

# Experimental Evidence on Large Language Models

By ALI ZARIFHONARVAR\*

*This study investigates the use of Large Language Models in inflation expectation formation by giving them text data from the Federal Open Market Committee. Also, by employing a novel experimental design, this research integrates generative AI with economic analysis to explore how different types of information treatments influence LLMs' responses. Using six different type of knowledge source it appears that there is an impact of the type of the retrieval each agent does. The variation in the one that has irrelevant document is higher than those that have access to related document. Also, similar to human subject the information treatment regarding mortgage rate has the highest impact on updating prior inflation expectations.*

JEL: C9, E31, D84

Keywords: Large Language Models, Inflation Expectation, Information, GPT, Survey Experiment

## I. Introduction

The use of Generative Artificial Intelligence (GAI), particularly with Large Language Models (LLMs) like ChatGPT, signals a shift in the landscape of economic research. These models have the potential to assist in a range of tasks, including idea generation, feedback, data analysis, writing, programming, and mathematical calculations (Korinek, 2023). As we begin to integrate LLMs into economics, it has become clear that their impact extends beyond only supporting various tasks; they can enhance the methodologies used in economic research, suggesting new approaches for economists to conduct studies and develop insights.

It goes without saying that people have very different views about how powerful LLMs are. Some view them as an advanced autocomplete feature, emphasizing their limitations and questioning their potential impact. Others see them as nearing Artificial General Intelligence (AGI), focusing on their capabilities and potential for complex tasks. This dichotomy shows the need for more comprehensive understanding of their behaviors and cognitive processes. The way LLMs think and operate differs from human, and conducting experiments with AI agents or surveying these agents could offer a novel method to bridge these perspectives. By deploying LLMs in structured surveys, particularly in

\* Indiana University Bloomington - alizarif@iu.edu

economic research, we can obtain a more clear, empirical understanding of their strengths and limitations.

Additionally, the study of Large Language Models is crucial not only for their potential applications but also for the insights they can provide into human cognition and decision-making. LLMs are designed to process and generate natural language in a manner that closely mimics human language. By understanding how these models interpret information and reach conclusions, we can gain valuable insights. This understanding could potentially help us better understand how humans make decisions, form opinions, and respond to information in various contexts. Moreover, by comparing the performance of LLMs with human subjects in experiments, we can identify the strengths and limitations of these models and assess their potential to augment or even replace human decision-making in certain domains.

In the landscape of generative AI and economics, this paper presents an exploration into the intersection of Large Language Models and inflation expectations surveys. The primary goal of this research is twofold: first, to delve into the decision-making processes of LLMs, particularly in the area of inflation expectation formation, and second, to assess the influence of different knowledge domains and information treatments on their responses.

To achieve these objectives, I introduce a new experimental design that involves custom-built LLMs on a GPT<sup>1</sup> model, specifically with access to data from the Federal Open Market Committee (FOMC). I will use different text corpora for the retrieval capability of these custom GPTs. This paper not only contributes to the growing body of literature on AI and economics but also offers practical insights into the future of AI-assisted economic forecasting and its potential impact on policy-making and individual financial decisions.

## II. Literature Review

Recent developments in the field of Generative Artificial Intelligence, especially in the context of Large Language Models, have showed a significant turning point in economics, introducing innovative tools for research. This paper intersects two main strands of literature: the experimental study of AI agents and large language models in economics, and the formation of inflation expectations. While there exist studies that attempt to connect these two domains ([Faria-e-Castro and Leibovici, 2023](#); [Bybee, 2023](#)), this paper goes deeper by delving more into the decision-making processes of large language models, and assessing the impact of information on AI agents in an experimental setup.

### A. Economics and LLMs

The introduction of large language models, such as ChatGPT, has significantly impacted various sectors, including the field of economics research. Recent stud-

<sup>1</sup>Generative Pre-trained Transformer

ies have showed the potential of LLMs in simulating complex economic scenarios and decision-making processes (Akata et al., 2023; Heydari and Lorè, 2023). Additionally, there is growing interest in understanding the broader implications of generative AI on economics research (Charness, Jabarian and List, 2023; Korinek, 2023).

For instance, Horton (2023) delves into how large language models can function as simulated economic agents, drawing parallels to the concept of '*homo economicus*' in economic theories. These LLMs, equipped with specific preferences, information, and endowments, are tested in a variety of scenarios to examine their behavior. Inspired by classic economic studies, these experiments have revealed that LLMs can imitate human-like behaviors, producing results that are qualitatively similar to those of human subjects. In a related study, Brookins and DeBacker (2023) explores the tendencies of GPT-3.5, towards fairness and cooperation. This research engages GPT model in classic games like the dictator game and the prisoner's dilemma, finding that the AI often chooses strategies that prioritize fairness over optimal outcomes, displaying human-like behavior.

Another aspect of AI research focuses on its potential as an equalizer in society. However, varying adaptation rates among different groups seem to prevent this potential. Haslberger, Gingrich and Bhatia (2023) conducted a study examining the impact of ChatGPT on worker productivity and inequality within a representative segment of the UK's working-age population. The findings suggest that while ChatGPT can enhance productivity, especially in complex tasks, its impact on reducing performance inequality is limited to within occupational groups. The study also indicates that ChatGPT does not significantly alter educational or occupational inequalities and might even create age-based disparities.

Furthermore, the effects of AI on workforce productivity and team dynamics have got attention. Dell'Acqua et al. (2023) conducted a lab in the field experiment in collaboration with Boston Consulting Group, investigates the impact of Large Language Models on the productivity and quality of work in knowledge-intensive tasks. Involving 758 consultants, the experiment assessed their performance across 18 tasks with varying access to GPT-4. Consultants using AI showed a notable increase in productivity (12.2% more tasks completed) and quality (40% higher quality results) for tasks within AI's capabilities. The study also observed distinct human-AI integration strategies among consultants: some acted as "Centaur," effectively dividing tasks between themselves and AI, while others integrated more deeply with AI, resembling "Cyborgs" in their approach.

Recent studies also have shown significant advances in the application of LLMs in the realm of macroeconomics. For example, Li et al. (2023) introduce a framework using large language models to create agents for macroeconomic simulations. These agents are uniquely designed to exhibit human-like decision-making and adaptability in complex economic scenarios.

Faria-e Castro and Leibovici (2023) explore the proficiency of LLMs in generating inflation forecasts, showing that these AI tools can outperform traditional

methods in accuracy and efficiency. They examine the capacity of Google AI's PaLM, to produce conditional inflation forecasts during 2019-2023. Comparing these forecasts with those from the Survey of Professional Forecasters, the study finds LLM forecasts generally more accurate, with lower mean-squared errors. Similar to this research, [Bybee \(2023\)](#) presents a novel approach to generating economic expectations by applying large language models to historical news data. It shows that this method closely aligns generated expectations with existing survey measures for different indicators.

### *B. Inflation Expectation Formation*

Inflation expectations play a crucial role in shaping economic behavior and policy. Understanding how these expectations are formed and modified is important for both policymakers and economists. This area of study delves into the psychological and informational aspect on how people and companies understand and predict future inflation. For instance, [Candia, Coibion and Gorodnichenko \(2020\)](#) examine the impact of how economic agents, especially households and firms, form their expectations and how these expectations influence their decisions. Their findings suggest that information about inflation can sometimes lead to unintended consequences, depending on how the news is interpreted. This aligns with [D'Acunto et al. \(2021\)](#) who show that consumers' inflation expectations are significantly influenced by the prices of frequently purchased grocery items, highlighting the role of personal consumption experiences in shaping these expectations. Additionally, [Coibion et al. \(2023\)](#) delve into the effectiveness of forward guidance communication, finding that households' expectations are responsive to near-term interest rate information but less so to longer-term forecasts.

The formation and impact of inflation expectations are further complicated when considering the perspective of firms. [Coibion and Gorodnichenko \(2015\)](#) present evidence of considerable diversity in firms' expectations about inflation and other macroeconomic conditions. This diversity is attributed to the varying incentives firms have to collect and process information. Similarly, [Coibion, Gorodnichenko and Ropele \(2020\)](#) reveal that higher inflation expectations among firms can lead to significant changes in their pricing, borrowing, and employment decisions. These findings are somehow the same as [Candia, Coibion and Gorodnichenko \(2021\)](#), that introduce new insights into U.S. firms' inflation expectations, noting their lack of anchoring and the disparity from those of households and professional forecasters.

Also, relationship between inflation expectations and monetary policy is a crucial area of study. [Coibion et al. \(2020\)](#) assess the potential of using inflation expectations as a policy tool, highlighting the challenges faced by policymakers in managing these expectations, particularly in low-inflation environments. The subjective nature of inflation expectations, as discussed by [Weber et al. \(2022\)](#) and [Weber \(2022\)](#), points to a consistent upward bias and strong influence from

prices of specific goods. These findings suggest that central bank communication strategies need to be carefully crafted to effectively manage public inflation expectations, considering the diverse ways in which different economic agents interpret and react to inflation-related information.

