Do LLM Agents Really Mimic Humans? Diagnosing and Aligning Microeconomic Behaviors in Macro-ABMs

Guangya Liu[1,2], **Cheng Wang**[1,2], **Jiangtong Li**[1,2], **Huafei Wu**[1,2], **Changjun Jiang**[1,2]*
[1]Key Laboratory of Embedded System and Service Computing,
Ministry of Education, Tongji University
[2]School of Computer Science and Technology, Tongji University
{2410928, cwang, jiangtongli, hfwu, cjjiang}@tongji.edu.cn

Abstract

Large Language Models (LLMs) are increasingly adopted in macroeconomic agent-based modeling (ABM). However, existing research focuses on replicating macro-level stylized facts while often neglecting verification of micro-level decision-making. We investigate this gap by comparing LLM agents to human responses from the Survey of Consumer Expectations (SCE) dataset. Our empirical analysis identifies specific limitations: weak trend responsiveness, mode collapse, and a potential data leakage. We propose the Heterogeneous Shock-Response Causal Transmission Framework to tackle these issues. To ensure theoretical consistency, we use LLMs to build a literature-verified causal graph in which macroeconomic shocks influence decisions via generated **mediator nodes**, while agent profiles serve as edge **moderators**. Building on this, during inference, we perform a path search to retrieve relevant causal chains and inject them as an explicit **Chain-of-Thought (CoT)**, prioritizing mechanistic logic over statistical pattern matching. To evaluate the effectiveness of our inference approach, we validate it via a two-stage process that combines micro-level dataset testing and macro-level simulation in the *EconAgent* system. Results from these experiments indicate that our framework improves alignment with human trends and effectively captures behavioral heterogeneity. Overall, this work contributes to the development of reliable and grounded economic simulations.

1 Introduction

The application of Large Language Models (LLMs) has expanded the methodology of Computational Social Science, particularly within economic ABM (Gao et al., 2024a; Li et al., 2024; Tang et al., 2024). In contrast to rule-based agents (Brock and Hommes, 1998; Tesfatsion and Judd, 2006), LLM agents process economic signals (*e.g.*, inflation, unemployment) and individual profiles to formulate microeconomic decisions, including consumption levels and labor supply. Aggregating these individual behaviors enables the simulation of macroeconomic dynamics.

Existing studies (Li et al., 2024; Piatti et al., 2024; Wang et al., 2025) primarily validate these systems by evaluating if aggregated results replicate established "stylized facts," such as the Phillips Curve (Lucas and Rapping, 1969) and Okun's Law (Gordon, 2010). However, these works rest on an implicit, unverified assumption: *that LLM agents adjust their behaviors under changing economic conditions consistent with actual human decision-making.* The absence of micro-level validation undermines the reliability of such simulations. Verifying this behavioral consistency is a foundational requirement for deploying these models in real-world policy analysis, such as evaluating interest rate adjustments.

To investigate the validity of these microfoundations, our research framework is structured as illustrated in Fig. 1. We first conduct an empirical comparison between LLM agents and human respondents using the SCE dataset (Armantier et al., 2017), a nationally representative source containing detailed demographic profiles and structured responses across core economic domains such as spending and labor. Our analysis identifies three critical misalignments. First, regarding **trend alignment**, LLMs fail to adjust their behavior in response to macroeconomic indicators, diverging from the reaction patterns observed in empirical data. Second, with respect to **heterogeneity**, we observe severe **mode collapse**: despite distinct persona profiles, LLM agents converge toward an "average persona," exhibiting significantly lower behavioral variance than human populations (Anthis et al., 2025). Third, we attribute apparent **data leakage** to the finding that model alignment is of-

* Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 36102–36122

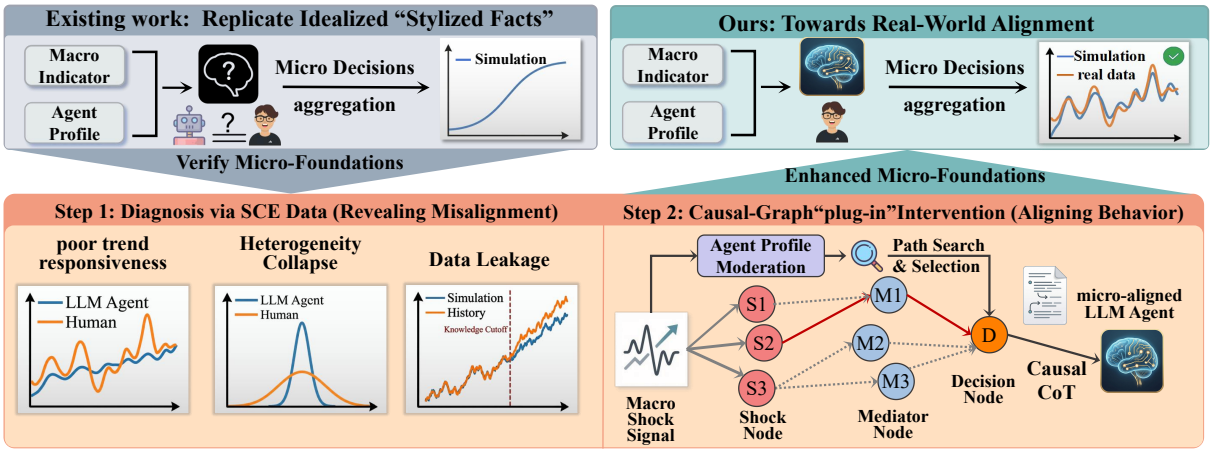


Figure 1: **To ground macro-simulations in micro-foundations, our workflow transitions from unverified aggregate modeling to individual-level analysis.** We **diagnose** and **enhance** agent behaviors (Step 1 & 2), ensuring that the final **macro-level aggregation** is built upon empirically aligned micro-decisions.

ten an artifact of memorization. While including specific dates in prompts artificially inflates correlations with human data, performance degrades significantly post-training cutoff.

To address these limitations, we propose the **H**eterogeneous **S**hock-**R**esponse **C**ausal **T**ransmission (**HSRCT**) framework, a modular intervention that aligns LLM agents with human behavior without relying on data leakage. Grounded in the economic theory of *mediation and moderation* (Baron and Kenny, 1986), we structure the reasoning process into a "Shock-Response" graph. Specifically, we use advanced LLMs to construct a causal graph in which macroeconomic shocks influence decisions via generated **mediator nodes** (transmission mechanisms). To ensure accuracy, we prompt the LLM to verify the economic grounding of each edge by retrieving corroborating academic literature via the Google Scholar API and performing iterative validation. Furthermore, agent profiles serve as edge moderators, explicitly accounting for individuals' heterogeneous sensitivities to macroeconomic shocks. During inference, we execute a path search on this graph to retrieve relevant causal chains, which are then incorporated into the prompt as an explicit CoT (Wei et al., 2022). This method decouples macro signal interpretation from decision-making, ensuring mechanistic reasoning takes precedence over statistical pattern matching.

We validate our method through a two-stage evaluation strategy. First, we use the SCE dataset to verify our framework's capacity to reproduce human decision-making dynamics. Second, to assess generalization capabilities in a dynamic setting, we utilize the *EconAgent* (Li et al., 2024) simulation system as a testbed. Extensive experiments on the SCE dataset indicate that our framework improves Spearman correlations with human trends and effectively restores behavioral heterogeneity. Furthermore, at the macro level, the simulation replicates real-world historical economic trends, surpassing the reproduction of simple stylized facts. Our contributions are summarized as

- We conduct a systematic diagnosis of micro-level alignment between LLM agents and human economic behaviors using SCE data, identifying distinct failures in trend alignment and heterogeneity.
- We propose the HSRCT framework, a modular intervention that structures reasoning via mediator-moderator logic to reduce mode collapse while improving interpretability.
- We demonstrate that our method achieves significant improvements in both micro-level behavioral alignment and system-level macroeconomic simulation, establishing a robust foundation for realistic economic modeling.

2 Related Work

Our study spans two research domains: (1) LLM-based economic simulation, and (2) behavioral alignment evaluation for LLMs.

2.1 LLM-Based Economic Simulation

The advancement of LLMs has catalyzed the deployment of generative agents to capture emergent aggregate behaviors in economic simulations (Gao et al., 2024a). Early works position LLM-based

simulations as "in silico" pilot studies to generate hypotheses for real-world testing (Horton, 2023). Notably, Li et al. (2024) simulates microeconomic labor and consumption decisions to generate macroeconomic indicators such as inflation. Subsequent studies have extended these simulators to varied contexts (Mi et al., 2025; Lopez-Lira, 2025; Yuzhe et al.). However, this research predominantly prioritizes reproducing macro-level *stylized facts* (*e.g.*, the Phillips Curve (Lucas and Rapping, 1969) and Okun's Law (Gordon, 2010)), often neglecting real-world data grounding at both micro and macro levels. In this work, we bridge the gap between idealized economic simulations and realistic human behavior by diagnosing and enhancing the micro-foundations of LLM-based ABM.

2.2 Behavioral Alignment Evaluation

Assessing the fidelity of LLM simulations is essential for ensuring their reliability (Gao et al., 2024b; Xie et al., 2024). Existing works often utilize the Turing Experiment method (Aher et al., 2023; Argyle et al., 2023) but are typically restricted to highly controlled, game-theoretic settings. For instance, recent studies examine LLM strategies in the Iterated Prisoner's Dilemma (Fontana et al., 2025) or assess rationality via Rock-Paper-Scissors games (Fan et al., 2024). Despite these advances (Arora et al., 2025; Gui and Toubia, 2023; Gao et al., 2025), there is a scarcity of research examining alignment grounded in *real-world economic data*, particularly concerning how LLM behaviors adjust to *changing economic environments*.

