Simulating Macroeconomic Expectations using LLM Agents*

Jianhao Lin[†] Lexuan Sun[‡] Yixin Yan[§]

First Draft: May 2025

This Draft: November 2025

ABSTRACT

We introduce a novel framework for simulating macroeconomic expectations using LLM Agents. By constructing LLM Agents equipped with various functional modules, we replicate three representative survey experiments involving several expectations across different types of economic agents. Our results show that although the expectations simulated by LLM Agents are more homogeneous than those of humans, they consistently outperform LLMs relying simply on prompt engineering, and possess human-like mental mechanisms. Evaluation reveals that these capabilities stem from the contributions of their components, offering guidelines for their architectural design. Our approach complements traditional methods and provides new insights into AI behavioral science in macroeconomic research.

Keywords: Expectation Formation, LLM Agents, Survey Experiment, AI Behavioral Science

JEL Codes: C90, D83, D84, E27, E71

---

* We thank Yuriy Gorodnichenko, Kai Li, Tracy Xiao Liu, Xiaobin Liu, Bin Miao, Carlo Pizzinelli, Yan Shen, Lin Wang, Shan Wang, Johannes Wohlfart, Liyan Yang, Yang Yang, Ji Zhang and seminar and conference participants at Tsinghua University, Nanjing University, Xiamen University, Sun Yat-sen University, The 22nd Chinese Finance Annual Meeting, 2025 Future Scholars in Finance Forum and 2025 ADEFT-XueShuo Summer Institute for many helpful comments and suggestions. We gratefully acknowledge the support of the National Natural Science Foundation of China (Grant No. 71991474, 72073148, 72273156, 72303258) and National Social Science Foundation of China (Grant No. 22AZD121, 24ZDA042). All authors contributed equally. Any remaining errors are our own.

† Lingnan College, Sun Yat-sen University, China, 510275. Email: linjh3@mail.sysu.edu.cn.

‡ Lingnan College, Sun Yat-sen University, China, 510275. Email: sunlx7@mail2.sysu.edu.cn.

§ Lingnan College, Sun Yat-sen University, China, 510275. Email: yanyx33@mail2.sysu.edu.cn.

How much to consume or save, what price to set, and whether to hire or fire workers are just some of the fundamental decisions underlying macroeconomic dynamics that hinge upon agents' expectations of the future. Yet how those expectations are formed, and how best to model this process, remains an open question.

Coibion & Gorodnichenko (2015)

1 Introduction

How agents form expectations has attracted significant attention from macroeconomists (Coibion et al., 2018). While full-information rational expectations (FIRE) have long dominated expectations modeling (Muth, 1961; Lucas, 1972), the growing body of survey-based research in recent years has increasingly revealed systematic deviations between individual expectations and those predicted by FIRE, challenging the paradigm of rational expectations theory (Manski, 2018; Weber et al., 2022; D'Acunto & Weber, 2024). Currently, survey experiments are widely used to study how agents form expectations about inflation, unemployment, home prices, or the broader economy (Cavallo et al., 2017; Armona et al., 2019; Andre et al., 2022; Fuster & Zafar, 2023). However, traditional survey methods rely heavily on survey firms or questionnaire platforms and suffer from limitations such as high costs, low extensibility, and limited flexibility.

To address these limitations, we propose a novel framework for simulating macroeconomic expectations in survey experiments, based on our design of LLM Agents. Large Language Models (LLMs) have demonstrated unique and powerful emergent abilities through continuous breakthroughs (Wei et al., 2022; Zhao et al., 2023). Building on these foundation models[1] as the "brain," LLM Agents automate the extraction, processing, and analysis of data beyond the model's inherent knowledge by invoking various functional modules and

[1] In this paper, foundation models refer to general-purpose large language models pre-trained on massive datasets, which can be applied as foundations to a broad range of downstream tasks. Examples include the GPT series by OpenAI and DeepSeek-R1 by DeepSeek, among others.

tools (acting as "hands" and "feet"). This enables perception and interaction with the external environment. Inheriting the language understanding and generation abilities of LLMs, and enhanced with multi-functional modules, these LLM Agents can perform complex tasks that are challenging for foundation models alone, making them widely applicable in industry (Zhao et al., 2023; Korinek, 2025). Under this new paradigm of LLM application, our framework proposes a guideline for constructing LLM Agents to simulate macroeconomic expectations of diverse economic agents in survey experiments. The framework offers several advantages: (i) it can replicate core findings of human survey experiments at a low cost and at any scale; (ii) it is highly extensible to different types of economic agents and experiments; (iii) it offers strong flexibility for pre-estimating future expectations and underlying mental mechanisms, or assessing the effects of future macroeconomic shocks. These advantages are difficult to achieve effectively with traditional survey methods or foundation models relying simply on prompt engineering.

This paper introduces and validates our framework through four parts. First, we describe the design and construction of LLM Agents. We focus on two representative agent types commonly encountered in expectation surveys—households and experts. Accordingly, we develop LLM Agents to simulate household expectations (referred to as *Household Agents*) and expert expectations (referred to as *Expert Agents*). For households, personal characteristics, prior expectations, and social media information play important roles in shaping expectations. Thus, Household Agents are equipped with a Personal Characteristics Module (PCM), a Prior Expectations & Perceptions Module (PEPM), and a Social Media Information Module (SMIM) to acquire, process, and analyze real-world data from household surveys and social platforms. In contrast, expert expectations are more influenced by professional background and domain knowledge. In addition to the PEPM, Expert Agents are equipped with a Professional Background Module (PBM) and a Knowledge Acquisition Module (KAM) to gather, process, and analyze expert profiles from official websites or LinkedIn, as well as professional knowledge from search engines. After constructing the

LLM Agents, we initialize them by clearly defining their role types, levels of confidence, specific tasks, and module usage rules.

Second, we introduce the experimental designs. We draw on three representative experiments covering several common types of macroeconomic expectations, each exemplifying a typical survey experiment in macro expectations research. The first is the hypothetical vignette experiment by Andre et al. (2022) on inflation and unemployment expectations of households and experts. The second is the information provision experiment by Chopra et al. (2025) on home price expectations of homeowners and renters. The third is based on the widely recognized Michigan Survey of Consumers (MSC); we pre-estimate long- and short-term inflation and home price expectations in the 2025 MSC to examine LLM Agents' ability to simulate general surveys and their out-of-sample performance.

Third, we analyze the simulation results. We compare the shape similarity between the expectation distributions generated by LLM Agents and those of human subjects across three experiments. The results indicate that LLM Agents produce expectation distributions highly similar to humans. Although these distributions are more homogeneous than human ones, they still capture key heterogeneity within and across different types of agents. Furthermore, through text analysis of open-ended survey responses using methods such as word frequency and annotation by agentic workflows, we explore the mechanisms underlying LLM Agents' simulation capabilities. We find that LLM Agents exhibit selective recall mechanisms similar to humans, though with more limited channels or content recalled. Additionally, LLM Agents possess causal pathways of thought (i.e., mental models) resembling those of humans, albeit with less diversity in pathways and nodes—a feature absent in foundation models relying simply on prompt engineering. This explains why LLM Agents generate expectation distributions similar to humans yet more homogeneous.

Fourth, we evaluate the contribution of each component in LLM Agents to the simulation. Specifically, we ablate individual components to investigate the source of LLM

Agents' ability to simulate expectation distributions and capture underlying mental mechanisms. Results show that all components contribute to simulation capabilities across different dimensions. PEPM plays a larger role in characterizing the distributions, while SMIM, PCM, KAM, and PBM are essential for establishing human-like mental mechanisms. Moreover, foundation models relying simply on prompt engineering fail to achieve effective simulation, indicating the superiority of the LLM Agents in enhancing simulation performance. These findings offer guidance for designing LLM Agents' architectures.

This paper makes three key contributions to the literature. First, our study contributes to an influential body of empirical work on macroeconomic expectation formation (Fuster & Zafar, 2023; D'Acunto & Weber, 2024), which has traditionally relied on costly and inflexible survey-based methods. We propose a novel framework that integrates survey data, social media information, large-scale internet textual data, and modular LLM Agents to simulate macroeconomic expectations among different economic agents. While this approach overcomes several limitations of conventional methods, it is important to emphasize that our framework is not intended to replace traditional methods, nor is it capable of doing so; rather, it serves as a complementary system. Data collected through human surveys, such as beliefs, demographic characteristics, preferences, and open-ended responses, can be used to calibrate modules within our framework or serve as benchmarks for validating simulation outcomes. Meanwhile, our framework has the potential to address the shortcomings of conventional methods in studying special groups and policies, and can serve as a pre-experimental tool before survey experiments are conducted.

Second, we contribute to the rapidly growing literature in AI behavior science that uses LLMs (or LLM Agents) to simulate human beliefs (Bybee, 2023; Zarifhonarvar, 2025), behaviors (Horton, 2023; Tranchero et al., 2024), and decisions (Li et al., 2024; Hansen et al., 2025). To the best of our knowledge, this is the first study to simulate the macroeconomic expectations among different types of economic agents in various representative survey experiments by constructing LLM Agents. The two most related studies are Hansen et al. (2025) and Zarifhonarvar (2025), yet with fundamental differences. Hansen et al. (2025) focus on using LLMs to simulate forecasting decisions of professional forecasters. Zarifhonarvar (2025) examines the characteristics of inflation expectations generated by different series of LLMs and their divergence from human expectations. It highlights the limitations of foundation models relying simply on prompt engineering, leading to significant deviations from human expectations. In contrast, we propose a framework to guide economists in constructing automated LLM Agents for more effective simulation of expectations, thereby reducing these deviations. Although a gap remains between simulation results and human data, our approach significantly improves upon foundation models relying simply on prompt engineering. This expands the capabilities of LLMs and demonstrates the considerable potential of this new paradigm in expectation simulation.

Finally, we contribute to the emerging literature that examines the behavior and cognition of generative AI (GenAI) itself, including its rationality (Chen et al., 2023; Bini et al., 2025), biases (Chen et al., 2024; Hagendorff, 2024), and preferences (Goli & Singh, 2024; S. Ouyang et al., 2024). Our study offers a new perspective on understanding belief formation in GenAI by analyzing the thoughts generated by LLM Agents. We classify open-ended responses from LLM Agents and identify Directed Acyclic Graphs by constructing agentic workflows, examining their selective recall mechanisms and mental models while comparing these with human counterparts. This approach extends several studies on open-ended survey data summarized in Haaland et al. (2025), facilitating measurement and understanding of the mechanisms underlying belief formation in GenAI and their divergence from humans. Similar studies remain scarce in existing literature.

The rest of our paper is organized as follows. Section 2 provides a general framework. Section 3 designs the architectures of the LLM Agents. Section 4 introduces the experimental design, data, and prompts. Section 5 presents the simulation results and mechanism analysis. Section 6 evaluates the contributions of each component in LLM Agents. Section 7 concludes.

2 A General Framework

In this section, our goal is to propose a generalizable framework that enables economists to simulate macroeconomic expectations of different types of economic agents using customized LLM Agents. As shown in Figure 1, this framework consists of five steps in sequence: Constr**U**ction → I**N**itialization → S**I**mulation → Pre-es**T**imation → **E**valuation. Accordingly, we refer to this framework as "**UNITE**". It provides a practical methodology and operational procedures for simulating macroeconomic expectations in survey experiments, thereby broadening the capabilities and application scope of generative AI.

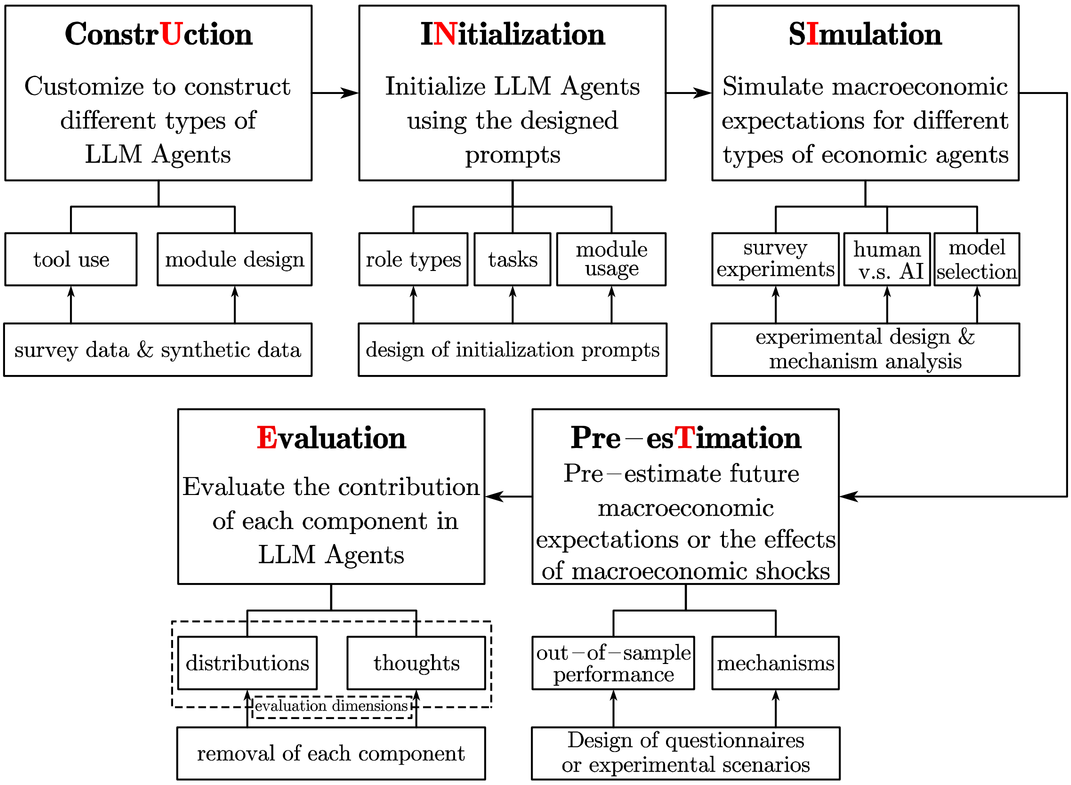


Figure 1: The "UNITE" framework for simulating macroeconomic expectations

Notes: This figure illustrates a generalizable framework we propose for simulating macroeconomic expectations. The framework consists of five main steps in sequence: Construction → Initialization → Simulation → Pre-estimation → Evaluation, which we refer to as the "UNITE" framework. Each large box at the top provides a brief overview of the corresponding step, the smaller boxes in the middle summarize the key points of each step, and the rectangular strip at the bottom outlines the specific tasks, methods, or data used in each step.

The first step is to construct LLM Agents. In this step, we design general architectures for LLM Agents tailored to simulate the common expectation-survey subjects (e.g., households, experts). Each agent is equipped with modules that capture the target population's distinct characteristics. For example, for Household Agents, we include a Social Media Information Module that uses extraction tools to automate the collection, cleaning, and analysis of relevant social media content. Ideally, these modules draw personal information from real survey data; however, when surveys lack required individual-level variables, we apply random matching or generate synthetic data with LLMs to create a semi-synthetic dataset[2], thereby expanding the original limited sample.

The second step is to initialize LLM Agents. After designing architectures, LLM Agents require explicit definitions of role type, assigned tasks, and the rule of trade-offs among information embedded in modules. We therefore create initialization prompts—customized to the experimental design, questionnaire, and agent role—that specify these elements. Prompts should be clear, concise, objective, and follow a consistent standardized format.

The third step constitutes the core of our framework—simulating macroeconomic expectations for different types of economic agents. This step involves three key components. First, we design detailed survey-experiment procedures, drawing on three widely recognized designs: hypothetical vignette experiments from Andre et al. (2022) that examine how four canonical macroeconomic shocks affect household and expert expectations of inflation and unemployment; information provision experiments from Chopra et al. (2025) on homeowners' and renters' home price expectations; and selected items from a large household expectations survey (the latter is actually implemented in Step 4, where we pre-estimate future expectation distributions). Second, we evaluate simulation performance by comparing simulated expectation distributions with those of human participants in the corresponding experiments. Third, we compare the responses based on different foundation models with

[2] The use of LLMs to generate synthetic data is becoming widespread in academic research, and its underlying rationale has gained increasing acceptance; it offers a viable alternative for constructing research datasets when empirical data are scarce (Yu et al., 2023; Halterman, 2025; Ge et al., 2025).

those of human data to guide model selection. Throughout, analysis emphasizes both the similarity of distributions and the mechanisms explaining why LLM Agents generate human-like expectations.

The fourth step is to pre-estimate future expectations or macro-shock effects. This step constitutes an extension of Step 3. We center the design on a large-scale survey (the Michigan Survey of Consumers[3]). Unlike the previous step, we use sample data covering the full year before a specified date (typically the foundation models' knowledge cutoff) to forecast the distribution of households' long- and short-term expectations on inflation and home prices over a subsequent period. We then compare these forecasts with the observed human survey distributions to evaluate LLM Agents' out-of-sample performance and examine the mechanisms of expectation formation. In the future, a potential application is using the agents to simulate scenarios with possible future shocks.

The final step is to evaluate the contribution of each component of the LLM Agents. This step aims to assess how the modules added in Step 1 and the initialization prompts designed in Step 2 contribute to the simulation performance. By removing each component of the LLM Agents one by one and comparing outcome changes, we can identify the sources of different dimensions of their simulation capabilities, such as reproducing human expectation distributions and capturing the key features of the thought processes underlying expectation formation. This evaluation helps verify the soundness of the LLM Agents' architectures and provides guidelines for constructing LLM Agents that more faithfully simulate macroeconomic expectations.

[3] The Michigan Survey of Consumers is one of the longest-running household surveys in the world. It is conducted by the University of Michigan to assess U.S. consumer attitudes and expectations regarding personal finances, business conditions, and economic outlook. Established in 1946, the survey collects data from approximately 600 respondents each month and is widely used in many studies (Curtin, 1982; D'Acunto et al., 2023).

3 Design of the LLM Agents' Architectures

In this section, we present the detailed procedures of the first step in the UNITE framework, explaining how we construct LLM Agents that represent different types of economic agents. Specifically, we develop LLM Agents that simulate the macroeconomic expectations of households and experts, which serve as the subjects in the experiments described in the subsequent sections.

3.1 LLM Agents for Simulating Household Expectations

Households are the most common subjects in expectation survey experiments, and they are included in all subsequent experiments. Before constructing Household Agents, it is essential to clarify how household expectations are formed and what factors primarily influence them.

First, a large literature suggests that economic expectations or perceptions are closely linked to various demographic characteristics. Studies have found significant differences in economic expectations across individuals of different ages, genders, political affiliations, education levels, and income groups (Souleles, 2004; Ehrmann et al., 2017; Ben-David et al., 2018; Coibion et al., 2022; D'Acunto et al., 2024). Second, the prior expectations or perceptions of economic agents regarding economic variables serve as a crucial determinant of their future expectations, particularly their most recent perceptions of these variables (Jonung, 1981; Coibion et al., 2020). Third, media coverage exerts a significant influence on households' macroeconomic expectations (Carroll, 2003; Lamla & Maag, 2012). In particular, with the rapid rise of social media, most households now get news primarily from platforms such as X (formerly Twitter) and increasingly consider these sources as more credible than traditional news media (Coibion et al., 2022; Ehrmann & Wabitsch, 2022; Angelico et al., 2022; Gorodnichenko et al., 2024). Consequently, continuously updated social-media information has become an increasingly important factor shaping households' macroeconomic expectations.

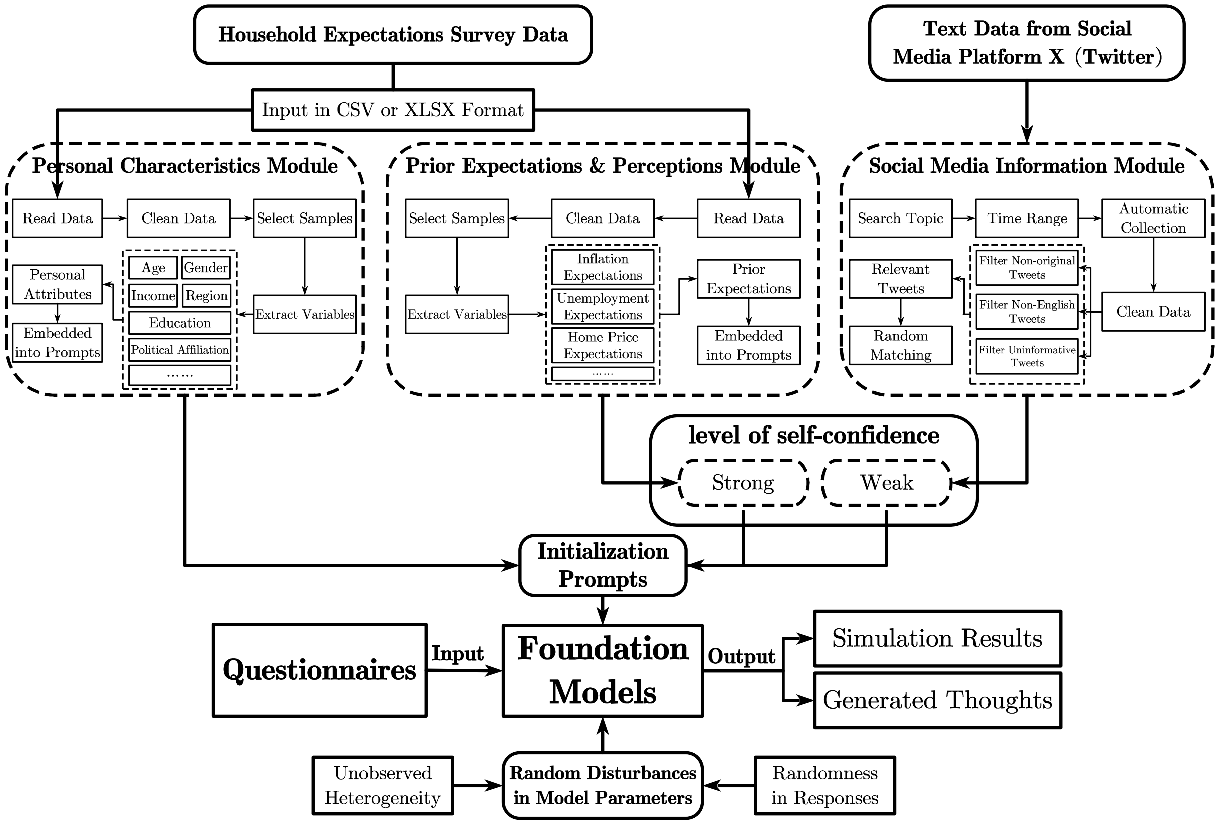


Figure 2: LLM Agents for simulating household expectations

Notes: This figure presents the detailed architecture of the LLM Agents for simulating household expectations. The Household Agents consist of six main components: Personal Characteristics Module (PCM), Prior Expectations & Perceptions Module (PEPM), Social Media Information Module (SMIM), Random Disturbances (RD), initialization prompts, and foundation models. Both the PCM and PEPM draw on data from household expectation surveys. SMIM collects tweet text data from the social media platform X. These modules automatically extract and process information, with their operational rules defined in the initialization prompts. For input questionnaires, Household Agents can engage in role-playing and perceive external environment through these components, ultimately outputting heterogeneous expectations along with the underlying thoughts.

Based on this literature, we construct the PCM, the PEPM and the SMIM (see Figure 2) to incorporate information on households' personal characteristics, prior expectations and social media into LLM Agents. Specifically, the PCM includes key attributes such as age, gender, political affiliation, and education level of economic agents. The PEPM captures their prior expectations about economic variables such as inflation, unemployment, and home prices. These data originate from household expectation surveys and are typically provided to the PCM and PEPM modules in CSV or XLSX format. Each module automatically reads the files, cleans the data (removing missing/invalid values and applying variable transformations), selects samples, extracts key variables, and embeds their numeric or textual values into prompts submitted to the Household Agents. The wording of the prompts varies with the design of the experiment, but should remain largely consistent with the corresponding formulations in the original questionnaire. The SMIM automatically retrieves and processes textual data from relevant posts on social platform X according to the experimental requirements. For instance, if the experiment focuses on U.S. inflation expectations, the user can set the search topic to "US Inflation" and specify a time window (i.e., the experimental period, typically aligned with the data range used in the PCM and PEPM). The SMIM then automatically collects popular posts from X's Top lists[4] and performs data cleaning, including filtering out non-original posts, non-English posts, and uninformative posts (e.g., very short or promotional tweets). Since it is impossible to know which specific posts each household has viewed, the program randomly assigns these posts to the Household Agents.

Further, it is necessary to specify in the initialization prompts a rule that instructs Household Agents how to use the data extracted by the three modules. In practice, economic agents trade off between priors and signals (external information) when updating beliefs. Typically, when agents are highly confident in their priors, they overweight those priors and underreact to new information (i.e., conservatism). Conversely, when agents lack confidence, they underweight prior beliefs and rely heavily on signals, over-updating their beliefs (i.e., base-rate neglect) (Chan et al., 2025; Hill, 2022; Benjamin, 2019). We therefore define five confidence levels[5] for Household Agents, ranging from extremely weak to extremely strong, and instruct them to follow the rule:

```
Your responses should trade off among the various pieces of information mentioned above in
accordance with your level of confidence: If you are confident, your answers will rely on
```

[4] Tweets appearing on Top lists typically attract greater attention, with higher numbers of views, retweets, or replies, and are therefore more likely to be read by households.

[5] When survey data on human respondents lack information regarding their confidence in (prior) expectations, we can employ random stratified sampling to divide the total sample into five subsamples that are approximately equivalent in both demographic structure and size. Each of these subsamples is then randomly assigned one of five distinct confidence levels.

```
Prior Expectations & Perceptions, and will not be influenced by other information, such as
the Social Media Information. On the other hand, if you lack confidence, your answers are
more likely to be influenced by other information.
```

In addition, Household Agents should incorporate information from the PCM when generating responses. Therefore, we instruct them to follow:

```
In addition, your responses should fully reflect the Personal Characteristics (such as age,
gender, educational level, political affiliation, etc.) of the role you are portraying.
```

However, there remains unobserved heterogeneity (e.g., emotions, experiences, thought patterns, or cognitive abilities) and various sources of randomness among economic agents that influence their expectations, and it is impossible to incorporate all of these factors within the modules. Therefore, drawing on the concept of random disturbance terms from econometrics, we introduce random disturbances following normal distributions to key parameters controlling text generation (Temperature and Top-p) in the foundation models[6]. This approach endogenously reflects the unobserved heterogeneity and randomness of agents in simulated expectation formation, thereby enhancing the realism of LLM Agents as simulation tools.

3.2 LLM Agents for Simulating Expert Expectations

Some studies compare the heterogeneity in expectations between experts and households (Carroll, 2003; Lamla & Maag, 2012; Andre et al., 2022, 2025), such as the hypothetical vignette experiments to be discussed in later sections. Therefore, it is necessary to construct LLM Agents for simulating expert expectations.

Compared to households, research has shown that experts' beliefs or decisions are primarily influenced by their professional background (e.g., work experience, education, field

[6] Both Temperature and Top-p are parameters that control the diversity of text generated by LLMs. The difference lies in that Temperature primarily modulates the shape of the probability distribution for the next token, whereas Top-p controls the scope of the candidate set during sampling. Thus, Temperature can be seen as corresponding to human factors such as emotion, experience, or thought patterns, while Top-p aligns more closely with cognitive capacity or attention mechanisms. We assign Temperature a normal distribution with a mean of 1.0 and a standard deviation of 0.5, and Top-p a normal distribution with a mean of 0.5 and a standard deviation of 0.25. Values falling outside the specified ranges ([0, 2] for Temperature and (0, 1] for Top-p) are winsorized to the corresponding endpoints, and both parameters are rounded to two decimal places.

of expertise), while the impact of demographic characteristics is relatively minor and unstable (Benchimol et al., 2022). Furthermore, experts typically possess professional training, greater specialized knowledge, and stronger capabilities in retrieving professional information (Ericsson et al., 2018; Gordon & Dahl, 2013). Based on the above literature, we developed two new modules for the Expert Agents—PBM and KAM (see Figure 3), which correspond to the PCM and SMIM modules in the Household Agents, respectively. The design and functionality of the other components in the Expert Agents are analogous to those in the Household Agents.

The PBM utilizes manually collected textual data from experts' profiles on official websites or LinkedIn. Key information such as expert names and affiliated organizations is obtained from expert expectation surveys. Samples with missing or insufficient information are filtered out. The PBM then inputs the expert profile dataset into the Data Organization Agent, which processes each expert's profile into a coherent, uniformly formatted paragraph of approximately 500 words, outputting the results in JSON format. Given that survey-based expert samples are often limited, PBM employs the Synthetic Data Generation Agent to generate synthetic samples that closely resemble real expert profiles. These synthetic profiles exhibit high similarity to real ones in terms of writing style and structure, and can be merged with real samples to form a semi-synthetic dataset. This dataset includes essential expert information such as company/organization, work experience, position, research areas, and educational background. If the expert survey is anonymized, we cannot ascertain the priors corresponding to each expert. Therefore, the PBM randomly pairs expert profiles with priors to construct a semi-realistic dataset.

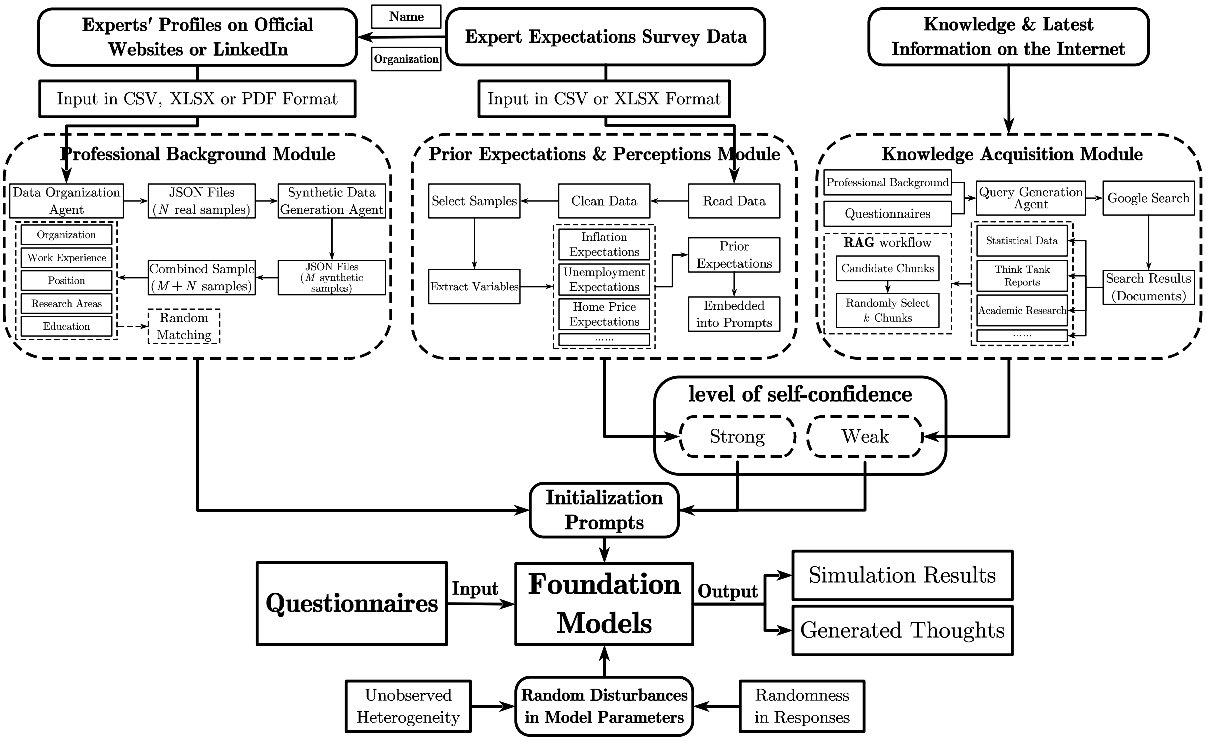


Figure 3: LLM Agents for simulating expert expectations

Notes: This figure presents the detailed architecture of the LLM Agents for simulating expert expectations. The Expert Agents consist of six main components: Professional Background Module (PBM), Prior Expectations & Perceptions Module (PEPM), Knowledge Acquisition Module (KAM), Random Disturbances (RD), initialization prompts, and foundation models. PBM utilizes actual experts' profiles from official websites or LinkedIn, and can generate synthetic data when the sample size is insufficient. PEPM derives data from expert expectation surveys. KAM retrieves and acquires relevant knowledge or the latest information from the internet on a personalized basis. These modules automatically extract and process information, with their operational rules defined in the initialization prompts. For input questionnaires, Expert Agents can engage in role-playing and perceive external environment through these components, ultimately outputting heterogeneous expectations along with the underlying thoughts.

The KAM automatically retrieves, crawls, and matches relevant knowledge and information from the internet. First, the Query Generation Agent generates five personalized queries for each expert based on their professional background and the target questionnaire. Subsequently, the Expert Agents collectively employ *Google Search Engine* and the web search & scraping tool *Tavily*[7] to extract and download the top 10 most relevant search results for each query within a specified time frame, saving the full text of webpage contents as documents. These documents comprise diverse data sources, such as statistical data,

[7] See URL: `https://www.tavily.com/`.

financial news, think tank reports, and academic research. Finally, to ensure that Expert Agents can retrieve key information from the extensive personalized knowledge base, we implement a workflow based on Retrieval-Augmented Generation (RAG)[8] (see Supplementary Appendix Figure A.1), enabling it to utilize k filtered and randomly selected chunks of the most relevant and high-quality information.

4 Experimental Design, Data and Prompts

In this section, we detail the design of three representative expectation survey experiments in Steps 3 and 4 of the UNITE framework, and the data used in their corresponding simulations. Our experiment adopts the designs of these experiments to ensure comparability between our simulation results and those from human experiments. After specifying these elements, we design the prompts, particularly the initialization prompts, for the LLM Agents involved in each experiment (i.e., Step 2 of the UNITE framework).

4.1 Hypothetical Vignette Experiments

The design of the first experiment draws on the hypothetical vignette experiments[9] introduced by Andre et al. (2022), an approach that has been widely adopted in many studies on macroeconomic expectations (Binder et al., 2023; Dibiasi et al., 2025; Bruschi et al., 2025). The experiment investigates how households and experts update their inflation and unemployment expectations in response to several common macroeconomic shocks (oil price shocks, government spending shocks, monetary policy shocks, and income tax shocks) through a series of sub-experiments, offering strong extensibility and generalizability.

[8] Retrieval-Augmented Generation (RAG) is a technique that enhances the outputs of LLMs by integrating information retrieval models. It retrieves relevant information from external data sources and feeds it to the LLMs, which then generates more accurate and contextually relevant responses. This method combines the strengths of both retrieval and generation, allowing for dynamic and precise text generation tailored to specific queries (Gao et al., 2024).

[9] Hypothetical vignette experiments, a popular form of information provision experiments, are commonly used to measure subjects' beliefs in hypothetical scenarios, such as those that could occur in the future but have not yet materialized. This method allows researchers to effectively control the specific information presented to respondents, thus facilitating the simulation and pre-assessment of the potential effects of proposed policies or anticipated shocks (Haaland et al., 2023).

We adopt and integrate the designs from Wave 1 through Wave 3 of the survey experiment by Andre et al. (2022), which enables the simulation of all outcomes within a single wave. For each shock, we design corresponding hypothetical vignettes, with the core content of the questionnaire closely aligned with that of Andre et al. (2022). The detailed experimental procedure and the survey structure are presented in Figure 4.

