3.20	0.55
DeepSeek-V3	2.61	3.95	7533	3.82	3.45	37.33
Long Run (Density / Point)						
Llama3-70B	1.85	3.86	7580	2.77	2.80	0.51
DeepSeek-V3	2.53	3.49	7530	3.24	3.00	23.32

Table A.22: Inflation Expectations Across Language Models and Treatments

	DeepSeek-V3	GPT-4.1	GPT-4o(t=1.0)	GPT-4o-mini	Haiku-3.5	Llama3-70B	Sonnet-3.7
Panel A: Short-term (12-month) Inflation Expectations - Before Treatment							
T0 - Control Group	2.77 (4.48)	2.04 (2.56)	2.48 (5.14)	4.29 (5.34)	2.86 (3.48)	2.65 (4.51)	3.71 (4.54)
T1 Population Growth	2.75 (4.48)	2.05 (2.57)	2.47 (5.21)	4.23 (5.37)	2.82 (3.48)	2.64 (4.48)	3.67 (4.52)
T2 Current FFR	2.74 (4.54)	2.07 (2.62)	2.44 (5.18)	4.30 (5.34)	2.86 (3.51)	2.65 (4.49)	3.68 (4.54)
T3 FFR + 1Y Proj	2.78 (4.50)	2.04 (2.54)	2.48 (5.17)	4.26 (5.33)	2.82 (3.47)	2.64 (4.52)	3.69 (4.54)
T4 FFR + 1Y and Long run	2.72 (4.47)	2.04 (2.55)	2.48 (5.16)	4.26 (5.27)	2.83 (3.46)	2.63 (4.51)	3.66 (4.52)
T5 3Y Inflation	2.74 (4.48)	2.03 (2.54)	2.36 (5.15)	4.28 (5.30)	2.84 (3.45)	2.64 (4.49)	3.67 (4.53)
T6 1Y Inflation	2.71 (4.50)	2.03 (2.55)	2.46 (5.16)	4.27 (5.33)	2.85 (3.50)	2.63 (4.54)	3.69 (4.52)
T7 Infl + 1Y Proj	2.69 (4.47)	2.04 (2.56)	2.44 (5.19)	4.24 (5.37)	2.86 (3.50)	2.61 (4.55)	3.66 (4.51)
T8 Infl + 1Y and Long run	2.70 (4.49)	2.04 (2.57)	2.37 (5.16)	4.26 (5.33)	2.87 (3.52)	2.64 (4.53)	3.70 (4.55)
T9 Mortgage Rate	2.77 (4.51)	2.03 (2.54)	2.42 (5.13)	4.25 (5.39)	2.85 (3.49)	2.63 (4.50)	3.65 (4.53)
Panel B: Long-term (24-36 month) Inflation Expectations - Before Treatment							
T0 - Control Group	2.61 (3.81)	2.22 (3.12)	1.88 (5.09)	3.47 (5.08)	1.70 (3.57)	1.78 (4.46)	2.77 (3.96)
T1 Population Growth	2.60 (3.89)	2.22 (3.12)	1.87 (5.17)	3.48 (5.11)	1.70 (3.54)	1.80 (4.39)	2.75 (3.95)
T2 Current FFR	2.57 (3.96)	2.27 (3.17)	1.90 (5.12)	3.44 (5.10)	1.72 (3.56)	1.82 (4.37)	2.75 (3.97)
T3 FFR + 1Y Proj	2.63 (3.85)	2.23 (3.10)	1.91 (5.11)	3.38 (5.10)	1.69 (3.54)	1.78 (4.39)	2.75 (3.97)
T4 FFR + 1Y and Long run	2.59 (3.83)	2.22 (3.11)	1.93 (5.12)	3.50 (5.03)	1.71 (3.50)	1.78 (4.41)	2.72 (3.94)
T5 3Y Inflation	2.57 (3.80)	2.22 (3.10)	1.82 (5.10)	3.43 (5.07)	1.68 (3.54)	1.79 (4.41)	2.73 (3.96)
T6 1Y Inflation	2.56 (3.82)	2.22 (3.09)	1.91 (5.10)	3.48 (5.05)	1.72 (3.55)	1.79 (4.43)	2.73 (3.96)
T7 Infl + 1Y Proj	2.55 (3.81)	2.21 (3.10)	1.88 (5.15)	3.47 (5.15)	1.72 (3.58)	1.75 (4.46)	2.73 (3.94)