As we can see by the review of current literature, the intersection of AI and inflation expectation formation presents a promising avenue for future research. In this domain, my contribution is to expand the understanding of how AI-driven models can enhance the prediction and analysis of inflation expectations. By integrating large language models with economic data, my research aims to understand how these models make their decisions and also how human subjects react to their decisions.

### III. Natural Language Processing and Transformers

Scientists have always been curious about how human language works and have wanted to understand it better. They have attempted to create machines that can understand and use language just as humans do. In the 1980s and 1990s, they focused on basic models and then more advanced models like recurrent neural networks (RNNs), which predicted the next word in a sentence to find fundamental grammar and semantics (Schmidhuber, 2022).

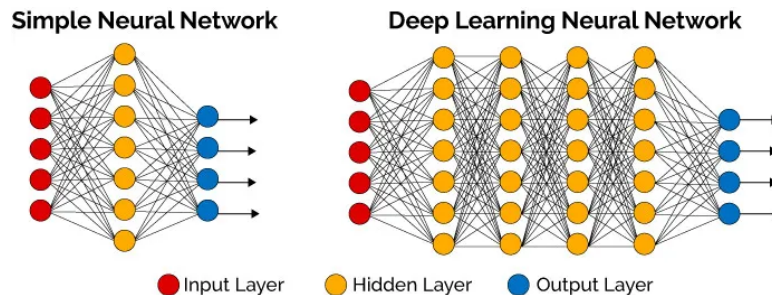


FIGURE 1. ILLUSTRATION OF NEURAL NETWORK AND DEEP LEARNING NETWORK

The transition to deep learning created a significant shift in language processing. Unlike simpler neural networks, deep learning models feature multiple hidden layers<sup>2</sup>(Figure 1), leading to higher accuracy in tasks. However, the introduction of the Transformer architecture in 2017 was a revolutionary development in the field of NLP (Vaswani et al., 2017). Its self-attention mechanism<sup>3</sup> significantly enhanced language models, enabling the capture of long-range de-

<sup>2</sup>Hidden layers in a neural network are layers of mathematical functions each designed to process certain aspects of the input data. These layers are not directly exposed to the input or output, but they play a crucial role in helping the network learn complex patterns.

<sup>3</sup>In transformers, the self-attention mechanism allows the model to analyze different parts of the input sequence independently and compute attention scores. These scores determine how much focus to put on other parts of the input when processing a specific part, enabling the model to dynamically prioritize which information is most relevant during tasks.

dependencies and the parallelization of computations. Models like BERT<sup>4</sup> and GPT demonstrated the profound capabilities of Transformers in language understanding, performing tasks such as translation, summarization, sentiment analysis, and text generation. These models analyze extensive text corpora to identify patterns, thereby enhancing their language processing abilities. Figure 2 shows the general architecture of a transformer network.

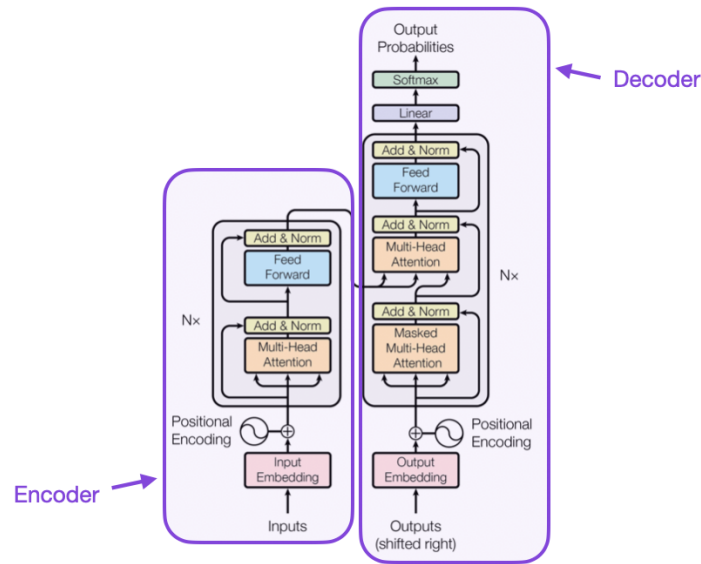


FIGURE 2. GENERAL TRANSFORMERS ARCHITECTURE

For example, GPT models as auto-regressive language models calculate the probability of a word sequence  $W = w_1, w_2, \dots, w_n$  as follows:

$$(1) \quad P(W) = \prod_{i=1}^n P(w_i | w_1, w_2, \dots, w_{i-1})$$

These improvements have made it easier for computers to understand and use human language, making our conversations with them feel more natural. However, training these advanced computer programs requires a lot of computer power and specific data, which can be limiting. To solve these problems, there are two main solutions: fine-tuning and retrieval-augmented generation

<sup>4</sup>Bidirectional Encoder Representations from Transformers

(RAG). Fine-tuning adjusts a pre-trained program to do a new task better, while RAG helps the program pull in extra information from outside sources when it needs to.

#### A. Retrieval Augmented Generation vs. Fine Tuning

Fine-tuning involves further training the pre-trained language model on a smaller, task-specific dataset. This allows the model to specialize in a particular domain or task, using the knowledge acquired during pre-training. Fine-tuning LLMs for specific topics, like economics, involves adjusting their 'thinking process' and it changes the model's internal weights to reduce errors in predictions.

$$(2) \quad \min_{\theta} \sum_{(x,y) \in D_{econ}} L(f(x; \theta), y)$$

On the other hand, RAG introduces a new approach by dynamically augmenting the model's capabilities with information retrieved from an external knowledge source. During the generation process, relevant information is retrieved based on the current context and integrated into the model's decision-making process. This method not only gives the model's access to information beyond its initial training data but also enhances its ability to generate informed and accurate outputs. The RAG process can be showed as follows (Gao et al., 2023):

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

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 3, 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.

TABLE 1—RAG VS FINE-TUNING. BALAGUER ET AL. (2024); GAO ET AL. (2024)

Model	RAG	Fine-tuning
Cost – input token size	Increased Prompt Size	Minimal
Cost – output token size	Potentially High	Precise and Potentially Lower
Initial cost	Low – creating embeddings	High – fine-tuning
Accuracy	Effective	Effective
New Knowledge	New Domain	New skill in domain
Data Processing	Minimal	Needs high-quality datasets
Interpretability	Traceable to sources	Often like a black box

In this study, RAG has several advantages over fine-tuning. It allows for the integration of new knowledge, ensuring that the model's responses are more

accurate. Additionally, the ability to trace the source of the retrieved information enhances the model’s interpretability (Table 1). Also, in many models like GPT-4, there are a lot of restrictions on AI Hallucinations<sup>5</sup>, which prevents the model from creating a response if it does not have relevant information.

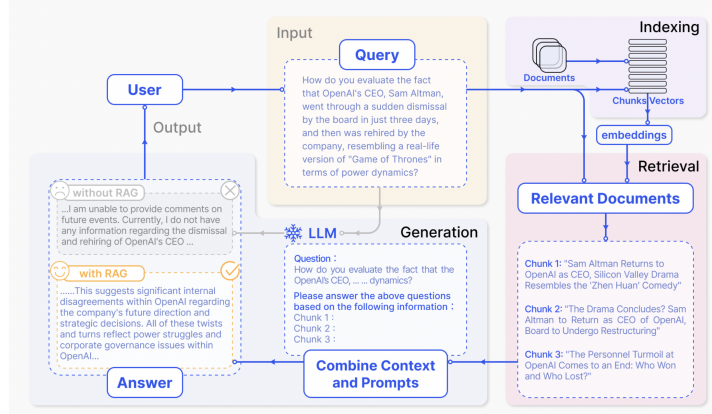


FIGURE 3. AN ILLUSTRATION OF HOW RAG WORKS GAO ET AL. (2023)

### B. Approximate Nearest Neighborhood

In the domain of large-scale data retrieval, Approximate Nearest Neighbor (ANN) is a technique that enables efficient information retrieval by approximating the closest matches to a query within large datasets. ANN algorithms aim to minimize the distance between the query point and the data points in the dataset, according to a specific metric, typically in a high-dimensional space. ANN is one of the main pillars of a Retrieval-Augmented Generation (RAG) pipeline when interacting with Large Language Models (LLMs), providing a mechanism for finding relevant information in handling and processing vast amounts of data. Generally speaking, every ANN search is captured by the inequality (Liu et al., 2004):

$$\min_{z \in D} \text{dist}(x, z) \leq \text{dist}(x, y) \leq C \cdot \min_{z \in D} \text{dist}(x, z),$$

where  $x$  represents the query point,  $y$  is the approximate nearest neighbor found by the algorithm,  $D$  is the dataset,  $\text{dist}(\cdot, \cdot)$  denotes the distance metric, and  $C > 1$  is the approximation factor.