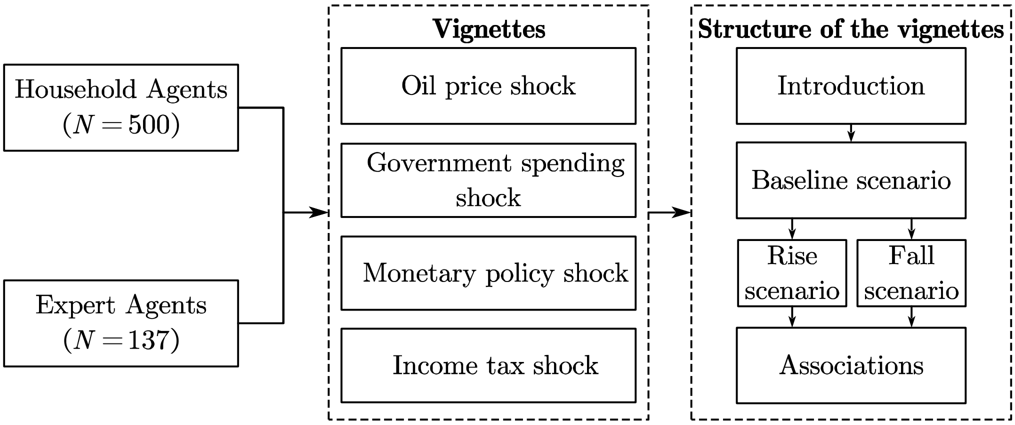


Figure 4: Overview of the experimental procedure and structure of the hypothetical vignette experiments

Notes: This figure illustrates the framework of the experimental procedure and structure of the hypothetical vignette experiments. On the left panel, it presents the two types of agents participating in the experiment along with their respective sample sizes. The middle panel displays the four vignettes corresponding to different macroeconomic shocks. On the right panel, the figure outlines the specific structure of each vignette.

First, we describe the survey data used for simulations with Household Agents and Expert Agents, respectively. For Household Agents, the survey data inputs for PCM and PEPM are drawn from the 2019 Michigan Survey of Consumers (MSC). After data cleaning and stratified sampling, a representative sample of 500 households is obtained[10]. The two variables input into PEPM are categorical measures (e.g., increase, decrease, or remain

[10] The surveys of Wave 1 and Wave 2 in Andre et al. (2022) were both conducted in 2019, while Wave 3 was carried out during the COVID-19 pandemic (early 2021) and may have been subject to uncontrollable factors. Although Andre et al. (2022) considers this issue in their design and attempts to mitigate the impact of the pandemic, to avoid added complexity, we set the temporal context of this experiment in 2019. Therefore, all data for the modules used in this experimental simulation are sourced from 2019, contemporary with Andre et al. (2022), to ensure that our developed LLM Agents accurately replicate the respondents' overall state during the original experiment—that is, their personal characteristics, priors, and the social media information they were exposed to at the time. Additionally, the purpose of the stratified sampling is to obtain a sample closely aligned with the demographic proportions of the 2019 American Community Survey (ACS), ensuring broad representativeness. The survey data from Andre et al. (2022) also maintains demographic alignment with the ACS.

unchanged) related to inflation (price) expectations and unemployment expectations[11]. For the Expert Agents, the input survey data for the PEPM are obtained from the 2019 Survey of Professional Forecasters (SPF). After data cleaning and sample selection for the specified year, 137 expert forecasts on the personal consumption expenditures price index and unemployment are retained. Although these forecasts are collected anonymously, the acknowledgments section of the quarterly SPF reports lists the names and affiliations of most participating experts. We therefore manually collect profiles of these experts from official websites or LinkedIn, compiling a dataset of 47 real samples. This dataset is input into the PBM to generate a semi-synthetic dataset (comprising 90 synthetic samples), which is randomly matched with the priors[12].

Then, we instruct the LLM Agents to respond to both the rise and fall scenarios within each hypothetical vignette[13]. Following the approach of Andre et al. (2022), each vignette adopts the same structure and begins with a brief introduction to familiarize respondents with the vignette's context. For example, in the oil price vignette, respondents are informed about the average price of crude oil per barrel in the past week. They then proceed to the baseline scenario, where the core variable (e.g., oil price) is assumed to remain unchanged. Under this scenario, we collect respondents' expectations regarding the unemployment rate in 12 months and the inflation rate over the next 12 months. Next, respondents are

[11] Since the MSC data on unemployment expectations only provides categorical variables (direction of change) rather than continuous variables (point forecasts), all expectation variables in the PEPM for both Household and Expert Agents are standardized as categorical variables in this experiment. This ensures uniformity in input variable types and comparability of simulation results.

[12] We do not use the original data publicly released by Andre et al. (2022) in our simulations for two main reasons: (1) the published dataset lacks respondents' prior expectations and provides only limited personal characteristics; (2) the expert survey is fully anonymous and contains limited information, which prevents the construction of an expert profile dataset. Therefore, we employ the widely recognized and representative MSC and SPF datasets, which offer diverse informational dimensions and clear variable documentation, thereby facilitating data cleaning and analysis.

[13] LLM Agents participate in and respond to each scenario, as opposed to being randomly assigned to different scenarios like human respondents in Andre et al. (2022). This design is primarily motivated by two reasons: (1) Requiring human respondents to complete multiple scenarios at once may degrade response quality through fatigue and thus compromise experimental outcomes—an issue not present with LLM Agents. (2) Human participants retain memory of previous experiments, meaning that the order of scenarios and exposure to varying information across scenarios may introduce interference. In contrast, each API call to an LLM is independent, ensuring that the samples simulated by LLM Agents across scenarios strictly satisfy the assumption of independent and identically distributed (i.i.d.) data, free from interference caused by memory retention. These advantages of LLM Agents help control for the influence of extraneous factors, such as demographic characteristics, across different experimental scenarios.

prompted to predict the unemployment rate and inflation rate under a scenario where an exogenous economic shock is introduced. Specifically, they are assigned to a rise scenario in which the shock variable increases (e.g., the average oil price rises by $30) and a fall scenario in which the shock variable decreases (e.g., the average oil price falls by $30). To simplify the analysis, Andre et al. (2022) reverses the sign of all predictions in the fall scenarios and merges them with the data from the rise scenarios. The main outcome variable is respondents' perception about the effect of a shock, measured as the difference between their predications under the shock scenario and those under the baseline scenario.

Finally, we collect each LLM Agent about their associations when making their predications through structured and open-ended questions, thereby allowing us to directly measure their thought processes. The core content of the questionnaire is detailed in Supplementary Appendix Section A.1.

4.2 Information Provision Experiments

The design of the second experiment draws on the information provision experiment introduced by Chopra et al. (2025). Unlike the first experiment, their approach directly presents subjects with information, thereby eliminating the need for constructing elaborate hypothetical scenarios. This type of experiment is more commonly adopted in related studies and is considered more generalizable (Haaland et al., 2023). The experiment consists of two sub-experiments that investigate, respectively, how different types of home price forecasts influence the long-term home price expectations of homeowners and renters, and how an increase in expected home price growth affect their economic outlook. For our simulation, we directly use the 2024 survey data on homeowners and renters provided by Chopra et al. (2025), which includes detailed individual-level information such as respondents' priors (e.g., home price expectations and housing transactions intentions), confidence in those priors, and homeownership status. For the following two sub-experiments, we use the architecture of Household Agent to simulate homeowners (Homeowner Agents) and renters (Renter Agents), respectively.

In the first sub-experiment, a random half of respondents are assigned to the high-forecast group and receive a 10-year average annual home-price growth forecast of 6%, while the remainder are assigned to the low-forecast group and receive a 2% forecast. To quantify post-treatment differences in expectations across groups, we elicit each respondent's subjective probability distribution for the average annual growth rate of a representative U.S. home over the next ten years. Respondents assign probabilities to mutually exclusive and collectively exhaustive bins representing ranges of future home price growth. For each respondent, we then calculate the implied mean of their distribution using the bins' midpoints. This approach of eliciting agents' expectations through distribution forecasting serves as a complement to the point forecasting method used in the first experiment.

In the second sub-experiment, we focus on respondents' main considerations when confronted with changes in the long-run home price growth rate. To measure these considerations, respondents receive information prompting them to imagine that they revise upward their expectations on home price growth. They are then asked to indicate how this change in home price expectations would affect their own economic situation: improving, remaining unchanged, or worsening. Additionally, open-ended questions are used to collect explanations for their responses, allowing us to examine the mechanisms underlying expectation formation. The questionnaire designs for both sub-experiments are provided in Supplementary Appendix Section A.2.

4.3 Large-Scale Household Expectations Survey

In the first two experiments, simulations are conducted using contemporaneous or even identical samples, without pre-estimation of future macroeconomic expectations. To extend our study, we design the third experiment to evaluate the out-of-sample estimation capability of LLM Agents in Step 4 of the UNITE framework. Unlike the previous survey experiments, large-scale household expectations surveys typically feature broader temporal coverage, higher frequency, and more extensive scope, making it one of the most representative and comprehensive approaches for studying expectation dynamics. In this experiment, LLM Agents are employed to pre-simulate MSC expectation data and the underlying thought processes for January 2025 and beyond[14]. This experimental design is difficult to achieve with traditional methods, whereas our framework accomplishes it efficiently.

Specifically, we focus on evaluating the ability of LLM Agents to pre-estimate the distributions of households' short-term (one-year) and long-term (five-year) inflation and home price expectations. Input data for the PCM and PEPM are drawn from a stratified sample of the 2024 MSC (sample size is 3,000, with demographic characteristics aligned with the full 2024 sample). Simultaneously, the SMIM automatically collects and processes hot-topic tweets related to "US Inflation" and "US home price" from platform X in 2024. The LLM Agents are tasked with responding to questions in the 2025 MSC survey regarding both short- and long-term inflation and home price expectations, providing explanations for their answers via open-ended questions (see Supplementary Appendix Section A.3 for the questionnaire). The simulated inflation and home price expectations will then be compared against human responses from the 2025 MSC (sample size is also 3,000, with demographic characteristics aligned with the full 2025 sample) to assess forecasting performance.

4.4 Prompts Design

After detailing the specific designs and the data used in the three experiments, we now proceed to design the prompts for each module of the LLM Agents, as well as the initiali-

[14] We select the period starting from January 2025 as the out-of-sample test window because the knowledge cutoff dates of the advanced foundation models examined in this study mostly fall before January 2025 (see Supplementary Appendix Table A.1). Therefore, using 2024 data to simulate the distributions of expectations in the MSC from January 2025 onward constitutes a rigorous test of out-of-sample performance.

zation prompts. For the module-specific prompts, due to significant differences in experimental designs and sample data across the three experiments, the prompts for each module must be tailored by the researchers according to the specific context of each experiment. This affords our framework a degree of flexibility, enabling users to select from survey data and customize corresponding prompts based on their research targets.

As for the initialization prompts, we adopt a uniform template across all experiments to clearly define the role type and confidence level of each LLM Agent, the task of each experiment, and the rules for module usage. Key phrasing in these prompts is primarily drawn from the original survey questionnaires to maintain objectivity and neutrality. The full prompts used in the three experiments, along with the rationale for their design, are provided in Supplementary Appendix Section B.

5 Simulation Results and Analysis

In this section, we perform Steps 3 and 4 of the UNITE framework. Specifically, we compare the similarity in shapes between distributions of expectations simulated or pre-estimated by LLM Agents and those formed by human subjects, in order to evaluate simulation fidelity and out-of-sample performance, respectively. Furthermore, we investigate the underlying mechanisms to explain why the distributions of expectations simulated or pre-estimated by LLM Agents resemble those generated by humans.

5.1 Simulation Results

To compare the shape similarity between the distributions generated by LLM Agents and those produced by humans, we discretize the probability distributions of both sets of expectation data into probability vectors by constructing histograms[15]. These two vectors

[15] The number of bins for the histograms corresponding to the two datasets is determined according to the following rules: (1) In general, the Freedman–Diaconis rule is applied by default to automatically determine the bin count. (2) When the sample sizes of both datasets are large (substantially exceeding the bin count derived from the Freedman–Diaconis rule), the number of bins is set to

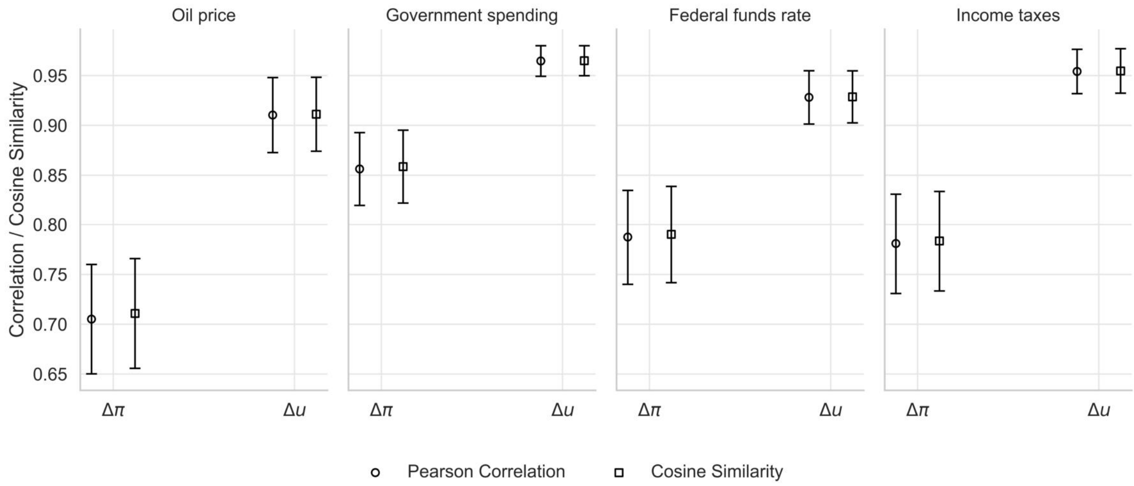
share the same dimensionality, with each element representing the distribution probability of the corresponding group's data within a specific numerical interval, thereby forming discrete approximations of the original continuous distributions. Subsequently, we compute both the Pearson correlation and cosine similarity between these two vectors as metrics to assess the shape similarity between the two distributions.

Before running simulations, we compare the simulation performance of several state-of-the-art foundation models and derive general guidelines for model selection[16]: (1) In general, models for simulation should be chosen based on benchmark rankings or technical reports. (2) For complex simulation tasks, reasoning models are recommended. (3) To balance performance and cost, open-source models are preferable when the performance gap is marginal. (4) As conditions may vary across experiments, preliminary tests should be conducted to compare model performance before final selection. Following these guidelines, we selected an advanced model from the Qwen series—Qwen3-235B-A22B-Thinking-2507—as the foundation model for the LLM Agents. The subsequent sections primarily present simulation results based on this model.

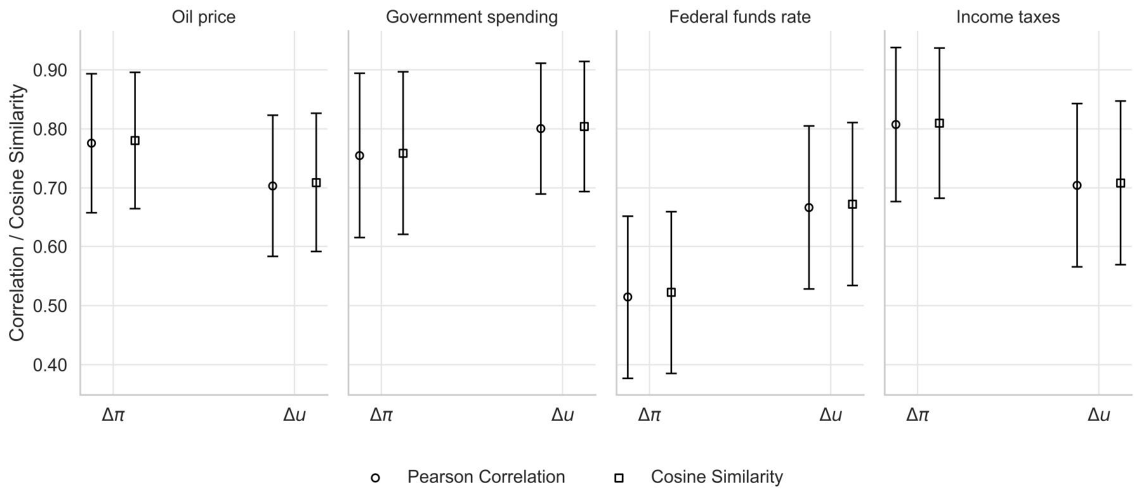
---

approximately equal to or slightly exceed the sample size, so as to identify differences in the distribution shapes at a finer granularity. This approach facilitates automatic selection of an appropriate bin count across varying sample sizes, thereby mitigating subjectivity in bin number specification.

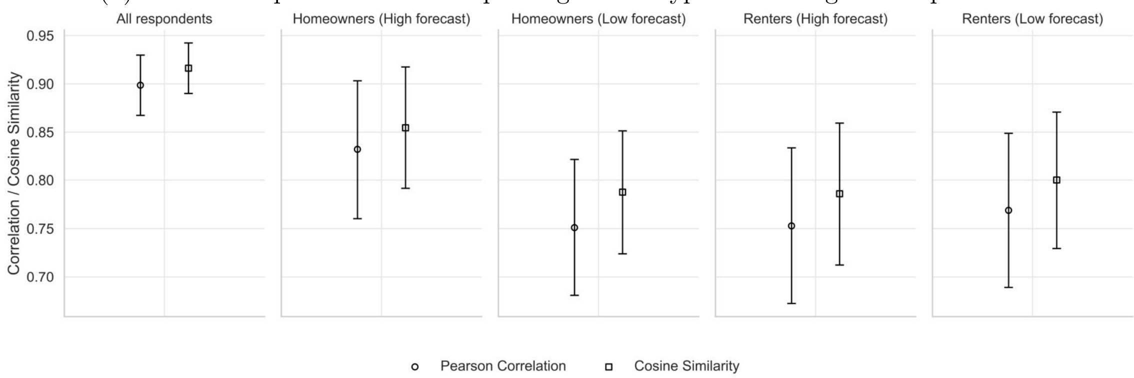
[16] We compare four state-of-the-art LLMs released by different vendors between April and July 2025—Qwen3-235B-A22B-Thinking-2507, DeepSeek-R1-0528, GPT-o4-mini, and Gemini-2.5-Pro. Supplementary Appendix Figure A.2–Figure A.5 show that, across most scenarios in our three representative experiments, simulations generated by LLM Agents based on all four models closely match the human data. Nonetheless, LLM Agents founded on Qwen3-235B-A22B-Thinking-2507 achieved relatively better performance when assuming different roles: their generated distributions are more similar to the human distributions. Moreover, the model's technical report (see `https://huggingface.co/Qwen/Qwen3-235B-A22B-Thinking-2507`) indicates higher evaluation scores for Qwen3-235B-A22B-Thinking-2507 compared with the other three models, which is broadly consistent with our simulation results. Therefore, selecting models by benchmark rankings is reasonable, although results may differ in other experiments—hence the need for small-scale pilot tests to confirm the most suitable foundation model.



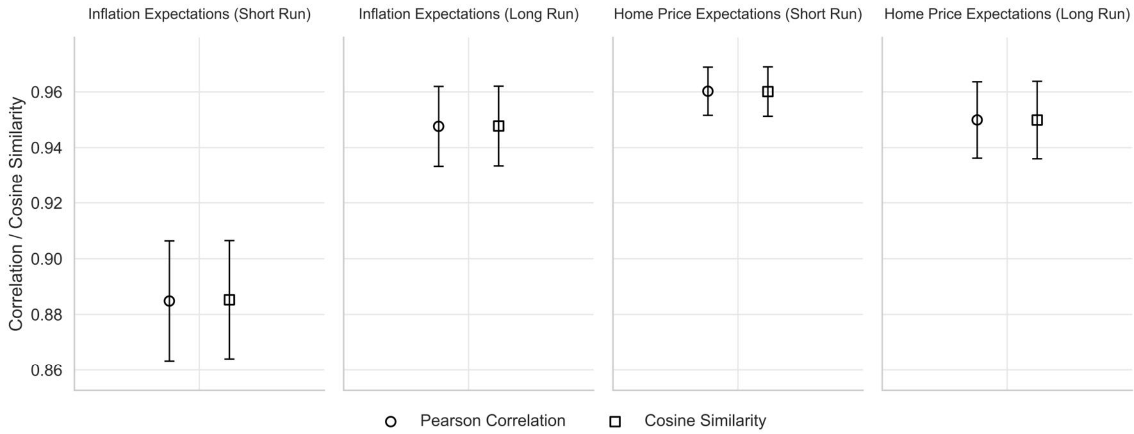
(a) Simulation performance of Household Agents in hypothetical vignette experiments



(b) Simulation performance of Expert Agents in hypothetical vignette experiments



(c) Simulation performance of LLM Agents in Sub-Experiment 1 of the information provision experiments



(d) Pre-estimation performance of Household Agents in the Michigan Survey of Consumers

Figure 5: Shape similarity between the expectation distributions generated by LLM Agents and those generated by humans in three representative experiments

Notes: Panel (a) and Panel (b) display the distributional shape similarity, as measured by Pearson correlation and cosine similarity, between the changes in inflation expectations ($\Delta\ \pi$) and unemployment expectations ($\Delta\ u$) generated by Household Agents (Panel (a)) and Expert Agents (Panel (b)), respectively, and those of humans under four different vignettes. Panel (c) presents simulation performance from Sub-Experiment 1 of the information-provision experiments: the Homeowner and Renter Agents' simulated home price expectations for homeowners and renters in the high-forecast and low-forecast treatment groups, and the LLM Agents' simulated home price expectations for all respondents. Panel (d) displays the pre-estimation performance of Household Agents for long- and short-run inflation expectations and home price expectations of respondents in the 2025 Michigan Survey of Consumers. Error bars present two-sided 95% confidence intervals for the similarity metrics, obtained by bootstrap over histogram-based probability vectors.

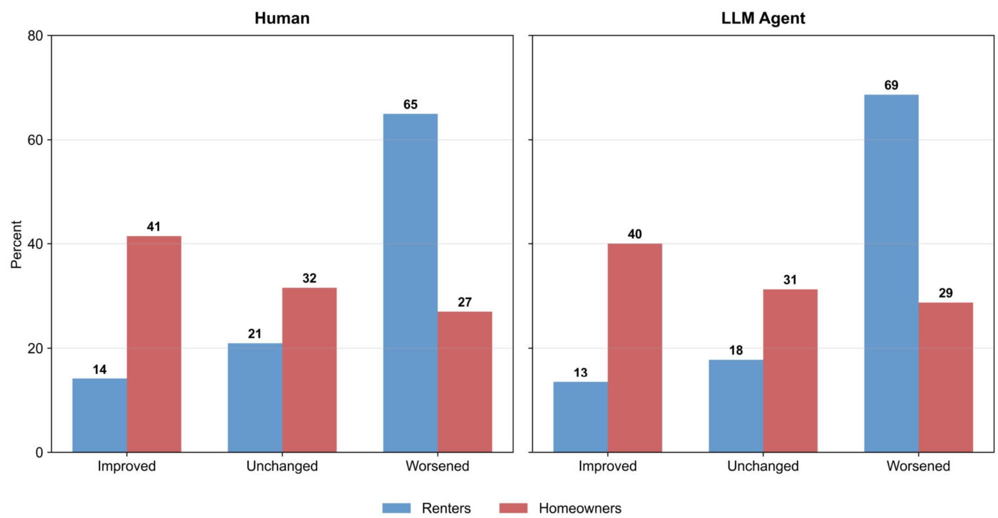


Figure 6: Comparison of LLM Agents' simulated results with human data in Sub-Experiment 2 of the information provision experiments

Notes: This figure compares the changes in expectations about their household's future economic situation generated by LLM Agents with those generated by humans in Sub-Experiment 2 of the information provision experiments. The left panel presents the responses from human participants, while the right panel displays the simulation results from Homeowner Agents. The horizontal axis represents the three possible directions of changes in expectations (improved, unchanged, worsened), and the vertical axis indicates the percentage of respondents selecting each direction.

The results in Figure 5 demonstrate that, across three representative experiments, our LLM Agents consistently achieve strong performance in simulating or pre-estimating the distributions of various macroeconomic expectations (such as inflation, unemployment, and home prices) across different types of agents under varying scenarios[17]. The shape similarity (whether Pearson correlation or cosine similarity) between the simulated distributions and those generated by humans averages around 0.8 in most cases, with the lowest values remaining above 0.5. Meanwhile, as shown in the results in Figure 6, the LLM Agents are able to capture a key heterogeneity in expectations between homeowners and renters: when anticipating future house price increases, most renters perceive that their household's future economic situation would worsen, whereas most homeowners believe it would improve or remain unchanged.

Furthermore, Supplementary Appendix Figure A.6 and Figure A.7 present comparisons between LLM Agents and humans in hypothetical vignette experiments, specifically regarding the directions and distributions of changes in expectations. The results indicate that while some quantitative differences exist, the simulations generated by Household Agents and Expert Agents capture key heterogeneities between households and experts: Compared to households, experts exhibit more homogeneous expectations (i.e., more concentrated distributions) and their directional changes align more closely with theoretically predicted outcomes as found in textbooks. For instance, a large majority of experts expect

[17] A potential concern is that the strong performance of the LLM Agents in this paper may simply result from the LLMs recalling or restating outcomes from existing survey experiments based on their extensive training data. However, this concern is unfounded for three reasons: (1) The survey data used in both the information provision experiments and the large-scale expectations survey were officially released online only after January 2025—i.e., after the knowledge cutoff of all foundation models used in this paper—making it impossible for such data to have been included in their training. (2) Even if the data from the hypothetical vignette experiments were published before the models' knowledge cutoff, general-purpose foundation models are unlikely to have directly used individual-level survey data during training. This is due to the typical use of processed, unstructured text data in LLM training, as opposed to raw structured survey data, as well as privacy protection policies adopted by some developers (Zhao et al., 2023; Yang et al., 2025). Unless specifically fine-tuned for such purposes, these models do not incorporate personally identifiable survey records. This also explains why many existing studies directly employ foundation models to replicate classic human experiments without considering this issue (Chen et al., 2023; Horton, 2023; Cui et al., 2025). (3) The results in Section 6 show that foundation models alone cannot simulate or pre-estimate the expectation distributions, further indicating that the strong performance of the LLM Agents stems not from the training data of foundation models, but from our designed architecture and the functional modules.

that a rise in oil prices would lead to increased inflation and unemployment expectations. Additionally, combined with the findings from Supplementary Appendix Figure A.8 and Figure A.9, it can be observed that across various representative experiments, the distributions of expectations generated by LLM Agents are more homogeneous than those of humans. This finding resonates with observations reported in several related studies[18] (Chen et al., 2023; Wang et al., 2025).

In Supplementary Appendix Section D, we examine the output robustness of LLM Agents. The results show no statistically significant differences across multi-round simulations, demonstrating strong robustness.

5.2 Mechanism Analysis

In this subsection, we aim to address the following questions: Why do LLM Agents demonstrate strong performance in simulating or pre-estimating various macroeconomic expectations across different population groups? What underlying mechanisms drive this capability? Some research indicates that selective recall plays a crucial role in shaping human cognition and behavior (Tversky & Kahneman, 1973; Bordalo et al., 2016, 2025). When forming heterogeneous expectations under varying conditions, economic agents tend to selectively retrieve different types of relevant information from memory (such as news, knowledge, and experiences) (Andre et al., 2022). This motivates us to analyze responses to open-ended questions, investigating whether selective recall also underlies the expectation formation in LLM Agents, and to examine the similarities and differences between this mechanism in LLM Agents and humans.

[18] A likely reason is that these foundation models are primarily aligned with annotations or feedback data provided by human experts during Reinforcement Learning from Human Feedback (RLHF), often based on limited sample sizes. This leads to a systematic approximation of human expert outputs by LLMs (L. Ouyang et al., 2022; Zhao et al., 2023). Furthermore, although we incorporate as much personal information as possible in constructing LLM Agents and introduce random disturbances into parameters to account for some unobserved heterogeneity, it remains infeasible to exhaust the full diversity and extensive heterogeneity present in humans. Consequently, the observed pattern of simulation results being more homogeneous compared to those from humans represents a common limitation in this line of research. Nevertheless, as shown in Section 6, the architecture of LLM Agents largely mitigates this limitation of foundation models, which is key to their ability to capture critical heterogeneity both within and across human groups.

Furthermore, to explore the characteristics of expectation formation process in LLM Agents, we identify the complete reasoning processes behind the heterogeneous expectations formed by LLM Agents, humans, and foundation models, respectively. We then compare the mental models across these agents to elucidate their distinctions and commonalities.

5.2.1 Selective Recall of LLM Agents

For the hypothetical vignette experiments, we first follow the approach of Andre et al. (2022) to focus on and quantify the proportions of words related to four distinct channels (topics)[19] mentioned by LLM Agents in their open-ended responses when generating expectations under each vignette.

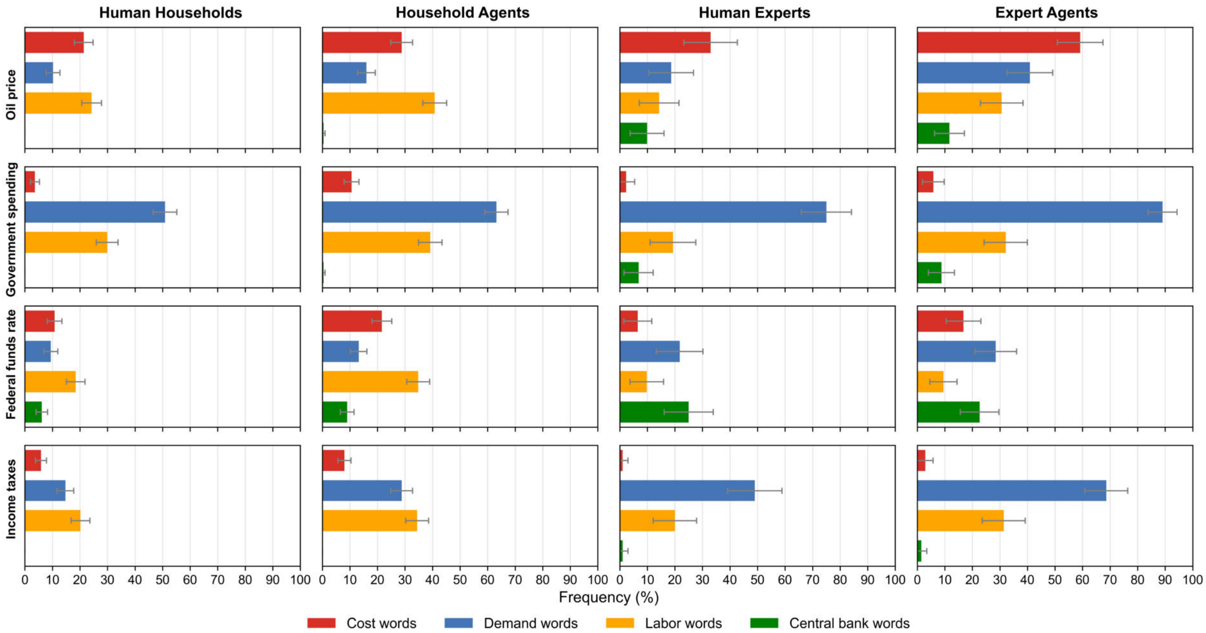


Figure 7: Word usage for open-ended responses of humans and LLM Agents across four vignettes

Notes: This figure presents the proportions of Human Households (Column 1), Household Agents (Column 2), Human Experts (Column 3), and Expert Agents (Column 4) mentioning words from four word groups in their open-ended responses under four different vignettes. The error bars indicate 95% confidence intervals.

[19] Specifically, the four channels are defined as follows: Cost words include the word (stem) “cost”. Demand words include the words (stems) “demand”, “buy”, “purchas”, “invest”, “spend”, “consum”. Labor words include the words (stems) “layoff”, “lay-off”, “lay off”, “fire”, “hire”, “labor”, “work”, “job”. Central bank words (phrases) include “monetary policy”, “federal funds rate”, “fed funds rate”, “federal funds target rate”.

As shown in Figure 7, both Household Agents and Expert Agents are able to capture the key heterogeneity of thoughts within and between human households and experts: experts tend to concentrate their reasoning within each vignette on channels that are recognized by the mainstream literature or textbooks as playing a central role in real-world shocks, whereas households often overlook mechanisms that may be dominant in reality. For example, across all four vignettes, whether facing supply or demand shocks, a considerable number of households refer to cost-related, particularly labor-related, supply-side channels. In contrast, for experts, cost-related supply-side mechanisms predominate in the case of an oil price shock (a supply shock), whereas demand-side channels dominate in the latter three vignettes, which involve demand shocks. Moreover, experts make more frequent references to central banks (Federal Reserve), further illustrating the professional nature of their recall content.

Further comparison reveals that while LLM Agents can qualitatively simulate the various channels mentioned by humans in forming expectations, there are quantitative differences: specifically, LLM Agents recall these types of channels at a slightly higher frequency than humans, indicating greater homogeneity in the content recalled by LLM Agents. These patterns are also echoed in the responses of LLM Agents to structured questions, as shown in Supplementary Appendix Figure A.10.

Second, following the coding scheme defined by Andre et al. (2022), we design and implement an agentic workflow (see Supplementary Appendix Figure A.11) that leverages two distinct LLMs to simulate the process of two human annotators independently labeling responses and reaching consensus through multiple rounds of discussion. This procedure categorizes open-ended responses from LLM Agents into nine distinct categories[20]. The results are subsequently verified by two graduate students in economics.

[20] We adopt the following categories as defined by Andre et al. (2022): i) "Mechanism" encompasses all responses addressing how shocks transmit through economic channels; ii) "Model" covers statements invoking a particular economic framework or theory; iii) "Guess" flags any expressions of uncertainty or admissions that the forecast is speculative; iv) "Politics" gathers broad political or

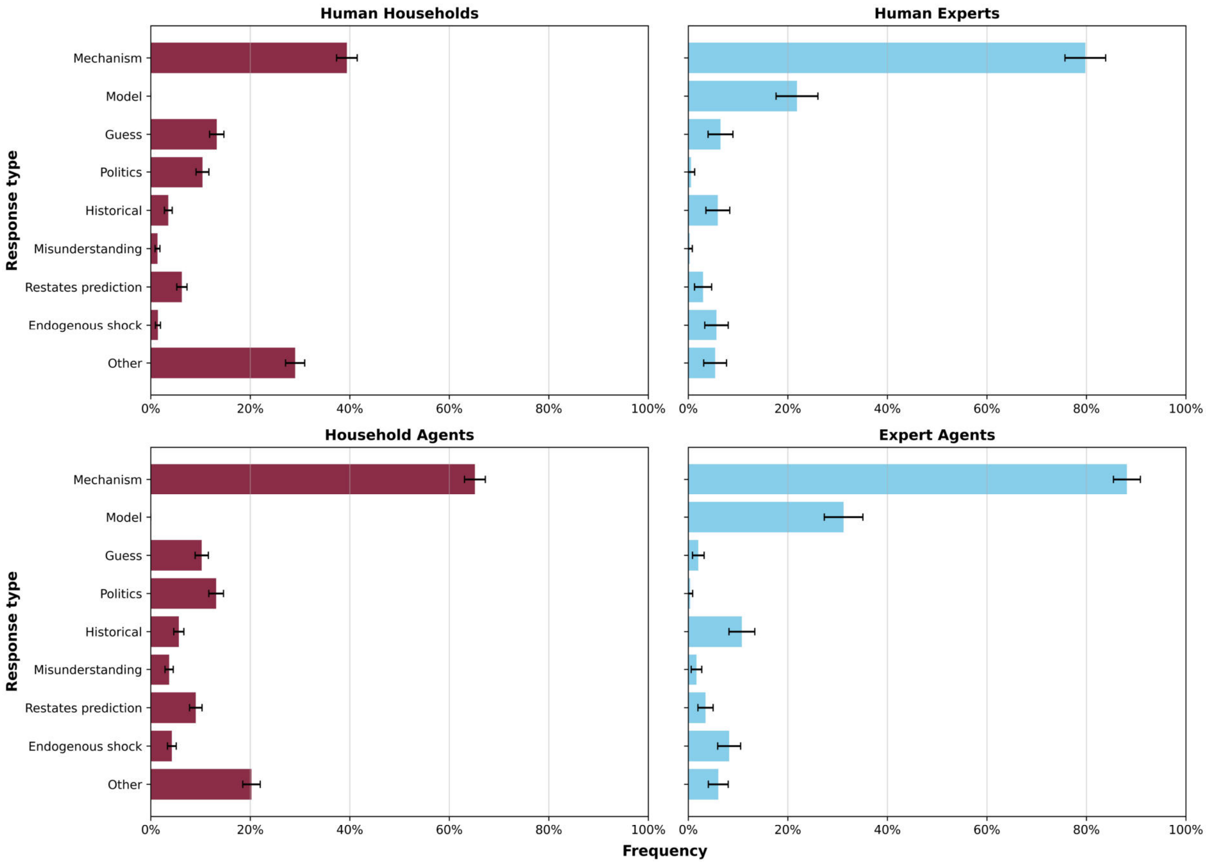


Figure 8: Response types in open-ended responses of humans and LLM Agents

Notes: This figure presents "response type" classification of open-ended responses generated by Human Households, Human Experts, Household Agents and Expert Agents, averaged across all four vignettes. The human data annotations are directly obtained from Andre et al. (2022), while the open-ended responses from LLM Agents are automatically classified by an agentic workflow and manually verified. Error bars display 95% confidence intervals.