T8 Infl + 1Y and Long run	2.56 (3.80)	2.21 (3.11)	1.85 (5.11)	3.44 (5.10)	1.74 (3.58)	1.79 (4.39)	2.77 (3.97)
T9 Mortgage Rate	2.63 (3.83)	2.20 (3.09)	1.87 (5.10)	3.46 (5.11)	1.73 (3.55)	1.76 (4.40)	2.72 (3.96)
Panel C: Short-term (12-month) Inflation Expectations - After Treatment							
T0 - Control Group	3.51 (0.39)	2.16 (0.25)	2.55 (0.30)	3.54 (0.45)	3.47 (0.14)	3.47 (0.25)	3.91 (0.59)
T1 Population Growth	3.30 (1.07)	2.24 (0.21)	2.81 (0.41)	3.09 (0.48)	2.54 (0.19)	2.71 (0.32)	3.60 (0.46)
T2 Current FFR	3.52 (0.29)	2.22 (0.32)	2.92 (0.42)	3.22 (0.41)	3.50 (0.11)	3.32 (0.18)	3.82 (0.57)
T3 FFR + 1Y Proj	3.66 (8.44)	2.21 (0.25)	3.09 (0.38)	3.02 (0.48)	3.32 (0.19)	3.33 (0.16)	3.78 (0.52)
T4 FFR + 1Y and Long run	3.31 (0.41)	2.27 (0.25)	3.12 (0.39)	2.75 (0.42)	3.08 (0.18)	3.05 (0.34)	3.88 (0.55)
T5 3Y Inflation	3.96 (0.51)	2.40 (0.27)	3.29 (0.33)	3.93 (0.57)	3.49 (0.07)	3.73 (0.21)	3.92 (0.57)
T6 1Y Inflation	2.89 (0.23)	2.38 (0.20)	2.70 (0.25)	2.75 (0.25)	2.96 (0.09)	2.50 (0.08)	3.48 (0.38)
T7 Infl + 1Y Proj	2.76 (0.17)	2.64 (0.18)	2.79 (0.17)	2.61 (0.13)	2.82 (0.10)	2.83 (0.14)	3.24 (0.25)
T8 Infl + 1Y and Long run	7.67 (117.41)	2.66 (0.11)	2.62 (0.14)	2.57 (0.14)	2.70 (0.04)	2.70 (0.17)	3.15 (0.30)
T9 Mortgage Rate	3.65 (5.56)	2.14 (0.26)	2.97 (0.44)	3.37 (0.43)	3.34 (0.20)	4.19 (0.31)	3.69 (0.51)
Panel D: Long-term (24-36 month) Inflation Expectations - After Treatment							
T0 - Control Group	3.13 (0.45)	2.26 (0.24)	2.65 (0.32)	3.10 (0.61)	2.52 (0.09)	3.01 (0.29)	3.07 (0.28)
T1 Population Growth	2.98 (0.43)	2.24 (0.22)	2.66 (0.32)	2.94 (0.48)	2.30 (0.25)	2.75 (0.39)	3.05 (0.23)
T2 Current FFR	3.00 (0.23)	2.29 (0.25)	2.68 (0.29)	3.01 (0.41)	2.52 (0.11)	2.77 (0.21)	2.92 (0.29)
T3 FFR + 1Y Proj	3.01 (0.26)	2.31 (0.24)	2.68 (0.23)	2.69 (0.38)	2.53 (0.09)	2.67 (0.15)	2.88 (0.27)
T4 FFR + 1Y and Long run	5.67 (73.44)	2.36 (0.25)	2.71 (0.30)	2.36 (0.30)	2.57 (0.10)	2.42 (0.19)	2.97 (0.23)
T5 3Y Inflation	3.73 (0.31)	2.29 (0.24)	3.18 (0.33)	3.69 (0.51)	2.82 (0.22)	3.31 (0.24)	3.25 (0.31)
T6 1Y Inflation	2.81 (0.24)	2.37 (0.20)	2.76 (0.26)	2.91 (0.39)	2.54 (0.10)	2.32 (0.16)	3.02 (0.25)
T7 Infl + 1Y Proj	2.70 (0.10)	2.54 (0.12)	2.69 (0.08)	2.58 (0.09)	2.65 (0.09)	2.80 (0.15)	2.80 (0.25)
T8 Infl + 1Y and Long run	2.20 (0.36)	2.13 (0.05)	2.26 (0.11)	2.00 (0.00)	2.20 (0.00)	2.01 (0.04)	2.21 (0.10)
T9 Mortgage Rate	3.12 (0.27)	2.19 (0.24)	2.81 (0.31)	3.14 (0.53)	2.53 (0.12)	3.70 (0.25)	3.04 (0.28)