Several algorithm and libraries are available that facilitate Approximate Nearest Neighbor (ANN) searches, each with different features. Here are the one that I used in this part:

<sup>5</sup>AI hallucinations occur when a model generates output that is uncorrelated with or unsupported by its training data or input, leading to inaccurate or wrong responses.



- **ANNOY (Approximate Nearest Neighbors Oh Yeah)**: This library employs trees constructed from random projections of data points to facilitate fast searches, enhancing the speed of data retrieval.
- **FAISS (Facebook AI Similarity Search)**: Developed by Facebook AI Research, FAISS uses quantization techniques and GPU acceleration to efficiently perform similarity search and clustering of dense vectors.
- **HSNW (Hierarchical Navigable Small World)**: HSNW constructs a graph that leverages the small-world structure inherent in high-dimensional data to ensure faster search times.

In the context of this research, the default RAG pipeline provided by OpenAI is used, which utilizes its own model for determining relevant information. The choice of ANN technique plays a crucial role in the RAG framework, as it directly influences the efficiency and effectiveness of the retrieval process.

To demonstrate how different models and methods highlight various aspects of the text that needs to be retrieved, I used eight embedding and tokenizing methods to find the nearest neighbors of the word "inflation" with three different approximate nearest method algorithm<sup>6</sup>. This practice illustrates the diversity in how these methods identify relevant information. As mentioned before in the survey experiment I do not define the method and Open AI uses its own method of retrieving knowledge. This is only a practice to show some aspect of the relationship between different part of the text to an example term like inflation.

The process in finding the approximate nearest neighbor is structured as follows:

- 1) Each set of meeting minutes is converted to text, and then all stopwords are removed.
- 2) A vector of tokens is created from the pre-processed text using different embedding models<sup>7</sup>.
- 3) Using three different embedding models—each with approximately 110 million parameters—a 768-dimensional vector representing each token in the document is created.
- 4) Three different indexing methods, based on three approximate neighborhood techniques, are used to find words that are contextually closest to the term "inflation expectation".
- 5) The vector is then unembedded to retrieve the actual words.

<sup>6</sup>I also employed two more traditional methods, Word2Vec and GloVe. The first one uses something called a context vector, but it still doesn't capture every aspect of the document. The second one utilizes the co-occurrence of different words to find similarity. The words in these two methods are: Word2Vec: [Inflation, inflationary, CPI, deflation, disinflation, GDP, deflationary, disinflationary, Inflationary, inflations] and GloVe: [rate, rates, rise, slowing, inflationary, unemployment, rising, growth, recession, fed]. These results show that the contextual understanding of these methods has some disadvantages, indicating the need to transition to more advanced models like transformers.

<sup>7</sup>These models 'gpt2', 'gpt2-medium', 'gpt2-large', 'gpt2-xl', 'roberta-base', 'roberta-large', 'bert-base-uncased', 'distilbert-base-uncased', 'bart-base', 'bart-large', 'bert-large-uncased', 'xlnet-base-cased' and 't5-base'. A short introduction to these models is provided in Appendix A.A3.

Figure 4 shows how different methods of finding nearest neighbor and also embedding models could find potentially different part of a text more close to some words. However, that are a lot of similarity as well. Interestingly, the world "deflation" coming up more in years that was and issue and not in recent years.

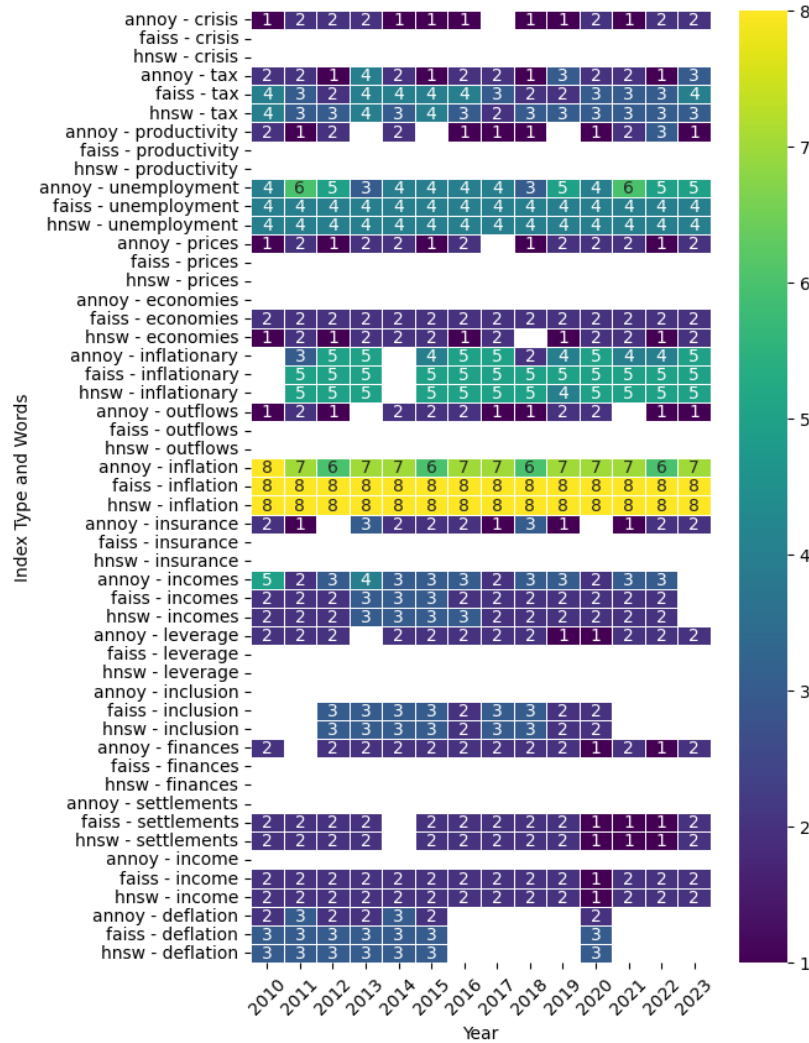


FIGURE 4. COMPARISON OF NEAREST NEIGHBORS FOR "INFLATION"

#### IV. Experimental Design

In the survey experiment, I conducted a randomized survey experiment on GPT-4 Turbo<sup>8</sup> with retrieval API assistants<sup>9</sup>, aiming to assess the influence of different knowledge domains on inflation expectation formation and also see the impact of information treatment. To achieve this, as we can see in Table 2 I categorized the knowledge domains into five distinct types placebo (P), using general information from Wikipedia texts as a baseline; economic reports of the president (E), providing a broad view of the US economic status; FOMC minutes from the year 2018 (M0) and FOMC minutes from the year 2023 (M1), to compare the difference between old and recent monetary policy status; and FOMC minutes from 2020 to 2023 (M2), offering a wider range of data on monetary policy over the past years. This structure allows me to dissect the impact of varying levels and types of economic knowledge on agents' expectations.

TABLE 2—KNOWLEDGE DOMAINS IN SURVEY EXPERIMENT

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

In the experiment, I first asked AI agents a series of questions to understand their initial economic expectations. These questions were based on the Survey of Consumer Expectations (SCE) conducted by the Federal Reserve Bank of New York, covering topics such as unemployment rates, interest rates on savings accounts, stock prices, and inflation over various periods. After these initial questions, the AI agents were randomly assigned into one of six groups. The first group acted as a control, receiving no additional information, while the remaining five groups each were provided with some pieces of information as new context to consider for their next questions. This setup tries to examine how specific information might influence the AI agents' economic outlooks. The list of treatments including are provided in Table 3. The initial questions and follow-up questions are shown in Table 4.

<sup>8</sup>GPT-4 Turbo is an upgrade to GPT-4 that's better at handling long texts and following detailed instructions, making it great for tasks like finding and using information in large documents. The cut-off time for it's training set of it is Dec 2023.

<sup>9</sup>OpenAI's API assistants are a feature that lets us create specialized GPTs with different capabilities, including document retrieval.

TABLE 3—TREATMENT GROUPS

Group	Title	Information
T.1	Control group	No additional information
T.2	Placebo group	Population of the U.S. grew by 1.2% between 2020 and 2022.
T.3	Current rate	FFR The interest rate set by the Federal Reserve, known as the Federal Funds Rate, is currently at 5.25%.
T.4	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 3.2%.
T.5	Current Inflation + 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 3.2%. One forecast at the Federal Reserve is that this inflation rate will be 2.8% on average over the next few years and in the longer-run.
T.6	Current fixed-rate 30-year mortgage	The current average rate for fixed-rate 30-year mortgage is 7.5% per year.