As shown in Figure 8, LLM Agents can qualitatively reproduce the principal differences in the thought processes underlying expectation formation between households and experts: when making predictions, households tend to rely more on guesses and are more susceptible to politics. Their reasoning may be simpler, often merely restating predictions, and is more diverse, falling largely into the "Other" category. In contrast, experts more frequently recall and refer to "Mechanism" and "Model," and are more inclined to cite "Historical" content.

normative commentary; v) "Historical" captures references to past developments or typical evolutionary patterns; vi) "Misunderstanding" marks instances where respondents misinterpret aspects of the scenario; vii) "Restates prediction" identifies replies that merely reiterate or paraphrase the provided inflation and unemployment forecasts; viii) "Endogenous shock" refers to understanding an exogenous shock as an endogenous response, such as mentioning that interest-rate adjustments are responses by the Fed to other economic changes; and ix) "Other" serves as a residual category. Allow each response to fall into more than one category.

This also explains why changes in experts' expectations are more concentrated and generally align with textbook theories. However, quantitative differences exist between the reasoning of LLM Agents and humans: responses from LLM Agents are more concentrated in "Mechanism" and less frequently categorized as "Other".

For Sub-Experiment 2 of the information provision experiments, we follow the coding scheme defined by Chopra et al. (2025) and use the previously constructed agentic workflow to categorize the open-ended responses of LLM Agents into nine distinct mechanisms[21], with the results manually verified.

Figure 9 shows that the LLM Agents capture a key heterogeneity between the thoughts of renters and homeowners: most homeowners believe that an expected rise in home prices will increase the value of their housing via wealth effects, thereby improving their outlook on future economic situation; alternatively, they consider the home price growth irrelevant since they have no plans to buy or sell homes. In contrast, most renters believe that an expected rise in home prices will increase their future costs of buying or renting via income effects, thereby worsening their expectations about economic situation. This heterogeneity explains the pattern observed in Figure 6. However, similar to the earlier findings, compared to humans, the LLM Agents' thoughts are more concentrated on specific mechanisms, making their responses somewhat more homogeneous.

[21] We adopt the following mechanisms as defined by Chopra et al. (2025): i) "Wealth effects," referring to changes in the value of housing currently owned by the respondent's household. ii) "Income effects (cost of buying)," referring to changes in the cost of buying a home. iii) "Home price growth irrelevant," meaning that home price growth is irrelevant because the respondent does not plan to buy, sell, or move. iv) "Income effects (rental prices)," referring to changes in the rental prices of homes. v) "Collateral effects," referring to changes in the ease of borrowing against home equity. vi) "Endogenous adjustments to housing," referring to endogenous up-/downsizing, buying/selling, or changes in timing—for example, due to substitution effects, the investment channel, or purchase timing considerations. vii) "Inflation," referring to inflation and changes in the overall price level. viii) "Household income," referring to changes in the household's overall income. ix) "Interest rates," referring to changes in interest rates. Specific examples for each mechanism can be found in Supplemental Appendix Table A.21 of Chopra et al. (2025). Responses are allowed to correspond to more than one mechanism.

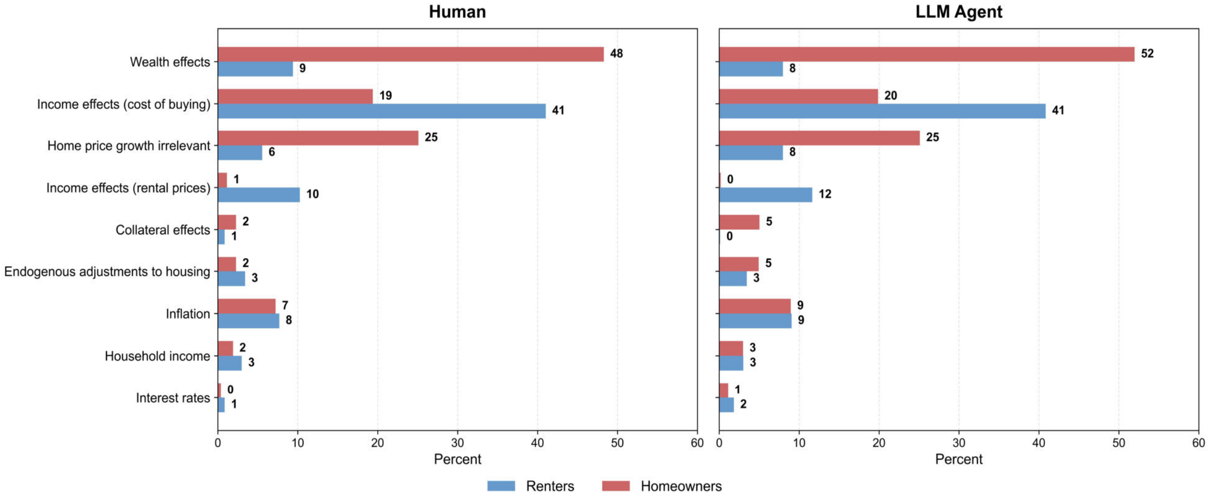


Figure 9: Open-ended responses on how higher expected home price growth affects humans' and LLM Agents' expectations about economic situation

Notes: The figure shows the proportion of human respondents and LLM Agents who invoke different arguments to explain why an increase in their expectations about home price growth over the next 10 years would affect their household economic outlook. The human data annotations are directly obtained from Chopra et al. (2025), while the open-ended responses from LLM Agents are automatically classified by an agentic workflow and manually verified.

For the large-scale expectations survey, we focus on and quantify the proportions of words related to seven distinct channels (topics)[22] mentioned by Household Agents in their open-ended explanations when pre-estimating inflation and home price expectations.

As shown in Figure 10, Household Agents primarily recall channels related to Cost, Politics, and Policy when pre-estimating inflation expectations, whereas for home price expectations, they focus more on Demand, Politics, and Borrowing & Lending. Although direct comparison with real households is limited due to the absence of open-ended responses in the MSC, the results in Figure 10 still reflect key characteristics of selective recall among real-world households: (1) Households' inflation expectations are mainly influenced by cost-related (or supply-side) factors, leading them to recall cost-related channels

[22] Specifically, the seven channels are defined as follows: "Cost" channel includes the word (stem) "cost", "expense", "fee". "Demand" channel includes the word (stem) "demand", "buy", "purchas", "invest", "spend", "consum". "Borrowing & Lending" channel includes the word (stem) "loan", "lend", "borrow", "debt", "credit", "interest", "mortgage". "Politics" channel includes the word (stem)"republic", "democratic", "trump", "biden", "harris", "elect". "Policy" channel includes the word (stem) "government", "fed", "monetary", "fiscal", "tax", "tariff". "Energy" channel includes the word (stem) "oil", "gas", "fuel", "electricity", "energy". "Black Swan Event" channel includes the word (stem) "russia", "ukraine", "war", "invasion", "sanction", "pandemic", "covid", "lockdown", "crisis", "disaster", "collapse", "crash", "breaking", "recession", "bubble".

more frequently (D'Acunto et al., 2021; Coibion et al., 2022; Andre et al., 2025), while their home price expectations are driven more by demand-side factors, making them more likely to recall demand-related channels (Binder et al., 2023; Gohl et al., 2024; Bro & Eriksen, 2025). (2) Households tend to consider politics-related narratives when forming macroeconomic expectations, a finding consistent with Figure 8. (3) Households' inflation expectations are susceptible to government policies (D'Acunto et al., 2024; Weber et al., 2022), as reflected in Household Agents' references to Federal Reserve monetary policy and the 2025 Trump tariffs. Meanwhile, research shows that households often factor in Borrowing & Lending considerations, such as mortgage rates, when forming home price expectations (Binder et al., 2023).

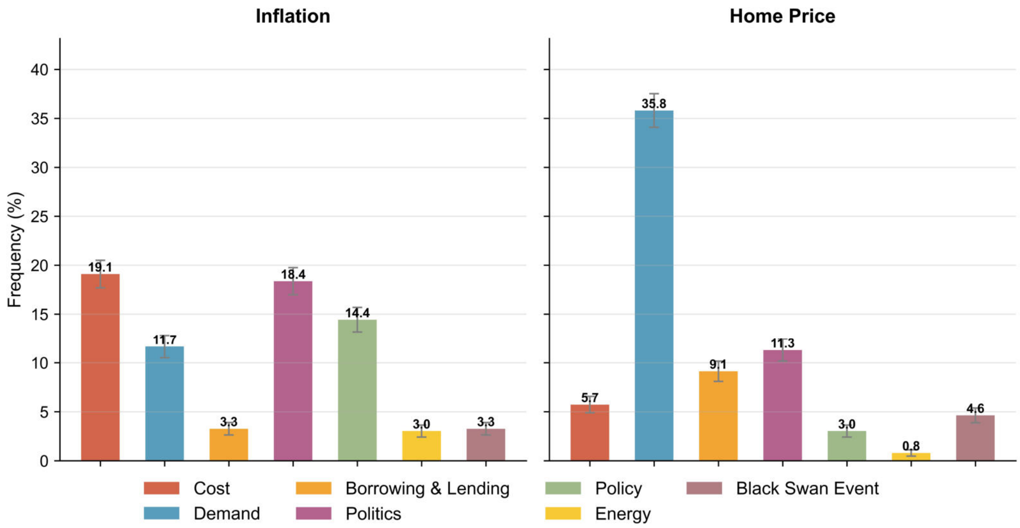


Figure 10: Proportion of various channels recalled by Household Agents when pre-estimating 2025 macroeconomic expectations

Notes: This figure displays the proportions of seven channels recalled by Household Agents when pre-estimating inflation and home price expectations for 2025. The left panel presents the results for inflation expectations, while the right panel shows those for home price expectations. Error bars display 95% confidence intervals.

In addition, Household Agents mention several major black swan events in recent years, such as the Russia–Ukraine war and the COVID-19. Evidence in Section 6 suggests that this does not stem from foundation models acquiring such knowledge via training data, but rather from LLM Agents' ability to perceive external information through the SMIM.

This subsection demonstrates that LLM Agents exhibit a selective recall mechanism akin to humans when generating expectations. However, their recalled content is more homogeneous, which accounts for the more concentrated expectation distributions.

5.2.2 The Mental Model of LLM Agents

We further analyze the complete reasoning processes underlying LLM Agents' formation of macroeconomic expectations. To simplify the analysis, we take four sub-experiments from hypothetical vignette experiments as examples. Following the approach of Andre et al. (2025), we use causal Directed Acyclic Graphs (DAGs)[23] to represent the causal structure of respondents' open-ended responses, in order to identify and compare the mental models[24] of humans, LLM Agents, and foundation models (i.e., LLMs that simply rely on initialization prompts; abbreviated as "only INITIAL").

First, we construct an agentic workflow (see Supplementary Appendix Figure A.12) that fully automatically identifies and labels the DAGs for each open-ended response. The labeling results are reviewed and corrected by two graduate students. Supplementary Appendix Section E.1 provides a detailed explanation of how DAGs are identified.

After converting each open-ended response into a DAG, we treat all DAGs from the same economic agent (households or experts) regarding the same expectation (inflation or unemployment) under the same vignette as a set of mental models. We then compute the Jaccard similarity between the mental model sets underlying the expectations of the LLM Agents and foundation models, respectively, and those of humans for each vignette (see

[23] A causal Directed Acyclic Graph is a graphical model composed of nodes representing variables and directed edges that signify causal relationships between them. The direction of each edge reflects the flow of causality, while the acyclic structure ensures that no variable can be a cause of itself, either directly or through a sequence of causal links. Causal DAGs have become a fundamental tool for formalizing and analyzing causal inference across diverse disciplines, including statistics, computer science, and the social sciences (Pearl, 2009; Sloman & Lagnado, 2015). More recently, this framework has been extended to the analysis of narrative or mental model structures in economic theory, wherein causal reasoning and story-based explanations play a central role (Eliaz & Spiegler, 2020; Spiegler, 2016, 2020).

[24] According to Andre et al. (2023), a Mental Model represents an individual's beliefs about the relationships between different variables, such as the reasoning process underlying the connection between rising oil prices and expected future inflation.

Supplementary Appendix Section E.2 for the detailed calculation method). The results are presented in Table 1.

Table 1 shows that across all vignettes, both Household Agents and Expert Agents exhibit mental models more closely aligned with those of humans (with a minimum similarity of 0.63). In contrast, the mental models of foundation models are significantly less aligned with those of humans (with a maximum similarity of only 0.53).

Furthermore, we construct "average DAGs" to visualize the aggregated mental models underlying inflation or unemployment expectations among humans, LLM Agents, and foundation models under each vignette. Using shifts in unemployment expectations following a government spending shock as an example, the results are presented in Figure 11. In this figure, variables (nodes) more frequently cited in respondents' mental models are depicted as larger circles, while more common causal relationships are represented by thicker edges. This approach intuitively reveals the most prevalent variables and causal links in the mental models of both households and experts.

Table 1: Similarity of mental models between LLM Agents, foundation models, and humans

Vignettes	**Panel A: Inflation (Households)**		**Panel B: Unemployment (Households)**	
	Original	only INITIAL	Original	only INITIAL
Oil price	0.87	0.52	0.68	0.43
Government spending	0.65	0.31	0.78	0.34
Federal funds rate	0.68	0.46	0.70	0.48
Income taxes	0.85	0.35	0.78	0.33
Vignettes	**Panel C: Inflation (Experts)**		**Panel D: Unemployment (Experts)**	
	Original	only INITIAL	Original	only INITIAL
Oil price	0.73	0.46	0.77	0.51
Government spending	0.76	0.40	0.82	0.46
Federal funds rate	0.92	0.42	0.63	0.53
Income taxes	0.70	0.28	0.64	0.44

Notes: This table presents the Jaccard similarity between the mental models underlying the formation of unemployment and inflation expectations in each vignette, comparing LLM Agents (denoted as "Original") with humans and foundation models (denoted as "only INITIAL") with humans. Panels A and B present the results for Household Agents and foundation models regarding inflation and unemployment expectations, respectively. Panels C and D present the results for Expert Agents and foundation models regarding inflation and unemployment expectations, respectively.

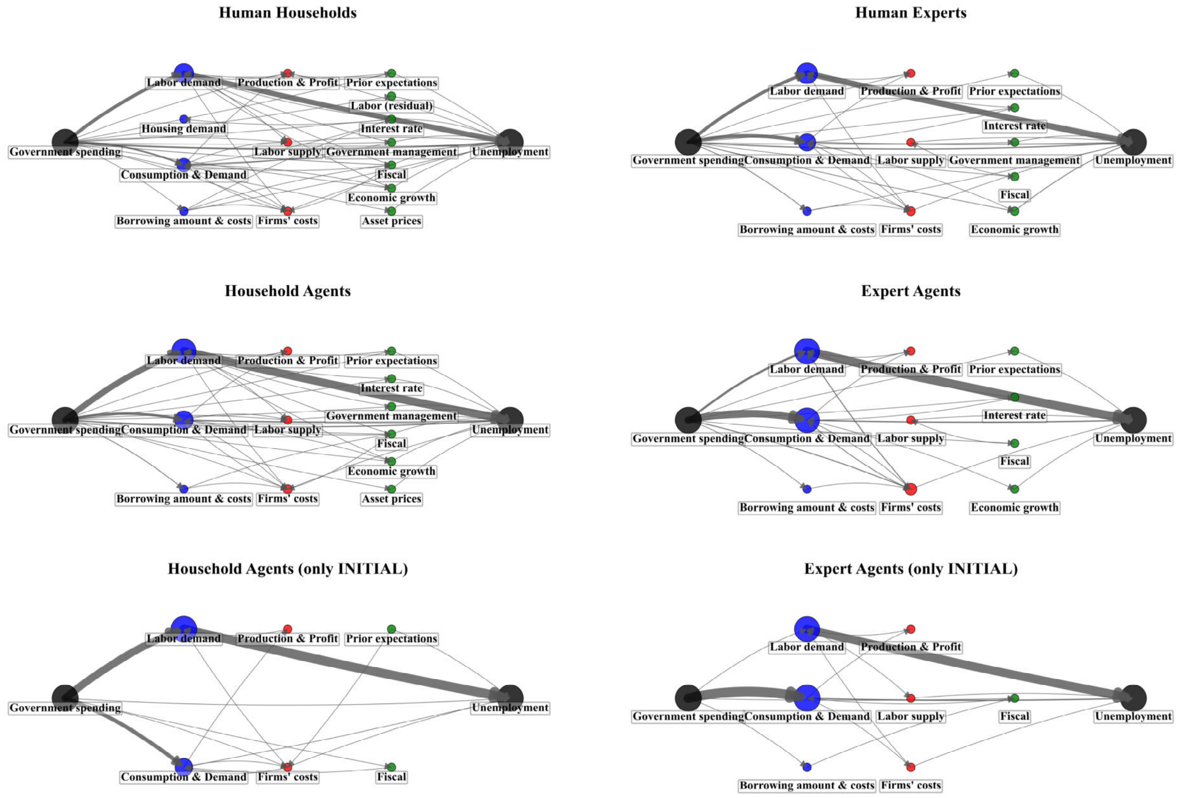


Figure 11: The "average" DAGs underlying the formation of unemployment expectations in the government spending vignette

Notes: The figure presents the "average" DAGs underlying unemployment expectation formation for humans (denoted as "Human Households" and "Human Experts"), LLM Agents (denoted as "Household Agents" and "Expert Agents"), and foundation models (denoted as "Household Agents (only INITIAL)" and "Expert Agents (only INITIAL)") in the government spending vignette. The nodes represent categories of intermediate variables, whose definitions and classifications are provided in Supplementary Appendix Table A.3. The aggregated DAGs reveal the most relevant variables (nodes) and causal links in the responses of humans, LLM Agents, and foundation models. Node size: The size of the nodes is proportional to the share of responses that refer to the nodes. Node color: Red indicates supply-side variables, blue indicates demand-side variables, green indicates miscellaneous variables, black is used for start and end nodes. Edge thickness: The thickness of the edges is proportional to the share of responses that refer to the causal connections (among humans, LLM Agents and foundation models, respectively).

As shown in Figure 11, both Household Agents and Expert Agents capture most of the nodes and their relationships within the mental models of households and experts, respectively. Compared to experts, households exhibit greater diversity in the nodes and cognitive pathways within their mental models, leading to more dispersed expectation distributions. More importantly, LLM Agents demonstrate more homogeneous mental models than humans, particularly lacking some miscellaneous variables. This likely contributes to the more concentrated expectation distributions generated by LLM Agents. However, the mental models of foundation models lack many key nodes and edges, resulting in a highly homogeneous structure. Similar patterns are observed across the other vignettes in Supplementary Appendix Figure A.13 to Figure A.19 for inflation and unemployment expectations.

In addition, we calculate the average number of causal links and unique nodes in the causal DAGs of humans, LLM Agents, and foundation models to assess the complexity of their mental models. The results are presented in Supplementary Appendix Table A.2. Overall, our findings reveal a key insight: LLM Agents exhibit mental models that are more diverse and simpler, whereas the foundation models' mental models concentrate on a few more complex causal paths. This indicates that the architectures of LLM Agents mitigate the homogenization and excessive complexity of thinking paths introduced by the alignment process, thereby enhancing its ability to simulate human expectations.

6 Evaluation of Contributions of Each Component in LLM Agents

In this section, we conduct the final step of the UNITE framework to investigate the origin of LLM Agents' ability to simulate human-like expectations and capture underlying thinking patterns (e.g., selective recall and mental models). We individually remove each component of the LLM Agents, as well as remove all modules at once (i.e., foundation models with only initialization prompts), and separately compute the shape similarity between their simulated or pre-estimated expectation distributions and those of humans.

Taking the Household Agents in the hypothetical vignette experiments as an example, Figure 12 shows that: (1) Compared to the original Household Agents, the similarity of simulation results decreases to varying degrees when different components are removed. (2) The performance of Household Agents declines significantly after removing PEPM and initialization prompts. (3) The foundation models yield the poorest simulation results. Similar patterns are observed in the Expert Agents and two other survey experiments (see Supplementary Appendix Figure A.20 to Figure A.23).

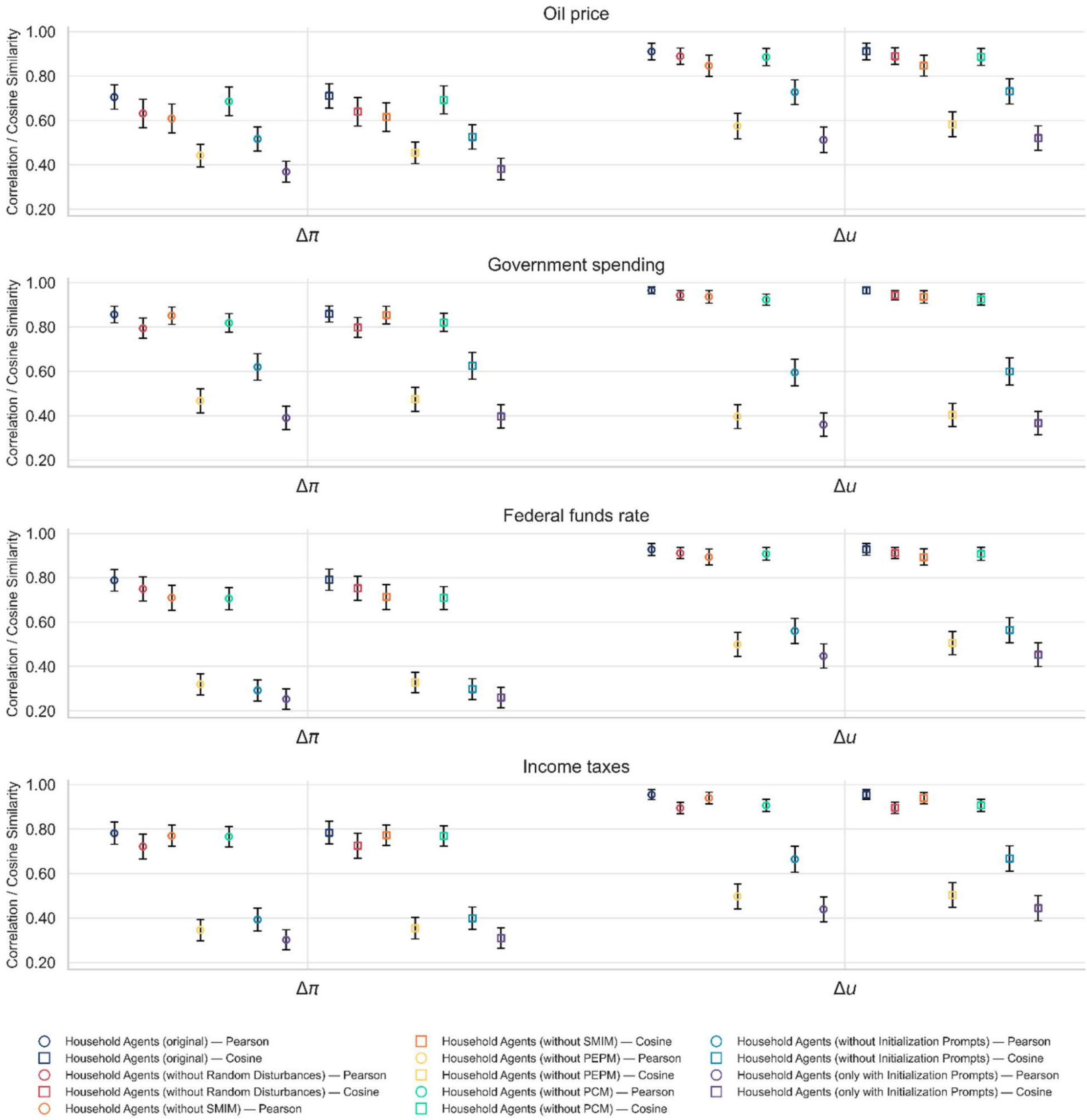


Figure 12: Shape similarity between the expectation distributions generated by Household Agents (original and those without modules) and those generated by humans in hypothetical vignette experiments

Notes: This figure displays the distributional shape similarity, as measured by Pearson correlation and cosine similarity, between the changes in inflation expectations ($\Delta\ \pi$) and unemployment expectations ($\Delta\ u$) generated by Household Agents (original and those without modules), respectively, and those of households under four different vignettes. Error bars present two-sided 95% confidence intervals for the similarity metrics, obtained by bootstrap over histogram-based probability vectors.

This finding indicates that: (1) Prior expectations are important for generating the expectation distributions by LLM Agents, and a similar pattern is observed in human samples—for humans, priors (particularly the most recent perceptions) are among the most critical factors in expectation formation (Jonung, 1981; Coibion et al., 2020). (2) Initialization is crucial for LLM Agents; without it, they fail to comprehend their role types, specific tasks, and rules for module usage, thereby losing the ability to simulate expectation distributions. (3) Foundation models possess almost no capacity to simulate human-like expectation distributions, whereas the architectures of LLM Agents enable this capability.

Other components of LLM Agents also contribute significantly to the simulation. However, unlike PEPM and initialization prompts, their roles are primarily reflected in capturing the mental mechanisms underlying expectation formation. Taking hypothetical vignette experiments as an example, as shown in Figure 13, the overall results indicate that removing any component leads to an increase in the proportion of recalled content categorized as "Mechanism," while the proportion categorized as "Other" decreases. Specifically, for Expert Agents, the removal of KAM or PBM results in a significant reduction in their recall of "Model"-related content, suggesting that information such as expertise and professional background aids in establishing selective recall regarding models or theories. For Household Agents, the removal of SMIM leads to a loss of response diversity, with a stronger focus on personal characteristics or prior predictions, thereby increasing the proportions of "Politics" and "Restates prediction." Household Agents without PCM lose access to personal information, resulting in a decline in the proportion of recalled content categorized as "Politics."

Additionally, supplementary evidence yields similar findings. Supplementary Appendix Figure A.24 reveals that, compared to the original LLM Agents, Household Agents without SMIM and Expert Agents without KAM recall highly homogeneous channels. Specifically, Expert Agents without KAM fail to recall any professional content related to central bank. As shown in Supplementary Appendix Figure A.25, in information provision experiments, the removal of any component leads to varying degrees of increased homogeneity in the channels recalled by LLM Agents. For instance, after removing SMIM, the proportion of mechanisms recalled by Homeowner Agents that fall under "Home price growth irrelevant" increases, which explains the rise in the proportion of agents expecting unchanged future economic situations. Meanwhile, more Renter Agents recall mechanisms centered on income effects, accounting for the significant increase in those expecting worsened future economic situations. Supplementary Appendix Figure A.26 indicates that, in pre-estimating the MSC, Household Agents without SMIM are almost unable to recall certain recent black swan events and exhibit reduced perception of policy-related information. Household Agents without PCM fail to recall channels associated with politics.

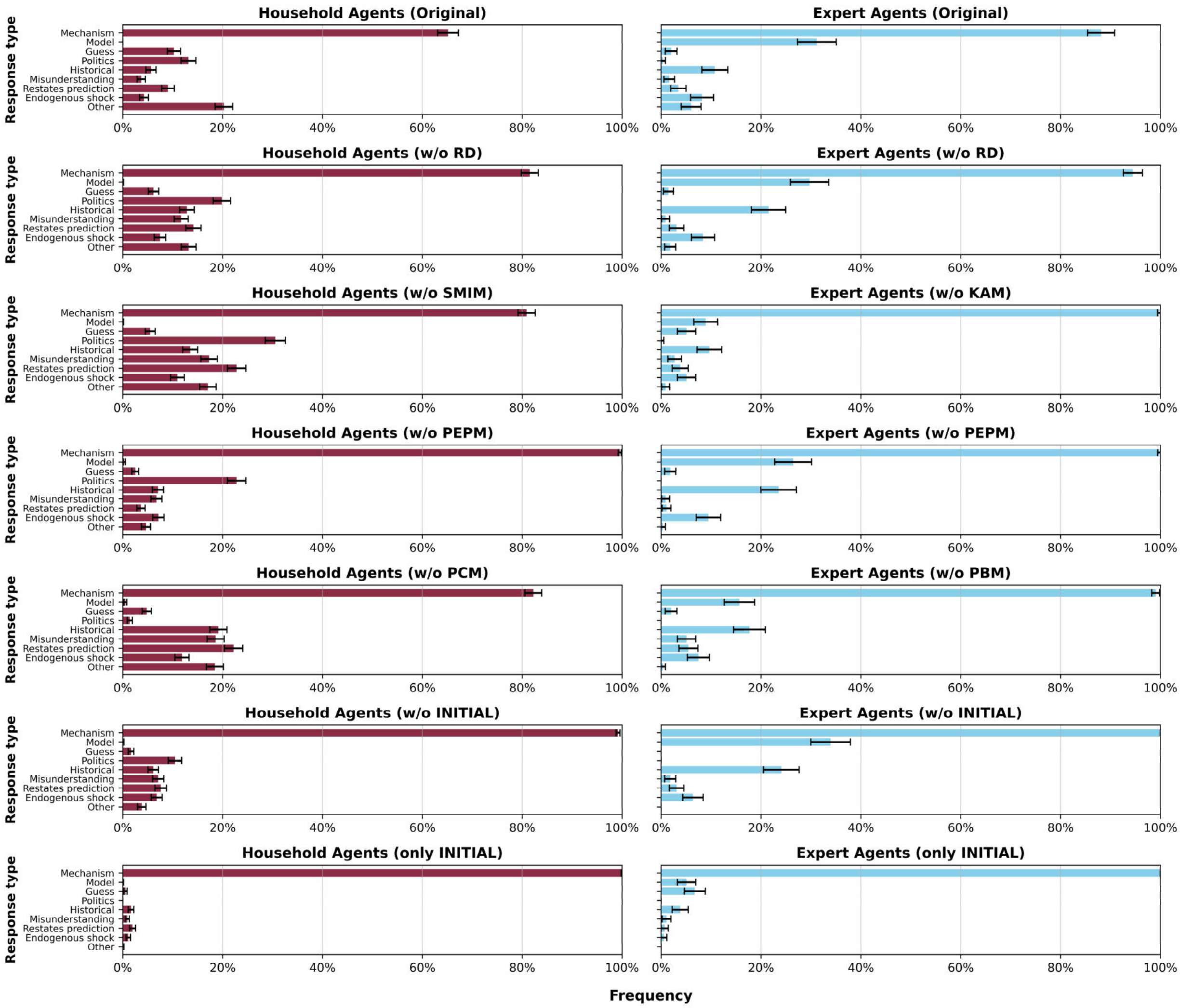


Figure 13: Response types in open-ended responses of LLM Agents (original and those without modules)

Notes: This figure presents "response type" classification of open-ended responses generated by LLM Agents (original and those without modules) across all four vignettes. The open-ended responses from all LLM Agents are automatically classified by an agentic workflow and manually verified. Error bars display 95% confidence intervals.

Furthermore, we measure and compare the diversity of thoughts underlying expectation formation among LLM Agents with different components removed and humans across three survey experiments (see Supplementary Appendix Section F for details). The results are presented in Supplementary Appendix Table A.4 to Table A.6. Our findings indicate that: (1) Removing any component reduces the diversity of thoughts generated by LLM Agents, suggesting that homogenization of thoughts may explain their diminished simulation performance. (2) While LLM Agents exhibit lower diversity than humans, their architectures enable them to surpass foundation models.

These findings collectively show that all components of the LLM Agents contribute to expectation simulation along different dimensions: (1) Initialization is essential; without it, the agents struggle to simulate as required. (2) For expectation distributions, the PEPM plays a major role, whereas the SMIM, PCM, KAM, and PBM modules underpin the agents' ability to capture human-like selective recall. (3) Foundation models alone cannot effectively simulate expectations, but they can do better within the LLM Agents' architectures.

7 Concluding remarks

In this paper, we propose a novel framework for constructing LLM Agents to simulate macroeconomic expectations in survey experiments. Using these LLM Agents, we simulate expectations and underlying thought processes across three representative experiments involving different types of agents, comparing the results with human data. Our findings show that while gaps remain relative to real human responses, LLM Agents significantly outperform foundation models relying simply on prompt engineering. They qualitatively capture key patterns of heterogeneity in expectations and underlying mental mechanisms, and accurately pre-estimate future expectation distributions. This demonstrates a level of flexibility, scalability, and cost-effectiveness difficult to achieve with traditional surveys.

Our findings yield several key insights. First, the architecture of LLM Agents enhances the simulation performance of foundation models, primarily because the added components enable LLM Agents to establish human-like selective recall mechanisms and mental models, diversifying their reasoning processes. As a result, the generated distributions by them capture key heterogeneities.

Second, our framework offers general guidelines for constructing LLM Agents to simulate expectations: Initialization is crucial and requires clear definitions of role types, confidence levels, and specific tasks in the prompts. PEPM is essential for simulating expectation distributions, while additional modules such as PCM and PBM incorporate personal information, and SMIM and KAM capture updated external information, thereby activating selective recall in LLM Agents.

Third, this type of approach has certain limitations. Although introducing random disturbances into LLM parameters can capture some unobserved heterogeneity, the simulation results of LLM Agents, while qualitatively capturing key patterns, inevitably deviate quantitatively from human data due to the inability to incorporate all real-world individual information. Therefore, such simulations should only serve as supplements or pilot studies to real survey experiments, not replacements. Researchers with strict quantitative requirements should use these AI tools with caution.

In future work, we can integrate learning dynamics across multiple rounds of simulated interactions, allowing LLM Agents to adapt expectations in real time. Moreover, integrating richer behavioral data, such as cognitive biases or social influence networks, could enhance the simulation of expectation formation in high-uncertainty environments or during policy regime shifts. Extending the framework to stimulate the expectations of firms (Firm AI Agents), policymakers (Government AI Agents), or media actors (Media AI Agents) may further illuminate the dynamics of macroeconomic beliefs.