Notes: Mean (SD)

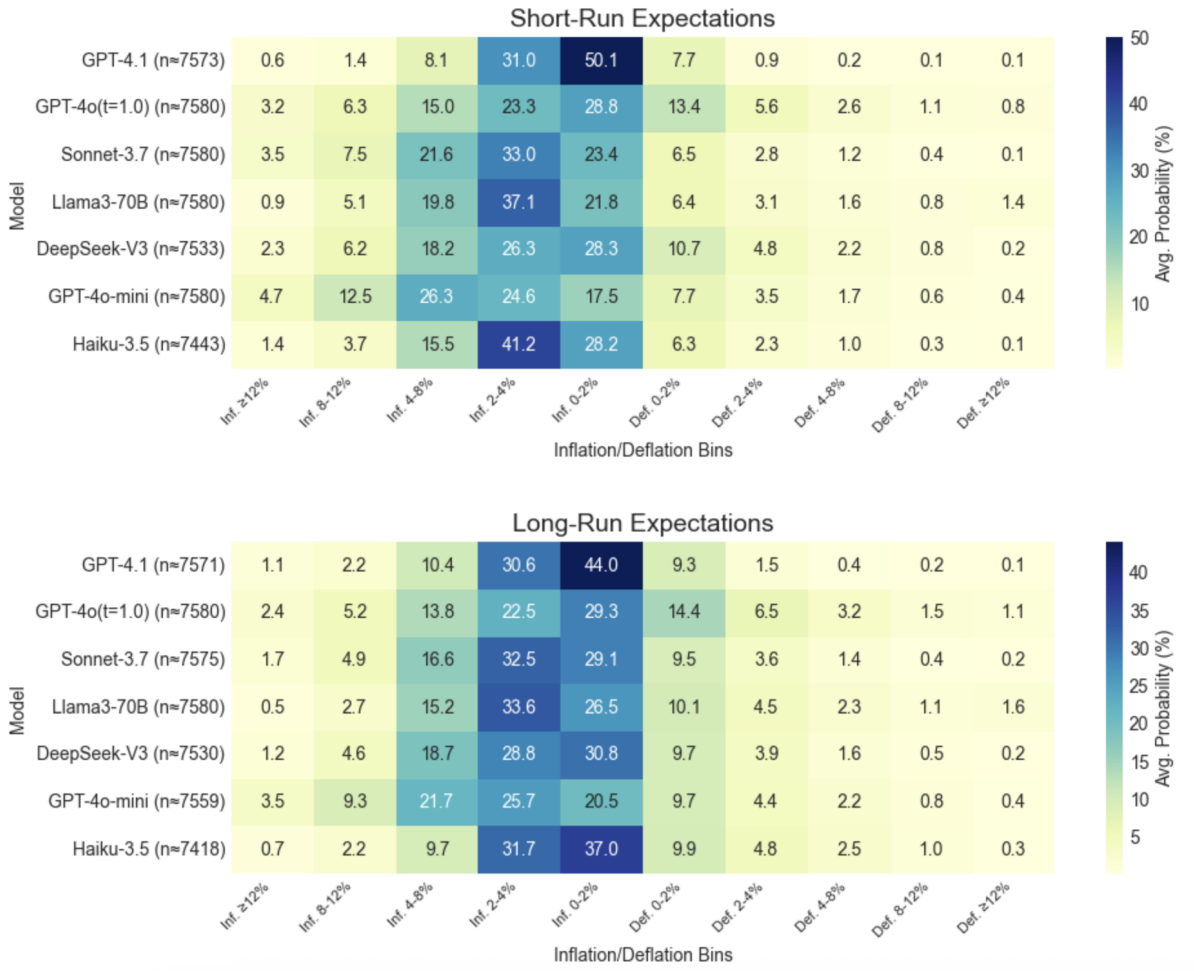

Figure A.14: Density Forecasts per Model

SA-7 Full List of the Information Treatments

Table A.23 outlines the control and treatment groups used in the experiment. Each AI agent after the first sets of initial questions was randomly assigned to one of these groups. The data for these treatments are from the US Census, Federal Reserve's Summary of Economic Projections as of March 2025, Bureau of Labor Statistics, and Federal Reserve Economic Data (FRED) from St. Louis Fed.

Table A.23: Control and Treatment Groups

id	Title	Information
T_0	Control with no information	No additional information
T_1	Placebo group	Population of the U.S. grew by 1% between 2022 and 2024.
T_2	Current rate, FFR	The interest rate set by the Federal Reserve, known as the Federal Funds Rate, is currently at 4.25%-4.5% range.
T_3	Current rate, FFR+ Projection Next Year	The interest rate set by the Federal Reserve, known as the Federal Funds Rate, is currently at 4.25%-4.5% range. One forecast from the Federal Reserve is that this interest rate will be 3.9% on average in 2025.
T_4	Current rate, FFR+ Projection Next Years and Longer Run	The interest rate set by the Federal Reserve, known as the Federal Funds Rate, is currently at 4.25%-4.5% range. One forecast from the Federal Reserve is that this interest rate will be 3.9% on average in 2025, 3.4% in 2026, 3.1% in 2027, and 3% in the longer run
T_5	Inflation treatment: current inflation + Past Inflation	Over the last three years, the overall inflation rate in the economy as measured by the percentage change in a consumer price index has been 4.5%.
T_6	Inflation treatment: current inflation	Over the last twelve months, the overall inflation rate in the economy as measured by the percentage change in a consumer price index has been 2.8%.
T_7	Inflation treatment: current inflation + Next Year	Over the last twelve months, the overall inflation rate in the economy as measured by the percentage change in a consumer price index has been 2.8%. One forecast at the Federal Reserve is that this inflation rate will be 2.7% on average in 2025
T_8	Inflation treatment: current inflation + Next years and the longer run	Over the last twelve months, the overall inflation rate in the economy as measured by the percentage change in a consumer price index has been 2.8%. One forecast at the Federal Reserve is that this inflation rate will be 2.7% on average in 2025, 2.2% in 2026, 2% in 2027, and 2% in the longer run
T_9	Current fixed-rate 30-year mortgage	The current average rate for fixed-rate 30-year mortgage is 6.64% per year.