3 Diagnosis of LLM Micro-Foundations

This section investigates the micro-foundations of LLM-based economic simulations.

3.1 Preliminaries

Data Source. The SCE is a panel survey of around $1,100$ U.S. household heads that queries spending and labor decisions every 4 months. We utilize public data spanning **24 waves from Mar. 2016 to Dec. 2023**, curating a dataset of $26,579$ spending and $26,803$ labor samples. We focus on two key microeconomic tasks:

- ***Spending Forecast***: Participants assign probabilities to potential household spending changes over the next 12 months.
- ***Offer Acceptance***: Participants report acceptance probabilities for sub-expected salary offers.

Further details about the data are provided in Appendix A. We evaluate GPT-3.5 Turbo, GPT-5, Gemini 2.5 Flash, Claude 3.5 Haiku. GPT-3.5 Turbo is chosen for its pre-September 2021 training-data cutoff, allowing evaluation on unseen data to bypass leakage risks. We include GPT-5 (Singh et al., 2025) as a SOTA representative, along with Gemini 2.5 Flash (Comanici et al., 2025) and Claude 3.5 Haiku (Anthropic, 2024) to ensure generalizability across model families.

Prompting Strategy. Each query includes four inputs: (1) respondent demographics; (2) historical macroeconomic indicators for this time step (unemployment rate, CPI, and federal funds rate from the preceding four months); (3) a memory component with responses and environment from the last time step; and (4) the specific domain question. To detect data leakage, we design a "w/ date" variant that explicitly embeds the real-world date into the prompt. To control for stochasticity, we query each prompt five times and utilize the average response ($r_{q,m}$) for analysis. Temperature and top-p are set to 1.0 to encourage diversity, following standard behavioral alignment protocols (Guo et al., 2023; Lorè and Heydari, 2023).

3.2 Evaluation Research Questions

We formulate three core research questions (RQs) to guide our evaluation:

- **RQ1.1 (Trend Alignment):** Do the aggregate behavioral adjustments of LLM agents correspond to human trends under varying economic conditions?
- **RQ1.2 (Heterogeneity):** Can LLM agents effectively capture the behavioral heterogeneity observed in empirical human populations?
- **RQ1.3 (Data Leakage):** To what extent does data leakage (rote memorization) undermine the validity of simulation outcomes?

3.3 Evaluation Result Analysis

RQ1.1 (Trend Alignment): As illustrated in Fig. 2 and Tab. 1, concerning temporal trend alignment, GPT-5 exhibits a trajectory closest to empirical human data, whereas Claude 3.5 Haiku demonstrates the lowest fidelity. However, most LLMs fail to track human behavioral trends over time consistently. We attribute this discrepancy to the inherent complexity of human decision-making mechanisms. In practice, macroeconomic indicators influence micro-behaviors via *mediation effects* (Mishkin, 1995; Kaplan et al., 2018). For instance, a rise in interest rates does not directly

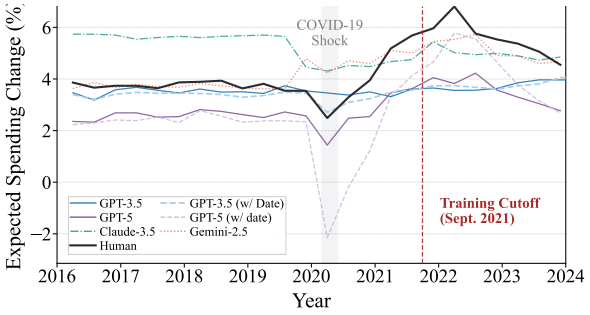


Figure 2: **Temporal Trend Alignment.** Comparison of forecast trajectories between humans and LLMs.

Model	Overall	Demographic Attributes			
		Age	Edu.	Home.	Income
GPT-3.5 Turbo	0.36	**-0.10**	-0.01	0.08	-0.21
GPT-5	**0.84**	-0.03	**0.19**	-0.10	-0.23
Claude 3.5 Haiku	-0.12	-0.08	0.18	-0.02	-0.21
Gemini 2.5 Flash	0.65	-0.03	0.12	**-0.17**	**-0.27**

Table 1: Spearman correlation coefficients (ρ) of different LLMs on Spending Forecast. Red and Blue indicate positive and negative correlations, respectively, with color intensity representing the magnitude.

dictate consumption; instead, it elevates the cost of borrowing, constraining disposable liquidity and subsequently suppressing consumption. Furthermore, multiple factors simultaneously impact decisions through distinct mediation channels, necessitating the identification of dominant causal pathways. This difficulty is further intensified during periods of extreme economic volatility, such as the time after the COVID-19 shock. Therefore, LLMs face difficulties in reasoning directly from macroeconomic indicators.