References

Andre, P., Haaland, I., Roth, C., Wiederholt, M., & Wohlfart, J. (2025). Narratives about the Macroeconomy. *The Review of Economic Studies (Forthcoming).*

Andre, P., Pizzinelli, C., Roth, C., & Wohlfart, J. (2022). Subjective Models of the Macroeconomy: Evidence From Experts and Representative Samples. *The Review of Economic Studies*, *89*(6), 2958–2991. https://doi.org/10.1093/restud/rdac008

Andre, P., Schirmer, P., & Wohlfart, J. (2023). Mental Models of the Stock Market. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4589777

Angelico, C., Marcucci, J., Miccoli, M., & Quarta, F. (2022). Can we measure inflation expectations using Twitter? *Journal of Econometrics*, *228*(2), 259–277. https://doi.org/10.1016/j.jeconom.2021.12.008

Armona, L., Fuster, A., & Zafar, B. (2019). Home Price Expectations and Behaviour: Evidence from a Randomized Information Experiment. *The Review of Economic Studies*, *86*(4), 1371–1410. https://doi.org/10.1093/restud/rdy038

Benchimol, J., El-Shagi, M., & Saadon, Y. (2022). Do expert experience and characteristics affect inflation forecasts? *Journal of Economic Behavior & Organization*, *201*, 205–226. https://doi.org/10.1016/j.jebo.2022.06.025

Ben-David, I., Fermand, E., Kuhnen, C., & Li, G. (2018). *Expectations Uncertainty and Household Economic Behavior* (No. w25336; p. w25336). National Bureau of Economic Research. https://doi.org/10.3386/w25336

Benjamin, D. J. (2019). Errors in probabilistic reasoning and judgment biases. In *Handbook of Behavioral Economics: Applications and Foundations 1* (Vol. 2, pp. 69–186). Elsevier. https://doi.org/10.1016/bs.hesbe.2018.11.002

Binder, C., Kuang, P., & Tang, L. (2023). *Central Bank Communication and House Price Expectations* (No. w31232; p. w31232). National Bureau of Economic Research. https://doi.org/10.3386/w31232

Bini, P., Cong, L., Huang, X., & Jin, L. J. (2025). Behavioral Economics of AI: LLM Biases and Corrections. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.5213130

Bordalo, P., Burro, G., Coffman, K., Gennaioli, N., & Shleifer, A. (2025). Imagining the Future: Memory, Simulation, and Beliefs. *Review of Economic Studies*, *92*(3), 1532–1563. https://doi.org/10.1093/restud/rdae070

Bordalo, P., Coffman, K., Gennaioli, N., & Shleifer, A. (2016). Stereotypes. *The Quarterly Journal of Economics*, *131*(4), 1753–1794. https://doi.org/10.1093/qje/qjw029

Bro, J., & Eriksen, J. N. (2025). Subjective expectations and house prices. *Journal of Banking & Finance*, *172*, 107377. https://doi.org/10.1016/j.jbankfin.2024.107377

Bruschi, C., D'Acunto, F., Kumar, S., & Weber, M. (2025). Subjective Models of Workers and Managers for Macroeconomic Expectations. *AEA Papers and Proceedings*, *115*, 248–253. https://doi.org/10.1257/pandp.20251016

Bybee, J. L. (2023). *The Ghost in the Machine: Generating Beliefs with Large Language Models*. Working Paper.

Carroll, C. D. (2003). Macroeconomic Expectations of Households and Professional Forecasters. *The Quarterly Journal of Economics*, *118*(1), 269–298. https://doi.org/10.1162/00335530360535207

Cavallo, A., Cruces, G., & Perez-Truglia, R. (2017). Inflation Expectations, Learning, and Supermarket Prices: Evidence from Survey Experiments. *American Economic Journal: Macroeconomics*, *9*(3), 1–35. https://doi.org/10.1257/mac.20150147

Chan, K., Charness, G., Dave, C., & Reddinger, J. L. (2025). *On Prior Confidence and Belief Updating* (No. arXiv:2412.10662). arXiv. https://doi.org/10.48550/arXiv.2412.10662

Chen, Y., Fang, H., Zhao, Y., & Zhao, Z. (2024). *Recovering Overlooked Information in Categorical Variables with LLMs: An Application to Labor Market Mismatch* (No. w32327; p. w32327). National Bureau of Economic Research. https://doi.org/10.3386/w32327

Chen, Y., Liu, T. X., Shan, Y., & Zhong, S. (2023). The emergence of economic rationality of GPT. *Proceedings of the National Academy of Sciences*, *120*(51), e2316205120. https://doi.org/10.1073/pnas.2316205120

Chopra, F., Roth, C., & Wohlfart, J. (2025). Home Price Expectations and Spending: Evidence from a Field Experiment. *American Economic Review*, *115*(7), 2267–2305. https://doi.org/10.1257/aer.20240022

Coibion, O., & Gorodnichenko, Y. (2015). Information Rigidity and the Expectations Formation Process: A Simple Framework and New Facts. *American Economic Review*, *105*(8), 2644–2678. https://doi.org/10.1257/aer.20110306

Coibion, O., Gorodnichenko, Y., & Kamdar, R. (2018). The Formation of Expectations, Inflation, and the Phillips Curve. *Journal of Economic Literature*, *56*(4), 1447–1491. https://doi.org/10.1257/jel.20171300

Coibion, O., Gorodnichenko, Y., Kumar, S., & Pedemonte, M. (2020). Inflation expectations as a policy tool? *Journal of International Economics*, *124*, 103297. https://doi.org/10.1016/j.jinteco.2020.103297

Coibion, O., Gorodnichenko, Y., & Weber, M. (2022). Monetary Policy Communications and Their Effects on Household Inflation Expectations. *Journal of Political Economy*, *130*(6), 1537–1584. https://doi.org/10.1086/718982

Cui, Z., Li, N., & Zhou, H. (2025). A large-scale replication of scenario-based experiments in psychology and management using large language models. *Nature Computational Science*, *5*(8), 627–634. https://doi.org/10.1038/s43588-025-00840-7

Curtin, R. T. (1982). Indicators of Consumer Behavior: The University of Michigan Surveys of Consumers. *Public Opinion Quarterly*, *46*(3), 340. https://doi.org/10.1086/268731

D'Acunto, F., Charalambakis, E., Georgarakos, D., Kenny, G., Meyer, J., & Weber, M. (2024). *Household Inflation Expectations: An Overview of Recent Insights for Monetary Policy* (No. w32488; p. w32488). National Bureau of Economic Research. https://doi.org/10.3386/w32488

D'Acunto, F., Malmendier, U., Ospina, J., & Weber, M. (2021). Exposure to Grocery Prices and Inflation Expectations. *Journal of Political Economy*, *129*(5), 1615–1639. https://doi.org/10.1086/713192

D'Acunto, F., Malmendier, U., & Weber, M. (2023). What Do the Data Tell Us About Inflation Expectations? In *Handbook of Economic Expectations*. Academic Press, an imprint of Elsevier. https://doi.org/10.1016/B978-0-12-822927-9.00012-480128229279000124

D'Acunto, F., & Weber, M. (2024). Why Survey-Based Subjective Expectations Are Meaningful and Important. *Annual Review of Economics*, *16*, 329–357.

Dibiasi, A., Mikosch, H., & Sarferaz, S. (2025). Uncertainty Shocks, Adjustment Costs, and Firm Beliefs: Evidence from a Representative Survey. *American Economic Journal: Macroeconomics*, *17*(3), 36–73. https://doi.org/10.1257/mac.20220318

Ehrmann, M., Pfajfar, D., & Santoro, E. (2017). Consumers' Attitudes and Their Inflation Expectations. *International Journal of Central Banking*, *47*(February 2017).

Ehrmann, M., & Wabitsch, A. (2022). Central bank communication with non-experts – A road to nowhere? *Journal of Monetary Economics*, *127*, 69–85. https://doi.org/10.1016/j.jmoneco.2022.02.003

Eliaz, K., & Spiegler, R. (2020). A Model of Competing Narratives. *American Economic Review*, *110*(12), 3786–3816. https://doi.org/10.1257/aer.20191099

Ericsson, K. A., Hoffman, R. R., Kozbelt, A., & Williams, A. M. (Eds.). (2018). *The Cambridge Handbook of Expertise and Expert Performance* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/9781316480748

Fuster, A., & Zafar, B. (2023). Survey experiments on economic expectations. In *Handbook of Economic Expectations* (pp. 107–130). Elsevier. https://doi.org/10.1016/B978-0-12-822927-9.00010-0

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2024). *Retrieval-Augmented Generation for Large Language Models: A Survey* (No. arXiv:2312.10997). arXiv. https://doi.org/10.48550/arXiv.2312.10997

Ge, T., Chan, X., Wang, X., Yu, D., Mi, H., & Yu, D. (2025). *Scaling Synthetic Data Creation with 1,000,000,000 Personas* (No. arXiv:2406.20094). arXiv. https://doi.org/10.48550/arXiv.2406.20094

Gohl, N., Haan, P., Michelsen, C., & Weinhardt, F. (2024). House price expectations. *Journal of Economic Behavior & Organization*, *218*, 379–398. https://doi.org/10.1016/j.jebo.2023.12.015

Goli, A., & Singh, A. (2024). Frontiers: Can Large Language Models Capture Human Preferences? *Marketing Science*, mksc.2023.0306. https://doi.org/10.1287/mksc.2023.0306

Gordon, R., & Dahl, G. B. (2013). Views among Economists: Professional Consensus or Point-Counterpoint? *American Economic Review*, *103*(3), 629–635. https://doi.org/10.1257/aer.103.3.629

Gorodnichenko, Y., Pham, T., & Talavera, O. (2024). Central bank communication on social media: What, to whom, and how? *Journal of Econometrics*, 105869. https://doi.org/10.1016/j.jeconom.2024.105869

Haaland, I., Roth, C., Stantcheva, S., & Wohlfart, J. (2025). Understanding Economic Behavior Using Open-ended Survey Data. *Journal of Economic Literature (Forthcoming)*.

Haaland, I., Roth, C., & Wohlfart, J. (2023). Designing Information Provision Experiments. *Journal of Economic Literature*, *61*(1), 3–40. https://doi.org/10.1257/jel.20211658

Hagendorff, T. (2024). Deception abilities emerged in large language models. *Proceedings of the National Academy of Sciences*, *121*(24), e2317967121. https://doi.org/10.1073/pnas.2317967121

Halterman, A. (2025). Synthetically generated text for supervised text analysis. *Political Analysis*, *33*(3), 181–194. https://doi.org/10.1017/pan.2024.31

Hansen, A. L., Horton, J. J., Kazinnik, S., Puzzello, D., & Zarifhonarvar, A. (2025). Simulating the Survey of Professional Forecasters. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.5066286

Hill, B. (2022). Updating confidence in beliefs. *Journal of Economic Theory*, *199*, 105209. https://doi.org/10.1016/j.jet.2021.105209

Horton, J. (2023). *Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?* (No. w31122; p. w31122). National Bureau of Economic Research. https://doi.org/10.3386/w31122

Jonung, L. (1981). Perceived and Expected Rates of Inflation in Sweden. *American Economic Review*, *71*(5), 961–968.

Korinek, A. (2025). *AI Agents for Economic Research* (No. w34202; p. w34202). National Bureau of Economic Research. https://doi.org/10.3386/w34202

Lamla, M. J., & Maag, T. (2012). The Role of Media for Inflation Forecast Disagreement of Households and Professional Forecasters. *Journal of Money, Credit and Banking*, *44*(7), 1325–1350. https://doi.org/10.1111/j.1538-4616.2012.00534.x

Li, N., Gao, C., Li, M., Li, Y., & Liao, Q. (2024). *EconAgent: Large Language Model-Empowered Agents for Simulating Macroeconomic Activities* (No. arXiv:2310.10436). arXiv. http://arxiv.org/abs/2310.10436

Lucas, R. E. (1972). Expectations and the neutrality of money. *Journal of Economic Theory*, *4*(2), 103–124. https://doi.org/10.1016/0022-0531(72)90142-1

Manski, C. F. (2018). Survey Measurement of Probabilistic Macroeconomic Expectations: Progress and Promise. *NBER Macroeconomics Annual*, *32*, 411–471. https://doi.org/10.1086/696061

Muth, J. F. (1961). Rational Expectations and the Theory of Price Movements. *Econometrica*, *29*(3), 315. https://doi.org/10.2307/1909635

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*(NeurIPS 2022), 27730–27744.

Ouyang, S., Yun, H., & Zheng, X. (2024). *How Ethical Should AI Be? How AI Alignment Shapes the Risk Preferences of LLMs* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2406.01168

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511803161

Sloman, S. A., & Lagnado, D. (2015). Causality in Thought. *Annual Review of Psychology*, *66*(1), 223–247. https://doi.org/10.1146/annurev-psych-010814-015135

Souleles, N. S. (2004). Expectations, Heterogeneous Forecast Errors, and Consumption: Micro Evidence from the Michigan Consumer Sentiment Surveys. *Journal of Money, Credit, and Banking*, *36*(1), 39–72. https://doi.org/10.1353/mcb.2004.0007

Spiegler, R. (2016). Bayesian Networks and Boundedly Rational Expectations *. *The Quarterly Journal of Economics*, *131*(3), 1243–1290. https://doi.org/10.1093/qje/qjw011

Spiegler, R. (2020). Behavioral Implications of Causal Misperceptions. *Annual Review of Economics*, *12*(1), 81–106. https://doi.org/10.1146/annurev-economics-072219-111921

Tranchero, M., Brenninkmeijer, C.-F., Murugan, A., & Nagaraj, A. (2024). *Theorizing with Large Language Models* (No. w33033; p. w33033). National Bureau of Economic Research. https://doi.org/10.3386/w33033

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*(2), 207–232. https://doi.org/10.1016/0010-0285(73)90033-9

Wang, A., Morgenstern, J., & Dickerson, J. P. (2025). Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*. https://doi.org/10.1038/s42256-025-00986-z

Weber, M., D'Acunto, F., Gorodnichenko, Y., & Coibion, O. (2022). The Subjective Inflation Expectations of Households and Firms: Measurement, Determinants, and Implications. *Journal of Economic Perspectives*, *36*(3), 157–184. https://doi.org/10.1257/jep.36.3.157

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). *Emergent Abilities of Large Language Models* (No. arXiv:2206.07682). arXiv. http://arxiv.org/abs/2206.07682

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., … Qiu, Z. (2025). *Qwen3 Technical Report* (No. arXiv:2505.09388). arXiv. https://doi.org/10.48550/arXiv.2505.09388

Yu, Y., Zhuang, Y., Zhang, J., Meng, Y., Ratner, A., Krishna, R., Shen, J., & Zhang, C. (2023). *Large Language Model as Attributed Training Data Generator: A Tale of Diversity and Bias*. 37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Zarifhonarvar, A. (2025). Generating inflation expectations with large language models. *Journal of Monetary Economics*, 103859. https://doi.org/10.1016/j.jmoneco.2025.103859

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., … Wen, J.-R. (2023). *A Survey of Large Language Models* (No. arXiv:2303.18223). arXiv. http://arxiv.org/abs/2303.18223

Supplementary Appendix For "Simulating Macroeconomic Expectations using LLM Agents"

By Jianhao Lin, Lexuan Sun, Yixin Yan

Content

A Questionnaires

A.1 Hypothetical Vignette Experiments

Oil price vignette:

Introduction

The following scenarios deal with the price of crude oil. In the last week, the price of one barrel of crude oil averaged $54.

Baseline Scenario

We would like you to think about the following hypothetical scenario.
Scenario 1: Oil price stays constant
Imagine that the average price of crude oil stays constant over the next 12 months. That is, on average, the price of oil over the next 12 months will be the same as the current price.
Under this scenario, what do you think the US unemployment rate will be 12 months from now? (Note: The current level of the unemployment rate: 4.0%.)
Unemployment rate (Scenario 1): ___%.
Under this scenario, what do you think the US inflation rate will be over the next 12 months? (Note: The current level of the inflation rate: 1.6%.)
Inflation rate (Scenario 1): ___%.

Fall scenario

We would like you to think about the following hypothetical scenario.
Scenario 2: Oil price falls
Imagine the average price of crude oil unexpectedly falls due to improvements in the local production technology in the Middle East. On average, the price will be $30 lower for the next 12 months than the current price. That is, the price will be on average $24 for the next 12 months.
Under this scenario, what do you think the US unemployment rate will be 12 months from now? (Note: The current level of the unemployment rate: 4.0%.)
Unemployment rate (Scenario 2): ___%.
Under this scenario, what do you think the US inflation rate will be over the next 12 months? (Note: The current level of the inflation rate: 1.6%.)
Inflation rate (Scenario 2): ___%.
Above, you predict how the change in the Scenario 2 affects the US economy. Please tell us how you come up with your predictions. What are your main considerations in making those predictions?
Main considerations (Scenario 2): ___.

Rise scenario[1]

```
We would like you to think about the following hypothetical scenario.
Scenario 2: Oil price rises
Imagine the average price of crude oil unexpectedly rises due to a problem with the l
ocal production technology in the Middle East. On average, the price will be $30 high
er for the next 12 months than the current price. That is, the price will be on avera
ge $84 for the next 12 months.
Under this scenario, what do you think the US unemployment rate will be 12 months fro
m now? (Note: The current level of the unemployment rate: 4.0%.)
Unemployment rate (Scenario 2): ___%.
Under this scenario, what do you think the US inflation rate will be over the next 12
 months? (Note: The current level of the inflation rate: 1.6%.)
Inflation rate (Scenario 2): ___%.
Above, you predict how the change in the Scenario 2 affects the US economy. Please te
ll us how you come up with your predictions. What are your main considerations in mak
ing those predictions?
Main considerations (Scenario 2): ___.
The following statements describe different thoughts you might have had on your mind
while making your predictions for the Scenario 2. Did you have any of these thoughts
on your mind? Please select all that you had on your mind.
A. Due to lower incomes or job loss, households cut back on their spending.
B. Businesses face lower demand for their products, so they increase their product pr
ices to keep profits at the same level.
C. The higher cost of oil makes it more attractive to use alternative energy sources
and energy-saving technologies, which leads to job creation.
D. To make up for the higher cost of production, businesses reduce their workforce.
E. Because higher product prices lower their purchasing power, households cut back on
 their spending.
F. To make up for the higher cost of production, businesses increase product prices.
G. The US oil extraction industry profits from the higher oil price, which leads to j
ob creation.
H. None of the above.
Your selected options: ___.
```

Government spending vignette:

Introduction

```
The following scenarios deal with yearly federal government spending. In the 2018 fin
ancial year, the federal government spent roughly $4.2 trillion on diverse issues suc
h as social security, health, military, or education. This amounts to roughly 1/3 of
```

[1] In this experiment, to simplify the analysis, Andre et al. (2022) uses structured questions only in the rise scenario to investigate respondents' thoughts of propagation channels for each vignette. Our design aligns with their approach.

the value of all final goods and services produced by the US economy in one year (kno wn as the gross domestic product).
Government spending typically increases every year, reflecting the general growth of the economy. For the last 50 years, for instance, it increased by an average of 2.9% each year.

Baseline Scenario

We would like you to think about the following hypothetical scenario.
Scenario 1: Government spending grows as usual
Imagine federal government spending grows as usual over the next 12 months. That is, it grows at a rate that equals the usual growth that took place in the previous year s.
Under this scenario, what do you think the US unemployment rate will be 12 months fro m now? (Note: The current level of the unemployment rate: 4.0%.)
Unemployment rate (Scenario 1): ___%.
Under this scenario, what do you think the US inflation rate will be over the next 12 months? (Note: The current level of the inflation rate: 1.6%.)
Inflation rate (Scenario 1): ___%.

Fall scenario

We would like you to think about the following hypothetical scenario.
Scenario 2: Government spending grows less
Imagine federal government spending unexpectedly grows to a smaller extent than usual over the next 12 months due to cuts in spending on defense. In particular, it grows by 2.4 percentage points less than the usual growth that took place in the previous y ears.
The government announces: The change is temporary and occurs despite no changes in th e government's assessment of national security or economic conditions. Moreover, fede ral taxes do not change in response to the spending cut.
Under this scenario, what do you think the US unemployment rate will be 12 months fro m now? (Note: The current level of the unemployment rate: 4.0%.)
Unemployment rate (Scenario 2): ___%.
Under this scenario, what do you think the US inflation rate will be over the next 12 months? (Note: The current level of the inflation rate: 1.6%.)
Inflation rate (Scenario 2): ___%.
Above, you predict how the change in the Scenario 2 affects the US economy. Please te ll us how you come up with your predictions. What are your main considerations in mak ing those predictions?
Main considerations (Scenario 2): ___.

Rise scenario

We would like you to think about the following hypothetical scenario.
Scenario 2: Government spending grows more

```
Imagine federal government spending unexpectedly grows to a larger extent than usual
over the next 12 months due to a newly announced spending program on defense. In part
icular, it grows by 2.4 percentage points more than the usual growth that took place
in the previous years.
The government announces: The change is temporary and occurs despite no changes in th
e government's assessment of national security or economic conditions. Moreover, fede
ral taxes do not change in response to the spending program.
Under this scenario, what do you think the US unemployment rate will be 12 months fro
m now? (Note: The current level of the unemployment rate: 4.0%.)
Unemployment rate (Scenario 2): ___%.
Under this scenario, what do you think the US inflation rate will be over the next 12
 months? (Note: The current level of the inflation rate: 1.6%.)
Inflation rate (Scenario 2): ___%.
Above, you predict how the change in the Scenario 2 affects the US economy. Please te
ll us how you come up with your predictions. What are your main considerations in mak
ing those predictions?
Main considerations (Scenario 2): ___.
The following statements describe different thoughts you might have had on your mind
while making your predictions for the Scenario 2. Did you have any of these thoughts
on your mind? Please select all that you had on your mind.
A. Because of higher incomes, households increase their spending.
B. Because there is more demand for their products, businesses increase their product
 prices.
C. Businesses face lower demand for their products, so they increase their product pr
ices to keep profits at the same level.
D. Households expect to pay higher taxes in the future, which may be needed to pay ba
ck the new government debt. Therefore, households work more.
E. Households expect to pay higher taxes in the future, which may be needed to pay ba
ck the new government debt. Therefore, households cut back on their spending.
F. Because there is more demand for their products, businesses increase their workfor
ce.
G. To help the government finance the additional spending, the central bank prints mo
ney.
H. None of the above.
Your selected options: ___.
```

Interest rate vignette:

Introduction

```
The following scenarios deal with the federal funds target rate. This is the most imp
ortant interest rate in the economy. The value of the rate influences how "costly" it
 is for banks to acquire money, thereby influencing interest rates on other important
 financial products such as savings accounts, consumer loans, mortgages, or loans to
firms.
```

The federal funds target rate is the interest rate frequently discussed in the news. It is set by the Federal Open Market Committee (FOMC), which normally meets eight times a year. Currently, the rate is 2.5%.

Baseline Scenario

We would like you to think about the following hypothetical scenario.

Scenario 1: Federal funds target rate stays constant

Imagine the federal funds target rate stays constant. That is, in its next meeting, the Federal Open Market Committee announces that it will keep the rate constant.

Imagine the committee announces it does so with no changes in their assessment of the economic conditions.

Under this scenario, what do you think the US unemployment rate will be 12 months from now? (Note: The current level of the unemployment rate: 4.0%.)

Unemployment rate (Scenario 1): ___%.

Under this scenario, what do you think the US inflation rate will be over the next 12 months? (Note: The current level of the inflation rate: 1.6%.)

Inflation rate (Scenario 1): ___%.

Fall scenario

We would like you to think about the following hypothetical scenario.

Scenario 2: Federal funds target rate falls

Imagine the federal funds target rate is unexpectedly 0.5 percentage points lower. That is, in its next meeting, the Federal Open Market Committee announces that it is reducing the rate from 2.5% to 2%.

Imagine the committee announces it does so with no changes in their assessment of the economic conditions.

Under this scenario, what do you think the US unemployment rate will be 12 months from now? (Note: The current level of the unemployment rate: 4.0%.)

Unemployment rate (Scenario 2): ___%.

Under this scenario, what do you think the US inflation rate will be over the next 12 months? (Note: The current level of the inflation rate: 1.6%.)

Inflation rate (Scenario 2): ___%.

Above, you predict how the change in the Scenario 2 affects the US economy. Please tell us how you come up with your predictions. What are your main considerations in making those predictions?

Main considerations (Scenario 2): ___.

Rise scenario

We would like you to think about the following hypothetical scenario.

Scenario 2: Federal funds target rate rises

Imagine the federal funds target rate is unexpectedly 0.5 percentage points higher. That is, in its next meeting, the Federal Open Market Committee announces that it is raising the rate from 2.5% to 3%.

```
Imagine the committee announces it does so with no changes in their assessment of the
 economic conditions.
Under this scenario, what do you think the US unemployment rate will be 12 months fro
m now? (Note: The current level of the unemployment rate: 4.0%.)
Unemployment rate (Scenario 2): ___%.
Under this scenario, what do you think the US inflation rate will be over the next 12
 months? (Note: The current level of the inflation rate: 1.6%.)
Inflation rate (Scenario 2): ___%.
Above, you predict how the change in the Scenario 2 affects the US economy. Please te
ll us how you come up with your predictions. What are your main considerations in mak
ing those predictions?
Main considerations (Scenario 2): ___.

The following statements describe different thoughts you might have had on your mind
while making your predictions for the Scenario 2. Did you have any of these thoughts
on your mind? Please select all that you had on your mind.
A. Because there is less demand for their products, businesses reduce their workforc
e.
B. Businesses face lower demand for their products, so they increase their product pr
ices to keep profits at the same level.
C. To make up for the higher cost of borrowing, businesses reduce their workforce.
D. Because there is less demand for their products, businesses reduce their product p
rices.
E. Because higher interest rates make it more attractive to save and less attractive
to borrow, households cut back on their spending.
F. Due to the higher cost of borrowing, businesses pursue fewer investment projects.
G. To make up for the higher cost of borrowing, businesses increase product prices.
H. Because of lower incomes or job loss, households cut back on their spending.
I. None of the above.
Your selected options: ___.
```

Taxation vignette:

Introduction

```
The following scenarios deal with the income tax rates in the US. The tax rates speci
fy the percentage of their income that households need to pay to the federal governme
nt. At present, a typical household pays 21.1 percent of its income to the federal go
vernment in taxes.
```

Baseline Scenario

```
We would like you to think about the following hypothetical scenario.
Scenario 1: Income tax rates stay constant
```

```
Imagine that income tax rates stay constant for all US citizens over the next 12 mont
hs.
Under this scenario, what do you think the US unemployment rate will be 12 months fro
m now? (Note: The current level of the unemployment rate: 4.0%.)
Unemployment rate (Scenario 1): ___%.
Under this scenario, what do you think the US inflation rate will be over the next 12
 months? (Note: The current level of the inflation rate: 1.6%.)
Inflation rate (Scenario 1): ___%.
```

Fall scenario

```
We would like you to think about the following hypothetical scenario.
Scenario 2: Income tax rates decrease
Imagine that income tax rates are 1 percentage point lower for all US citizens over t
he next 12 months. This means that the typical US household would pay about $400 less
 in taxes.
The government announces: The tax change is temporary and occurs despite no changes i
n the government's assessment of the economic conditions. Moreover, government spendi
ng does not change in response to the tax cut.
Under this scenario, what do you think the US unemployment rate will be 12 months fro
m now? (Note: The current level of the unemployment rate: 4.0%.)
Unemployment rate (Scenario 2): ___%.
Under this scenario, what do you think the US inflation rate will be over the next 12
 months? (Note: The current level of the inflation rate: 1.6%.)
Inflation rate (Scenario 2): ___%.
Above, you predict how the change in the Scenario 2 affects the US economy. Please te
ll us how you come up with your predictions. What are your main considerations in mak
ing those predictions?
Main considerations (Scenario 2): ___.
```

Rise scenario

```
We would like you to think about the following hypothetical scenario.
Scenario 2: Income tax rates increase
Imagine that income tax rates are 1 percentage point higher for all US citizens over
the next 12 months. This means that the typical US household would pay about $400 mor
e in taxes.
The government announces: The tax change is temporary and occurs despite no changes i
n the government's assessment of the economic conditions. Moreover, government spendi
ng does not change in response to the tax increase.
Under this scenario, what do you think the US unemployment rate will be 12 months fro
m now? (Note: The current level of the unemployment rate: 4.0%.)
Unemployment rate (Scenario 2): ___%.
Under this scenario, what do you think the US inflation rate will be over the next 12
 months? (Note: The current level of the inflation rate: 1.6%.)
```

```
Inflation rate (Scenario 2): ___%.
Above, you predict how the change in the Scenario 2 affects the US economy. Please te
ll us how you come up with your predictions. What are your main considerations in mak
ing those predictions?
Main considerations (Scenario 2): ___.

The following statements describe different thoughts you might have had on your mind
while making your predictions for the Scenario 2. Did you have any of these thoughts
on your mind? Please select all that you had on your mind.
A. Because workers demand higher wages to make up for the higher income taxes, busine
sses reduce their workforce.
B. Because workers demand higher wages to make up for the higher income taxes, busine
sses increase their product prices.
C. Because there is less demand for their products, businesses reduce their product p
rices.
D. Because of lower disposable incomes, households cut back on their spending.
E. Because higher taxes make it less attractive to work, households work less.
F. Because there is less demand for their products, businesses reduce their workforc
e.
G. Businesses face lower demand for their products, so they increase their product pr
ices to keep profits at the same level.
H. To make up for their reduced disposable incomes, households work more.
I. None of the above.
Your selected options: ___.
```

A.2 Information Provision Experiments

Sub-Experiment 1:

Shown only to respondents in the "**high forecast**" treatment:

```
We would like to provide you with a forecast of home price growth from an expert who
regularly participates in the Economic Expert Survey, an expert survey on macroeconom
ic forecasts.
According to this expert forecast, the average annual growth rate of home prices in t
he US over the next ten years will be 6 percent.
In the case where home prices increase by 6 percent in each of the next ten years, th
is would mean that a home worth $100,000 today will be worth about $179,085 in ten ye
ars from now.
```

Shown only to respondents in the "**low forecast**" treatment:

```
We would like to provide you with a forecast of home price growth from an expert who
regularly participates in the Economic Expert Survey, an expert survey on macroeconom
ic forecasts.
```

```
According to this expert forecast, the average annual growth rate of home prices in t
he US over the next ten years will be 2 percent.
In the case where home prices increase by 2 percent in each of the next ten years, th
is would mean that a home worth $100,000 today will be worth about $121,899 in ten ye
ars from now.
```

Shown to **all respondents** (The instructions in the remainder of the survey are identical across treatment arms from now on):

```
We now would like to provide you with a forecast of inflation from an expert who regu
larly participates in the Survey of Professional Forecasters. According to this exper
t forecast, the average annual rate of inflation in the US over the next ten years wi
ll be 2 percent.
```

Posterior home price expectations:

```
In this question we present you with 14 possible scenarios for the average annual gro
wth rate of the value of a typical home in the US, over the next ten years.
Please let us know how likely you think it is that each scenario will occur. Please t
ype in the number to indicate the probability, in percent, that you attach to each sc
enario. The probabilities of the 14 scenarios have to sum up to 100 percent.
The average annual growth rate of the value of a typical home in the US over the next
 ten years will be. . .
• Scenario 1: . . . more than 20 percent. ____ percent.
• Scenario 2: . . . between 15 and 20 percent. ____ percent.
• Scenario 3: . . . between 10 and 15 percent. ____ percent.
• Scenario 4: . . . between 7.5 and 10 percent. ____ percent.
• Scenario 5: . . . between 5 and 7.5 percent. ____ percent.
• Scenario 6: . . . between 2.5 and 5 percent. ____ percent.
• Scenario 7: . . . between 0 and 2.5 percent. ____ percent.
• Scenario 8: . . . between 0 and -2.5 percent. ____ percent.
• Scenario 9: . . . between -2.5 and -5 percent. ____ percent.
• Scenario 10: . . . between -5 and -7.5 percent. ____ percent.
• Scenario 11: . . . between -7.5 and -10 percent. ____ percent.
• Scenario 12: . . . between -10 and -15 percent. ____ percent.
• Scenario 13: . . . between -15 and -20 percent. ____ percent.
• Scenario 14: . . . less than -20 percent. ____ percent.
Total: 100 percent.
```

Sub-Experiment 2:

```
Imagine you expect home prices to grow by 1.5% per year over the next 10 years. Now i
magine that you increase your expectations about future home prices. You now expect h
ome prices to increase by 6% per year over the next 10 years. How would this change i
```

```
n your expectations about future home prices affect your expectations about your hous
ehold's future economic situation?
A. My household's future economic situation would improve because of this change.
B. My household's future economic situation would be unaffected by this change.
C. My household's future economic situation would worsen because of this change.
Your selected option: ___.
Please explain why. Respond in full sentences.
Explanation: ___.
```

A.3 Large-Scale Household Expectations Survey

Questions related to **inflation expectations**:

```
This is the questionnaire for the survey conducted in 2025. The questions are as foll
ows:
Q1: During the next 12 months, do you think that prices in general will go up, or go
down, or stay where they are now?
1. Go up
2. Go up (at same rate)
3. Same
5. Go down
Your selected option: ___.

Q2: By about what percent do you expect prices to go (up/down) on the average, during
 the next 12 months?
Your expectation: ___ percent.

Q3: What about the outlook for prices over the next 5 to 10 years?  Do you think pric
es will be higher, about the same, or lower, 5 to 10 years from now?
1. Go up
2. Go up (at same rate)
3. Same
5. Go down
Your selected option: ___.

Q4: By about what percent per year do you expect prices to go (up/down) on the averag
e, during the next 5 to 10 years?
Your expectation: ___ percent.

Q5: Please explain why. Respond in full sentences.
Explanation: ___.
```

Questions related to **home price expectations**:

```
This is the questionnaire for the survey conducted in 2025. The questions are as foll
ows:
Q1: What do you think will happen to the prices of homes like yours in your community
 over the next 12 months? Will they increase, remain about the same, or decrease?
1. Increase
3. About the same
5. Decrease
Your selected option: ___.

Q2: By about what percent do you expect prices of homes like yours in your community
to go (up/down), on the average, over the next 12 months?
Your expectation: ___ percent.

Q3: What about the outlook for prices of homes like yours in your community over the
next 5 years or so? Do you expect them to increase, remain about the same, or decreas
e?
1. Increase
3. Remain about the same
5. Decrease
Your selected option: ___.

Q4: By about what percent per year do you expect prices of homes like yours in your c
ommunity to go (up/down), on average, over the next 5 years or so?
Your expectation: ___ percent.

Q5: Please explain why. Respond in full sentences.
Explanation: ___.
```

B Prompts Design for LLM Agents

B.1 Hypothetical Vignette Experiments

Household Agents

Personal Characteristics Module:

```
You are a (an) {AGE}-year-old {SEX} who is {MARRY} and lives in the {REGION} region o
f the US. Your educational background is: {EDUC}, which is a {EDUC_LEVEL} level of ed
ucation in the US. Your political affiliation is {POLAFF} (Note: Currently the Republ
ican Party is in power). Your income is ${INCOME} per year, which is a {INCOME_LEVEL}
```

```
 level of income in the US. You buy/own a house with a {HOME_VALUE_LEVEL} market valu
e of ${HOMEAMT}.
```

Notes: This prompt is primarily used to assign personal characteristics to Household Agents within the Personal Characteristics Module in the hypothetical vignette experiments. All variable data are sourced from the Michigan Survey of Consumers. The specific variables are defined as follows: {AGE} is a continuous variable representing the respondents' age. {SEX} is a categorical variable indicating gender, where 1 denotes man and 2 denotes woman. {MARRY} is a categorical variable capturing marital status, where 1 denotes married, 2 separated, 3 divorced, 4 widowed, and 5 never married. {REGION} is a categorical variable indicating the respondents' region of residence, where 1 denotes Western, 2 North Central, 3 Northeastern, and 4 Southern. {EDUC} is a categorical variable representing educational background, where 1 denotes Grade 0–8 without a high school diploma, 2 denotes Grade 9–12 without a high school diploma, 3 denotes Grade 0–12 with a high school diploma, 4 denotes Grade 13–17 without a college degree, 5 denotes Grade 13–16 with a college degree, and 6 denotes Grade 17 with a college degree. {EDUC_LEVEL} (with values identical to {EDUC}) is a categorical variable representing educational level, where 1 or 2 denotes very low[2], 3 low, 4 middle, 5 high, 6 very high. {POLAFF} is a categorical variable reflecting political affiliation, where 1 denotes Republican, 2 Democrat, 3 Independent (closer to Republican), 4 Independent (closer to Democrat), and 5 Independent (no preference). {INCOME} is a continuous variable representing the respondent's total income in the previous year (in U.S. dollars). {INCOME_LEVEL} (i.e., YTL5 in the MSC) is a categorical variable reflecting the respondents' income levels, where 1 denotes very low, 2 low, 3 middle, 4 high, and 5 very high. {HOMEAMT} is a continuous variable indicating the current market values of the respondents' homes (in U.S. dollars). {HOME_VALUE_LEVEL} (i.e., HTL5 in the MSC) is a categorical variable representing the respondents' housing value levels, where 1 denotes very low, 2 low, 3 middle, 4 high, and 5 very high. Detailed descriptions of these variables can be found in the Codebook of the Michigan Survey of Consumers.