For instance, I assigned different groups specific types of economic data before reassessing their expectations. For example, one group was informed about the current Federal Funds Rate set by the Federal Reserve, while another received detailed information on both current inflation and future forecasts. After providing these details, I again asked the groups expectation-based questions to gauge how the provided information influenced their views on the economy. This experimental design helps to determine not only the direct effects of specific economic data on expectations but also how varying types of information shape agents' perceptions of economic conditions. Visually, the structure of the experiment can be shown in Figure 5.

TABLE 4—SURVEY QUESTIONS

Q_id	Question
Q1.I	Last 12 month Point Estimate (Pre-Treatment)
Q2.I	Next 12 month Distribution Estimate (Pre-Treatment)
Q3.I	Longer-run Distribution Estimate (Pre-Treatment)
Q1.F	Next 12 month Point Estimate (Post-Treatment)
Q2.F	Longer-run Point Estimate (Post-Treatment)

Note: "I" indicates initial questions and "F" indicates follow-up questions. The full table of the questions is presented at Appendix A.A4.

In this version of the experiment, that I call it **Run\_V1**, I surveyed all six assistants (without any followup or information treatments) to assess their comprehension and approach to different knowledge sets. Meanwhile, for the information treatment, I only used **Assistant M2** that has the FOMC minutes data over the past four years.

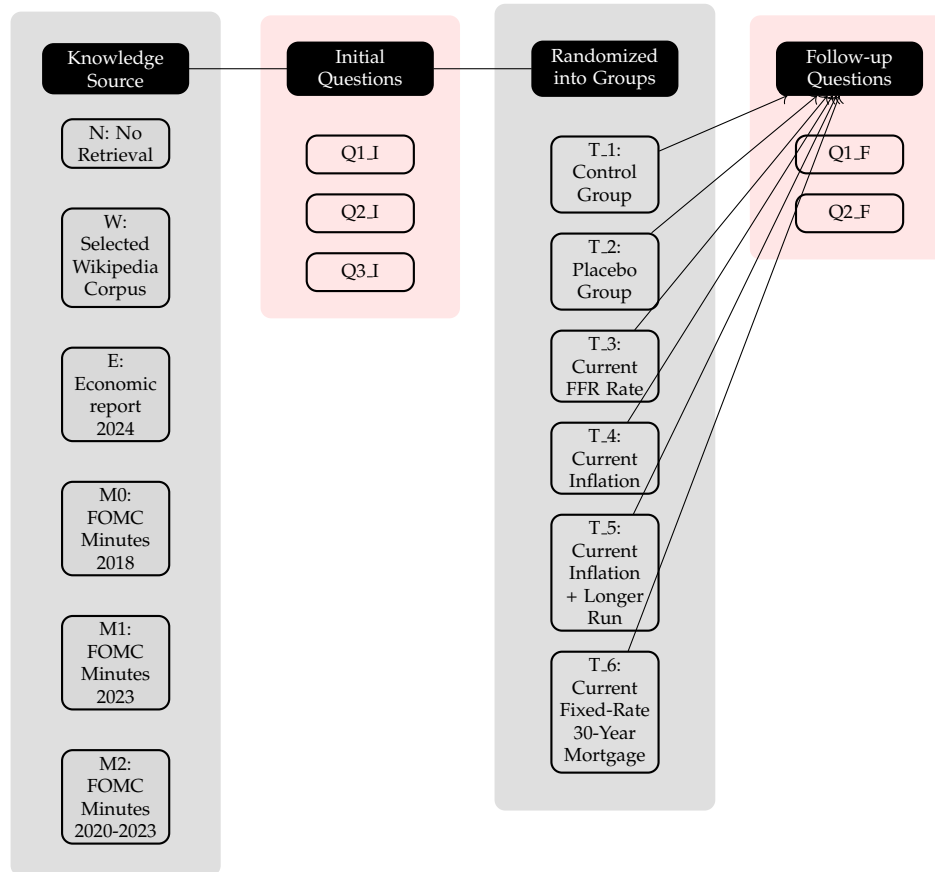


FIGURE 5. AN ILLUSTRATION OF THE STAGES OF THE EXPERIMENT

## V. Results

### A. Preliminary Observations

This experiment consisted of two phases. The first phase is the no-information treatment. In this phase, I had 600 subjects across 6 categories of assistants. Naturally, because I didn't have information treatment, there are only 3 main questions. The first is the inflation for the past 12 months in point estimate, the second is the distribution estimate of inflation for the next 12 months, and the third is the inflation for the period between 2 years from now and 3 years from now<sup>10</sup>.

Table 5 shows the summary statistics of these three questions. For question 1, I report the point estimate, and for questions 2 and 3, I used the midpoint formula to calculate the mean inflation in two ways. One considering the values the subjects put for deflation and one without considering the deflation values and adding them to the lowest inflation bin.

TABLE 5—SUMMARY STATISTICS FOR NO INFORMATION TREATMENT

Deflation	Assistant	$\pi_{t-1} : \pi_{t_0}$				$\pi_{t_0} : \pi_{t_1}$				$\pi_{t_2} : \pi_{t_3}$			
		min	max	mean	std	min	max	mean	std	min	max	mean	std
Yes	N	2	3.5	2.43	0.26	3.24	7.5	5.31	1	2.49	7.3	4.68	1.08
	W	0	11.58	2.72	2.4	0.94	14	5.96	2.27	1.7	14	5.17	1.92
	E	0	7.1	3.01	1.29	-1	14	4.93	2.2	-2.25	11	4.84	1.98
	M0	1.7	3.6	2.07	0.21	1.08	14	4.36	1.9	1.2	10.12	4.54	1.8
	M1	2	8	3.85	1.1	1	14	4.67	2.37	1.09	14	4.55	2.23
	M2	1.7	6.85	3.86	0.86	1.55	8	4.72	1.5	1.32	11.66	4.97	1.8
	All	0	11.58	2.99	1.40	-1	14	4.99	2	-2.25	14	4.79	1.84
No	N	2	3.5	2.43	0.26	3.24	7.5	5.34	0.97	2.79	7.3	4.84	0.97
	W	0	11.58	2.72	2.4	1.04	14	5.99	2.23	1.7	14	5.2	1.9
	E	0	7.1	3.01	1.29	1	14	5.07	2.06	1.1	11	4.95	1.84
	M0	1.7	3.6	2.07	0.21	1.1	14	4.41	1.87	1.2	10.12	4.58	1.78
	M1	2	8	3.85	1.1	1.09	14	4.68	2.37	1.09	14	4.58	2.23
	M2	1.7	6.85	3.86	0.86	1.32	11.66	4.72	1.5	1.32	11.66	4.99	1.79
	All	0	11.58	2.99	1.40	1	14	5.04	1.96	1.08	14	4.86	1.80

Figure 6 shows the response to the question about past year inflation per assistants. As expected there is variation within each assistant. Also, Figure 8 represents the response to the distribution estimate questions. As expected deflation in the short-run and long-run doesn't have that much frequency. In the short run the bin with 4 and 8 has the highest share and in the longer run the bin with inflation between 2 and 4 has the highest share. We can also look at the distribution of next year (short-run) inflation expectation that is calculated by

<sup>10</sup>The full list of questions is provided in Appendix A.A4

midpoint formula in Figure 8. As expected assistant W has the highest variance and assistant M2 has the one of the lowest variance.