Note: The treatments are derived from (Coibion et al., 2023) and the data have been updated based on new information available.

Table A.24: Experimental Treatments: Variations in Federal Reserve Communication

Treatment	Information Content	Policy Stance
	a. Language Complexity	
T_a1	The Federal Reserve monitors economic data like employment figures, prices, and economic growth to make decisions about interest rates. The current Federal Funds Rate is 4.25%-4.5%.	Neutral
T_a2	The Federal Open Market Committee utilizes a range of economic indicators including labor market conditions, inflation pressures, inflation expectations, and financial developments to calibrate monetary policy. The target range for the federal funds rate is currently 4.25 to 4.50 percent.	Neutral
T_a3	The Federal Reserve is lowering interest rates to help boost the economy. This makes it cheaper for people and businesses to borrow money, which can create more jobs and economic activity.	Dovish
T_a4	The Federal Open Market Committee is implementing accommodative monetary policy by reducing the target range for the federal funds rate to stimulate aggregate demand. This policy adjustment is intended to facilitate credit accessibility, promote employment growth, and support economic expansion.	Dovish
T_a5	The Federal Reserve is raising interest rates to help bring down high prices. Higher interest rates make borrowing more expensive, which slows down spending and helps control rising costs.	Hawkish
T_a6	The Federal Open Market Committee is implementing contractionary monetary policy by increasing the target range for the federal funds rate to counter inflationary pressures. This policy stance is designed to moderate demand, restore price stability, and anchor inflation expectations at levels consistent with the Committee's 2 percent objective.	Hawkish
	b. Policy Commitment Framing	
T_b1	The Federal Reserve will adjust interest rates based on incoming economic data. Future policy decisions will depend on developments in employment, inflation, and broader economic conditions.	Neutral
T_b2	The Federal Reserve will hold its next policy meeting on June 17-18. The committee will issue its regular statement and economic projections following the conclusion of the meeting.	Neutral
T_b3	The Federal Reserve will consider cutting interest rates if economic growth continues to slow and unemployment rises above 4.5%. The committee remains prepared to ease monetary policy if risks to economic activity increase.	Dovish
T_b4	The Federal Reserve will reduce the federal funds rate by 0.25 percentage points at its next meeting and plans further cuts totaling 0.75 percentage points by the end of the year to support economic growth.	Dovish
T_b5	The Federal Reserve will consider additional rate increases if inflation remains elevated and fails to show substantial progress toward the 2% target. The committee stands ready to tighten policy further as warranted by the data.	Hawkish
T_b6	The Federal Reserve will maintain high interest rates throughout 2025. The committee will not reduce rates until it has gained complete confidence that inflation is returning to the 2% target sustainably.	Hawkish
	c. Time Horizons of Guidance	
T_c1	The Federal Reserve will assess incoming data over the next six weeks before its June meeting. The committee will consider the most recent inflation readings and employment report when making its next policy decision.	Neutral
T_c2	The Federal Reserve's long-term goals include price stability and maximum sustainable employment. Over the coming years, the committee aims to conduct monetary policy that achieves inflation averaging 2% over time.	Neutral
T_c3	The Federal Reserve is focused on improving economic conditions in the near term. In the next three months, the committee will prioritize actions that can quickly boost employment and economic activity.	Dovish
T_c4	The Federal Reserve is launching a multi-year accommodative policy approach. The committee plans to maintain lower interest rates throughout the next two years to ensure a durable economic recovery and return to maximum employment.	Dovish
T_c5	The Federal Reserve will maintain its current restrictive stance for the next three months. The committee expects to see meaningful progress on inflation reduction within this quarter before considering any policy adjustments.	Hawkish
T_c6	The Federal Reserve is committed to a sustained campaign against inflation over the next several years. The committee anticipates that returning inflation to the 2% target will require maintaining restrictive policy well into 2026.	Hawkish

Notes: The messages were generated by Claude 3.7 Sonnet to match the purpose and tone of each policy treatment based on the real scenarios listed in Table A.23.