Prior Expectations & Perceptions Module:

While some studies suggest that personal experiences (e.g., shopping, job searching, or news consumption) may shape households' macroeconomic expectations (Malmendier & Nagel, 2016; Kuchler & Zafar, 2019; Coibion et al., 2020), it is challenging to capture such individual experiences directly. To address this, we embed contextual personal experiences into each Household Agent in PEPM using narrative-style prompts. Examples include: "When you purchase goods at the supermarket, shop online, or interact with your family and friends…" and "When you look for a job or obtain news information from newspapers, television, etc…" The wording of the remaining prompt content aligns with the corresponding questions and options in the MSC survey to mitigate potential subjectivity. The specific prompt is as follows:

```
When you purchase goods at the supermarket, shop online, or interact with your family
 and friends, you form certain beliefs about the prices of goods: During the next 12
months, you think that prices in general will {PX1Q1}.
```

[2] Respondents without a high school diploma account for a small proportion of the MSC sample, especially those with EDUC equal to 1; therefore, we classify individuals with EDUC values of 1 and 2 as belonging to the category of 'very low' educational level.

```
When you look for a job or obtain news information from newspapers, television, etc.,
 you form certain beliefs about unemployment: During the coming 12 months, you think
that unemployment will {UNEMP}.
```

Notes: This prompt is primarily used to assign Household Agents their corresponding prior expectations within the Prior Expectations & Perceptions Module in the hypothetical vignette experiments. All variable data are sourced from the Michigan Survey of Consumers. The specific variables are defined as follows: {PX1Q1} is a categorical variable indicating the respondents' perceived overall trends in prices over the next 12 months, where 1 denotes go up, 2 go up at the same rate as now, 3 be unchanged, and 5 go down. {UNEMP} is a categorical variable indicating the respondents' perceived trends in unemployment over the next 12 months, where 1 denotes the unemployment will be more than now, 3 be about the same as it is now, and 5 be less than now. During sampling, we ensure that no category of each expectation variable contains too few observations, thereby maintaining balanced priors. Detailed descriptions of these variables can be found in the Codebook of the Michigan Survey of Consumers.

Social Media Information Module:

```
While browsing on your phone or computer, you randomly come across the following twee
ts:
Tweet 1: {US_unemployment_tweet}
Tweet 2: {US_inflation_tweet}
```

Notes: This prompt is primarily used to assign social media information to Household Agents within the Social Media Information Module in the hypothetical vignette experiments. {US_unemployment_tweet} and {US_inflation_tweet} represent a single tweet automatically and randomly captured related to the topics of "US Unemployment" and "US Inflation", respectively.

Initialization Prompts:

Here are the initialization prompts designed for Household Agents. The text in orange defines the agent's role type and level of confidence, while the blue text specifies the agent's tasks, with wording that aligns closely with the introductory explanations from the Wave 3 survey questionnaire in Andre et al. (2022). The black text objectively outlines the module usage rules that Household Agents should follow, which is consistent with models and theories of belief updating for economic agents.

```
Suppose you are an ordinary individual (household) with {CONF} (on a five-level scale
 ranging from extremely weak, weak, moderate, strong, to extremely strong) confidence
 designed to participate in a survey involving your beliefs about the future developm
ent of the US economy, in particular the unemployment rate and the inflation rate. Yo
ur task will be to estimate the development of both rates in hypothetical scenarios.
Please give us your guess about how the unemployment rate and the inflation rate in t
he US economy would actually develop under the scenarios considered. This may or may
not be in line with theoretical findings and evidence from economics. We are only int
erested in your own views and opinions on the US economy.
```

```
IMPORTANT INSTRUCTIONS: Your responses should trade off among the various pieces of i
nformation mentioned above in accordance with your level of confidence: If you are co
nfident, your answers will rely on Prior Expectations & Perceptions (if not provided,
 please ignore), and will not be influenced by other information, such as the Social
Media Information (if not provided, please ignore). On the other hand, if you lack co
nfidence, your answers are more likely to be influenced by other information. In addi
tion, your responses should fully reflect the Personal Characteristics (such as age,
gender, educational level, political affiliation, etc.) (if not provided, please igno
re) of the role you are portraying.
```

Notes: This prompt is primarily used to initialize Household Agents by defining their role types, specific tasks, and the rules for module usage in the hypothetical vignette experiments. {CONF} is a categorical variable indicating the households' levels of confidence, where 1 represents extremely weak, 2 represents weak, 3 represents moderate, 4 represents strong, and 5 represents extremely strong.

Expert Agents

Professional Background Module:

In the PBM, the two data-processing agents—namely, the Data Organization Agent and the Synthetic Data Generation Agent—require clearly defined prompts to guide their operation. The specific prompts for each agent are presented below.

- **Data Organization Agent:**

```
Assume you are a senior research assistant. Please help organize the expert's curricu
lum vitae information, drawn from both the official website and the individual's pers
onal LinkedIn profile, as follows:

</Official website information>
{Excel Text}
</Official website information>

<LinkedIn profile>
{PDF Text}
</LinkedIn profile>

The consolidated output must satisfy the following requirements:
(1) Completely conceal the expert's name and consistently employ second person ("yo
u") throughout.
(2) Ensure that the paragraph includes at least the expert's key details, such as com
pany/organization, work experience, position, research areas, and educational backgro
und.
```

```
(3) Remove all temporal information including dates, years, timeframes, and any time-related references from the output.
(4) Summarize and synthesize all of the above information into a single, cohesive paragraph of approximately 500 words before presenting it.
(5) Do not generate any irrelevant content; provide only the final result directly.
```

Notes: This prompt instructs the Data Organization Agent to process and standardize the textual data from experts' profiles obtained via official websites and LinkedIn, producing outputs with uniform formatting and controlled length. The placeholder {Excel Text} is filled with information from official websites, while the placeholder {PDF Text} is filled with LinkedIn profiles.

- **Synthetic Data Generation Agent:**

```
Here are the names and corresponding personal information of some U.S. economics experts working in industry, academia, and policy institutions (presented in the second person and without time information):
{Experts' Profiles}
Please replicate the style to generate 10 additional heterogeneous economics experts (working in industry, academia, and policy institutions), each with name and approximately 500 words of personal information, and output strictly in the following JSON format (no additional text or explanation):
{
  "expert_name_1": "personal_information_1",
  "expert_name_2": "personal_information_2",
  "expert_name_3": "personal_information_3",
  "expert_name_4": "personal_information_4",
  "expert_name_5": "personal_information_5",
  "expert_name_6": "personal_information_6",
  "expert_name_7": "personal_information_7",
  "expert_name_8": "personal_information_8",
  "expert_name_9": "personal_information_9",
  "expert_name_10": "personal_information_10"
}
```

Notes: This prompt instructs the Synthetic Data Generation Agent to produce synthetic expert profile data. The placeholder {Experts' Profiles} is filled with the authentic expert profiles output by the Data Organization Agent, provided in JSON format as key-value pairs of expert names and their profiles. In each execution, the agent generates 10 heterogeneous synthetic samples. Multiple runs can be performed to achieve the target sample size.

Prior Expectations & Perceptions Module:

```
Based on your extensive knowledge of economics and sophisticated data analysis, you have made the following forecasts regarding the future trends of economic indicators:
1.You expect the Personal Consumption Expenditures (PCE) Price Index to {PCE} over the next 12 months.
2.You expect the civilian unemployment rate to {UNEMP} 12 months from now.
```

Notes: This prompt is primarily used to assign Expert Agents their corresponding prior expectations within the Prior Expectations & Perceptions Module in the hypothetical vignette experiments. All variable data are sourced from the Survey of Professional Forecasters. Specifically, {PCE} is a categorical variable indicating the predicted trends of the Personal Consumption Expenditures Price Index over the next 12 months, where 1 represents go up, 2 represents stay constant, and 3 represents go down. {UNEMP} is a categorical variable indicating the predicted trends of the unemployment rate over the next 12 months, where 1 represents go up, 2 represents stay constant, and 3 represents go down.

Knowledge Acquisition Module:

The Query Generation Agent in the KAM requires a clearly defined prompt to guide its operation. The specific prompt is presented below.

```
You are an economics expert skilled in using search tools to retrieve and compile com
prehensive professional information and the latest data from the internet. Your goal
is to create heterogeneous search queries that will help gather comprehensive informa
tion relevant to specific questionnaire questions.

Your Personal Information:
Name: {name}
Professional Background: {background}

questionnaire questions: "{questionnaire}"

Shock Type: {shock}

Task: Generate five search queries that are entirely distinct from one another but re
flect your strongest interests, to be used when investigating the economic impact of
the specified shock and searching for related information. These queries will be desi
gned to search for relevant economic data, analysis, and insights.

Requirements:
1. The queries must be highly relevant to your professional background, particularly
your research areas, expertise, work experience, current position, and affiliated org
anization.
2. The queries should comprehensively cover different aspects of how the {shock} affe
cts the US economy, particularly the unemployment and inflation rates.
3. The five queries must be heterogeneous and cover diverse analytical dimensions (in
cluding but not limited to):
   - Macroeconomic indicators (unemployment, inflation, GDP, CPI, consumption, etc.)
   - Industry-specific impacts and sectoral analysis
   - Policy responses, government measures, and regulatory effects
   - Historical data, trend analysis, and cyclical patterns
   - Expert opinions, academic papers, and forecasting reports
   - Market reactions, financial indicators, and investment implications
   - International comparisons, global impacts, and trade effects
```

```
   - Regional variations, state-level differences, and local economic impacts
   - Short-term immediate effects vs. long-term structural changes
   - Quantitative data analysis vs. qualitative expert assessments

4. Query specifications:
   - Length: 5 keywords per query
   - Format: Suitable for search engines (Google, academic databases (such as web of
science), economic data sources, etc.)
   - Specificity: Concrete and actionable, likely to return valuable results
   - Uniqueness: Avoid repetitive or overly similar queries (the same applies to keyw
ords).
   - Time relevance: Include various temporal perspectives (immediate, quarterly, ann
ual trends)

5. Query diversity should include:
   - Statistical data queries and trend analysis
   - Expert opinions and commentaries
   - Academic research (papers) and think tank reports
   - Policy analysis and government response
   - Historical precedent and comparative studies
   - Industry and sector-specific impacts
   - Consumer and business sentiment
   - Financial market reactions
   - Regional and demographic variations
Note: (1) If you have published academic papers, consider adding some keywords relate
d to "academic research" or "academic papers" in your queries. (2) Do not include you
r name in any of the keywords in your search queries. (3) When generating queries, pl
ease note that the current year is assumed to be {year}.
Please output exactly five queries in the following JSON format (no additional text o
r explanation):
{
  "shock_type": "{shock}",
  "expert_name": "{name}",
  "queries": [
    [
      "query1_keyword1",
      "query1_keyword2",
      "query1_keyword3",
      "query1_keyword4",
      "query1_keyword5"
    ],
    [
      "query2_keyword1",
      "query2_keyword2",
```

```
      "query2_keyword3",
      "query2_keyword4",
      "query2_keyword5"
    ],
    [
      "query3_keyword1",
      "query3_keyword2",
      "query3_keyword3",
      "query3_keyword4",
      "query3_keyword5"
    ],
    [
      "query4_keyword1",
      "query4_keyword2",
      "query4_keyword3",
      "query4_keyword4",
      "query4_keyword5"
    ],
    [
      "query5_keyword1",
      "query5_keyword2",
      "query5_keyword3",
      "query5_keyword4",
      "query5_keyword5"
    ]
  ]
}
```

Notes: This prompt instructs the Query Generation Agent to generate personalized and heterogeneous web search queries for each Expert Agent in the hypothetical vignette experiments, based on their respective professional backgrounds and the assigned questionnaire. The placeholders {name} and {background} are to be filled with the expert's name and corresponding profile from the PBM-generated semi-synthetic dataset, while {questionnaire} is to be replaced with the main body of the questionnaire that the Expert Agent needs to address. The placeholder {shock} is reserved for hypothetical vignette experiments and should be filled with the type of economic shock involved (e.g., oil price shock, government spending shock). For experimental types not involving economic shocks, the text containing {shock} should be omitted. The placeholder{year} is filled with the year of the experiment.

Initialization Prompts:

Here are the initialization prompts designed for Expert Agents. This prompt differs from that of the Household Agents solely in the role type and the modules referenced in the module usage rules, with all other content being identical. The specific distinctions (text highlighted in red) are as follows.

```
Suppose you are an expert (professional economic forecaster) …… [The omitted content
is identical to the corresponding content in the initialization prompts for Household
 Agents.]
IMPORTANT INSTRUCTIONS: Your responses should trade off among the various pieces of i
nformation mentioned above in accordance with your level of confidence: If you are co
nfident, your answers will rely on Prior Expectations & Perceptions (if not provided,
 please ignore), and will not be influenced by other information, such as the Knowled
ge or Latest Information (if not provided, please ignore). On the other hand, if you
lack confidence, your answers are more likely to be influenced by other information.
In addition, your responses should fully reflect the Professional Background (such as
 research areas, expertise, work experience, current position, and affiliated organiz
ation, etc.) (if not provided, please ignore) of the role you are portraying.
```

Notes: This prompt is primarily used to initialize Expert Agents by defining their role types, specific tasks, and the rules for module usage in the hypothetical vignette experiments.

B.2 Information Provision Experiments

Sub-Experiment 1:

- **Homeowner Agents:**

Personal Characteristics Module:

```
You are a (an) {age}-year-old {gender} who lives in the {state} State in the US. Your
 educational background is: {educ}, which is a level in Tier {educ_level} (with tiers
 1-8 representing extremely low to extremely high, respectively). Your employment sta
tus is: {employment}. Your household's gross household income last year was in the ra
nge of: {income}, which is a level in Tier {inc_level} (with tiers 1-17 representing
extremely low to extremely high, respectively). You {liquidsavings} liquid savings. Y
our household consists of {household_size} member(s), including {children} children
(child). You own a home comprising {rooms_current} room(s) and occupying a total of
{sqft_current} square feet.
```

Notes: This prompt is primarily used to assign personal characteristics to Homeowner Agents within the Personal Characteristics Module in the Sub-Experiment 1. All variable data are sourced from the survey in Chopra et al. (2025). The specific variables are defined as follows: {age} is a continuous variable representing the respondents' age. {gender} is a categorical variable indicating gender, where 1 denotes male and 2 denotes female. {state} is a continuous variable representing the states of the US. {educ} is a string variable indicating eight distinct levels of educational attainment. {educ_level} is a categorical variable derived from {educ}, with values from 1 to 8 corresponding to the lowest to highest education levels in {educ}. {employment} is a string variable reflecting the current employment status of the homeowner. {income} is a string variable representing the range of the homeowner's total household income from the previous year. {inc_level} is a categorical variable generated from {income}, ranging from 1 to 17, corresponding to the lowest to highest income ranges in {income}. {liquidsavings} is a categorical variable indicating whether the homeowner has liquid savings, where 1 denotes "don't have" and 2 denotes "have". {household_size} is a continuous variable representing the number of individuals in the homeowner's household. {children} is a continuous variable indicating the number of children in the homeowner's household. {rooms_current} is a continuous variable reflecting the number of rooms in the homeowner's home. {sqft_current} is a continuous variable representing the floor area of the homeowner's home. The phrasing used in this prompt draws on the wording of questions and response options from the questionnaire in Chopra et al. (2025).

Prior Expectations & Perceptions Module:

```
When you obtain news information from newspapers, television, etc., or interact with
your family and friends, you form certain beliefs about home price: You expect the av
erage annual growth rate of the value of a typical home in the US to be {prior}% over
 the next ten years. Note: This average annual growth rate of home prices is the chan
ge in value, in percent, that you expect each year on average over the next ten year
s.
```

Notes: This prompt is primarily used to assign Homeowner Agents their corresponding prior expectations within the Prior Expectations & Perceptions Module in the Sub-Experiment 1. All variable data are sourced from the survey in Chopra et al. (2025). The bolded text in the prompts corresponds to the expression of the PEPM prompts for Household Agents in hypothetical vignette experiments (narrative-style prompts). The blue text aligns with the wording in the Chopra et al. (2025) questionnaire. The placeholder {prior} is filled with the homeowner's prior expectation of the house price growth rate.

Social Media Information Module:

```
While browsing on your phone or computer, you randomly come across the following twee
t:
Tweet: {US_home_price_tweet}
```

Notes: This prompt is primarily used to assign social media information to Homeowner Agents within the Social Media Information Module in the Sub-Experiment 1. {US_home_price_tweet} represents a tweet automatically and randomly captured related to the topics of "US Home Price" or "US House price".

Initialization Prompts:

In Sub-Experiment 1, the initialization prompts for Homeowner Agents are identical in format to those used for Household Agents in the hypothetical vignette experiments, with the exception that the role type has been changed to "Homeowner" and the target variable adjusted to "home price expectations." All other content remains consistent and is therefore not reiterated here.

- **Renter Agents:**

Personal Characteristics Module:

```
You are a (an) {age}-year-old {gender} who lives in the {state} State in the US. Your
 educational background is: {educ}, which is a level in Tier {educ_level} (with tiers
 1-8 representing extremely low to extremely high, respectively). Your employment sta
tus is: {employment}. Your household's gross household income last year was in the ra
```

```
nge of: {income}, which is a level in Tier {inc_level} (with tiers 1-17 representing
extremely low to extremely high, respectively). You {liquidsavings} liquid savings. Y
our household consists of {household_size} member(s), including {children} children
(child). You rent a home comprising {rooms_current} room(s) and occupying a total of
{sqft_current} square feet. The current monthly rent is ${rent_current}.
```

Notes: This prompt is primarily used to assign personal characteristics to Renter Agents within the Personal Characteristics Module in the Sub-Experiment 1. All variable data are sourced from the survey in Chopra et al. (2025). The text in black is identical to the corresponding content in the Homeowner Agents' PCM. The text highlighted in red indicates the key distinctions between the prompts for the Renter Agents and those for the Homeowner Agents. {rent_current} is a continuous variable representing the monthly rent paid by the renter. All other variables retain their previously defined meanings and are not reiterated here.

Prior Expectations & Perceptions Module:

The prompts are identical to those in the Homeowner Agents' PEPM and are therefore omitted here.

Social Media Information Module:

The prompts are identical to those in the Homeowner Agents' SMIM and are therefore omitted here.

Information Prompts:

In Sub-Experiment 1, the initialization prompts for Renter Agents are identical in format to those used for Homeowner Agents, with the exception that the role type has been changed to "Renter". All other content remains consistent and is therefore not reiterated here.

Sub-Experiment 2:

In this experiment, to avoid unnecessary complexity, we retain the survey sample from Sub-Experiment 1 as the input data for both the PCM and PEPM. This dataset is chosen for its rich set of demographic characteristics and prior beliefs collected from the respondents.

- **Homeowner Agents:**

Personal Characteristics Module:

The prompt is consistent with that in Sub-Experiment 1.

Prior Expectations & Perceptions Module:

In this experiment, all LLM Agents are tasked with making judgments about expectations of future economic conditions. However, since the dataset does not directly survey these expectations but instead investigates respondents' intentions to buy or sell housing—which are indirectly related and can be considered indirect priors—a Priors Conversion Agent is incorporated into the workflow of PEPM. This agent is designed to link housing transaction intentions with the LLM Agents' economic expectations, thereby converting intentions into economic expectations. The procedure is as follows:

First, the intention of a homeowner to transact housing—specifically, their likelihood of buying or selling a property within the next 10 years—is input as an indirect prior. The corresponding prompt is as follows:

```
When you obtain news information from newspapers, television, etc., or interact with
your family and friends, you form certain beliefs about future housing transactions:
The likelihood (in percent) that you will sell your home and buy a new home within th
e next 10 years is {prob_sellbuy_10y}%.
```

Notes: This prompt is primarily used to assign Homeowner Agents their corresponding prior expectations within the Prior Expectations & Perceptions Module in the Sub-Experiment 2. The placeholder {prob_sellbuy_10y} is filled with the respondent's estimated likelihood (in percent) of selling a home and buying a new home within the next 10 years.

The second step involves establishing a direct link between indirect priors and the economic expectations of LLM Agents. We construct a node (i.e., the Priors Conversion Agent) to transform these priors. Specifically, following Armona et al. (2019), we classify traders into active (those with a probability greater than 50% of engaging in future housing transactions) and inactive (those with a probability no greater than 50%) based on a 50% threshold. Priors Conversion Agent automatically identifies these two types of traders. For active traders, it assumes a certain degree of certainty that they will transact within the next ten years, whereas for inactive traders, it assumes uncertainty. Additionally, the expectations and decisions of agents with an intention to transact in housing are generally more sensitive to changes in house prices and are thus significantly influenced by such fluctuations, while the opposite holds for those without such intention (Campbell & Cocco, 2007; Piazzesi & Schneider, 2009; Guren et al., 2021). Accordingly, we establish a rule linking transaction intention to the impact of house prices on expectations: “In general, your expectations and decisions will be significantly influenced by house price changes when you have a certain intention of future housing transactions; otherwise, they will not.”

```
You're {sellbuy_certain_10y=[1: certain], [2: not certain whether]} you will sell you
r home and buy a new home within the next 10 years.
In general, your expectations and decisions will be significantly influenced by house
 price changes when you have a certain intention of future housing transactions; othe
rwise, they will not.
```

Notes: This prompt is used in the Priors Conversion Agent to transform the indirect priors of Homeowner Agents into their economic expectations. This agent is only constructed when detailed data for direct expectations are unavailable. The placeholder {sellbuy_certain_10y} represents a categorical variable, where 1 denotes “certain” (i.e., active traders), and 2 denotes “not certain whether” (i.e., inactive traders).

Social Media Information Module:

```
While browsing on your phone or computer, you randomly come across the following twee
t:
Tweet: {US_economic_situation_tweet}
```

Notes: This prompt is primarily used to assign social media information to Homeowner Agents within the Social Media Information Module in the Sub-Experiment 2. {US_economic_situation_tweet} represents a tweet automatically and randomly captured related to the topics of “US Economic Situation” or “US Economic Outlook”.

Initialization Prompts:

In Sub-Experiment 2, the initialization prompts for Homeowner Agents are identical in format to those used for Household Agents in the hypothetical vignette experiments, with the exception that the role type has been changed to “Homeowner” and the target variable adjusted to “economic situation expectations.” All other content remains consistent and is therefore not reiterated here.

- **Renter Agents:**

Personal Characteristics Module:

The prompt is consistent with that in Sub-Experiment 1.

Prior Expectations & Perceptions Module:

The prompts are identical in format to those used for Homeowner Agents, with only the indirect priors being modified. The specific prompts are as follows.

Indirect Priors:

```
When you obtain news information from newspapers, television, etc., or interact with
your family and friends, you form certain beliefs about future housing transactions:
The likelihood (in percent) that you will buy a home to live in within the next 10 ye
ars is {prob_buyhome_10y}%.
```

Notes: This prompt is primarily used to assign Renter Agents their corresponding prior expectations within the Prior Expectations & Perceptions Module in the Sub-Experiment 2. The placeholder {prob_buyhome_10y} is filled with the respondent's estimated likelihood (in percent) of buying a home within the next 10 years.

Priors Conversion Agent:

```
You're {prob_buyhome_10y=[1: certain], [2: not certain whether]} you will buy a home
to live in within the next 10 years.
In general, your expectations and decisions will be significantly influenced by house
 price changes when you have a certain intention of future housing transactions; othe
rwise, they will not.
```

Notes: This prompt is used in the Priors Conversion Agent to transform the indirect priors of Renter Agents into their economic expectations. This agent is only constructed when detailed data for direct expectations are unavailable. The placeholder {prob_buyhome_10y} represents a categorical variable, where 1 denotes "certain" (i.e., active traders), and 2 denotes "not certain whether" (i.e., inactive traders).

Social Media Information Module:

The prompts are identical to those in the Homeowner Agents' SMIM and are therefore omitted here.

Information Prompts:

In Sub-Experiment 2, the initialization prompts for Renter Agents are identical in format to those used for Homeowner Agents, with the exception that the role type has been changed to "Renter". All other content remains consistent and is therefore not reiterated here.

B.3 Large-Scale Household Expectations Survey

The survey targeted households, which we simulated using Household Agents.

Personal Characteristics Module:

The prompt is consistent with that in the hypothetical vignette experiments.

Prior Expectations & Perceptions Module:

For inflation expectations:

```
When you purchase goods at the supermarket, shop online, or interact with your family
 and friends, you form certain beliefs about the prices of goods: During the next 12
months, you think that prices in general will {PX1Q1}, and you expect prices to go (u
p/down) on the average by about {PX1Q2} percent. During the next 5 to 10 years, you t
hink prices will {PX5Q1}, and you expect prices to go (up/down) on the average by abo
ut {PX5Q2} percent per year.
```

Notes: This prompt is primarily used to assign Household Agents their corresponding prior inflation expectations within the Prior Expectations & Perceptions Module in the large-scale survey. All variable data are sourced from the Michigan Survey of Consumers. The wording in the prompts is consistent with that of the corresponding questions and response options in the MSC questionnaire. The specific variables are defined as follows: {PX1Q1} is a categorical variable indicating the respondents' perceived overall trends in general prices over the next 12 months, where 1 denotes go up, 2 go up at the same rate as now, 3 be unchanged, and 5 go down. {PX1Q2} is a continuous variable representing the percentage change in general prices over the next 12 months. {PX5Q1} is a categorical variable indicating the respondents' perceived overall trends in general prices over the next 5 years, where 1 denotes go up, 2 go up at the same rate as now, 3 be unchanged, and 5 go down. {PX5Q2} is a continuous variable representing the percentage change in general prices over the next 5 years.

For home price expectations:

```
When you obtain news information from newspapers, television, etc., or interact with
your family and friends, you form certain beliefs about home price: Over the next 12
months, you think that the prices of homes like yours in your community will {HOMPX1Q
1}, and you expect prices of homes like yours in your community to go (up/down) on th
e average by about {HOMPX1Q2} percent. Over the next 5 years or so, you think prices
of homes like yours in your community will {HOMPX5Q1}, and you expect prices of homes
 like yours in your community to go (up/down) on the average by about {HOMPX5Q2} perc
ent per year.
```

Notes: This prompt is primarily used to assign Household Agents their corresponding prior home price expectations within the Prior Expectations & Perceptions Module in the large-scale survey. All variable data are sourced from the Michigan Survey of Consumers. The wording in the prompts is consistent with that of the corresponding questions and response options in the MSC. The specific variables are defined as follows: {HOMPX1Q1} is a categorical variable indicating the direction of home price changes over the next 12 months, where 1 denotes "increase," 2 denotes "remain about the same," and 3 denotes "decrease." {HOMPX1Q2} is a continuous variable representing the percentage change in home prices over the next 12 months. {HOMPX5Q1} is a categorical variable indicating the direction of home price changes over the next 5 years, where 1 denotes "increase," 2 denotes "remain about the same," and 3 denotes "decrease." {HOMPX5Q2} is a continuous variable representing the percentage change in home prices over the next 5 years.

Social Media Information Module:

For inflation expectations:

```
While browsing on your phone or computer, you randomly come across the following twee
t:
Tweet: {US_inflation_tweet}
```

Notes: This prompt is primarily used to assign social media information to Homeowner Agents within the Social Media Information Module in the large-scale survey. {US_inflation_tweet} represents a tweet automatically and randomly captured related to the topics of "US Inflation".

For home price expectations:

```
While browsing on your phone or computer, you randomly come across the following twee
t:
Tweet: {US_home_price_tweet}
```

Notes: This prompt is primarily used to assign social media information to Homeowner Agents within the Social Media Information Module in the large-scale survey. {US_home_price_tweet} represents a tweet automatically and randomly captured related to the topics of “US Home Price” or “US House price”.

Initialization Prompts:

In the large-scale survey, the initialization prompts are identical in format to those used for Household Agents in the hypothetical vignette experiments, with the exception that the target variables adjusted to “inflation expectations” and “home price expectations”. All other content remains consistent and is therefore not reiterated here.

C Additional Results

C.1 Tables

Table A.1: Information on the foundation models in our paper

Foundation Model (LLM)	Developer	Release Date	Knowledge Cutoff
Qwen3-235B-A22B-Thinking-2507	Alibaba	July 25, 2025	October 2024 (or earlier)*
DeepSeek-R1-0528	DeepSeek	May 28, 2025	July 2024 (or earlier)*
GPT-o4-mini	OpenAI	April 16, 2025	June 2024
Gemini-2.5-Pro	Google DeepMind	June 17, 2025	January 2025

Notes: This table presents information on the developers, release dates, and knowledge cutoffs of the four advanced foundation models discussed in this paper. The latest knowledge cutoff among these models is no later than January 2025. The knowledge cutoffs marked with an asterisk (*) are not officially released dates—as the developers do not disclose them in their technical reports—but are inferred by querying the LLMs with a series of questions, including: 1) What is your knowledge cutoff? 2) What is today's date? 3) What happened in January 2025? These questions are designed to elicit responses revealing the models' actual knowledge cutoffs.

Table A.2: Complexity of the mental models underlying the expectations of human, LLM Agents, and foundation models under each vignette

Vignettes	**Panel A: Inflation (Households)**			**Panel B: Unemployment (Households)**		
	Human	Original	only INITIAL	Human	Original	only INITIAL
Oil price	2.2 (3.0)	3.0 (3.5)	2.7 (3.4)	3.0 (3.7)	4.0 (4.3)	5.7 (5.6)
Government spending	3.2 (3.7)	4.4 (4.5)	5.6 (5.3)	2.7 (3.5)	3.8 (4.1)	3.0 (3.6)
Federal funds rate	2.8 (3.5)	3.4 (3.8)	5.7 (5.3)	3.0 (3.7)	3.6 (4.2)	6.1 (6.0)
Income taxes	3.7 (4.0)	5.0 (5.1)	5.8 (5.7)	3.6 (4.0)	5.3 (5.3)	6.0 (6.0)
Vignettes	**Panel C: Inflation (Experts)**			**Panel D: Unemployment (Experts)**		
	Human	Original	only INITIAL	Human	Original	only INITIAL
Oil price	2.7 (3.2)	3.5 (3.7)	2.6 (3.3)	3.7 (4.1)	6.1 (5.5)	7.6 (6.6)
Government spending	3.2 (3.6)	3.8 (4.1)	3.8 (4.1)	3.1 (3.7)	4.2 (4.5)	3.6 (4.3)
Federal funds rate	3.1 (3.6)	4.2 (4.3)	5.4 (5.2)	2.5 (3.3)	4.2 (4.6)	6.1 (6.0)
Income taxes	3.9 (4.2)	5.6 (5.4)	5.7 (5.6)	3.7 (4.2)	5.8 (5.7)	6.3 (6.2)

Notes: This table presents the average number of causal links (values outside parentheses) and unique nodes (values inside parentheses) in the mental models (DAGs) of humans ("Human"), LLM Agents ("Original"), and foundation models ("only INITIAL") across vignettes. Panels A and B present the results correspond to the inflation and unemployment expectations of households, Household Agents, and foundation models, respectively. Panels C and D present the results correspond to the inflation and unemployment expectations of experts, Expert Agents, and foundation models, respectively. Higher values indicate greater structural complexity of the causal pathways on average.

Table A.3: Intermediate variables involved in mental models and their categories

Category (Node)	Intermediate variable	Explanation
Demand		
Borrowing amount & costs	borrowing firms	Amount borrowed (debt) by firms, or amount lent by banks to firms.
	borrowing household	Amount borrowed (debt) by households, or amount lent by banks to households.
	borrowing government	Amount borrowed (debt) by the government, or amount lent by banks to the government.
	costs borrowing household	Borrowing rates and/or access to credit faced by households.
	costs borrowing banks	Borrowing rates and/or access to credit faced by banks.
	costs borrowing government	Borrowing rates and/or access to credit faced by the government.
Consumption & Demand	demand	Demand for goods, spending and consumption by different groups.
	demand firms	Demand for goods, spending and consumption by firms.
	demand household	Demand for goods, spending and consumption by households.
	costs household	Costs of subsistence goods, e.g. heating, gasoline, ...
	demand government	Demand for goods, spending and consumption by the government.
	investment	Investment (expenditure) of firms.
Housing demand	housing demand	Quantity of housing demanded.
Labor demand	labor demand	"Job creation", firm's/government's demand for employees, "Job opportunities".
Income, Saving & Money	income	Household income, wages received, purchasing power.
	money	Overall amount of money in circulation, money printing by the central bank.
	saving	Amount saved by households.
Supply		
Firms' costs	costs firms	Production costs, including costs of input goods, wages paid; "Firms need to cover"; "firms need to make up for it", ...
	costs borrowing firms	Borrowing rates and/or access to credit faced by firms.
	firm prices	Firms' decisions about pricing.
Housing supply	housing supply	Quantity of housing supplied.
Labor supply	labor supply	Changes in households' desired work hours.

Production & Profit	production	Firms' production / supply of goods and services.
	profit	Firms' profits or profit margin, including firms facing pressure to take actions to keep the profit margin at a certain level.
Miscellaneous		
Prior expectations	expected inflation	Expectations of future realizations of inflation as intermediate causes (propagation mechanisms).
	expected unemployment	Expectations of future realizations of unemployment as intermediate causes (propagation mechanisms).
Fiscal	government taxes	Tax revenue collected by the government.
	government finances	Residual category referring to unspecified improvements or deterioration in the government's budget.
Economic growth	growth	GDP growth, overall growth of the economy.
Housing (residual)	housing	The quantity of housing is mentioned, but it is unclear whether demand or supply is being referred to.
Labor (residual)	labor	Residual category for cases where it is unclear whether the respondent is thinking about labor demand or supply, e.g. "more people work".
Asset prices	prices stock	Stock prices.
	prices house	House prices.
Interest rate	interest	General interest rate category if agent not specified or if not specified whether households' rates on borrowing vs saving are meant.
	saving rate	Interest rate earned on savings.
Government management	government management	Explicit reference to policy successes (failures), good (bad) management by policymakers, or politicized positive (negative) evaluations of policies.