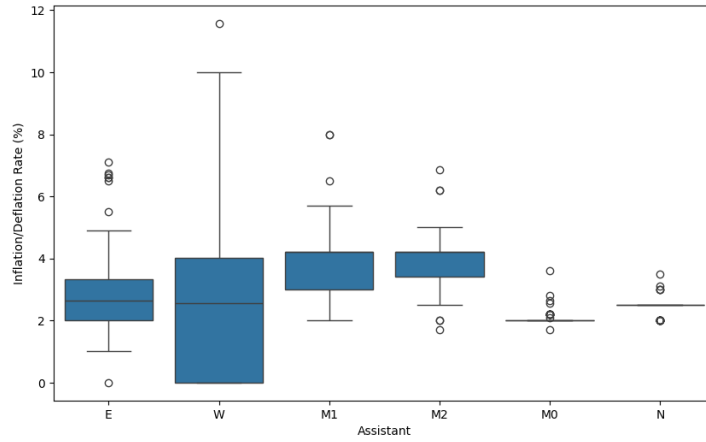


FIGURE 6. INFLATION/DEFLATION RATE OVER THE LAST 12 MONTHS BY ASSISTANT

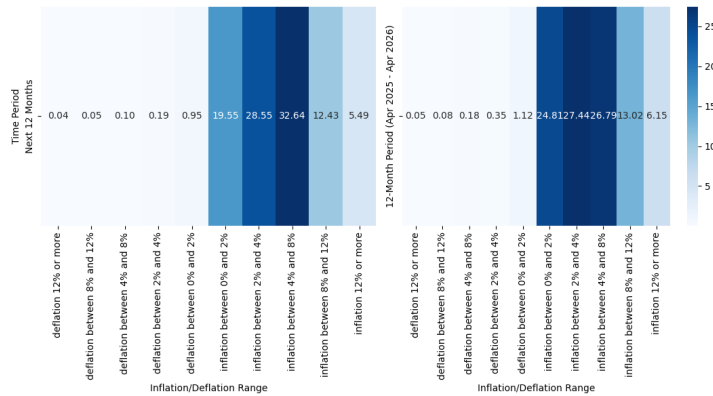


FIGURE 7. AVERAGE PROBABILITY DISTRIBUTION FOR ALL SUBJECTS

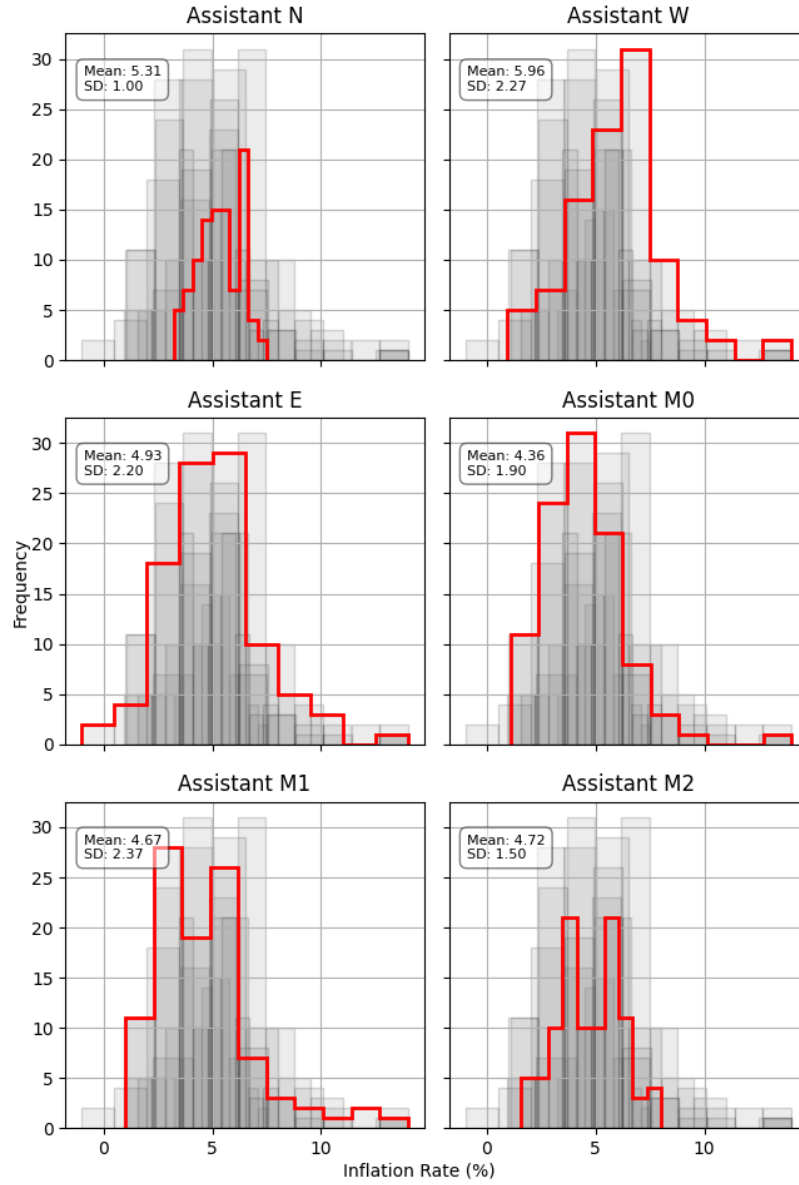


FIGURE 8. DISTRIBUTION OF MEAN OF NEXT 12 MONTH INFLATION EXPECTATION



In the second phase of the experiment that was run separately, I have used two different assistants: one without retrieval power (N) and the other with 4 years of information(M2). All the processes of the experiment, as shown in Figure 5, were implemented for these subjects. As a result, we have 1200 subjects divided by 600 for each assistant and also 100 for each treatment.

Table 6 and Table 8 have the details about the mean and standard deviation of control group and the other five treatment groups for Assistant N and M2.

TABLE 6—DESCRIPTIVE STATISTICS OF ASSISTANT N

	T1		T2		T3	
	Mean	SD	Mean	SD	Mean	SD
Pre-Treatment						
Perceived inflation, previous 12 months	2.40	0.27	2.45	0.50	2.41	0.25
Expected inflation, 12-month ahead	5.60	0.98	5.38	1.02	5.51	0.96
Expected inflation, longer period	4.85	1.08	4.94	1.07	4.94	1.03
Post-Treatment						
Expected inflation, 12-month ahead	3.35	0.24	3.40	0.20	3.30	0.26
Expected inflation, longer period	3.89	0.55	3.72	0.47	3.76	0.39
	T4		T5		T6	
	Mean	SD	Mean	SD	Mean	SD
Pre-Treatment						
Perceived inflation, previous 12 months	2.46	0.29	2.42	0.21	2.49	0.41
Expected inflation, 12-month ahead	5.47	1.14	5.20	1.02	3.36	0.25
Expected inflation, longer period	5.03	1.16	5.14	1.18	3.88	0.47
Post-Treatment						
Expected inflation, 12-month ahead	3.38	0.22	2.80	0.00	5.50	1.02
Expected inflation, longer period	3.66	0.41	2.80	0.00	5.31	1.32

To see the comparison of short-run and long-run inflation expectations across different treatments for two assistants we can look at Figure 9. For both short-run and long-run expectations, the prior values (blue bars) remain relatively consistent across treatments for each assistant type. However, the posterior values (orange bars) show more variation, suggesting that the treatments have an effect on updating inflation expectations. The differences between prior and posterior values appear more pronounced for the M2 assistant compared to the N assistant, particularly for treatments T.2, and T.4 in the short-run expectations.

TABLE 7—DESCRIPTIVE STATISTICS OF ASSISTANT M2

	T1		T2		T3	
	Mean	SD	Mean	SD	Mean	SD
<b>Pre-Treatment</b>						
Perceived inflation, previous 12 months	5.10	1.73	5.17	1.56	4.88	1.69
Expected inflation, 12-month ahead	4.88	1.54	5.14	1.79	5.08	1.90
Expected inflation, longer period	4.97	2.06	4.76	1.93	4.72	2.02
<b>Post-Treatment</b>						
Expected inflation, 12-month ahead	3.29	0.80	3.61	0.62	3.57	0.63
Expected inflation, longer period	3.51	0.82	4.19	0.90	3.85	0.77

	T4		T5		T6	
	Mean	SD	Mean	SD	Mean	SD
<b>Pre-Treatment</b>						
Perceived inflation, previous 12 months	5.10	1.57	5.16	1.61	5.01	1.69
Expected inflation, 12-month ahead	5.07	2.00	5.01	2.11	4.91	1.88
Expected inflation, longer period	4.34	1.47	4.24	1.61	4.54	2.07
<b>Post-Treatment</b>						
Expected inflation, 12-month ahead	3.33	0.21	2.80	0.02	4.03	0.92
Expected inflation, longer period	3.74	0.60	2.80	0.00	4.01	0.93

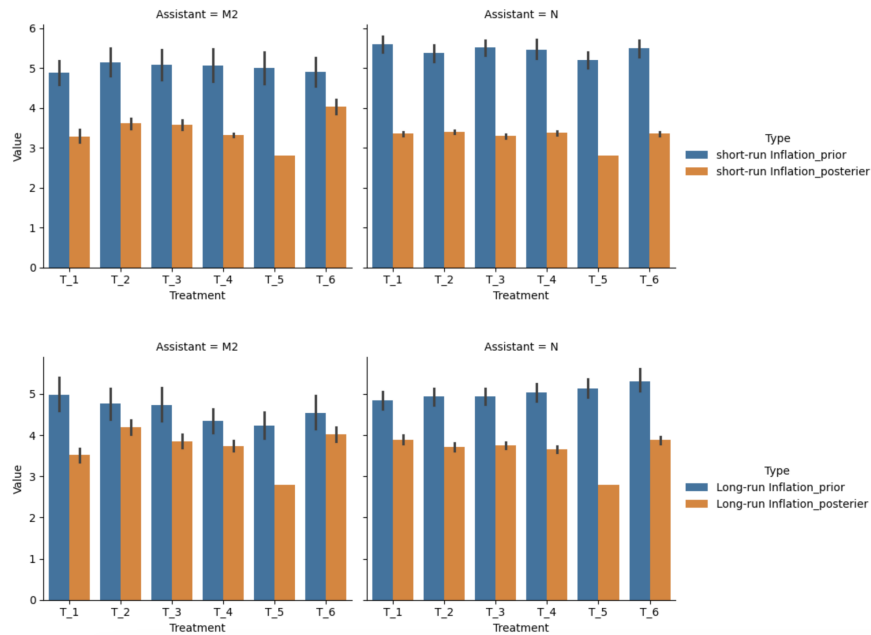


FIGURE 9. COMPARISON OF PRIOR AND POSTERIOR INFLATION EXPECTATIONS

### B. Empirical Model

To empirically assess the impact of information treatments on AI agents' economic expectations, I follow [Coibion, Gorodnichenko and Weber \(2022\)](#); [Coibion et al. \(2023\)](#) specification that allows to capture the impact of new information. The model is formulated as follows:

$$(3) \quad X_j^{\text{post}} = \alpha X_j^{\text{pre}} + \sum_{k=2}^6 \beta_k \text{Treatment}_j^{(k)} + \sum_{k=2}^6 \gamma_k \text{Treatment}_j^{(k)} \times X_j^{\text{pre}} + \mathbf{W}_j \Psi + \epsilon_j$$