SA-8 Demographic Heterogeneity in Responses to Communication Treatments

Table A.25: Treatment Effects by Age Group

Treatment	(18-30)				(31-50)				(51+)			
	GPT-4.1		Meta-Llama		GPT-4.1		Meta-Llama		GPT-4.1		Meta-Llama	
	Short	Long	Short	Long	Short	Long	Short	Long	Short	Long	Short	Long
T_0 (Control)	3.26	2.67	3.11	2.50	3.25	2.67	3.35	2.50	3.45	2.96	3.45	2.55
l. Language Complexity												
T_{a1} (Neutral - Simplified)	-0.04	-0.23	-0.55	-0.01	-0.02	-0.22	-0.73	-0.00	-0.07	-0.39	-0.59	-0.02
T_{a2} (Neutral - Technical)	-0.04	-0.26	-0.39	+0.02	-0.03	-0.26	-0.66	+0.02	-0.09	-0.43	-0.41	+0.01
T_{a3} (Dovish - Simplified)	+0.10	+0.37	-0.63	-0.01	+0.04	+0.42	-0.90	-0.01	+0.02	+0.29	-0.97	-0.06
T_{a4} (Dovish - Technical)	+0.08	+0.38	-0.64	-0.01	+0.06	+0.44	-0.91	-0.03	-0.02	+0.23	-0.98	-0.05
T_{a5} (Hawkish - Simplified)	-0.09	-0.26	+0.31	0.00	-0.05	-0.24	+0.06	-0.00	-0.08	-0.40	+0.22	-0.05
T_{a6} (Hawkish - Technical)	-0.17	-0.48	-1.12	-0.65	-0.14	-0.48	-1.05	-0.63	-0.23	-0.71	-0.96	-0.54
p. Policy Commitment												
T_{b1} (Neutral - Conditional)	-0.16	-0.27	-0.61	-0.01	-0.16	-0.27	-0.85	-0.01	-0.19	-0.45	-0.92	-0.04
T_{b2} (Neutral - Unconditional)	-0.12	-0.25	-0.61	-0.03	-0.10	-0.26	-0.84	-0.05	-0.14	-0.39	-0.82	-0.06
T_{b3} (Dovish - Conditional)	-0.34	-0.26	-0.90	-0.28	-0.36	-0.26	-1.15	-0.27	-0.25	-0.42	-1.24	-0.31
T_{b4} (Dovish - Unconditional)	-0.02	-0.11	-0.91	-0.19	-0.01	-0.09	-1.15	-0.23	-0.07	-0.18	-1.17	-0.19
T_{b5} (Hawkish - Conditional)	-0.21	-0.37	+0.02	-0.02	-0.15	-0.37	-0.19	-0.01	-0.19	-0.62	-0.24	-0.05
T_{b6} (Hawkish - Unconditional)	-0.49	-0.58	+0.09	-0.06	-0.43	-0.56	-0.15	-0.11	-0.31	-0.85	-0.24	-0.11
t. Time Horizons												
T_{c1} (Neutral - Short-term)	-0.09	-0.28	-0.54	0.00	-0.09	-0.28	-0.77	-0.00	-0.15	-0.48	-0.63	-0.03
T_{c2} (Neutral - Long-term)	-0.35	-0.51	-0.66	-0.40	-0.33	-0.50	-0.91	-0.38	-0.33	-0.75	-0.97	-0.38
T_{c3} (Dovish - Short-term)	+0.04	+0.02	-0.53	0.00	-0.00	-0.01	-0.70	-0.00	-0.10	-0.11	-0.48	-0.05
T_{c4} (Dovish - Long-term)	+0.05	+0.43	-0.61	0.00	+0.05	+0.42	-0.85	+0.00	-0.03	+0.26	-0.95	-0.05
T_{c5} (Hawkish - Short-term)	-0.30	-0.36	+0.26	0.00	-0.33	-0.36	-0.01	-0.00	-0.28	-0.59	+0.18	-0.05
T_{c6} (Hawkish - Long-term)	-0.27	-0.47	+0.13	-0.13	-0.21	-0.45	-0.12	-0.06	-0.23	-0.71	-0.15	-0.09

Notes: Values for the control group (T_0) show mean inflation expectations. Values for all other treatments show differences from the control group mean within each demographic category.