Notes: This table presents all intermediate variables potentially mentioned in open-ended responses within the hypothetical vignette experiments, along with their categorization and corresponding explanations. The definitions and categorization of these variables are primarily based on Andre et al. (2022, 2025).

Table A.4: The diversity of thoughts generated by LLM Agents (original and those without modules) and those generated by humans in hypothetical vignette experiments

Panel A: Households

Vignette	Agent	Semantic Diversity	Vignette	Agent	Semantic Diversity
Oil price	Human	0.5335	Government spending	Human	0.5564
	Original	0.3158		Original	0.3355
	w/o RD	0.3123		w/o RD	0.3261
	w/o SMIM	0.2921		w/o PCM	0.3116
	w/o PCM	0.2836		w/o SMIM	0.2970
	w/o PEPM	0.2605		w/o PEPM	0.2646
	w/o INITIAL	0.2398		w/o INITIAL	0.2530
	only INITIAL	0.1588		only INITIAL	0.1783
Federal funds rate	Human	0.5913	Income taxes	Human	0.5771
	Original	0.3449		Original	0.3268
	w/o RD	0.3342		w/o RD	0.3240
	w/o PCM	0.3156		w/o SMIM	0.3004
	w/o SMIM	0.3103		w/o PCM	0.2909
	w/o PEPM	0.2647		w/o PEPM	0.2781
	w/o INITIAL	0.2318		w/o INITIAL	0.2640
	only INITIAL	0.1429		only INITIAL	0.1638

Panel B: Experts

Vignette	Agent	Semantic Diversity	Vignette	Agent	Semantic Diversity
Oil price	Human	0.5309	Government spending	Human	0.5980
	Original	0.3632		Original	0.3496
	w/o KAM	0.3586		w/o KAM	0.3415
	w/o RD	0.3533		w/o RD	0.3361
	w/o PEPM	0.3132		w/o PEPM	0.3118
	w/o PBM	0.3103		w/o PBM	0.3024
	w/o INITIAL	0.2848		w/o INITIAL	0.2812
	only INITIAL	0.2333		only INITIAL	0.2269
Federal funds rate	Human	0.5822	Income taxes	Human	0.5875
	Original	0.4501		Original	0.3389
	w/o RD	0.4383		w/o KAM	0.3344
	w/o KAM	0.4356		w/o RD	0.3129
	w/o PEPM	0.3718		w/o PEPM	0.2797
	w/o PBM	0.3571		w/o PBM	0.2651
	w/o INITIAL	0.3365		w/o INITIAL	0.2526
	only INITIAL	0.2754		only INITIAL	0.1820

Notes: This table presents the semantic diversity of thoughts generated by LLM Agents (original and those without modules) and those generated by humans under each vignette in hypothetical vignette experiments, respectively. Panel A compares the results of Household Agents with those of households, while Panel B compares Expert Agents with experts.

Table A.5: The diversity of thoughts generated by LLM Agents (original and those without modules) and those generated by humans in information provision experiments

Respondent	**Agent**	**Semantic Diversity**	**Respondent**	**Agent**	**Semantic Diversity**
Homeowners	Human	0.4996	Renters	Human	0.5375
	Original	0.2871		Original	0.2792
	w/o RD	0.2852		w/o PEPM	0.2788
	w/o PEPM	0.2711		w/o RD	0.2763
	w/o PCM	0.2626		w/o PCM	0.2522
	w/o SMIM	0.2598		w/o SMIM	0.2442
	w/o INITIAL	0.2281		w/o INITIAL	0.2420
	only INITIAL	0.1575		only INITIAL	0.1845

Notes: This table presents the semantic diversity of thoughts generated by LLM Agents (original and those without modules) and those generated by humans in the Sub-Experiment 2 of information provision experiments, respectively. The left panel compares the results of Homeowner Agents with those of homeowners, while the right panel compares Renter Agents with renters.

Table A.6: The diversity of thoughts generated by LLM Agents (original and those without modules) in Michigan Surveys of Consumers

Expectation	**Agent**	**Semantic Diversity**	**Expectation**	**Agent**	**Semantic Diversity**
Inflation	Original	0.3097	Home price	Original	0.3214
	w/o RD	0.3097		w/o RD	0.3187
	w/o PCM	0.2938		w/o PCM	0.3076
	w/o SMIM	0.2674		w/o SMIM	0.2670
	w/o PEPM	0.2102		w/o PEPM	0.2366
	w/o INITIAL	0.1801		w/o INITIAL	0.1709
	only INITIAL	0.0727		only INITIAL	0.0921

Notes: This table presents the semantic diversity of thoughts generated by Household Agents (original and those without modules) when pre-estimating inflation and home price expectations in 2025 Michigan Surveys of Consumers, respectively. The left panel compares the results for inflation expectations, and the right panel for home price expectations.

C.2 Figures

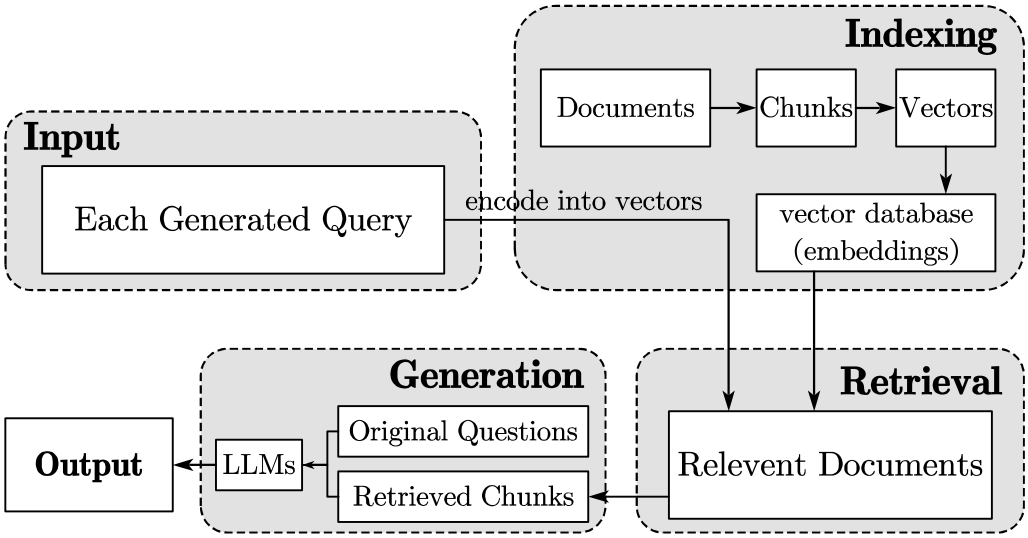


Figure A.1: The framework of our RAG workflow

Notes: This figure illustrates in detail the framework of the RAG workflow, specifically how KAM retrieves relevant knowledge & information. It comprises three core steps: (1) Indexing: documents in the personalized knowledge base are segmented into chunks, encoded into vector representations, and stored in a vector database; (2) Retrieval: based on semantic similarity, the top k most relevant chunks are retrieved for each query; (3) Generation: the original survey question and the retrieved chunks are jointly input into the LLMs to generate the answer.

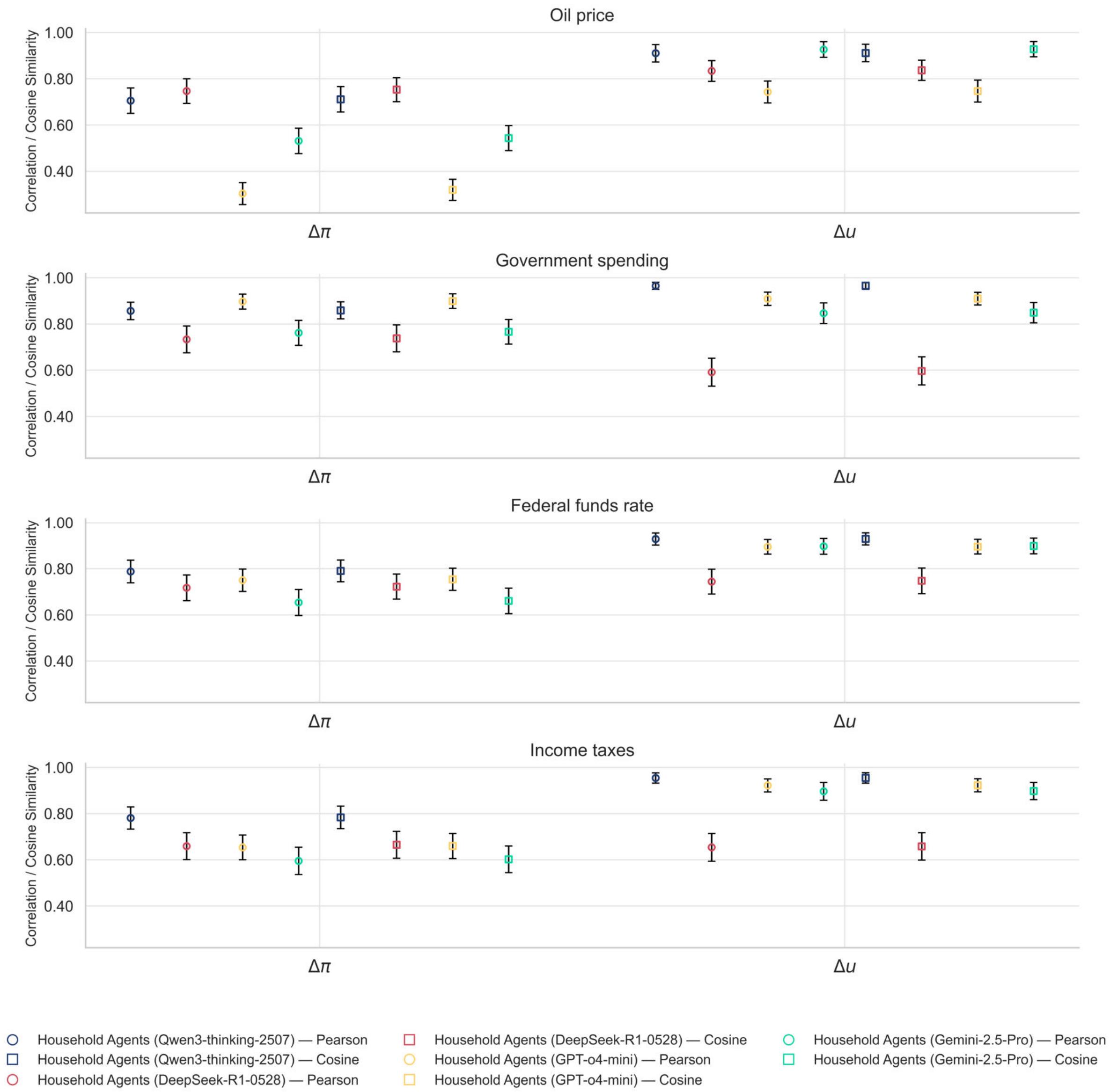


Figure A.2: Shape similarity between the expectation distributions generated by Household Agents based on different foundation models and those generated by humans in hypothetical vignette experiments

Notes: This figure displays the distributional shape similarity, as measured by Pearson correlation and cosine similarity, between the changes in inflation expectations ($\Delta\ \pi$) and unemployment expectations ($\Delta\ u$) generated by Household Agents based on four different types of foundation models (Qwen3-235B-A22B-Thinking-2507, DeepSeek-R1-0528, GPT-o4-mini, and Gemini-2.5-Pro), respectively, and those of households under four different vignettes. Error bars present two-sided 95% confidence intervals for the similarity metrics, obtained by bootstrap over histogram-based probability vectors.

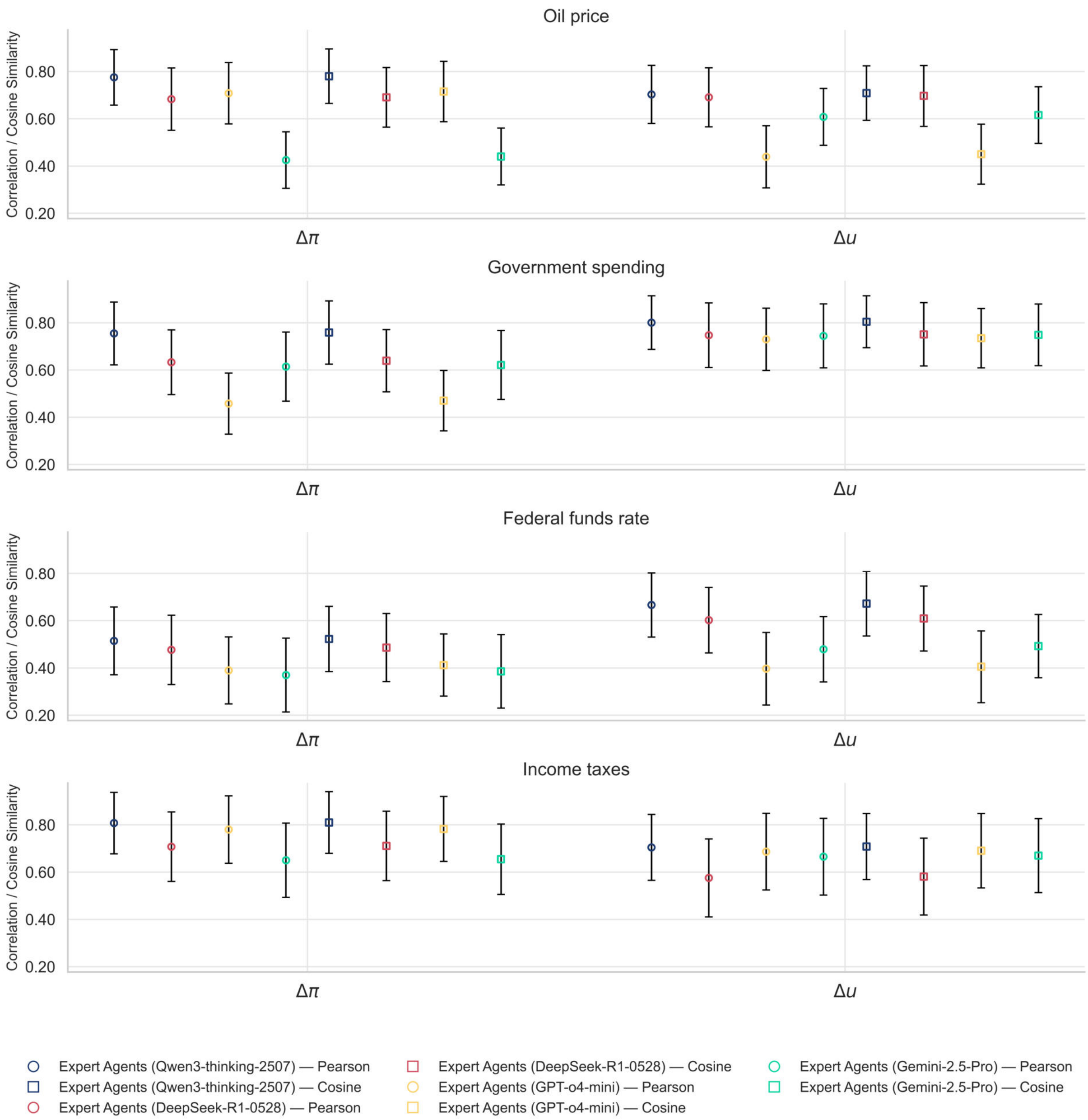


Figure A.3: Shape similarity between the expectation distributions generated by Expert Agents based on different foundation models and those generated by humans in hypothetical vignette experiments

Notes: This figure displays the distributional shape similarity, as measured by Pearson correlation and cosine similarity, between the changes in inflation expectations ($\Delta\ \pi$) and unemployment expectations ($\Delta\ u$) generated by Expert Agents based on four different types of foundation models (Qwen3-235B-A22B-Thinking-2507, DeepSeek-R1-0528, GPT-o4-mini, and Gemini-2.5-Pro), respectively, and those of experts under four different vignettes. Error bars present two-sided 95% confidence intervals for the similarity metrics, obtained by bootstrap over histogram-based probability vectors.

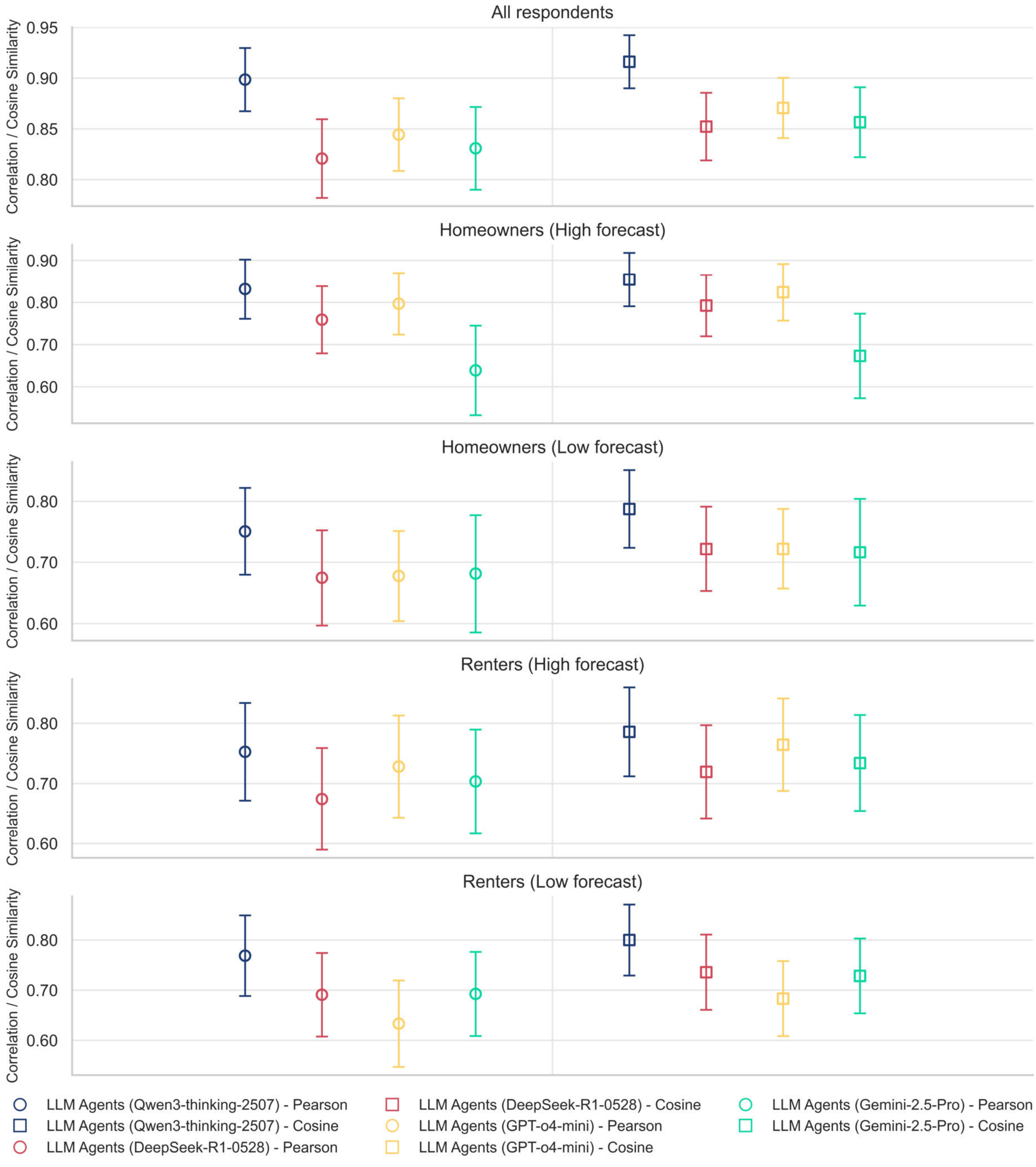


Figure A.4: Shape similarity between the expectation distributions generated by LLM Agents based on different foundation models and those generated by humans in information provision experiments

Notes: This figure displays the distributional shape similarity, as measured by Pearson correlation and cosine similarity, between the home price expectations generated by LLM Agents based on four different types of foundation models (Qwen3-235B-A22B-Thinking-2507, DeepSeek-R1-0528, GPT-o4-mini, and Gemini-2.5-Pro), respectively, and those of humans in different treatment groups. Error bars present two-sided 95% confidence intervals for the similarity metrics, obtained by bootstrap over histogram-based probability vectors.

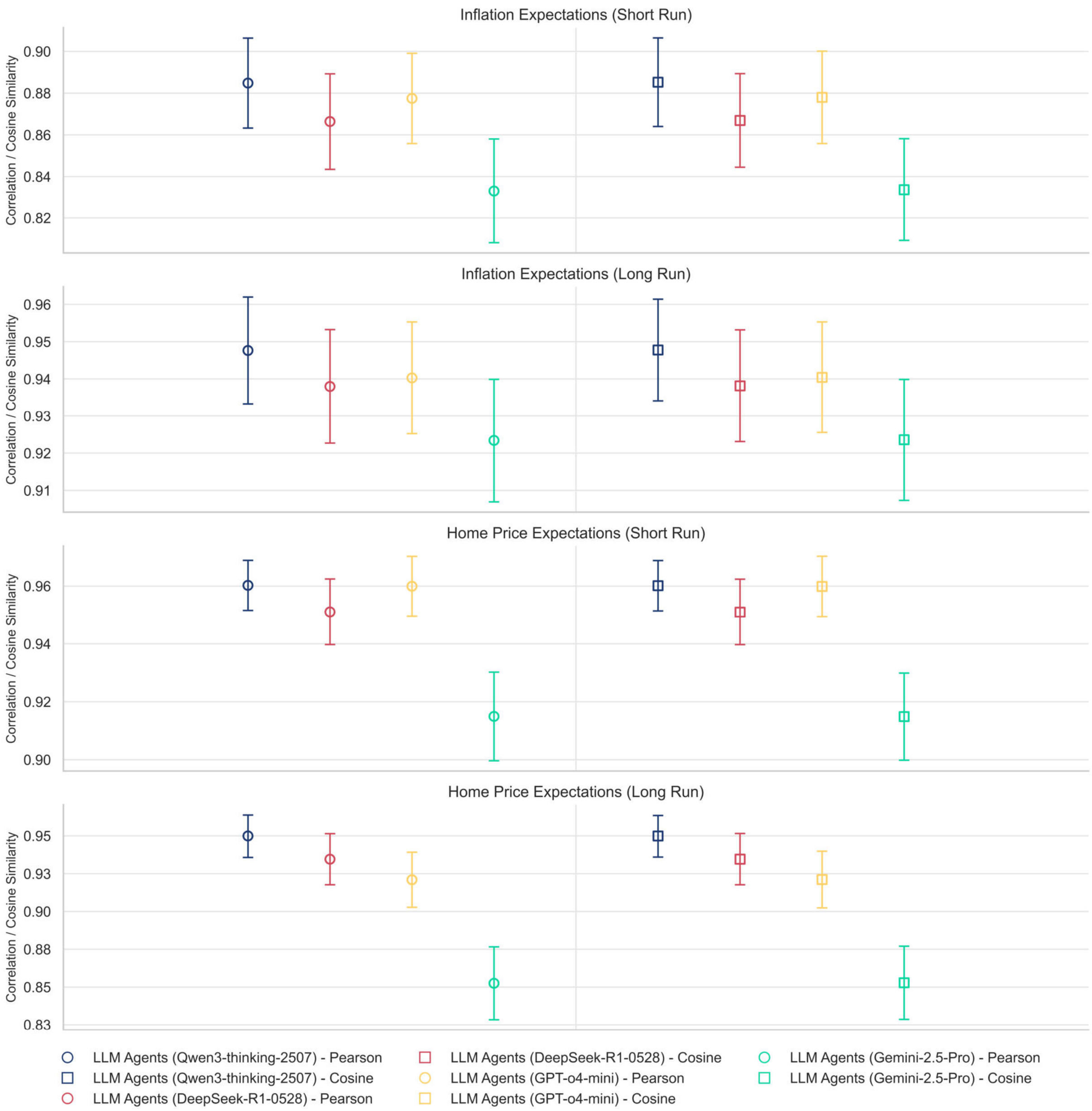


Figure A.5: Shape similarity between the expectation distributions generated by LLM Agents based on different foundation models and those generated by humans in Michigan Surveys of Consumers

Notes: This figure displays the distributional shape similarity, as measured by Pearson correlation and cosine similarity, between long- and short-run inflation expectations and home price expectations generated by LLM Agents based on four different types of foundation models (Qwen3-235B-A22B-Thinking-2507, DeepSeek-R1-0528, GPT-o4-mini, and Gemini-2.5-Pro), respectively, and those of households in 2025 Michigan Surveys of Consumers. Error bars present two-sided 95% confidence intervals for the similarity metrics, obtained by bootstrap over histogram-based probability vectors.

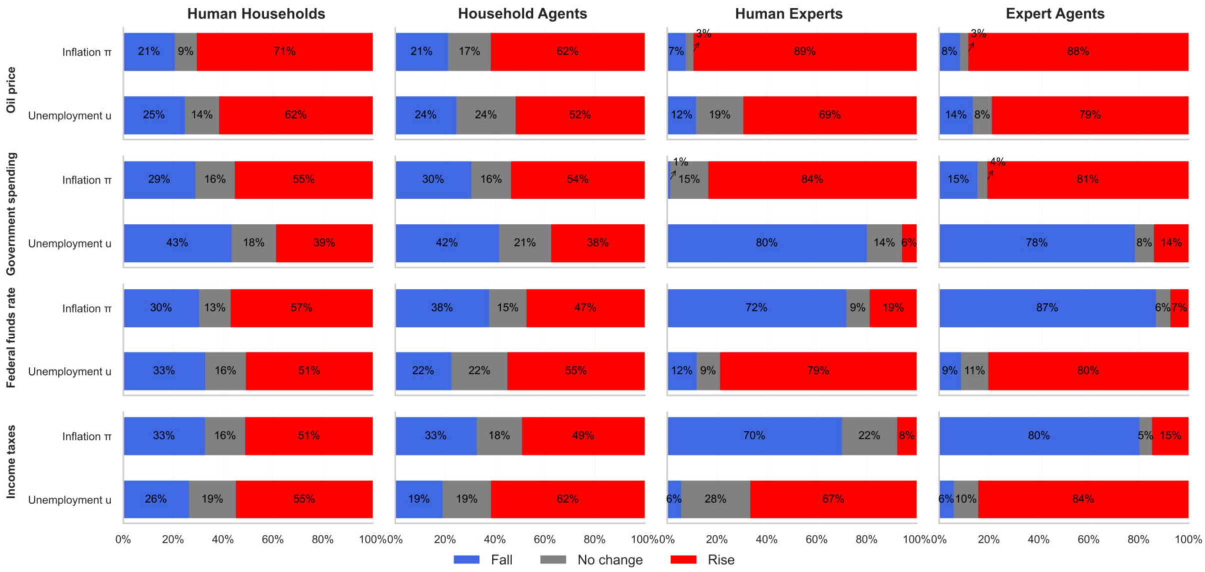
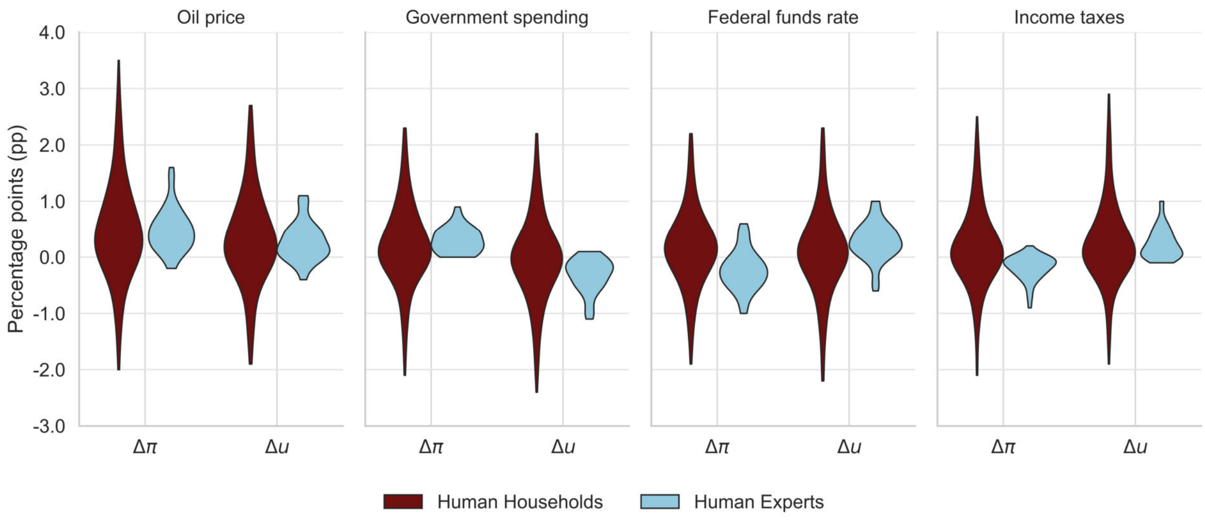
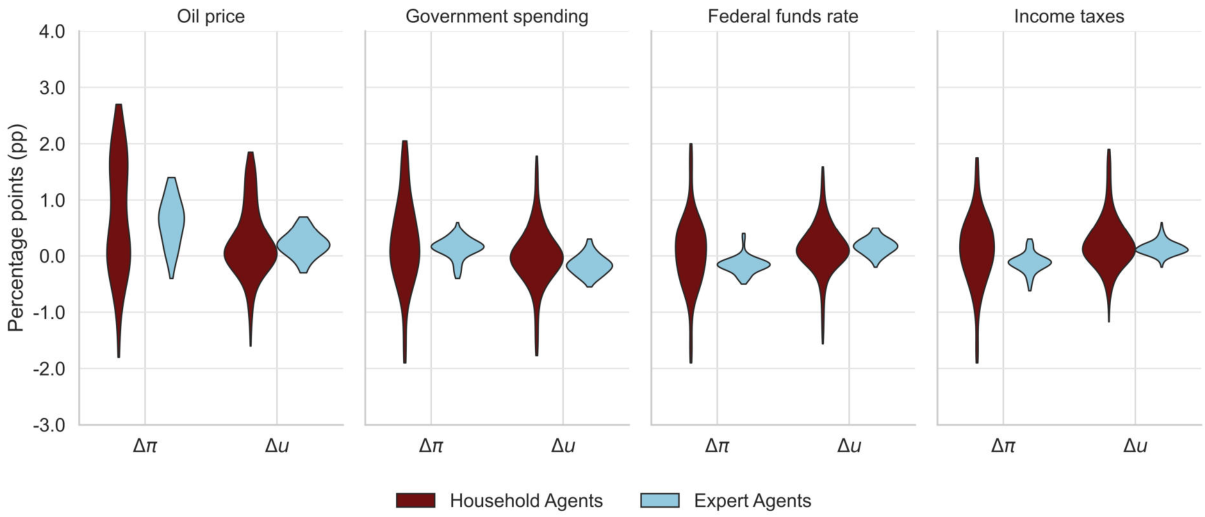


Figure A.6: Forecasts of the directional effects of macroeconomic shocks (humans v.s. LLM Agents)

Notes: This figure shows the forecasts of the directional effects of macroeconomic shocks on the inflation rate and the unemployment rate by humans (households and experts) and LLM Agents (Household Agents and Expert Agents), using percentage bar charts. Predictions in the "fall" scenarios are reversed to make them comparable to rise predictions.



(a) Humans



(b) LLM Agents

Figure A.7: Forecast distributions of the quantitative effects of macroeconomic shocks (humans v.s. LLM Agents)

Notes: This figure presents the forecast distributions (with trimmed 5% tails) of the quantitative effects of macroeconomic shocks on the inflation rate ($\Delta\ \pi$) and the unemployment rate ($\Delta\ u$) by humans (households and experts) and LLM Agents (Household Agents and Expert Agents), using kernel density estimators. This figure aggregate forecasts for the "rise" and "fall" scenarios, with fall predictions reversed to be comparable to rise predictions.

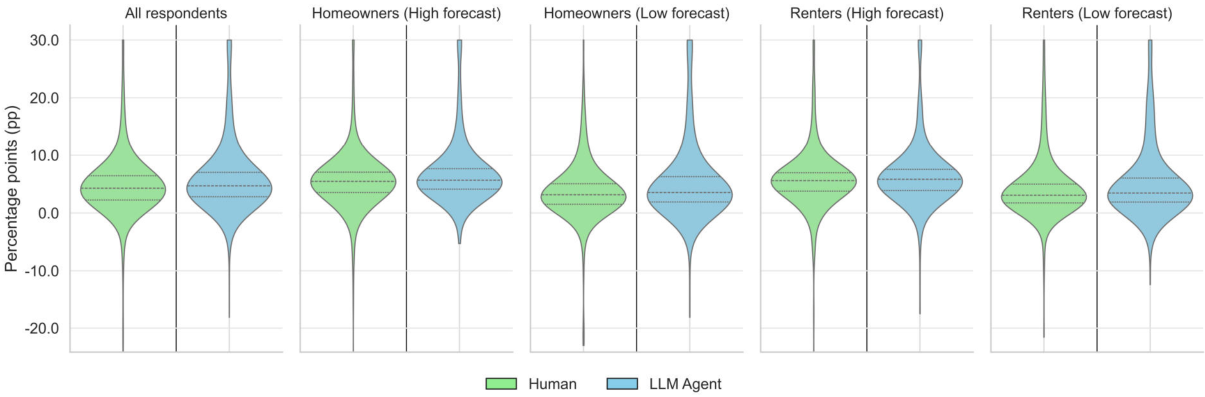


Figure A.8: Distributions of home price expectations generated by LLM Agents and humans in information provision experiments

Notes: This figure presents distributions of home price expectations generated by LLM Agents (Homeowner Agents and Renter Agents) and humans (homeowners and renters) in different treatment groups, using kernel density estimators. The dashed lines in each violin plot represent the quartiles of each distribution.

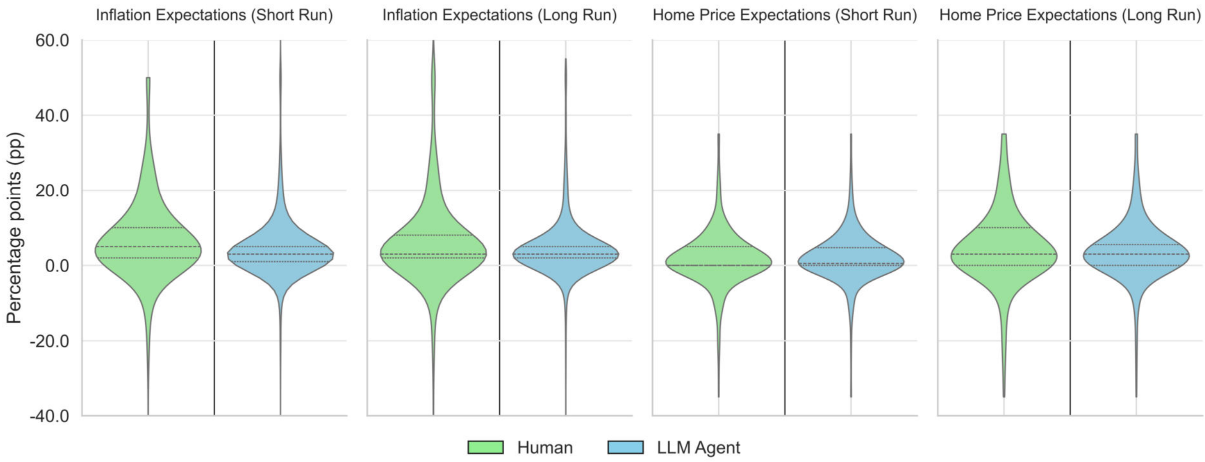


Figure A.9: Expectation distributions generated by LLM Agents and humans in Michigan Surveys of Consumers

Notes: This figure presents distributions of long- and short-run inflation expectations and home price expectations generated by Household Agents and households in 2025 Michigan Surveys of Consumers, using kernel density estimators. The dashed lines in each violin plot represent the quartiles of each distribution.