RQ1.2 (Heterogeneity): In Fig. 3, we visualize response distributions: while human data exhibits a flat distribution reflecting population heterogeneity, LLMs show a highly concentrated distribution, indicating severe **mode collapse**. From a theoretical perspective, human heterogeneity stems from *moderation effects*, where attributes (*e.g.*, income, savings) modulate responses to macroeconomic shocks (Kaplan and Violante, 2014)—a dynamic LLMs fail to replicate. We attribute this collapse to training objectives like Maximum Likelihood Estimation and Reinforcement Learning from Human Feedback (Ouyang et al., 2022; Casper et al., 2023), which favor probable or "safe" responses, thereby suppressing the variance needed to mimic diverse behaviors. Following standard heterogeneity analysis practices (Wooldridge, 2010), we examine demographic alignment by calculating time-series

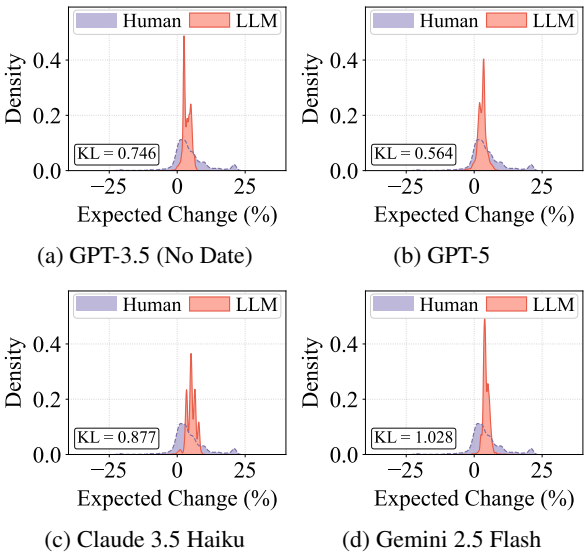


Figure 3: Comparison of response distributions with KL-divergence.

Model	Prompt	Overall	Time Period	
			Pre-Cutoff	Post-Cutoff
GPT-3.5 Turbo	w/ Date	0.71	0.36	-0.39
	w/o Date	0.36	0.00	-0.45
GPT-5	w/ Date	0.83	0.55	**0.96**
	w/o Date	**0.84**	**0.56**	0.86

Table 2: Spearman correlation coefficients (ρ) on the Spending Forecast task using prompts with (**w/ Date**) and without (**w/o Date**) explicit survey dates.

correlations between specific human profile groups and their corresponding LLM agents. In Tab. 1, we report the weighted average correlations, revealing that even for advanced models like GPT-5, alignment deteriorates at the granular level.

RQ1.3 (Data Leakage): In Tab. 2, we observe distinct divergences in model responses to temporal cues. For **GPT-3.5 Turbo**, results indicate an explicit dependency on memorization. In the pre-cutoff period, injecting dates boosts performance significantly (0.00 to 0.36), confirming that the model uses dates as retrieval cues for training memory. However, in the post-cutoff period, including dates fails to improve performance. This contrast suggests GPT-3.5's alignment relies on rote memorization; once data exceeds the training horizon, neither internal knowledge nor date cues support valid reasoning. Conversely, **GPT-5** exhibits a distinct pattern. Injecting dates yields a negligible performance impact (0.84 w/o Date vs. 0.83 w/ Date). This suggests GPT-5 may have implicitly internalized macroeconomic mappings (effectively "memorizing" economic laws), rendering date cues redundant. Consequently, distinguishing genuine

reasoning from implicit leakage in GPT-5 becomes nearly impossible. Therefore, we select GPT-3.5 Turbo as the backbone for the following experiments to minimize leakage. Refer to Appendix C.1 for analysis in the labor domain.

4 Intervention

As illustrated in Fig. 4, the proposed **HSRCT** framework comprises two core components: **(i) Causal Transmission Graph Modeling** and **(ii) Causal Transmission-Augmented Reasoning**.

4.1 Causal Transmission Graph Modeling

This module constructs a directed causal graph (Pearl, 2009; Kiciman et al., 2024) that explicitly models the propagation of macroscopic shocks through intermediate mechanisms to shape individual-level decisions. The graph encodes two economic dynamics: a *mediation* effect that defines causal transmission pathways, and a *moderation* effect that captures demographic heterogeneity.

4.1.1 Mediation Effect Layer Construction

Standard prompting strategies often map macroeconomic shocks directly to micro-responses, neglecting intermediate transmission mechanisms. To address this, we introduce a **Mediation Effect Layer** based on a structured $L_0 \rightarrow L_1 \rightarrow L_2$ architecture, where macroeconomic shocks (L_0) influence micro-outcomes (L_2) via specific mediators (L_1).
(1) Mediator Generation. We first define the node topology, comprising three categories:

• **Shock Nodes (L_0):** Pre-defined macroeconomic indicators (*e.g.*, *Interest Rate*).
• **Mediator Nodes (L_1):** Intermediate mechanisms synthesized by the LLM (*e.g.*, *Consumer Confidence*, *Liquidity Constraints*).
• **Decision Nodes (L_2):** Task-specific behavioral outcomes (*e.g.*, *Household Spending Forecast*).