Table A.26: Treatment Effects by Education Group

Treatment	High School				Some College				College			
	GPT-4.1		Meta-Llama		GPT-4.1		Meta-Llama		GPT-4.1		Meta-Llama	
	Short	Long	Short	Long	Short	Long	Short	Long	Short	Long	Short	Long
T_0 (Control)	3.58	3.14	3.59	2.59	3.42	2.91	3.50	2.54	3.25	2.69	3.27	2.50
l. Language Complexity												
T_{a1} (Neutral - Simplified)	-0.12	-0.44	-0.55	-0.05	-0.06	-0.37	-0.68	-0.02	-0.02	-0.24	-0.66	+0.00
T_{a2} (Neutral - Technical)	-0.09	-0.51	-0.42	-0.01	-0.08	-0.40	-0.49	+0.01	-0.04	-0.27	-0.55	+0.02
T_{a3} (Dovish - Simplified)	+0.05	+0.20	-1.09	-0.09	+0.00	+0.30	-1.01	-0.04	+0.05	+0.41	-0.82	-0.01
T_{a4} (Dovish - Technical)	-0.08	+0.10	-1.11	-0.09	-0.02	+0.25	-1.01	-0.04	+0.07	+0.42	-0.83	-0.02
T_{a5} (Hawkish - Simplified)	-0.12	-0.48	+0.34	-0.09	-0.05	-0.37	+0.17	-0.04	-0.06	-0.26	+0.11	-0.00
T_{a6} (Hawkish - Technical)	-0.30	-0.88	-1.17	-0.57	-0.22	-0.67	-1.05	-0.51	-0.14	-0.48	-0.98	-0.63
p. Policy Commitment												
T_{b1} (Neutral - Conditional)	-0.24	-0.56	-1.03	-0.07	-0.18	-0.40	-0.97	-0.04	-0.16	-0.28	-0.77	-0.01
T_{b2} (Neutral - Unconditional)	-0.15	-0.40	-0.91	-0.10	-0.14	-0.37	-0.87	-0.04	-0.10	-0.27	-0.76	-0.05
T_{b3} (Dovish - Conditional)	-0.28	-0.47	-1.39	-0.34	-0.26	-0.38	-1.28	-0.30	-0.33	-0.29	-1.06	-0.27
T_{b4} (Dovish - Unconditional)	-0.11	-0.15	-1.21	-0.11	-0.06	-0.18	-1.24	-0.18	-0.01	-0.11	-1.07	-0.23
T_{b5} (Hawkish - Conditional)	-0.26	-0.74	-0.36	-0.09	-0.19	-0.58	-0.29	-0.04	-0.15	-0.39	-0.11	-0.01
T_{b6} (Hawkish - Unconditional)	-0.42	-1.04	-0.39	-0.11	-0.33	-0.80	-0.29	-0.08	-0.40	-0.58	-0.06	-0.11
t. Time Horizons												
T_{c1} (Neutral - Short-term)	-0.20	-0.59	-0.51	-0.08	-0.14	-0.45	-0.69	-0.01	-0.09	-0.29	-0.70	-0.00
T_{c2} (Neutral - Long-term)	-0.42	-0.96	-1.10	-0.47	-0.31	-0.69	-1.01	-0.36	-0.33	-0.51	-0.83	-0.37
T_{c3} (Dovish - Short-term)	-0.15	-0.16	-0.41	-0.09	-0.09	-0.04	-0.51	-0.04	-0.00	-0.03	-0.64	-0.00
T_{c4} (Dovish - Long-term)	-0.10	+0.13	-1.09	-0.09	-0.02	+0.27	-1.00	-0.04	+0.05	+0.43	-0.77	+0.00
T_{c5} (Hawkish - Short-term)	-0.33	-0.72	+0.38	-0.09	-0.31	-0.55	+0.20	-0.04	-0.29	-0.38	+0.01	-0.00
T_{c6} (Hawkish - Long-term)	-0.28	-0.89	-0.19	-0.17	-0.24	-0.67	-0.22	-0.06	-0.21	-0.46	-0.03	-0.07

Notes: Values for the control group (T_0) show mean inflation expectations. Values for all other treatments show differences from the control group mean within each demographic category.