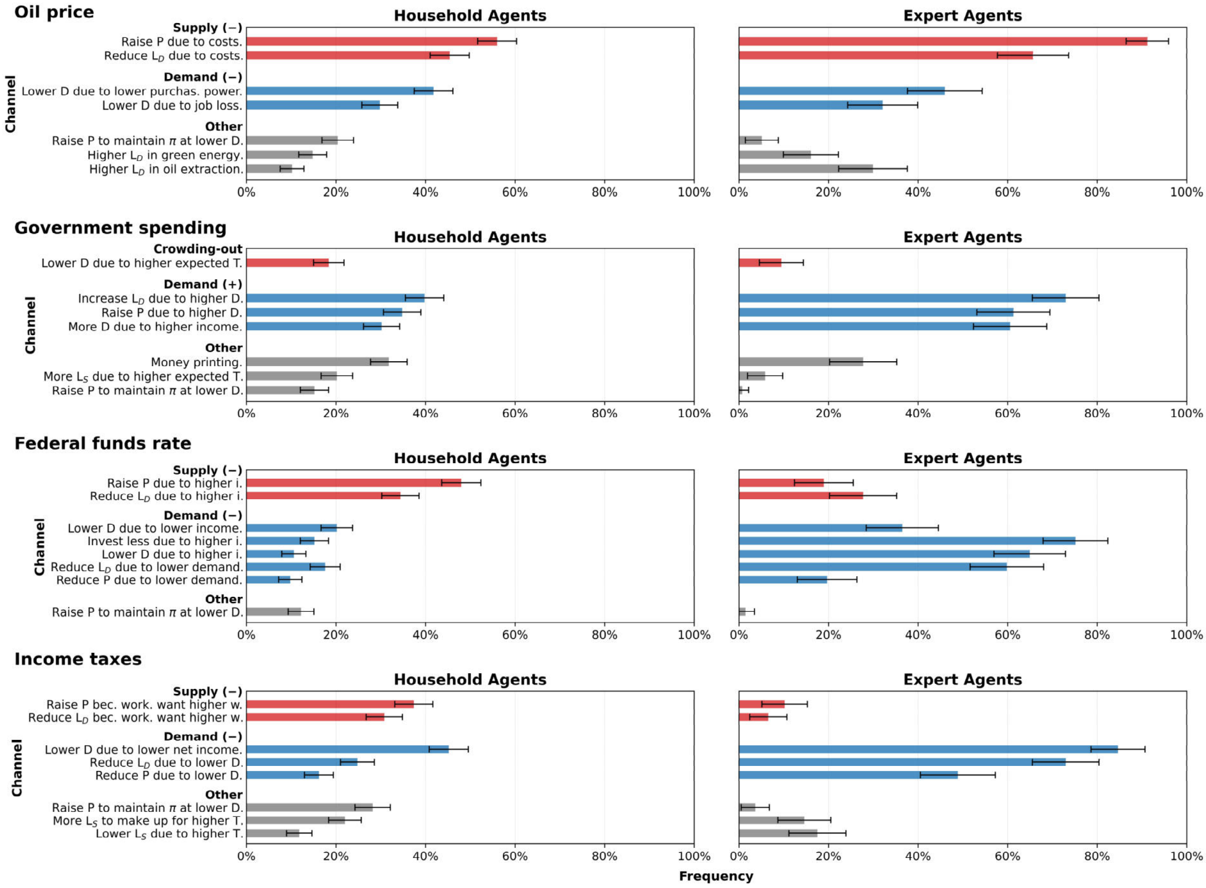


Figure A.10: LLM Agents' thoughts of propagation channels

Notes: This figure shows which propagation channels are selected by LLM Agents when they make their predictions. LLM Agents can select the channels from a list. The results are displayed separately for each vignette and for Household Agents(left panel) and Expert Agents (right panel). Error bars display 95% confidence intervals. P abbreviates "firm prices," L_D "labour demand," D "product demand," "firm profits," T "taxes," i "interest rates," w "wages," and L_S "labour supply." The format of this figure is consistent with Figure 3 in Andre et al. (2022), making them comparable.

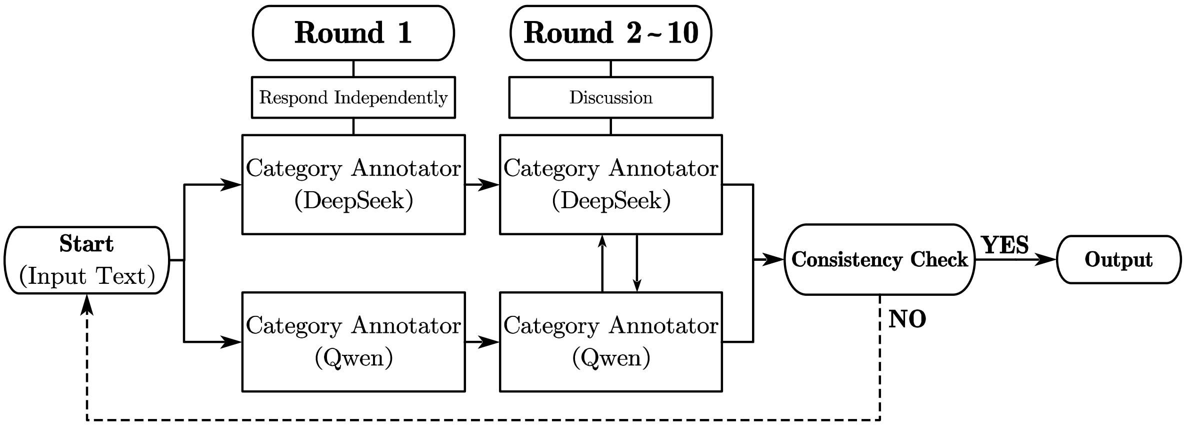


Figure A.11: Architecture of an agentic workflow for classifying open-ended responses

Notes: This figure illustrates the architecture of an agentic workflow for classifying or labeling open-ended responses. In Round 1, two distinct medium-scale LLMs (deepseek-r1-distill-qwen-32b and qwq-32b) serve as Category Annotators, independently labeling input text based on predefined criteria. After Round 1, the two Annotators discuss their initial results. When their outputs agree, the result passes consistency check and is output. If discrepancies arise, they engage in multiple discussion rounds until consensus is reached. Should no agreement be achieved by Round 10, the entire process repeats until consistent outputs are attained.

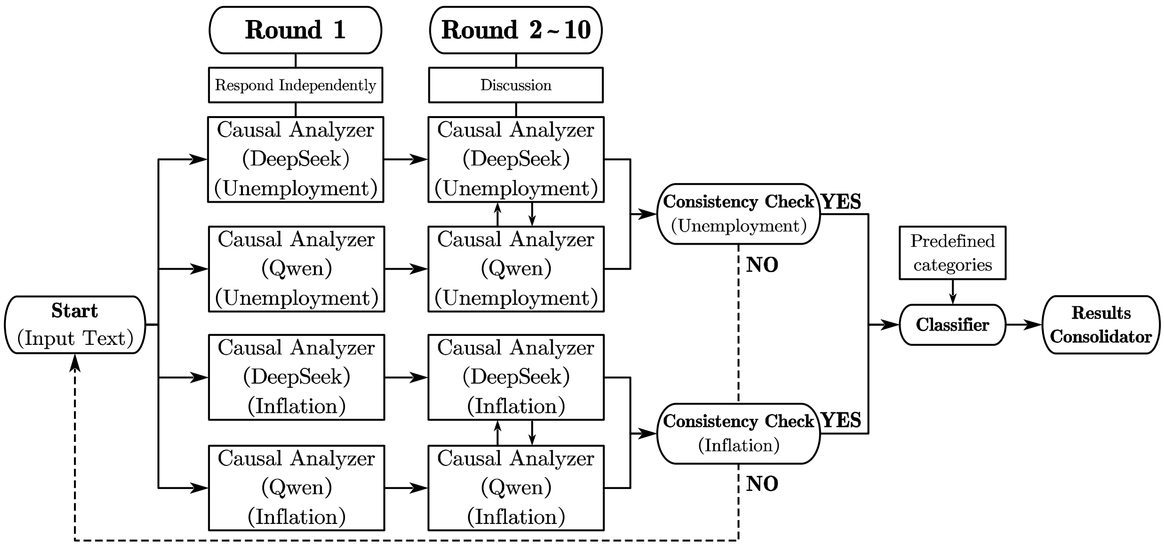


Figure A.12: Architecture of an agentic workflow for identifying DAGs in open-ended responses

Notes: This figure illustrates the architecture of an agentic workflow designed to identify Directed Acyclic Graphs (DAGs) in open-ended responses. Two medium-scale LLMs of different types—deepseek-r1-distill-qwen-32b and qwq-32b—serve as Causal Analyzers, each independently identifying causal pathways in open-text responses related to unemployment and inflation expectations. After Round 1, the two Analyzers discuss their initial results. If their outputs agree, they pass the consistency check; if not, they engage in multiple discussion rounds until consensus is reached. If no agreement is achieved after 10 rounds, the entire process is repeated. A Classifier then categorizes the intermediate variables in the causal pathways according to predefined categories. Finally, a Results Consolidator consolidates the categorized results. Further details are provided in Section E.1.

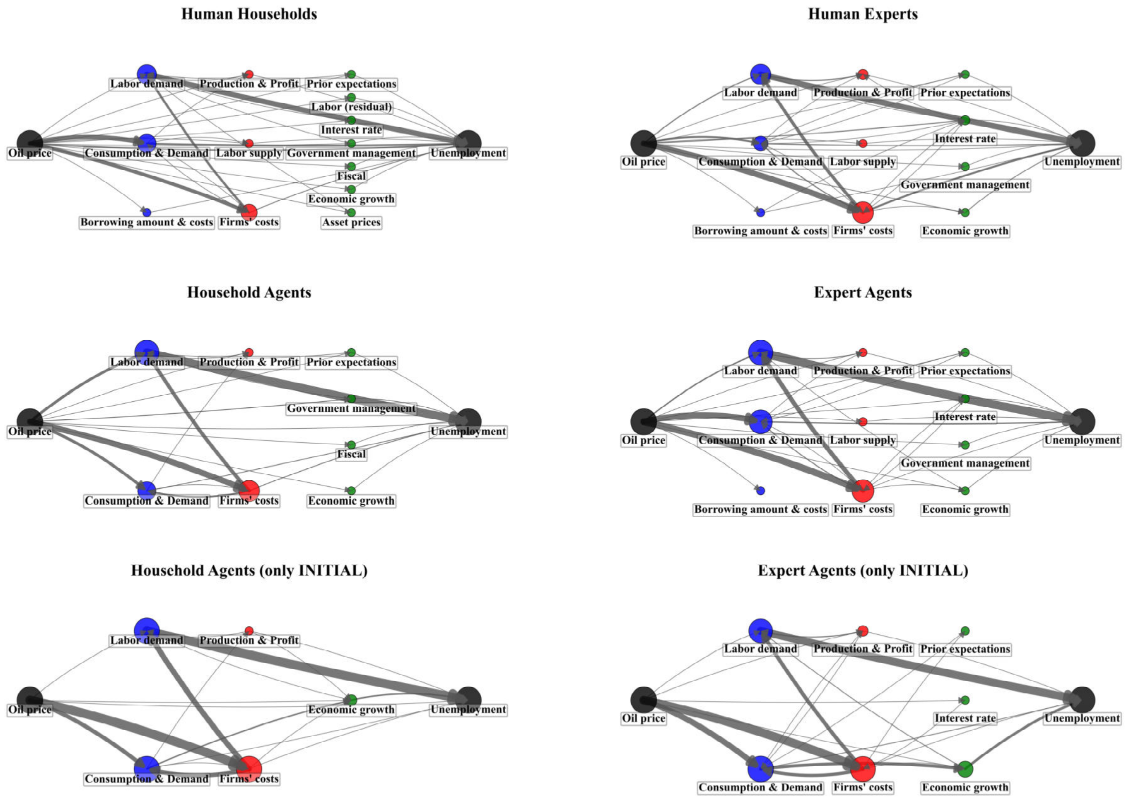


Figure A.13: The "average" DAGs underlying the formation of unemployment expectations in the oil price vignette

Notes: The figure presents the "average" DAGs underlying unemployment expectation formation for humans (denoted as "Human Households" and "Human Experts"), LLM Agents (denoted as "Household Agents" and "Expert Agents"), and foundation models (denoted as "Household Agents (only INITIAL)" and "Expert Agents (only INITIAL)") in the oil price vignette. The nodes represent categories of intermediate variables, whose definitions and classifications are provided in Supplementary Appendix Table A.3. The aggregated DAGs reveal the most relevant variables (nodes) and causal links in the responses of humans, LLM Agents, and foundation models. Node size: The size of the nodes is proportional to the share of responses that refer to the nodes. Node color: Red indicates supply-side variables, blue indicates demand-side variables, green indicates miscellaneous variables, black is used for start and end nodes. Edge thickness: The thickness of the edges is proportional to the share of responses that refer to the causal connections (among humans, LLM Agents and foundation models, respectively).

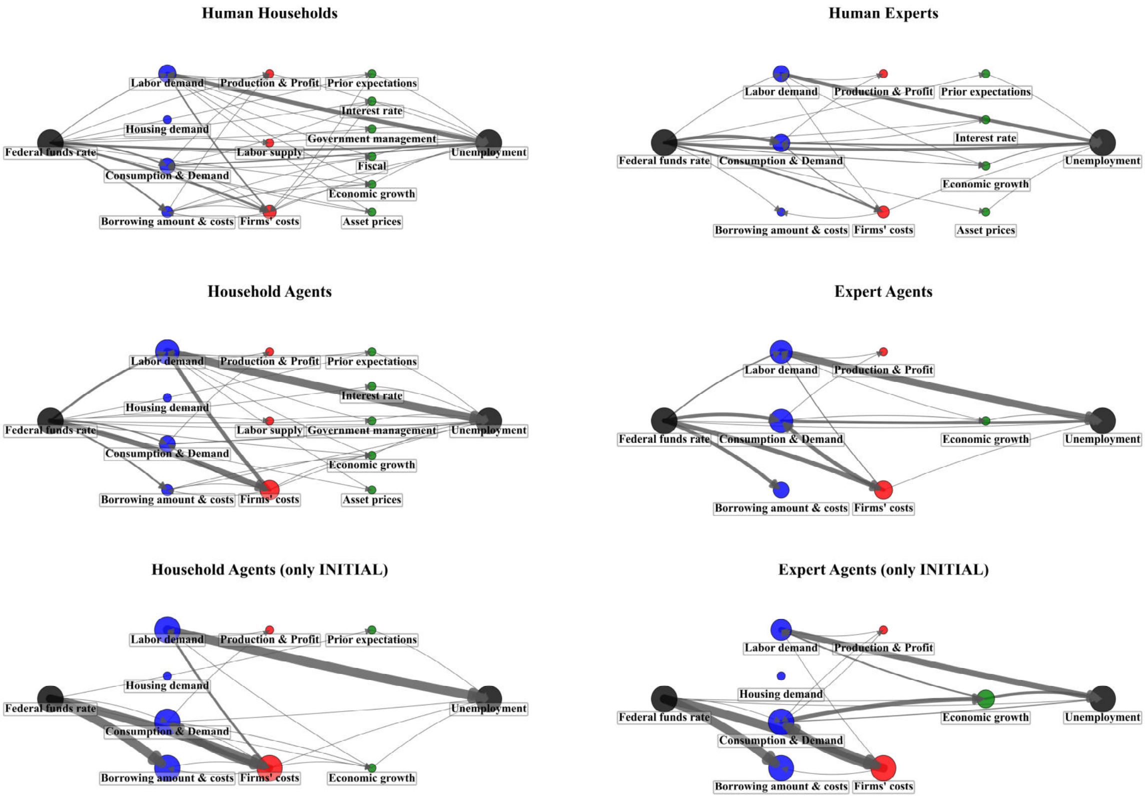


Figure A.14: The "average" DAGs underlying the formation of unemployment expectations in the interest rate vignette

Notes: The figure presents the "average" DAGs underlying unemployment expectation formation for humans (denoted as "Human Households" and "Human Experts"), LLM Agents (denoted as "Household Agents" and "Expert Agents"), and foundation models (denoted as "Household Agents (only INITIAL)" and "Expert Agents (only INITIAL)") in the interest rate vignette. The nodes represent categories of intermediate variables, whose definitions and classifications are provided in Supplementary Appendix Table A.3. The aggregated DAGs reveal the most relevant variables (nodes) and causal links in the responses of humans, LLM Agents, and foundation models. Node size: The size of the nodes is proportional to the share of responses that refer to the nodes. Node color: Red indicates supply-side variables, blue indicates demand-side variables, green indicates miscellaneous variables, black is used for start and end nodes. Edge thickness: The thickness of the edges is proportional to the share of responses that refer to the causal connections (among humans, LLM Agents and foundation models, respectively).

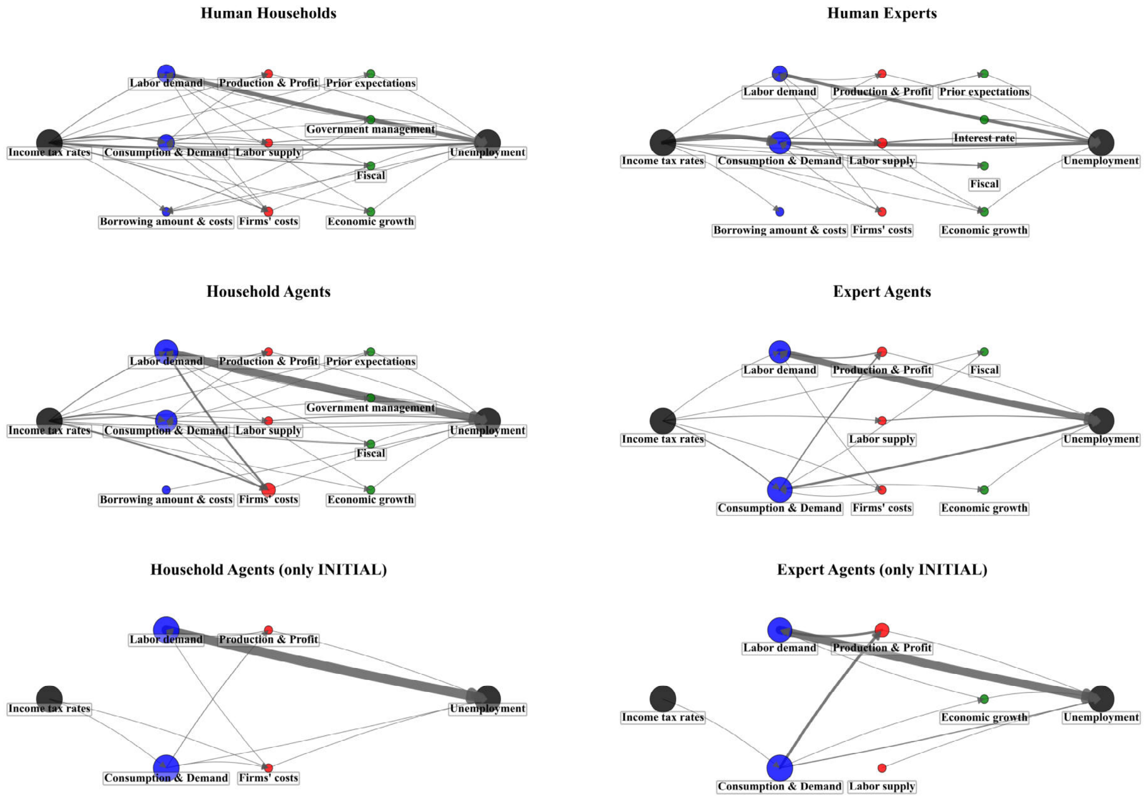


Figure A.15: The "average" DAGs underlying the formation of unemployment expectations in the taxation vignette

Notes: The figure presents the "average" DAGs underlying unemployment expectation formation for humans (denoted as "Human Households" and "Human Experts"), LLM Agents (denoted as "Household Agents" and "Expert Agents"), and foundation models (denoted as "Household Agents (only INITIAL)" and "Expert Agents (only INITIAL)") in the taxation vignette. The nodes represent categories of intermediate variables, whose definitions and classifications are provided in Supplementary Appendix Table A.3. The aggregated DAGs reveal the most relevant variables (nodes) and causal links in the responses of humans, LLM Agents, and foundation models. Node size: The size of the nodes is proportional to the share of responses that refer to the nodes. Node color: Red indicates supply-side variables, blue indicates demand-side variables, green indicates miscellaneous variables, black is used for start and end nodes. Edge thickness: The thickness of the edges is proportional to the share of responses that refer to the causal connections (among humans, LLM Agents and foundation models, respectively).

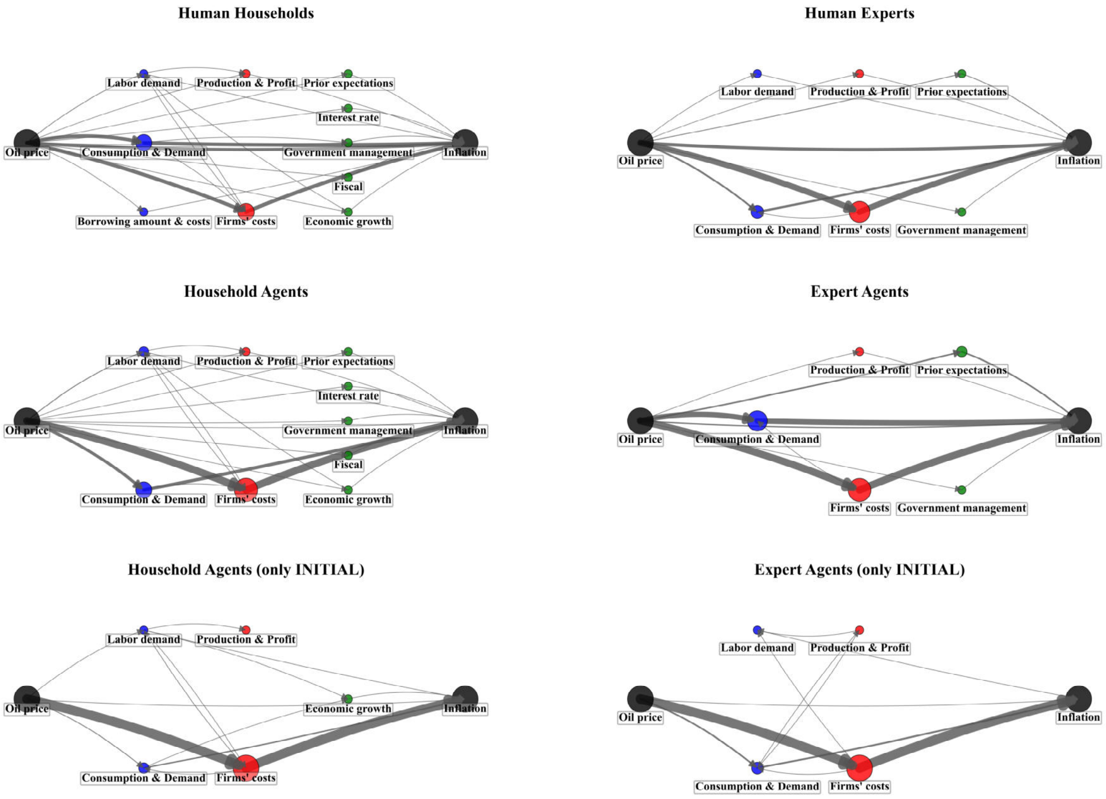


Figure A.16: The "average" DAGs underlying the formation of inflation expectations in the oil price vignette

Notes: The figure presents the "average" DAGs underlying inflation expectation formation for humans (denoted as "Human Households" and "Human Experts"), LLM Agents (denoted as "Household Agents" and "Expert Agents"), and foundation models (denoted as "Household Agents (only INITIAL)" and "Expert Agents (only INITIAL)") in the oil price vignette. The nodes represent categories of intermediate variables, whose definitions and classifications are provided in Supplementary Appendix Table A.3. The aggregated DAGs reveal the most relevant variables (nodes) and causal links in the responses of humans, LLM Agents, and foundation models. Node size: The size of the nodes is proportional to the share of responses that refer to the nodes. Node color: Red indicates supply-side variables, blue indicates demand-side variables, green indicates miscellaneous variables, black is used for start and end nodes. Edge thickness: The thickness of the edges is proportional to the share of responses that refer to the causal connections (among humans, LLM Agents and foundation models, respectively).

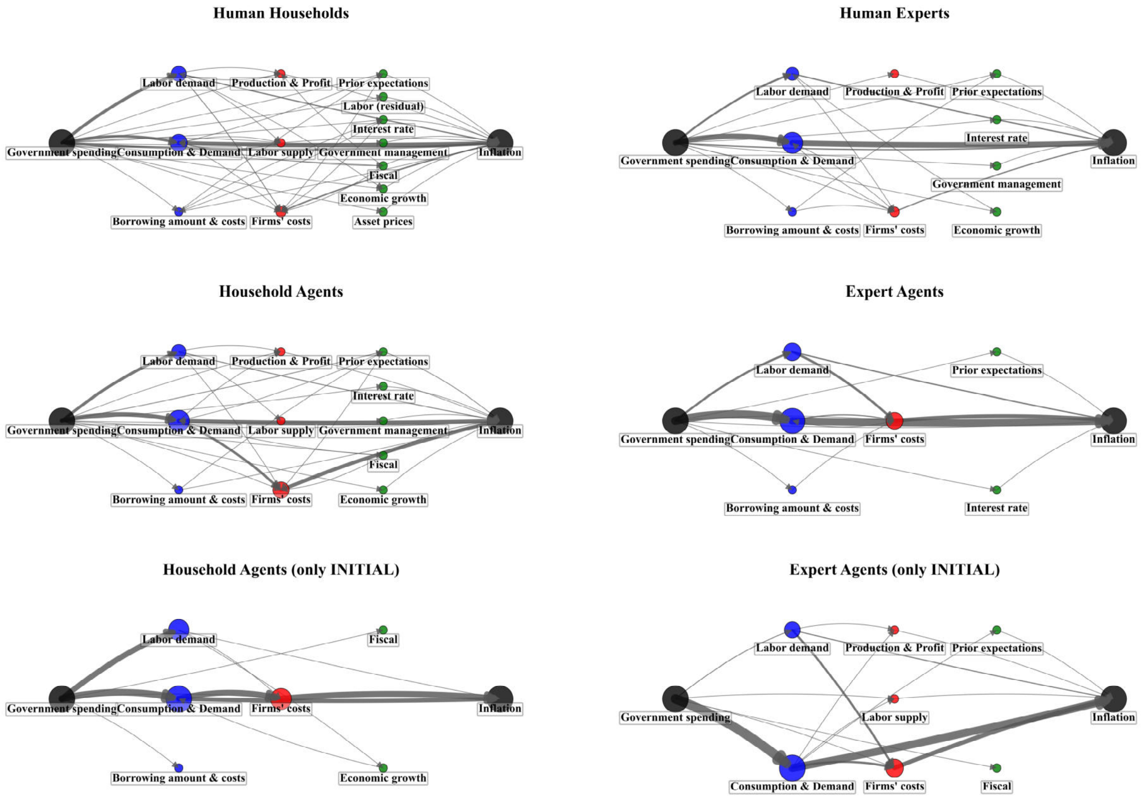


Figure A.17: The "average" DAGs underlying the formation of inflation expectations in the government spending vignette

Notes: The figure presents the "average" DAGs underlying inflation expectation formation for humans (denoted as "Human Households" and "Human Experts"), LLM Agents (denoted as "Household Agents" and "Expert Agents"), and foundation models (denoted as "Household Agents (only INITIAL)" and "Expert Agents (only INITIAL)") in the government spending vignette. The nodes represent categories of intermediate variables, whose definitions and classifications are provided in Supplementary Appendix Table A.3. The aggregated DAGs reveal the most relevant variables (nodes) and causal links in the responses of humans, LLM Agents, and foundation models. Node size: The size of the nodes is proportional to the share of responses that refer to the nodes. Node color: Red indicates supply-side variables, blue indicates demand-side variables, green indicates miscellaneous variables, black is used for start and end nodes. Edge thickness: The thickness of the edges is proportional to the share of responses that refer to the causal connections (among humans, LLM Agents and foundation models, respectively).

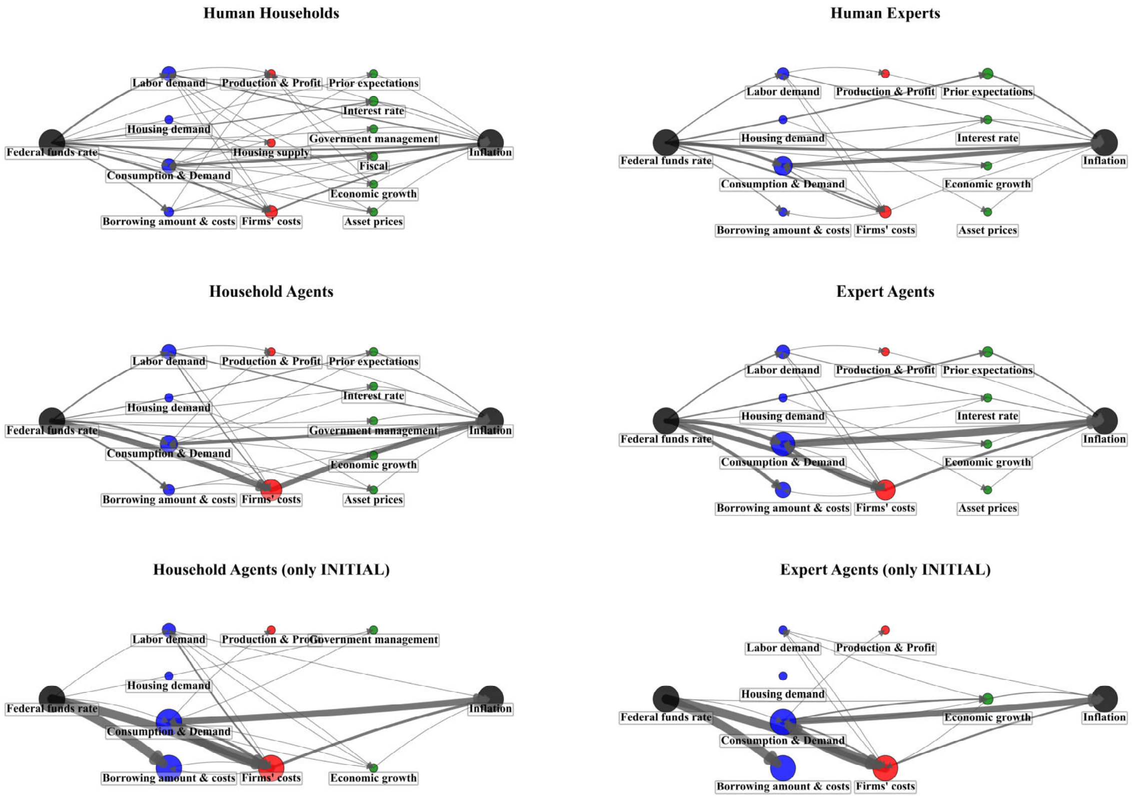


Figure A.18: The "average" DAGs underlying the formation of inflation expectations in the interest rate vignette

Notes: The figure presents the "average" DAGs underlying inflation expectation formation for humans (denoted as "Human Households" and "Human Experts"), LLM Agents (denoted as "Household Agents" and "Expert Agents"), and foundation models (denoted as "Household Agents (only INITIAL)" and "Expert Agents (only INITIAL)") in the interest rate vignette. The nodes represent categories of intermediate variables, whose definitions and classifications are provided in Supplementary Appendix Table A.3. The aggregated DAGs reveal the most relevant variables (nodes) and causal links in the responses of humans, LLM Agents, and foundation models. Node size: The size of the nodes is proportional to the share of responses that refer to the nodes. Node color: Red indicates supply-side variables, blue indicates demand-side variables, green indicates miscellaneous variables, black is used for start and end nodes. Edge thickness: The thickness of the edges is proportional to the share of responses that refer to the causal connections (among humans, LLM Agents and foundation models, respectively).

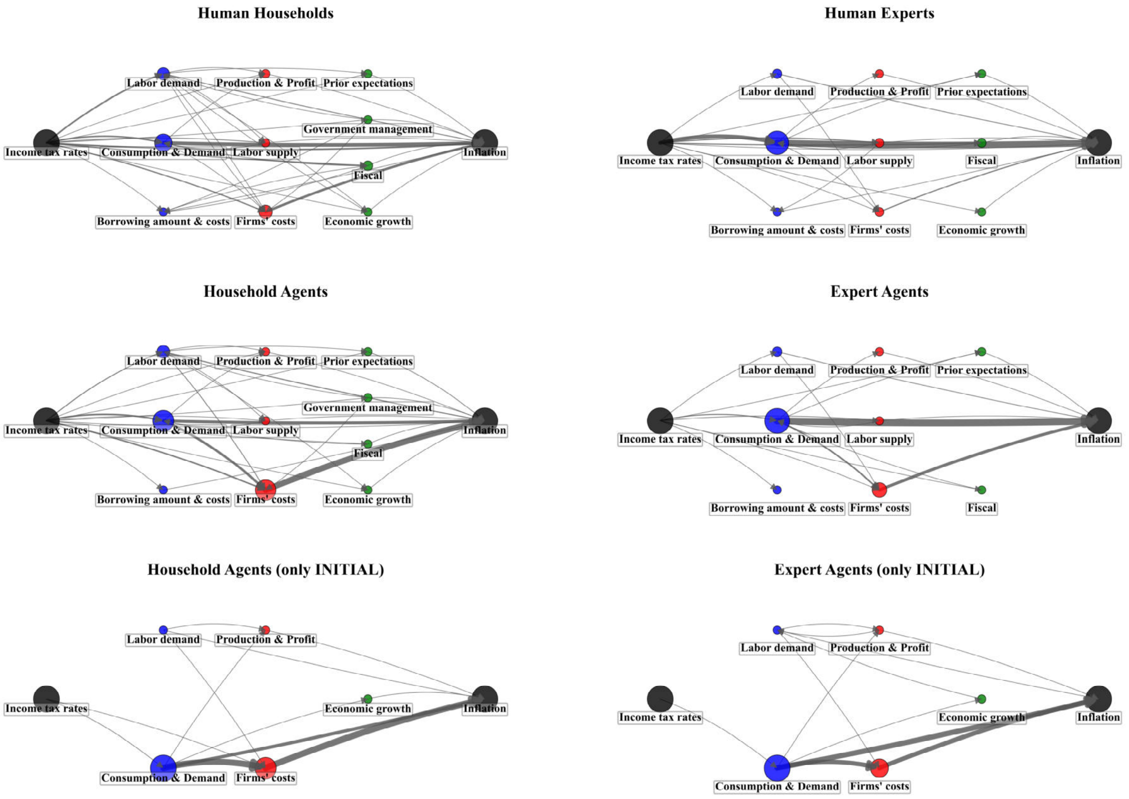


Figure A.19: The "average" DAGs underlying the formation of inflation expectations in the taxation vignette

Notes: The figure presents the "average" DAGs underlying inflation expectation formation for humans (denoted as "Human Households" and "Human Experts"), LLM Agents (denoted as "Household Agents" and "Expert Agents"), and foundation models (denoted as "Household Agents (only INITIAL)" and "Expert Agents (only INITIAL)") in the taxation vignette. The nodes represent categories of intermediate variables, whose definitions and classifications are provided in Supplementary Appendix Table A.3. The aggregated DAGs reveal the most relevant variables (nodes) and causal links in the responses of humans, LLM Agents, and foundation models. Node size: The size of the nodes is proportional to the share of responses that refer to the nodes. Node color: Red indicates supply-side variables, blue indicates demand-side variables, green indicates miscellaneous variables, black is used for start and end nodes. Edge thickness: The thickness of the edges is proportional to the share of responses that refer to the causal connections (among humans, LLM Agents and foundation models, respectively).