While L_0 and L_2 are fixed by task definitions, mediator nodes (L_1) are generated by the LLM to bridge the transmission gap between macro-shocks and micro-decisions.
(2)Causal Edge Specification. The LLM establishes layer-wise connectivity ($L_n \rightarrow L_{n+1}$) to form directed edges, including (i) the Directional Sign ($\text{sign}_{(i,j)} \in \{+, -\}$); (ii) the Mediation Effect Strength ($s^{\text{mediation}}_{(i,j)} \in [0, 1]$); and (iii) a mechanism Explanation.
(3) Evidence-Based Causal Verification. To ensure theoretical validity, we introduce an *automated literature verification* mechanism that grounds the causal graph in established economic research. Specifically, for each candidate edge, we deploy an LLM-based agent integrated with the **Google Scholar** (via SerpApi [1]) to perform rigorous verification (Yao et al., 2022; Asai et al., 2024). The verification protocol proceeds as follows: (i) **Query Generation:** The agent formulates a precise academic search query targeting the specific mechanism (*e.g.*, "impact of interest rate hikes on household liquidity"). (ii) **Evidence Retrieval:** The system retrieves abstracts from the top-5 most relevant papers. (iii) **Support Assessment:** The agent analyzes retrieved texts to evaluate whether consensus or substantial evidence supports the directional causality. Edges lacking support from literature are pruned, ensuring that retained mediation pathways are economically plausible.

4.1.2 Moderation Effect Layer Construction

Demographic groups show distinct sensitivities to identical transmission paths. We model this by applying moderation constraints to validated edges.
(1) Moderator Identification. For each edge, the model identifies the dominant profile attributes (see Appendix Tab. 9) governing heterogeneity. To balance complexity and nuance, we restrict selection to the **Top-2 significant moderators**. For instance, regarding *Interest Rate* $\rightarrow$ *Liquidity Constraints*, the model identifies **Income** and **Education** as primary moderators, as they influence sensitivity more profoundly than *gender*.
(2) Heterogeneous Treatment Specification. The model traverses the value space of identified moderators (*e.g.*, Income brackets) to quantify modulation effects. For each category, the system assigns a **Moderation Effect Strength** (scalar $[-1, 1]$), where positive scores (> 0) signify amplification, negative scores (< 0) denote attenuation, and zero indicates neutrality. *Example:* For the *Interest Rate* $\rightarrow$ *Liquidity Constraints*, the model assigns high positive scores to low-income brackets ($< \$10k$), reflecting binding constraints. Conversely, it assigns negative scores to high-income groups ($> \$200k$), reflecting wealth's buffering effect. This yields a profile-based causal graph.

4.2 Causal Transmission-Augm. Reasoning

This module utilizes the causal graph to guide LLM in formulating microeconomic decisions aligned with macroeconomic trends and individual profiles.

[1] https://serpapi.com/

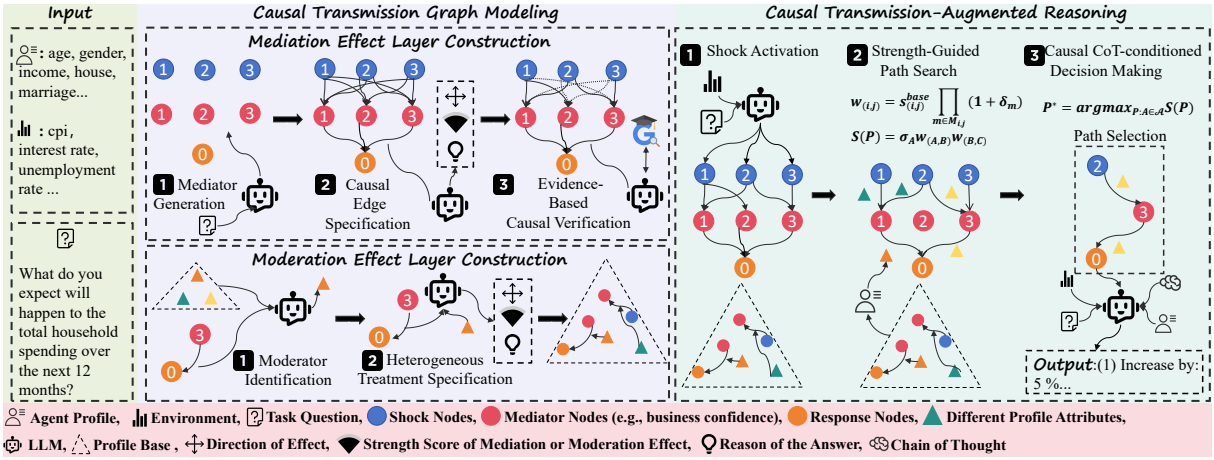


Figure 4: Overview of the HSRCT. The pipeline consists of two phases: (1) **Causal Transmission Graph Modeling**, where the graph is constructed based on mediation and moderation effects; and (2) **Causal Transmission-Augmented Reasoning**, where the dominant transmission path is selected to help the LLM make decisions.