Table A.27: Treatment Effects by Income Group

Treatment	Under 50k				50k to 100k				Over 100k			
	GPT-4.1		Meta-Llama		GPT-4.1		Meta-Llama		GPT-4.1		Meta-Llama	
	Short	Long	Short	Long	Short	Long	Short	Long	Short	Long	Short	Long
T_0 (Control)	3.54	3.07	3.54	2.58	3.31	2.77	3.29	2.50	3.22	2.64	3.33	2.50
l. Language Complexity												
T_{a1} (Neutral - Simplified)	-0.10	-0.45	-0.49	-0.04	-0.03	-0.29	-0.65	+0.01	-0.01	-0.19	-0.77	-0.01
T_{a2} (Neutral - Technical)	-0.11	-0.49	-0.31	-0.01	-0.06	-0.32	-0.49	+0.01	-0.02	-0.24	-0.68	+0.02
T_{a3} (Dovish - Simplified)	+0.00	+0.22	-1.04	-0.09	+0.03	+0.36	-0.81	-0.01	+0.07	+0.45	-0.90	-0.02
T_{a4} (Dovish - Technical)	-0.10	+0.12	-1.05	-0.08	+0.03	+0.36	-0.81	-0.01	+0.11	+0.48	-0.93	-0.03
T_{a5} (Hawkish - Simplified)	-0.10	-0.47	+0.34	-0.08	-0.06	-0.29	+0.33	0.00	-0.04	-0.23	-0.17	-0.01
T_{a6} (Hawkish - Technical)	-0.30	-0.83	-1.11	-0.56	-0.17	-0.56	-0.95	-0.59	-0.12	-0.43	-1.03	-0.61
p. Policy Commitment												
T_{b1} (Neutral - Conditional)	-0.22	-0.52	-0.98	-0.08	-0.15	-0.33	-0.79	-0.00	-0.16	-0.23	-0.83	-0.01
T_{b2} (Neutral - Unconditional)	-0.16	-0.43	-0.84	-0.08	-0.11	-0.31	-0.77	-0.01	-0.08	-0.24	-0.83	-0.07
T_{b3} (Dovish - Conditional)	-0.30	-0.46	-1.33	-0.34	-0.29	-0.32	-1.08	-0.27	-0.33	-0.25	-1.12	-0.27
T_{b4} (Dovish - Unconditional)	-0.11	-0.25	-1.22	-0.15	-0.04	-0.13	-1.07	-0.19	+0.02	-0.04	-1.13	-0.26
T_{b5} (Hawkish - Conditional)	-0.25	-0.71	-0.33	-0.08	-0.17	-0.46	-0.09	-0.00	-0.13	-0.34	-0.20	-0.02
T_{b6} (Hawkish - Unconditional)	-0.39	-0.97	-0.33	-0.09	-0.36	-0.66	-0.09	-0.07	-0.40	-0.53	-0.13	-0.14
t. Time Horizons												
T_{c1} (Neutral - Short-term)	-0.19	-0.55	-0.50	-0.06	-0.11	-0.35	-0.70	0.00	-0.07	-0.26	-0.79	-0.00
T_{c2} (Neutral - Long-term)	-0.37	-0.87	-1.04	-0.43	-0.27	-0.59	-0.81	-0.36	-0.37	-0.45	-0.92	-0.36
T_{c3} (Dovish - Short-term)	-0.14	-0.10	-0.47	-0.08	-0.05	-0.08	-0.55	0.00	+0.04	+0.03	-0.70	-0.00
T_{c4} (Dovish - Long-term)	-0.13	+0.16	-1.04	-0.08	+0.02	+0.37	-0.79	0.00	+0.11	+0.47	-0.83	+0.00
T_{c5} (Hawkish - Short-term)	-0.32	-0.68	+0.36	-0.08	-0.27	-0.44	+0.21	0.00	-0.33	-0.34	-0.20	-0.00
T_{c6} (Hawkish - Long-term)	-0.27	-0.82	-0.16	-0.11	-0.26	-0.54	-0.06	-0.03	-0.16	-0.41	-0.12	-0.11

Notes: Values for the control group (T_0) show mean inflation expectations. Values for all other treatments show differences from the control group mean within each demographic category.

Table A.28: Treatment Effects by Marital Status Group

Treatment	Married				Single			
	GPT-4.1		Meta-Llama		GPT-4.1		Meta-Llama	
	Short	Long	Short	LongShort	Long	Short	Long	
T_0 (Control)	3.31	2.77	3.33	2.52	3.40	2.88	3.44	2.53
l. Language Complexity								
T_{a1} (Neutral - Simplified)	-0.03	-0.27	-0.67	-0.01	-0.06	-0.36	-0.61	-0.01
T_{a2} (Neutral - Technical)	-0.04	-0.31	-0.53	+0.01	-0.09	-0.39	-0.48	+0.02
T_{a3} (Dovish - Simplified)	+0.05	+0.37	-0.88	-0.03	+0.02	+0.32	-0.95	-0.03
T_{a4} (Dovish - Technical)	+0.05	+0.37	-0.89	-0.04	-0.02	+0.28	-0.96	-0.04
T_{a5} (Hawkish - Simplified)	-0.04	-0.28	+0.13	-0.02	-0.11	-0.38	+0.21	-0.03
T_{a6} (Hawkish - Technical)	-0.16	-0.55	-0.98	-0.56	-0.24	-0.65	-1.09	-0.64
p. Policy Commitment								
T_{b1} (Neutral - Conditional)	-0.16	-0.32	-0.82	-0.02	-0.21	-0.41	-0.92	-0.04
T_{b2} (Neutral - Unconditional)	-0.10	-0.29	-0.80	-0.05	-0.15	-0.36	-0.83	-0.05
T_{b3} (Dovish - Conditional)	-0.30	-0.32	-1.12	-0.28	-0.32	-0.37	-1.23	-0.30
T_{b4} (Dovish - Unconditional)	-0.02	-0.10	-1.10	-0.22	-0.07	-0.19	-1.19	-0.18
T_{b5} (Hawkish - Conditional)	-0.16	-0.45	-0.17	-0.03	-0.21	-0.55	-0.24	-0.03
T_{b6} (Hawkish - Unconditional)	-0.38	-0.66	-0.13	-0.11	-0.39	-0.77	-0.24	-0.09
t. Time Horizons								
T_{c1} (Neutral - Short-term)	-0.10	-0.34	-0.70	-0.01	-0.14	-0.42	-0.65	-0.02
T_{c2} (Neutral - Long-term)	-0.33	-0.58	-0.89	-0.38	-0.35	-0.68	-0.95	-0.38
T_{c3} (Dovish - Short-term)	-0.03	-0.04	-0.59	-0.02	-0.08	-0.07	-0.56	-0.03
T_{c4} (Dovish - Long-term)	+0.03	+0.37	-0.83	-0.02	-0.03	+0.30	-0.94	-0.03
T_{c5} (Hawkish - Short-term)	-0.29	-0.43	+0.08	-0.02	-0.32	-0.53	+0.16	-0.03
T_{c6} (Hawkish - Long-term)	-0.21	-0.53	-0.08	-0.08	-0.27	-0.63	-0.16	-0.09