Here,  $X_j^{\text{post}}$  and  $X_j^{\text{pre}}$  represent the AI agent  $j$ 's expectations after and before receiving the treatment, respectively.  $\text{Treatment}_j^{(k)}$  indicates whether AI agent  $j$  was subjected to treatment  $k$ . The coefficients  $\beta_k$  and  $\gamma_k$  measure the level and interaction effects of the treatments on the expectations, respectively, with  $\gamma_k$  expected to be negative, indicating that more informative treatments lead agents to revise their expectations more significantly away from their prior beliefs. The vector  $\mathbf{W}_j$  includes control variables relevant to each agent, and  $\Psi$  denotes the parameters associated with these controls, with  $\epsilon_j$  capturing the error term<sup>11</sup>.

This specification helps us understand how AI agents integrate new information with their pre-existing beliefs, quantifying the adjustments made in their economic outlook in response to different types of information treatments ([Coibion, Gorodnichenko and Kumar, 2018](#)).

### C. Analysis

To evaluate the effects of the different information treatments, I estimate the specification in Equation (3) for both assistants N and M2. Table 8 and Table 9 shows the regression results for short-run (next 12 months) and long-run (2-3 years ahead) inflation expectations. Focusing first on the coefficients on the prior expectations ( $\alpha$ ), we see that for the no-retrieval assistant N, the weight placed on priors is quite high at around 0.8 for short-run expectations and 0.9 for long-run expectations. This indicates that in the absence of new information, assistant N exhibits persistent expectations that change little between time periods. In contrast, for the M2 assistant with access to FOMC minutes, the weight on priors is much lower at 0.4-0.5, suggesting that this assistant naturally revises its expectations more based on the core information it has access to.

Turning to the information treatment effects, I find significant differences across information types. The placebo treatment T2 has little impact on expectations for both assistants. Providing information about just the current federal funds rate

<sup>11</sup>In the the first phase of this experiment I only run the queries using a no persona approach and as a result I do not have covariance that are specific for each agent. In the next phase I will create a sample of AI agent that mimics the actual demographic features Survey of Consumer Expectations (SCE) conducted by the Federal Reserve Bank of New York

in T3 generates a small downward revision in short-run inflation expectations for N but no effect on M2. However, when the treatments involve not just the current FFR but also forward guidance about its future path, we see much larger effects. For example, treatments T4 (current inflation rate) and T5 (current and longer-run inflation forecasts) lead to large revisions in expectations for both assistants and horizons. The treatment providing information about current mortgage rates (T6) also has sizable effects, particularly for longer-run expectations which is compatible to [Coibion et al. \(2023\)](#).

TABLE 8—REGRESSION RESULTS FOR ASSISTANT N

Variable	Coef. (Significance)	Std Err	t	P>  t
<b>Short-Run N (R-squared: 0.494, N: 600)</b>				
const	3.55***	0.13	28.48	0.00
Treatment.T.2	-0.04	0.17	-0.24	0.81
Treatment.T.3	-0.26	0.18	-1.47	0.14
Treatment.T.4	-0.20	0.16	-1.24	0.22
Treatment.T.5	-0.75***	0.17	-4.49	0.00
Treatment.T.6	-0.27	0.17	-1.56	0.12
X_pre	-0.04	0.02	-1.60	0.11
Treatment.T.2.X_pre	0.01	0.03	0.45	0.65
Treatment.T.3.X_pre	0.04	0.03	1.17	0.24
Treatment.T.4.X_pre	0.04	0.03	1.38	0.17
Treatment.T.5.X_pre	0.04	0.03	1.16	0.25
Treatment.T.6.X_pre	0.05	0.03	1.60	0.11
<b>Long-Run N (R-squared: 0.631, N: 600)</b>				
const	2.15***	0.16	13.47	0.00
Treatment.T.2	0.16	0.23	0.71	0.48
Treatment.T.3	0.60**	0.23	2.59	0.01
Treatment.T.4	0.81***	0.22	3.65	0.00
Treatment.T.5	0.65***	0.22	2.90	0.00
Treatment.T.6	0.91***	0.22	4.24	0.00
X_pre	0.36***	0.03	11.13	0.00
Treatment.T.2.X_pre	-0.07	0.05	-1.63	0.10
Treatment.T.3.X_pre	-0.16***	0.05	-3.34	0.00
Treatment.T.4.X_pre	-0.22***	0.04	-5.03	0.00
Treatment.T.5.X_pre	-0.36***	0.04	-8.21	0.00
Treatment.T.6.X_pre	-0.20***	0.04	-4.92	0.00

Notes: \*\*\*, \*\*, and \* denote statistical significance at 1%, 5%, and 10% levels, respectively.

TABLE 9—REGRESSION RESULTS FOR ASSISTANT M2

Variable	Coef. (Significance)	Std Err	t	P>  t
<b>Short-Run M2 (R-squared: 0.270, N: 600)</b>				
const	3.53***	0.21	17.02	0.00
Treatment.T.2	0.01	0.28	0.04	0.97
Treatment.T.3	0.13	0.27	0.46	0.64
Treatment.T.4	-0.12	0.27	-0.45	0.65
Treatment.T.5	-0.73*	0.26	-2.79	0.01
Treatment.T.6	0.49	0.27	1.81	0.07
X_pre	-0.05	0.04	-1.26	0.21
Treatment.T.2.X_pre	0.06	0.05	1.19	0.24
Treatment.T.3.X_pre	0.03	0.05	0.65	0.52
Treatment.T.4.X_pre	0.03	0.05	0.66	0.51
Treatment.T.5.X_pre	0.05	0.05	1.03	0.31
Treatment.T.6.X_pre	0.05	0.05	0.99	0.32
<b>Long-Run M2 (R-squared: 0.540, N: 600)</b>				
const	3.05***	0.16	19.64	0.00
Treatment.T.2	-0.62**	0.22	-2.80	0.01
Treatment.T.3	-0.26	0.22	-1.18	0.24
Treatment.T.4	-0.49*	0.24	-2.02	0.04
Treatment.T.5	-0.25	0.23	-1.10	0.27
Treatment.T.6	-0.27	0.21	-1.29	0.20
X_pre	0.09***	0.03	3.22	0.00
Treatment.T.2.X_pre	0.28***	0.04	6.56	0.00
Treatment.T.3.X_pre	0.13***	0.04	3.18	0.00
Treatment.T.4.X_pre	0.18***	0.05	3.57	0.00
Treatment.T.5.X_pre	-0.09*	0.05	-1.98	0.05
Treatment.T.6.X_pre	0.18***	0.04	4.40	0.00

Notes: \*\*\*, \*\*, and \* denote statistical significance at 1%, 5%, and 10% levels, respectively.

## VI. Discussion and Concluding Remarks

This study presents a novel approach to investigating the formation of inflation expectations by integrating Large Language Models with economic and text data. One of the key findings is the impact of the type of knowledge source on the variation in responses. The results indicate that LLMs with access to relevant economic data, such as the Federal Open Market Committee minutes, exhibit lower variance in their inflation expectations compared to those with access to irrelevant information. This highlights the importance of providing LLMs with domain-specific knowledge to enhance the accuracy and consistency of their responses.

Furthermore, the information treatment regarding mortgage rates had the most significant impact on updating prior inflation expectations for LLMs, as evidenced by previous studies for human subjects. This finding underscores the potential effectiveness of communicating market-relevant information, such as mortgage rates, in shaping economic expectations and decision-making processes.

In future directions, it would be valuable to extend this research by incorporating a wider range of LLMs and exploring the impact of persona for AI agents. For instance, introducing demographic characteristics to the LLMs could provide insights into how different segments of the population might interpret and respond to economic information. Additionally, investigating the performance of LLMs in comparison to human subjects under similar experimental conditions could further show the strengths, limitations, and potentials of these AI models in economic forecasting.

Overall, this study contributes to the growing body of literature on the intersection of generative AI and economics, offering practical implications for the future of AI-assisted economic forecasting and its potential impact on policy-making and individual financial decisions.