(1) Shock Activation. This module utilizes the causal graph to guide LLM agents in formulating microeconomic decisions aligned with macroeconomic trends and individual profiles. The model generates a directional indicator ($\text{sign}_{shock} \in \{+1, -1\}$) and a normalized intensity score ($\sigma \in [0, 1]$). To mitigate ambiguity from unstructured text, our implementation strictly focuses on three objectives and key indicators: unemployment rate, inflation rate, and interest rate (Taylor, 1993; Bachmann et al., 2015; Armantier et al., 2017).

(2) Strength-Guided Path Search. Given activated shock nodes, we identify the dominant causal trajectory driving decisions via a *profile-conditioned path search*. Specifically, unlike static graphs, causal edge influence is individual-specific. We define the *effective edge weight* $w_{(i,j)}$ as the base causal strength modulated by the respondent's demographic profile:

$$w_{(i,j)} = s^{\text{mediation}}_{(i,j)} \cdot \prod_{m \in \mathcal{M}_{i,j}} (1 + \delta_m), \quad (1)$$

where $s^{\text{mediation}}_{(i,j)} \in [0, 1]$ is the base mediation strength, and $\delta_m \in [-1, 1]$ is the moderation score for the respondent's attribute m. This multiplicative formulation allows demographic traits to amplify (> 1) or attenuate (< 1) the transmission signal. Let $\mathcal{A}$ be the set of activated macroeconomic shocks. A transmission path $\mathcal{P} = A \to B \to C$ flows from shock A via mediator B to decision C. The cumulative strength $\mathcal{S}(\mathcal{P})$ is computed as:

$$\mathcal{S}(\mathcal{P}) = \sigma_A \cdot w_{(A,B)} \cdot w_{(B,C)}, \quad (2)$$

where σ_A denotes initial shock intensity. Beyond magnitude, directional impact is determined by sign propagation. The final direction $\mathcal{D}(\mathcal{P}) \in \{+1, -1\}$ is computed as the product of the initial shock direction and the signs of intermediate edges:

$$\mathcal{D}(\mathcal{P}) = \text{sign}_{shock} \cdot \text{sign}_{(A,B)} \cdot \text{sign}_{(B,C)}, \quad (3)$$

where, $+1$ signifies an *increase* in the decision variable, while -1 signifies a *decrease*.

(3) Causal CoT-conditioned Decision Reasoning. We identify the optimal path $\mathcal{P}^*$ maximizing cumulative strength across all potential shocks to capture dominant economic pressure:

$$\mathcal{P}^* = \arg\max_{\mathcal{P}: A \in \mathcal{A}} \mathcal{S}(\mathcal{P}), \quad (4)$$

This step filters noisy signals to select the macroeconomic narrative most relevant to the individual. Finally, $\mathcal{P}^*$ is translated into a natural language CoT rationale (*e.g.*, *"The increase of interest rates decreases your liquidity..."*), which is combined with the human profile and historical data to form the augmented prompt for LLM decision. Please refer to Figs. 11–17 in the Appendix for the prompts.

5 Intervention Experiments

This section evaluates the proposed framework through four core research questions:

- **RQ2.1 Micro-level Alignment:** Does HSRCT enhance temporal trend alignment with SCE data while restoring behavioral heterogeneity?
- **RQ2.2 Ablation Study:** How do individual intervention modules impact performance?
- **RQ2.3 Interpretability:** Do generated reasoning chains align with economic principles?

Model	Overall	Demographic Attributes			
		Age	Edu.	Home.	Income
GPT-3.5 Turbo	0.36	-0.10	-0.01	0.08	-0.21
GPT-5	0.84	-0.03	0.19	-0.10	-0.23
GPT-3.5 Turbo + **HSRCT**	**0.86***	**0.26***	**0.33***	**0.24***	**0.39***

Table 3: **Spearman correlation (ρ) on Spending Forecast.** The * denotes statistically significant improvement over baseline ($p < 0.05$) under paired t-test for 5 repeated runs.

• **RQ2.4 Macro-level Validity:** Does HSRCT enhance fidelity in replicating real-world macroeconomic trends within system-level simulations?

5.1 Micro-level Alignment

As illustrated in Tab. 3 and Fig. 5, the proposed **HSRCT** framework achieves significant improvements in aligning behavioral trends between LLM agents and human data. Specifically, our method outperforms the backbone baseline by a margin of 0.50 in Spearman correlation. Notably, applying our intervention to the GPT-3.5 Turbo backbone yields performance surpassing the GPT-5 model, highlighting the efficacy of structured causal reasoning over mere model scale.