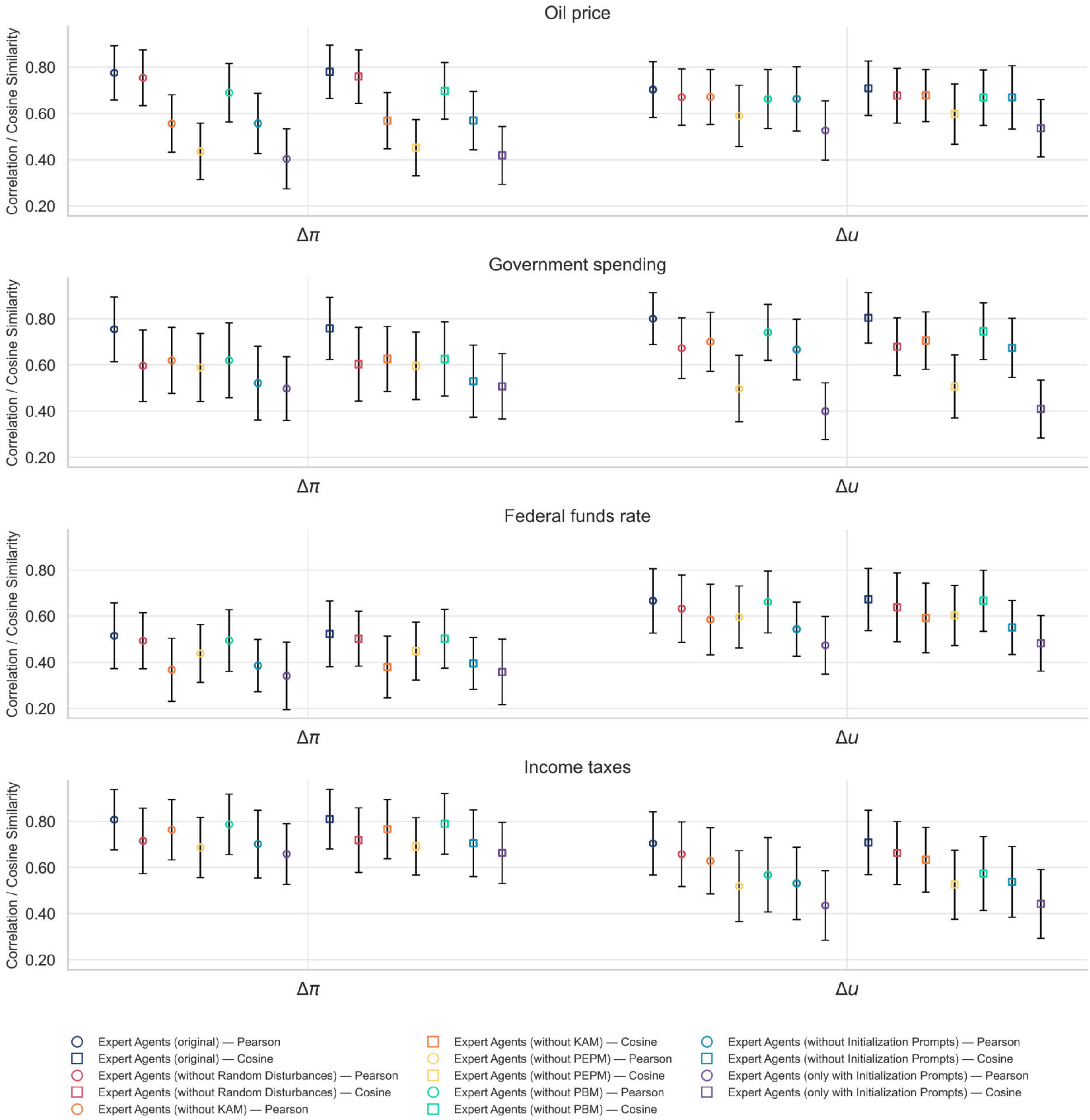


Figure A.20: Shape similarity between the expectation distributions generated by Expert Agents (original and those without modules) and those generated by humans in hypothetical vignette experiments

Notes: This figure displays the distributional shape similarity, as measured by Pearson correlation and cosine similarity, between the changes in inflation expectations ($\Delta\ \pi$) and unemployment expectations ($\Delta\ u$) generated by Expert Agents (original and those without modules), respectively, and those of experts under four different vignettes. Error bars present two-sided 95% confidence intervals for the similarity metrics, obtained by bootstrap over histogram-based probability vectors.

Figure A.21: Shape similarity between the expectation distributions generated by LLM Agents (original and those without modules) and those generated by humans in information provision experiments

Notes: This figure displays the distributional shape similarity, as measured by Pearson correlation and cosine similarity, between the home price expectations generated by LLM Agents (original and those without modules), respectively, and those of humans in different treatment groups. Error bars present two-sided 95% confidence intervals for the similarity metrics, obtained by bootstrap over histogram-based probability vectors.

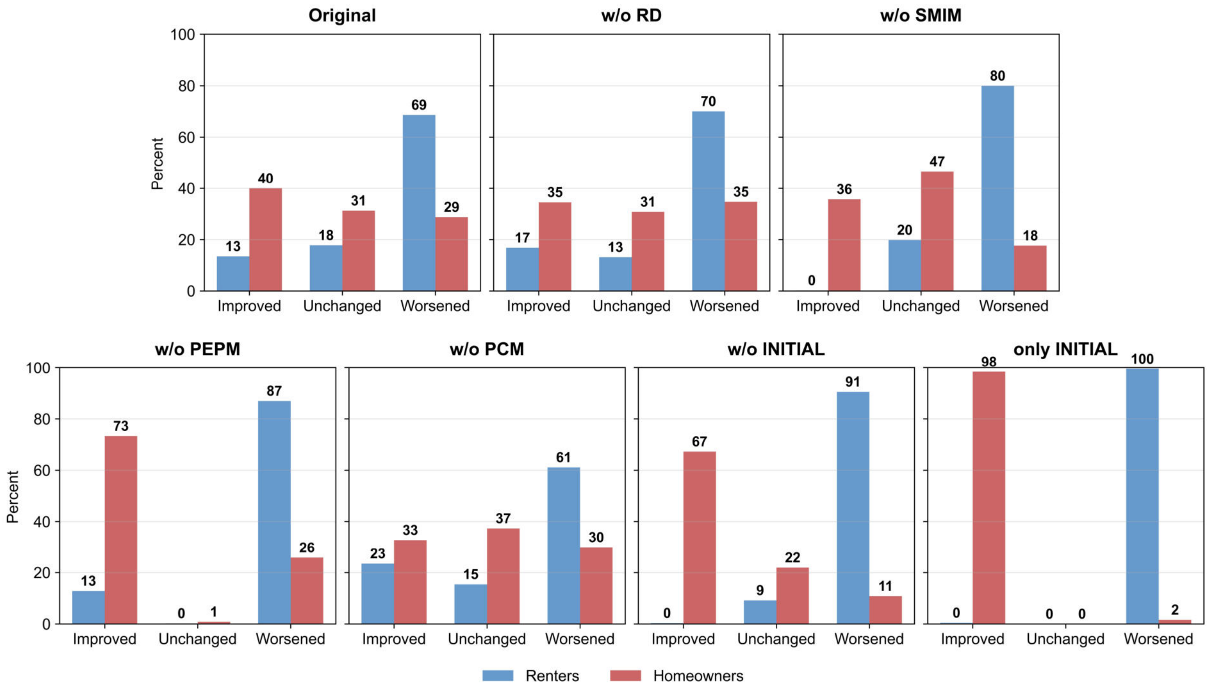


Figure A.22: Comparison of original LLM Agents' simulated results with those of LLM Agents without modules in Sub-Experiment 2 of the information provision experiments

Notes: This figure compares the changes in expectations about their household's future economic situation generated by original LLM Agents with those generated by LLM Agents without modules in Sub-Experiment 2 of the information provision experiments. The horizontal axis in each subplot represents the three possible directions of changes in expectations (improved, unchanged, worsened), and the vertical axis in each subplot indicates the percentage of respondents selecting each direction.

Figure A.23: Shape similarity between the expectation distributions generated by LLM Agents (original and those without modules) and those generated by humans in Michigan Surveys of Consumers

Notes: This figure displays the distributional shape similarity, as measured by Pearson correlation and cosine similarity, between long- and short-run inflation expectations and home price expectations generated by LLM Agents (original and those without modules), respectively, and those of households in 2025 Michigan Surveys of Consumers. Error bars present two-sided 95% confidence intervals for the similarity metrics, obtained by bootstrap over histogram-based probability vectors.

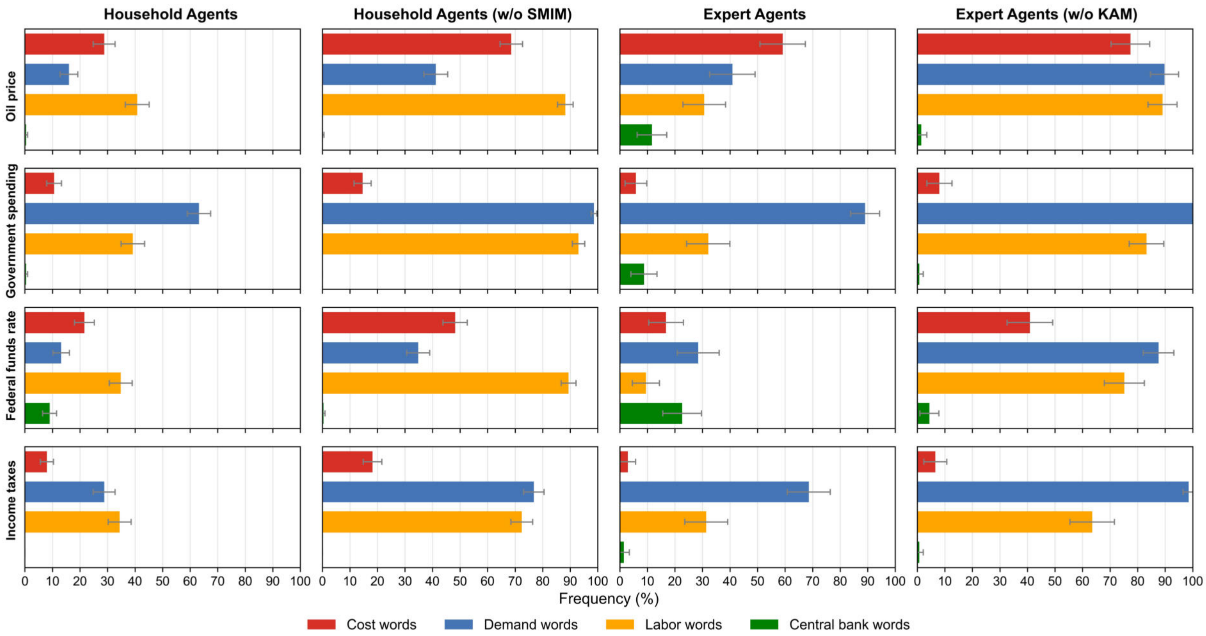


Figure A.24: Word usage for open-ended responses of LLM Agents (original and those without modules)

Notes: This figure presents the proportions of Household Agents (Column 1), Household Agents without SMIM (Column 2), Expert Agents (Column 3), and Expert Agents without KAM (Column 4) mentioning words from four word groups in their open-ended responses under four different vignettes. The error bars indicate 95% confidence intervals.

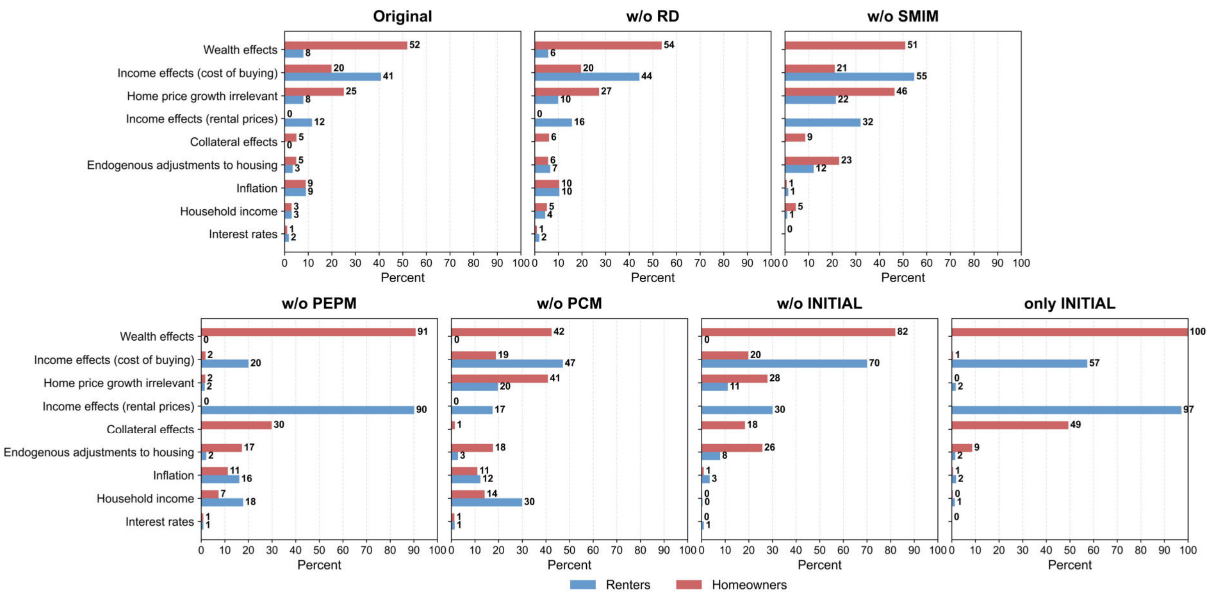
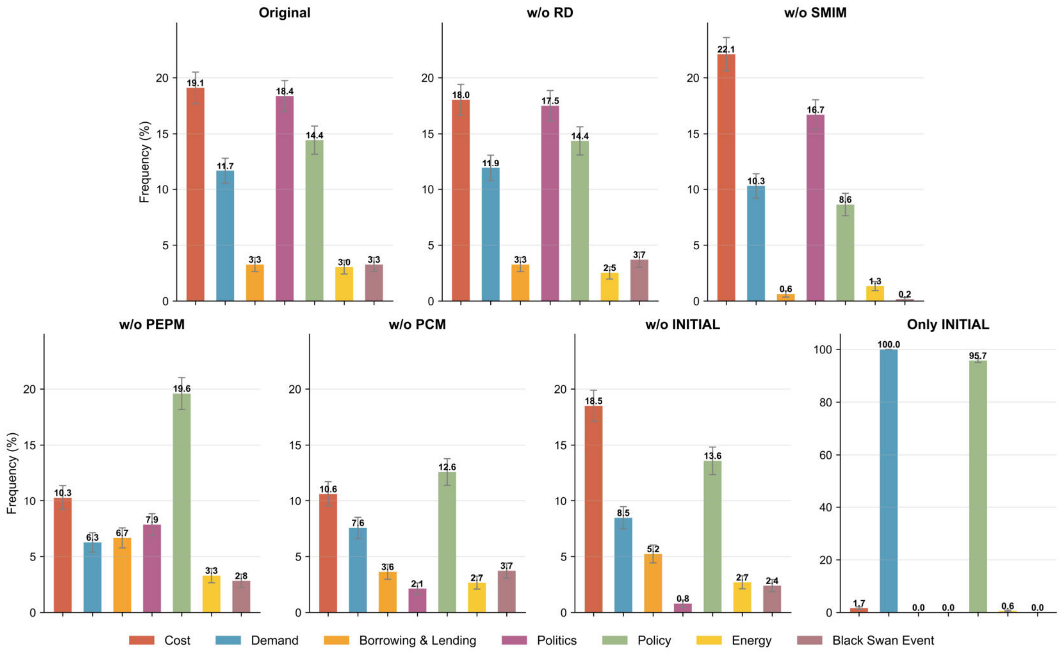
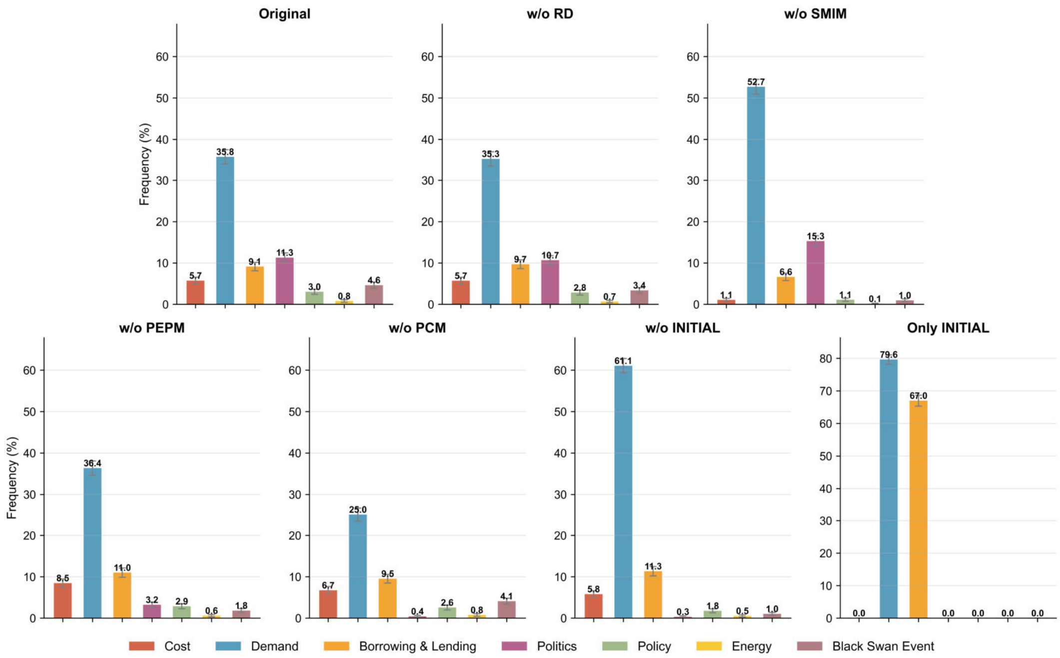


Figure A.25: Open-ended responses on how higher expected home price growth affects expectations about economic situation generated by LLM Agents (original and those without modules)

Notes: The figure shows the proportion of LLM Agents (original and those without modules) who invoke different arguments to explain why an increase in their expectations about home price growth over the next 10 years would affect their household economic outlook. The open-ended responses from all LLM Agents are automatically classified by an agentic workflow and manually verified.



(a) Inflation expectations



(b) Home price expectations

Figure A.26: Proportion of various channels recalled by Household Agents (original and those without modules) when pre-estimating 2025 macroeconomic expectations

Notes: This figure displays the proportions of seven channels recalled by Household Agents (original and those without modules) when pre-estimating inflation and home price expectations for 2025. Panel (a) presents the results for inflation expectations, while Panel (b) shows the results for home price expectations. Error bars display 95% confidence intervals.

D Robustness of Simulation Results

To account for stochastic elements in information acquisition and matching between Household and Expert Agents, as well as potential randomness in LLM outputs[3], we examine the robustness of simulation results. Considering cost and runtime, we repeatedly run the oil price vignette experiment 20 times, each with different random seeds. For Household Agents, we randomly sample 100 cases based on demographic characteristics due to the large original dataset, while for Expert Agents, we utilize all 137 synthetic samples but randomly reassign personal information in PBM to priors in PEPM.

Figure A.27 shows the distributions of unemployment and inflation expectations from 20 simulation runs. It can be observed that the expectation distributions generated in each run are closely aligned for both Household Agents and Expert Agents. Furthermore, we applied the Kruskal–Wallis test to rigorously examine whether significant differences exist among the distributions from these 20 simulations. As presented in Table A.7, the test results indicate that we cannot reject the null hypothesis (p-values are at least 0.85), so the distributions across all 20 simulations originate from the same population and show no statistically significant differences. This suggests that, although the simulation results of LLM Agents exhibit variability due to randomness, the differences are not statistically significant.

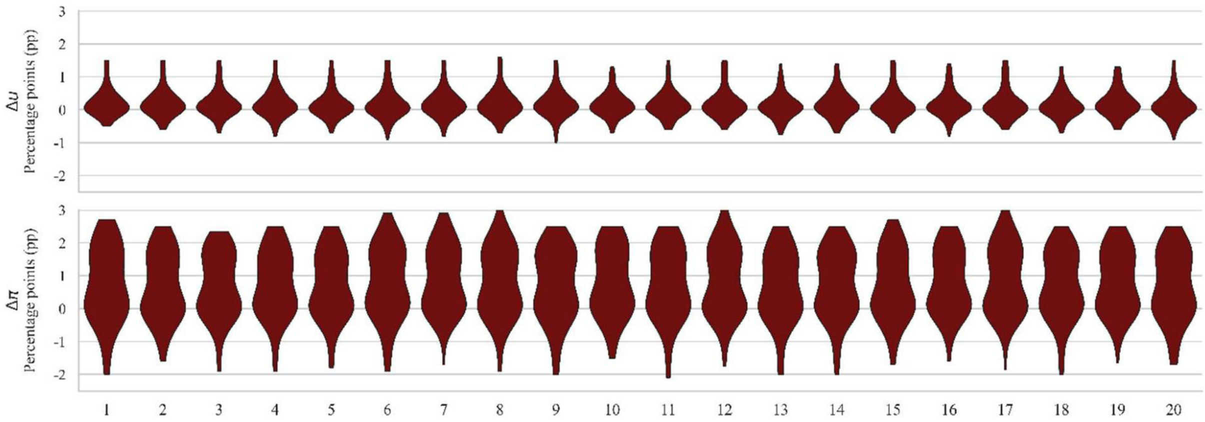
Additionally, we examine whether there are significant differences among the 20 reasoning processes generated by the LLM Agents. Specifically, we first use the Sentence-BERT (SBERT) model all-MiniLM-L6-v2[4] to obtain embeddings for each open-ended response from every run of the LLM Agents. We then aggregate the embeddings of all open-ended responses in each run into a global embedding—representing the overall semantic

---

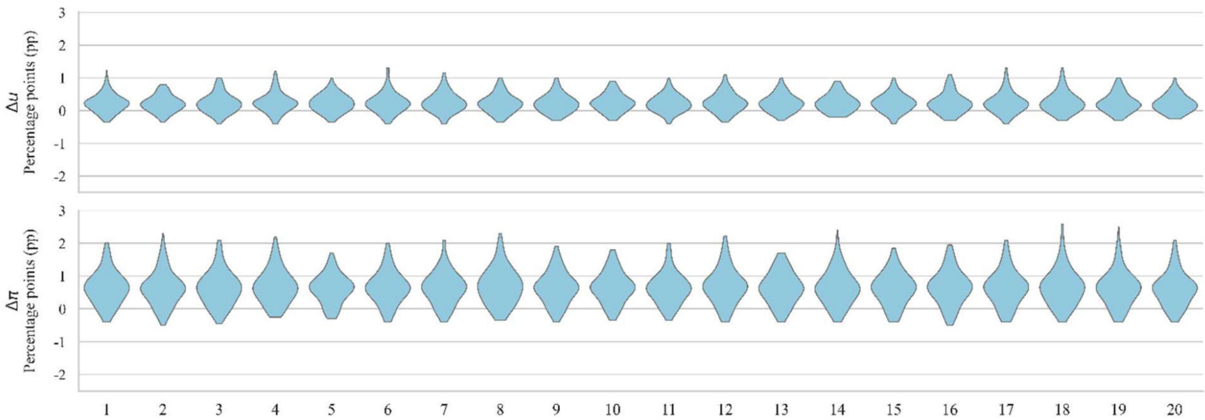
[3] Even if we set and keep a fixed random `'seed'` in the LLMs' parameters across all runs, this only partially reduces output randomness and cannot eliminate it entirely.

[4] Detailed information on the model is available at `https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2`.

information of that run—via four pooling methods: max pooling, mean pooling, min pooling, and weighted mean pooling based on text length. Next, we pair the global embeddings of all 20 runs and identify the lowest cosine similarity among all pairs. As shown in Table A.8, for both Household Agents and Expert Agents, the minimum semantic similarity between any two runs exceeds 0.97, regardless of the pooling method. These results collectively indicate strong robustness in the simulation outcomes of the LLM Agents, especially when the sample size is not small.



(a) Simulation results generated by Household Agents



(b) Simulation results generated by Expert Agents

Figure A.27: Expectation distributions simulated by LLM Agents under oil price shocks in 20 simulations

Notes: Panels (a) and (b) present the results of 20 simulations showing the distributions of changes in expectations (with trimmed 5% tails) generated by Household Agents and Expert Agents in the oil price vignette, respectively. The x-axis labels the simulations from 1 to 20. The y-axis represents the percentage range of changes in expectations.

Table A.7: Kruskal–Wallis test on expectation distributions across 20 simulations

Agent	**Variable**	**Total N**	**Mean N per Group**	**H-statistic**	**p-value**
Household Agents	$\Delta\ \pi$	3830	191.5	12.6594	0.8555
	$\Delta\ u$	3830	191.5	11.2398	0.9155
Expert Agents	$\Delta\ \pi$	5214	260.7	9.0924	0.9719
	$\Delta\ u$	5229	261.4	6.8884	0.9948

Notes: The table presents the Kruskal–Wallis test results on the distributions of changes in inflation ($\Delta\ \pi$) and unemployment expectations ($\Delta\ u$) generated by Household Agents and Expert Agents over 20 simulations. The null hypothesis states that all independent distributions have the same central tendency and thus originate from the same population; the alternative hypothesis posits that at least one distribution differs in central tendency and therefore originates from a different population.

Table A.8: The minimum semantic similarity between the thoughts of LLM Agents across 20 simulations

Agent	**Pooling Method**	**Minimum Similarity**
Household Agents	Max	0.979337
	Mean	0.997027
	Min	0.975266
	Weighted Mean	0.996773
Expert Agents	Max	0.977026
	Mean	0.997873
	Min	0.976205
	Weighted Mean	0.997891

Notes: This table presents the minimum pairwise consine similarity of LLM Agents' thought processes across 20 simulations, where the overall embeddings of open-ended responses in each simulation are aggregated using max pooling, mean pooling, min pooling, and length-weighted mean pooling, respectively.

E Analysis of Mental Models

E.1 How to Identify Directed Acyclic Graphs

We develop an agentic workflow to automatically identify and label DAGs in open-ended responses, with its architecture shown in Figure A.12. First, we employ two medium-scale LLMs of different types—deepseek-r1-distill-qwen-32b and qwq-32b—as Causal Analyzers. Each independently identifies causal pathways in open-text responses related to unemployment and inflation expectations. Given the vignette and expectation type, the start and end points of each causal pathway are predetermined; thus, the workflow focuses on identifying and categorizing all intermediate variables (see Table A.3) in these pathways. We illustrate the tasks of the Causal Analyzers in the first round using the example of processing open-text responses about inflation expectations under an oil price vignette, as detailed in the following prompt:

```
You are an expert economic analyst specializing in inflation causal structure identif
ication. Your task is to analyze survey responses and identify causal pathways from o
il price changes to inflation outcomes.
Note: The survey is conducted under a scenario of rising oil price and is intended to
 examine how respondents perceive the causal pathways of the impact of rising oil pri
ce on inflation.
For each text sample (i.e. response), you must identify causal structures in the foll
owing format:
{Oil price}→{Intermediate variable 1 (optional)}→{Intermediate variable 2 (optiona
l)}→...→{Inflation}
The intermediate variables must come from this predefined list:
[See Table A.3, which will not be elaborated here.]
It is important to note that if the text is complex and mentions multiple variables a
nd their causal relationships, there may be more than one causal path related to infl
ation. For example, the causal paths regarding inflation may include:
{Oil price} → {Inflation};
{Oil price} → {Intermediate variable 1} → {Inflation};
{Oil price} → {Intermediate variable 2} → {Intermediate variable 3} → {Inflation}.
Analyze the following text and identify causal pathways from oil price to inflation:
<text>{start}</text>
Identify all causal pathways that lead from oil price increases to inflation and prov
ide corresponding reasons. Focus specifically on inflation-related outcomes.  Please
```

```
make sure to identify all mentioned intermediate variables; even if these variables a
re considered unchanged, they should still be included in the causal pathways as long
 as they are mentioned. Output your response strictly in the following JSON format:
{
  "reasons": "Explain your reasoning for giving this identification. (about 3-4 sente
nces)",
  "pathways": [
    "pathway1",
    "pathway2",
    "..."
  ]
}
Note: If multiple pathways exist, include the most important ones (no more than thre
e) in the array. If only one pathway exists, include only one string. If no relevant
pathway to inflation is found, use an empty array []. Each pathway should be a string
 in the format: "{Oil price}→{intermediate variables}→{Inflation}".
```

Second, upon completion of Round 1, the two Analyzers discuss their initial findings. If their outputs align, they pass the consistency check directly. If not, they engage in iterative discussion until consensus is reached. Should consensus remain elusive after 10 rounds, the entire process is repeated until agreement is achieved.

Third, a Classifier categorizes the intermediate variables within the causal pathways, consolidating diverse variables into generalized nodes. The classification criteria are predefined, as detailed in Table A.3.

Finally, a Results Consolidator integrates the categorized outcomes. The consolidation rules are as follows: (1) If multiple intermediate variables from the same category (node) appear consecutively in a pathway, merge them into a single node. (2) If multiple pathways become identical after the previous step, merge them into a single pathway. (3) For other cases, simply update the pathways with the category information and output directly.

We demonstrate the output of the Classifier and Consolidator with a simple example: If the agreed-upon pathways after discussion are as follows:

{Oil price}→{cost firms}→{firm prices}→{Inflation} and {Oil price}→{costs borrowing firms}→{Inflation}.

After categorization: {Oil price}→{Firms' costs}→{Firms' costs}→{Inflation} and {Oil price}→{Firms' costs}→{Inflation}.
After consolidation (final result): {Oil price}→{Firms' costs}→{Inflation} (merged as identical).

E.2 Similarity of Mental Models

After completing the identification in Section E.1, each open-ended response is converted into a DAG (i.e., a mental model). First, we represent each DAG as an edge list—a set of unique causal connections.

For example, in a given vignette, suppose there are two human samples: Response 1 has a DAG "A→B→C" with an edge list E_1 = {A→B, B→C}; Response 2 has a DAG "A →C; B→C" with E_2 = {A→C, B→C}. The aggregated edge list for all human responses in this vignette (termed the "mental model set") is then E_{Human} = {A→B, B→C, A→ C}. Similarly, we assume the mental model set for LLM Agents in this vignette is E_{Agent} = {A→B, A→C, C→D}.

Next, we measure the similarity between the mental models of humans and LLM Agents in this vignette using the Jaccard similarity between their mental model sets

$$Sim(Human, Agent) = \frac{|E_{Human} \cap E_{Agent}|}{|E_{Human} \cup E_{Agent}|},$$

where $|\cdot|$ denotes the number of elements in a set.

The Jaccard similarity equals 1 (0) if and only if the two mental models are identical (completely different). This similarity increases as the number of common elements between the two sets grows. For example, in the case above, the similarity between the mental models of humans and LLM Agents is $Sim(Human, Agent) = \frac{2}{4} = \frac{1}{2}$.

F Measuring the Diversity of the Thoughts Underlying Expectations

In this section, we detail how to measure the diversity of thoughts underlying respondents' expectations, which essentially reflects the semantic diversity in their open-ended responses. Given n open-ended responses, we convert each into n embeddings using the all-MiniLM-L6-v2 model to capture their semantic information.

Next, we compute the pairwise cosine similarity between all embeddings to obtain a similarity matrix $\boldsymbol{S} \in \mathbb{R}^{n \times n}$, where element $S_{ij} = \frac{\mathbf{e}_i \cdot \mathbf{e}_j}{\|\mathbf{e}_i\| \|\mathbf{e}_j\|}$ denotes the semantic similarity between response i and response j, with $\mathbf{e}_i$ and $\mathbf{e}_j$ being their corresponding embeddings.

Then, we compute the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ of matrix $\boldsymbol{S}$. After filtering out the few negative values, these eigenvalues are normalized into a probability distribution $p_i = \frac{\lambda_i}{\sum_{j=1}^{k} \lambda_j}$, where k is the number of positive eigenvalues. The *Shannon entropy* is then calculated as

$$H = -\sum_{i=1}^{k} p_i ln\,(p_i).$$

A higher entropy value H indicates a more uniform eigenvalue distribution, reflecting greater dispersion of texts in the semantic space and thus higher semantic diversity, and vice versa. Finally, to eliminate the influence of text quantity on entropy, H is normalized to [0, 1] by the maximum entropy $\ln k$

$$D = \frac{H}{\ln k},$$

where D represents the metric of semantic diversity used in this paper.

This method is applied to the open-ended responses from three representative experiments, and the resulting diversity of underlying thoughts generated by LLM Agents with different components removed and humans is shown in Table A.4 to Table A.6, respectively.

References

Andre, P., Haaland, I., Roth, C., Wiederholt, M., & Wohlfart, J. (2025). Narratives about the Macroeconomy. *The Review of Economic Studies (Forthcoming).*

Andre, P., Pizzinelli, C., Roth, C., & Wohlfart, J. (2022). Subjective Models of the Macroeconomy: Evidence From Experts and Representative Samples. *The Review of Economic Studies*, *89*(6), 2958–2991. https://doi.org/10.1093/restud/rdac008

Armona, L., Fuster, A., & Zafar, B. (2019). Home Price Expectations and Behaviour: Evidence from a Randomized Information Experiment. *The Review of Economic Studies*, *86*(4), 1371–1410. https://doi.org/10.1093/restud/rdy038

Campbell, J. Y., & Cocco, J. F. (2007). How do house prices affect consumption? Evidence from micro data. *Journal of Monetary Economics*, *54*(3), 591–621. https://doi.org/10.1016/j.jmoneco.2005.10.016

Chopra, F., Roth, C., & Wohlfart, J. (2025). Home Price Expectations and Spending: Evidence from a Field Experiment. *American Economic Review*, *115*(7), 2267–2305. https://doi.org/10.1257/aer.20240022

Coibion, O., Gorodnichenko, Y., Kumar, S., & Pedemonte, M. (2020). Inflation expectations as a policy tool? *Journal of International Economics*, *124*, 103297. https://doi.org/10.1016/j.jinteco.2020.103297

Guren, A. M., McKay, A., Nakamura, E., & Steinsson, J. (2021). Housing Wealth Effects: The Long View. *The Review of Economic Studies*, *88*(2), 669–707. https://doi.org/10.1093/restud/rdaa018

Kuchler, T., & Zafar, B. (2019). Personal Experiences and Expectations about Aggregate Outcomes. *The Journal of Finance*, *74*(5), 2491–2542. https://doi.org/10.1111/jofi.12819

Malmendier, U., & Nagel, S. (2016). Learning from Inflation Experiences. *The Quarterly Journal of Economics*, *131*(1), 53–87. https://doi.org/10.1093/qje/qjv037

Piazzesi, M., & Schneider, M. (2009). Momentum Traders in the Housing Market: Survey Evidence and a Search Model. *American Economic Review*, *99*(2), 406–411. https://doi.org/10.1257/aer.99.2.406