## REFERENCES

- Akata, Elif, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz.** 2023. "Playing repeated games with Large Language Models." *arXiv preprint arXiv:2305.16867*.
- Balaguer, Angels, Vinamra Benara, Renato Luiz de Freitas Cunha, Roberto de M. Estevão Filho, Todd Hendry, Daniel Holstein, Jennifer Marsman, Nick Mecklenburg, Sara Malvar, Leonardo O. Nunes, Rafael Padilha, Morris Sharp, Bruno Silva, Swati Sharma, Vijay Aski, and Ranveer Chandra.** 2024. "RAG vs Fine-tuning: Pipelines, Tradeoffs, and a Case Study on Agriculture."
- Brookins, Philip, and Jason Matthew DeBacker.** 2023. "Playing games with GPT: What can we learn about a large language model from canonical strategic games?" *Available at SSRN 4493398*.
- Bybee, Leland.** 2023. "The Ghost in the Machine: Generating Beliefs with Large Language Models." *arXiv preprint arXiv:2305.02823*.
- Candia, Bernardo, Olivier Coibion, and Yuriy Gorodnichenko.** 2020. "Communication and the beliefs of economic agents."
- Candia, Bernardo, Olivier Coibion, and Yuriy Gorodnichenko.** 2021. "The Inflation Expectations of US Firms: Evidence from a new survey."
- Charness, Gary, Brian Jabarian, and John A List.** 2023. "Generation next: Experimentation with ai."
- Coibion, Olivier, and Yuriy Gorodnichenko.** 2015. "Information rigidity and the expectations formation process: A simple framework and new facts." *American Economic Review*, 105(8): 2644–2678.
- Coibion, Olivier, Dimitris Georgarakos, Yuriy Gorodnichenko, and Michael Weber.** 2023. "Forward guidance and household expectations." *Journal of the European Economic Association*, 2131–2171.
- Coibion, Olivier, Yuriy Gorodnichenko, and Michael Weber.** 2022. "Monetary policy communications and their effects on household inflation expectations." *Journal of Political Economy*, 130(6): 1537–1584.
- Coibion, Olivier, Yuriy Gorodnichenko, and Saten Kumar.** 2018. "How do firms form their expectations? New survey evidence." *American Economic Review*, 108(9): 2671–2713.
- Coibion, Olivier, Yuriy Gorodnichenko, and Tiziano Ropele.** 2020. "Inflation expectations and firm decisions: New causal evidence." *The Quarterly Journal of Economics*, 135(1): 165–219.

- Coibion, Olivier, Yuriy Gorodnichenko, Saten Kumar, and Mathieu Pedemonte.** 2020. "Inflation expectations as a policy tool?" *Journal of International Economics*, 124: 103297–103297.
- D'Acunto, Francesco, Ulrike Malmendier, Juan Ospina, and Michael Weber.** 2021. "Exposure to grocery prices and inflation expectations." *Journal of Political Economy*, 129(5): 1615–1639.
- Dell'Acqua, Fabrizio, Edward McFowland, Ethan R Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Kraymer, François Candelon, and Karim R Lakhani.** 2023. "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality." *Harvard Business School Technology & Operations Mgt. Unit Working Paper*, (24-013).
- Faria-e Castro, Miguel, and Fernando Leibovici.** 2023. "Artificial Intelligence and Inflation Forecasts."
- Gao, Yunfan, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang.** 2023. "Retrieval-augmented generation for large language models: A survey." *arXiv preprint arXiv:2312.10997*.
- Gao, Yunfan, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang.** 2024. "Retrieval-Augmented Generation for Large Language Models: A Survey."
- Haslberger, Matthias, Jane Gingrich, and Jasmine Bhatia.** 2023. "No great equalizer: experimental evidence on AI in the UK labor market." *Available at SSRN*.
- Heydari, Babak, and Nunzio Lorè.** 2023. "Strategic Behavior of Large Language Models: Game Structure vs. Contextual Framing." *Contextual Framing (September 10, 2023)*.
- Horton, John J.** 2023. "Large language models as simulated economic agents: What can we learn from homo silicus?"
- Korinek, Anton.** 2023. "Generative AI for economic research: Use cases and implications for economists." *Journal of Economic Literature*, 61(4): 1281–1317.
- Li, Nian, Chen Gao, Yong Li, and Qingmin Liao.** 2023. "Large Language Model-Empowered Agents for Simulating Macroeconomic Activities." *arXiv preprint arXiv:2310.10436*.
- Liu, Ting, Andrew Moore, Ke Yang, and Alexander Gray.** 2004. "An investigation of practical approximate nearest neighbor algorithms." *Advances in neural information processing systems*, 17.



- Schmidhuber, Juergen.** 2022. "Annotated history of modern ai and deep learning." *arXiv preprint arXiv:2212.11279*.
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin.** 2017. "Attention is all you need." *Advances in neural information processing systems*, 30.
- Weber, Michael.** 2022. "Subjective inflation expectations of households." *Business Economics*, 57(4): 217–221.
- Weber, Michael, Francesco D'Acunto, Yuriy Gorodnichenko, and Olivier Coibion.** 2022. "The subjective inflation expectations of households and firms: Measurement, determinants, and implications." *Journal of Economic Perspectives*, 36(3): 157–184.

## APPENDIX

### A1. Data

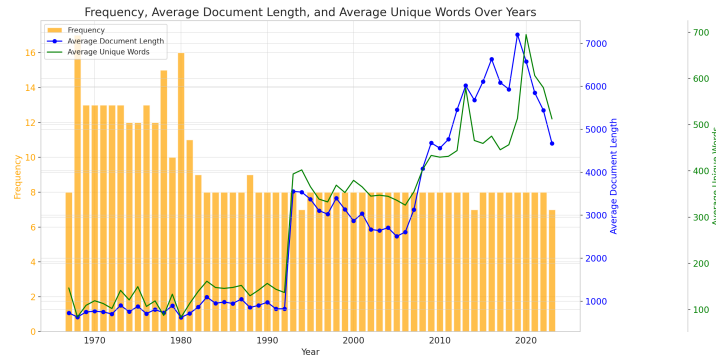
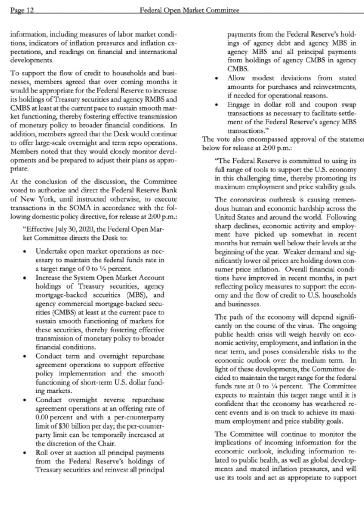


FIGURE A1. NUMBER OF MINUTES ISSUED BY FOMC, LENGTH AND UNIQUE WORDS PER DOCUMENT



2023—a resilient labor market and strong activity coupled with declining inflation—are consistent with a “soft landing” scenario.

But challenges remain. Elevated real interest rates compared with earlier during the pandemic—against the backdrop of a labor market that appears to have rebounded—could reduce investment in rate-sensitive sectors. In addition, the impact of geopolitical conflicts on markets and supply chains remains uncertain. To the extent that consumer attitudes respond to price levels rather than, or in addition to, inflation, consumer sentiment could remain weaker than economic data would predict, since prices are unlikely to broadly decline outright. However, recent real wage gains could potentially help support both confidence and consumer spending.

This chapter begins with a review of the economy in 2023. It first examines the acceleration in real GDP and its sources, and then surveys major labor market developments, highlighting their consistency with the “soft landing” scenario. Next, the chapter describes recent progress in disinflation. It then describes developments in financial markets, exploring both potential upside and downside risks. Finally, the chapter reviews the forecast underpinning the President’s Fiscal Year 2025 Budget and summarizes the near-term and long-term outlooks.

#### The Year in Review: The Continuing Recovery

This section describes the continued postpandemic recovery in 2023 and the easing of supply chain bottlenecks, explores the state of demand and supply rebalancing in the labor market, and provides updates on the progress of disinflation over the past year.

#### Output in 2023: A Return to Normal Growth

Real GDP accelerated to a pace of 3.1 percent over the four quarters of 2023, somewhat above the average growth of about 2.4 percent in the expansion period before the COVID-19 pandemic, and higher than the anemic 0.7 percent pace in 2022:Q4. Table 2-1 disaggregates real GDP growth into its major components.