Beyond aggregate trends, our method successfully recovers population heterogeneity, a dimension often compromised in standard LLM simulations by mode collapse. As shown in Tab. 3, alignment between distinct human demographic groups and their corresponding LLM agents improves significantly. For instance, the sample-size-weighted average correlation across income groups reaches 0.39, whereas GPT-5 attains only -0.23. Furthermore, as depicted in Fig. 6, the post-intervention LLM response distribution closely approximates human patterns. This confirms that our framework effectively translates profile attributes into heterogeneous economic decisions. We observe comparable improvements within the labor domain. Detailed results are provided in Appendix C.2.

5.2 Ablation Study

To validate the necessity of each HSRCT module, we conduct an ablation study with three variants: (1) **w/o Mediation Layer**, removing intermediate mediator nodes; (2) **w/o Moderation Layer**, excluding profile-conditional edge adjustments; and (3) **w/o Strength-Guided Path Search**, replacing the optimal path with a random valid alternative. As summarized in Tab. 4, eliminating any component leads to consistent performance degradation across both overall trend alignment and heterogeneity metrics. Specifically, the absence of Mediation disrupts explicit transmission logic, hindering the processing of environmental shifts. Removing Moderation renders the model agnostic to group-specific sensitivities, failing to distinguish how diverse populations (*e.g.*, wealthy vs. vulnerable) respond to identical shocks. Finally, omitting Strength-Guided Path Search causes reliance on plausible but irrelevant reasoning chains, leading the model to overlook the dominant economic driver for the specific individual.

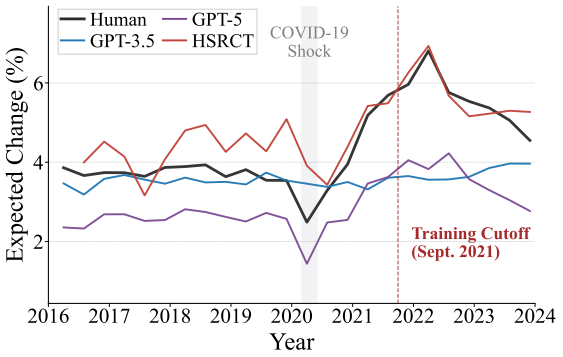
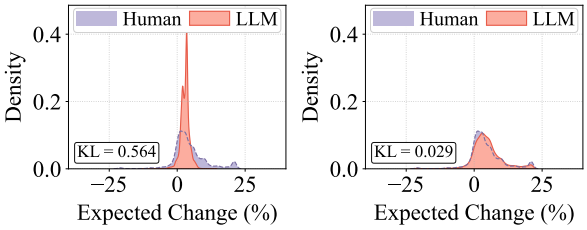


Figure 5: Comparison of time series trends between Human data, Standard LLMs, and the proposed HSRCT.



(a) GPT-5 (b) GPT-3.5 + HSRCT

Figure 6: Comparison of density distributions of responses generated by GPT-5 and the HSRCT.

5.3 Interpretability

To rigorously assess the interpretability and rationality of generated reasoning chains, we design a stratified human evaluation experiment. Instead of random sampling, we define three representative agent profiles: Group A (annual income $<$ \$30k), Group B (\$30k–\$100k), and Group C ($>$ \$100k). For each income group, we randomly select 50 individual cases from the spending and labor domains, respectively. We recruit two Ph.D. candidates in Economics from QS Top-100 universities to evaluate the reasoning chains. Presented with agent profiles and macroeconomic contexts, evaluators perform a pairwise comparison between reasoning

Setting	Overall	Demographic Attributes			
		Age	Edu.	Home.	Income
w/o Media.	0.58	0.15	0.21	0.17	0.28
w/o Moder.	0.55	0.15	0.18	0.15	0.16
w/o Search	0.51	0.16	0.13	0.18	0.25
HSRCT Framework	**0.86***	**0.34***	**0.33***	**0.28***	**0.39***

Table 4: **Ablation study.** Media., Moder., and Search represent Mediation Layer, Moderation Layer, and Strength-Guided Path Search, respectively.

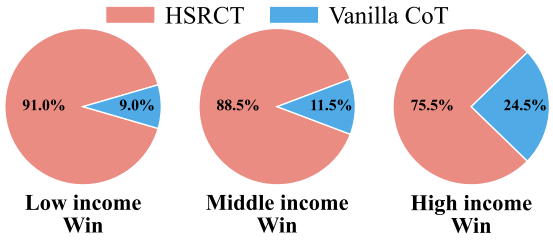


Figure 7: **Human Evaluation Results.** Pairwise comparison of win rates between HSRCT and Standard CoT across three income profiles.

generated by our HSRCT method and a baseline CoT from GPT-5, voting for the more economically rational option. Fig. 7 illustrates the results. The Inter-Annotator Agreement stands at 0.89 (Cohen's Kappa), indicating a high level of consensus among experts. Results indicate that our method outperforms the baseline CoT by a substantial margin. This demonstrates that our intervention generates decision-making processes that are more interpretable and aligned with experts compared to the black-box reasoning of standard LLMs.