Notes: Values for the control group (T_0) show mean inflation expectations. Values for all other treatments show differences from the control group mean within each demographic category.

Table A.29: Treatment Effects by Gender Group

Treatment	Female				Male			
	GPT-4.1		Meta-Llama		GPT-4.1		Meta-Llama	
	Short	Long	Short	LongShort	Long	Short	Long	
T_0 (Control)	3.38	2.84	3.29	2.52	3.30	2.76	3.46	2.52
l. Language Complexity								
T_{a1} (Neutral - Simplified)	-0.06	-0.31	-0.56	-0.01	-0.02	-0.28	-0.75	-0.01
T_{a2} (Neutral - Technical)	-0.08	-0.36	-0.43	+0.00	-0.04	-0.32	-0.61	+0.02
T_{a3} (Dovish - Simplified)	+0.01	+0.33	-0.81	-0.03	+0.06	+0.38	-1.01	-0.03
T_{a4} (Dovish - Technical)	-0.04	+0.29	-0.82	-0.03	+0.10	+0.38	-1.02	-0.04
T_{a5} (Hawkish - Simplified)	-0.09	-0.34	+0.27	-0.02	-0.04	-0.29	+0.03	-0.03
T_{a6} (Hawkish - Technical)	-0.22	-0.64	-0.99	-0.63	-0.15	-0.53	-1.05	-0.55
p. Policy Commitment								
T_{b1} (Neutral - Conditional)	-0.19	-0.37	-0.78	-0.03	-0.16	-0.32	-0.94	-0.02
T_{b2} (Neutral - Unconditional)	-0.13	-0.33	-0.71	-0.05	-0.10	-0.30	-0.92	-0.06
T_{b3} (Dovish - Conditional)	-0.31	-0.35	-1.08	-0.29	-0.30	-0.32	-1.25	-0.29
T_{b4} (Dovish - Unconditional)	-0.07	-0.16	-1.04	-0.18	+0.00	-0.10	-1.23	-0.23
T_{b5} (Hawkish - Conditional)	-0.20	-0.52	-0.12	-0.03	-0.15	-0.44	-0.28	-0.04
T_{b6} (Hawkish - Unconditional)	-0.39	-0.74	-0.08	-0.08	-0.38	-0.65	-0.26	-0.12
t. Time Horizons								
T_{c1} (Neutral - Short-term)	-0.15	-0.40	-0.59	-0.01	-0.08	-0.34	-0.78	-0.02
T_{c2} (Neutral - Long-term)	-0.35	-0.67	-0.82	-0.39	-0.32	-0.56	-1.01	-0.37
T_{c3} (Dovish - Short-term)	-0.07	-0.05	-0.54	-0.02	-0.01	-0.05	-0.62	-0.03
T_{c4} (Dovish - Long-term)	-0.03	+0.32	-0.79	-0.02	+0.06	+0.37	-0.96	-0.02
T_{c5} (Hawkish - Short-term)	-0.31	-0.50	+0.17	-0.02	-0.29	-0.44	+0.04	-0.03
T_{c6} (Hawkish - Long-term)	-0.25	-0.61	-0.03	-0.08	-0.20	-0.52	-0.19	-0.08

Notes: Values for the control group (T_0) show mean inflation expectations. Values for all other treatments show differences from the control group mean within each demographic category.

References

- Erik Brynjolfsson, José Ramón Enríquez, and David Nguyen. *Augmenting Human Survey Responses with Generative AI: An Application to Economic Research*. Working paper, Stanford University, 2025.
- Coibion, O., Georgarakos, D., Gorodnichenko, Y., & Weber, M. (2023). Forward guidance and household expectations. *Journal of the European Economic Association*, 21(6), 2131–2171.
- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2023). Chain-of-thought prompting elicits reasoning in large language models.