FIGURE A2. AN SCREENSHOT OF PRESIDENTIAL ECONOMICS REPORT AND FOMC MINUTES DOCUMENTS

## A2. Prompts

**System Prompt**

The GPT is designed to answer users' questions about the expectations for future inflation based on their understanding of the provided documents. It responds with numerical percentage estimates, reflecting its opinions and anticipations about inflation while acknowledging the inherent uncertainty in such forecasts. The GPT uses the information from the text to derive its responses but does not use exact inflation targets from the document. It presents answers in two formats: single percentage points for point estimates and a list of percentages for distribution estimates. The GPT ensures responses are solely numerical and not any written statements with the alphabets and formatted accordingly: for point estimates, it uses [\_\_\_ %] and for distribution estimates, it uses [\_\_\_ %,\_\_\_ %,\_\_\_ %,\_\_\_ %,\_\_\_ %,\_\_\_ %,\_\_\_ %,\_\_\_ %,\_\_\_ %].

**RAG Prompt or Prompt before Prompt**

Do not use the exact inflation if mentioned in the document. 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:** "Over the last 12 months, what do you think the overall rate of inflation/deflation has been in the economy? The rate of inflation/deflation was [Your answer as a number] percent over the last 12 months. If you think there was inflation, please enter a positive number. If you think there was deflation, please enter a negative number. If you think there was neither inflation nor deflation, please enter zero."

**Additional Context:** "The interest rate set by the Federal Reserve, known as the Federal Funds Rate, is currently at 5.25%."

**Follow-up:** "Considering this additional information, what do you now expect the rate of inflation to be over the next 12 months?"

**An Example of a Bad Prompt** You are a helpful assistant. Read the document and then answer the question you were asked. Please provide your response in a concise manner.

### A3. Embedding Models for Approximate Nearest Neighbors

The embedding, tokenization, and unembedding models used in ANN are based Hugging Face<sup>12</sup>.

TABLE A1—SUMMARY OF MODELS AND THEIR PARAMETERS

Model	Description	Parameters
<b>BERT</b>	Bidirectional Encoder Representations from Transformers. Designed for a wide range of NLP tasks.	Small: 110M, Large: 340M
<b>GPT-2</b>	Generative Pre-trained Transformer 2. Uses Transformer architecture to generate text.	Small: 120M, Medium: 380M, Large: 812M
<b>RoBERTa</b>	Robustly Optimized BERT Pretraining Approach. Trains longer with larger batches and more data.	Small: 110M, Large: 340M
<b>XLNet</b>	Generalized autoregressive pre-training. Maximizes likelihood over all permutations of the factorization order.	Base: 110M

Note: The vector dimension for each token in the models varies according to size: The vector dimensions are 768 for 'small', 1024 for 'medium', and 1280 for 'large'.

<sup>12</sup>Hugging Face is a platform for machine learning, deep learning, and transformers applications that we load their models locally and then use it

## A4. Survey Questions

---

Q_id	Question
Q1.I	Over the last 12 months, what do you think the overall rate of inflation/deflation has been in the economy? The rate of inflation/deflation was [Your answer as a number] percent over the last 12 months. If you think there was inflation, please enter a positive number. If you think there was deflation, please enter a negative number. If you think there was neither inflation nor deflation, please enter zero.
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: [ ____ %, ____ %, ____ %, ____ %, ____ %, ____ %, ____ %, ____ %, ____ %, ____ %]
Q3.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, 12-month period between April 2025 (24 months from now) and April 2026 (36 months from now)... [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: [ ____ %, ____ %, ____ %, ____ %, ____ %, ____ %, ____ %, ____ %, ____ %, ____ %]
Q1.F	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 ____ %.
Q2.F	What do you expect the rate of inflation to be over that period? Please give your best guess. Over the 12-month period between April 2025 (24 months from now) and April 2026 (36 months from now), I expect the rate of inflation to be ____ %.

---

## A5. API Calls Details

Due to the prompt engineering technique used in the API calls, the rate of correct answers for all questions is high. In the first phase of the experiment with 600 AI subjects and 3 sets of questions, there were only 9 missing values for question Q1.I. For the other two questions, I received a full response.

There was another issue with Q1.I where some agents didn't answer in point estimates, but they attempted to answer in distributions. This occurred because the distribution type of questions was also specified in the system prompter. However, it didn't have an impact on the result. Many of the agents simply repeated their answers 10 times. For example, instead of answering [2.5%], they responded [2.5%,2.5%,2.5%,2.5%,2.5%,2.5%,2.5%,2.5%,2.5%,2.5%]. Some, however, responded with 10 different numbers, from which I used their average (92 out of 600). This behavior varied among different assistants; however, assistant N did a perfect job and followed the prompts perfectly (see Table A3).

TABLE A3—COUNT OF SINGLE AND MULTI VALUES, EQUAL AND NOT EQUAL (PHASE 1)

Assistant	Single Value Count	Multi Value Count	Equal	Not Equal	Total
E	89	11	91	9	100
M0	73	27	96	4	100
M1	53	47	76	24	100
M2	66	34	80	20	100
N	100	0	100	0	100
W	59	41	65	35	100
Total	440	160	508	92	600

The other very minor issue was about the summation of distribution estimates for Q2.I and Q3.I. There were only 5 observations for each question where the sum was not exactly equal to 100 percent, but it was very close. The standard deviation of the answers was 0.08 and 0.58 respectively. However, rounding some of the bins resolved this issue. For the second phased on the experiment, as shown in A4 there was less issues and it was resolved the same way as for phase one.

TABLE A4—COUNT OF SINGLE AND MULTI VALUES, EQUAL AND NOT EQUAL (PHASE 2)

Assistant	Single Value Count	Multi Value Count	Equal	Not Equal	Total
N	595	5	600	0	600
M2	462	138	551	49	600
Total	1057	143	1151	49	1200

In terms of the detail for API calls they are provided in Table ???. Each token is approximately 4-5 characters per English word.

Date	Context Tokens	Generated Tokens
14 Apr	8,160,544	127,574
15 Apr	12,213,625	209,672
Total API Requests: 13,249		
Total Tokens: 20,782,730		

TABLE A5—API REQUESTS AND TOTAL TOKENS

### A6. Miscellaneous Responses

The following table shows some of the responses generated by the AI agent with different prompts. It also displays the results for prompts that were not implemented well, leading to inaccurate responses.

Treatment	Assistant	Response	Question
T_1	M0	The document outlines that the medians of participants' projections for both total and core Personal Consumption Expenditures (PCE) price inflation are expected to be around [1.9%] in 2018 and anticipated to rise to [2.1%] by 2020. From the description, it appears that inflation expectations are gradually increasing but remaining close to the Federal Reserve's target of around 2%. Based on these data points and trends reflected up to 2020, a reasonable estimate of the expected rate of inflation over the next 12 months could be projected within a similar range, around [2.0%]. This provides a baseline expectation given the historical data trends and projected figures available in the document.	Q1.F
T_1	W	Projecting the rate of inflation specifically between April 2025 and April 2026 requires considering potential long-term economic policies, global market trends, and other macroeconomic factors, such as energy prices or geopolitical stability. In the absence of specific forecasting data from within the document and under normal economic conditions, central bank inflation targets provide a useful reference. Central banks often aim for an ...	Q2.F
T_6	E	The information provided does not offer a direct basis from which to infer future inflation rates. Consequently, without specific economic analyses, projections or other related content, predicting the rate of inflation over the next 12 months based on the given mortgage rate alone is not feasible.	Q1.F
T_2	W	Based on the information provided in the document, I am unable to determine the overall rate of inflation or deflation over the last 12 months. The document does not contain specific economic data or indicators pertaining to inflation rates.	Q2.I
T_5	W	I'm unable to provide opinions, forecasts, or generate speculative data, including predicting future inflation or deflation rates.	Q1.F
T_1	M0	To provide a detailed and accurate answer to your question, I will need to analyze the document you've uploaded. I will begin by examining the document to understand its insights on inflation expectations or relevant economic outlooks for the given time frame. This will enable me to make informed estimations for each category you've asked about. I will now start analyzing the document	Q3.I
T_1	M0	As an AI, I'm not capable of predicting future economic indicators like inflation rates. If you're seeking expert analysis or forecasts regarding inflation rates over the next 12 months, I recommend consulting economic forecasts from reliable financial institutions, central banks, or government economic reports, which typically use economic models and current economic data to provide their inflation forecasts.	



T.4	M2	<p>Predicting the future unemployment rate involves complexities such as economic policy changes, technological advancements, global economic conditions, and market trends, among others. While inflation data provides insight into economic conditions, it does not directly predict unemployment rates due to the multifaceted relationship between inflation and unemployment, often discussed in terms of the Phillips curve.</p> <p>Given that I lack real-time data access and forecasting capabilities, it's difficult to ascertain the likelihood of a higher unemployment rate 12 months from now based on past inflation rates alone. For accurate predictions, analyzing a wide range of economic indicators alongside inflation rates is essential. Economists and financial analysts use various models to make such forecasts, and it's advisable to consult their analyses for predictions on economic conditions, including the unemployment rate.</p>	Q2.I
T.3	M1	<p>Over the 12-month period between April 2025 and April 2026, a reasonable estimate for the rate of inflation might be around 4% to 6%</p>	Q2.F

---