5.4 Macro-level Validity

To assess macro-level generalization, we integrate the *HSRCT* module into the state-of-the-art *EconAgent* simulation framework (Li et al., 2024). We adapt the simulation mechanism to conduct a monthly "historical re-enactment" experiment spanning April 2020 to December 2021; this period serves as a rigorous "stress test" by capturing the peak volatility and rapid policy shifts of the COVID-19 economic shock. In contrast to the original setting where agents perceive endogenous indicators from prior steps, we modify the **Perception Module** to incorporate real-world monthly data (Unemployment, Interest, and Inflation Rates) as the environmental context. Conditioned on this realistic context, agents formulate micro-level decisions (labor supply and consumption), which the simulation engine aggregates to derive current macroeconomic indicators. As illustrated in

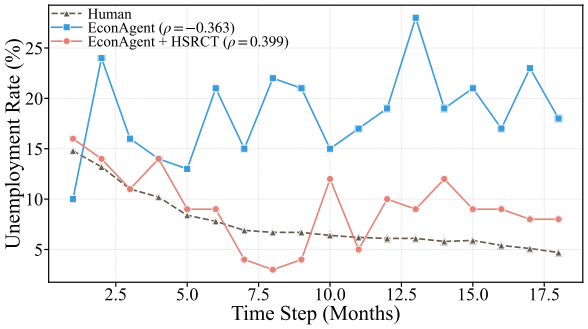


Figure 8: **Macro-level Simulation Validation.** Comparative analysis of simulated unemployment rates against empirical data.

Fig. 8, we compare the trajectories of simulated unemployment rates (Ours and GPT-3.5 Turbo) against the real unemployment rate (Human) in the U.S. Simulations driven by HSRCT achieve Spearman correlations of 0.39, significantly outperforming the original *EconAgent*, which yields -0.36. These results confirm that our method not only enhances micro-level rationality but also translates into more accurate macro-level forecasting capabilities. More experimental results can be found in the Appendix C.3, and C.4. Please refer to Appendix B for implementation details.

6 Conclusion

In this work, we narrow the gap between idealized economic simulations and realistic human behavior by investigating and refining the micro-foundations of LLM-based ABM. Our empirical diagnosis identifies misalignments with real-world data, characterized by weak trend responsiveness, mode collapse, and a reliance on data leakage over generalized reasoning. To mitigate these limitations, we propose the HSRCT Framework. By grounding decision-making in a structured causal graph that models economic transmission mechanisms (mediation) and individual sensitivities (moderation), our approach promotes mechanistic reasoning while reducing reliance on rote memorization. Extensive experiments on the micro-level SCE dataset and the macro-level *EconAgent* system indicate that our modular framework improves behavioral alignment and effectively recovers population heterogeneity. Ultimately, this study advances beyond the replication of stylized facts, offering a reliable, interpretable methodology for constructing credible computational social science simulations that support real-world policy analysis.

7 Limitations

While HSRCT effectively aligns microeconomic behaviors, this study is subject to two limitations that warrant future investigation: 1) **Geographic and Cultural Specificity:** Our micro-empirical validation relies on the SCE, representing U.S. households. Therefore, identified behavioral patterns may not generalize to economies with distinct structural characteristics or cultural norms governing savings and debt. 2) **Simplification of Macro-Dynamics:** To maintain computational tractability, the current framework restricts inputs to three core macroeconomic indicators: unemployment, inflation, and interest rates. Real-world decision-making involves a wider array of variables, including fiscal policies, stock market volatility, and geopolitical events, which are currently omitted from our shock nodes.

Ethical Considerations

Data Privacy and Usage. This study employs the SCE dataset, a publicly available resource standard in economic research. The dataset is strictly anonymized, containing neither Personally Identifiable Information nor offensive content. We comply with all data usage policies and terms of service established by the Federal Reserve Bank of New York. All statistical analyses and descriptive statistics presented herein derive exclusively from this aggregated, anonymized data.

Human Evaluation and Compensation. While primary experiments rely on public datasets, this study incorporates a human evaluation component to assess the economic interpretability of generated reasoning chains (see Section 5.3). We recruited two external evaluators, both Ph.D. candidates in Economics from QS Top-100 universities. These evaluators remain independent of the paper's authorship. To ensure fair labor practices, evaluators received compensation exceeding the local minimum wage (approximately $20/hour). Evaluators provided informed consent prior to participation. As the evaluation focused exclusively on model-generated text quality without collecting personal information or conducting psychological experimentation, formal Institutional Review Board approval was not required.

Use of AI Assistants. We employed AI assistants (*e.g.*, GPT-5) exclusively for linguistic polishing and formatting to enhance manuscript readability. The conceptualization, experimental design, data analysis, and scientific conclusions remain the original work of the authors.

Potential Risks. We do not anticipate immediate ethical risks or negative societal impacts resulting from this work. The proposed framework seeks to enhance the alignment of economic simulations with human behavior, supporting the development of reliable economic modeling tools.

Acknowledgments

This work is supported by the Discipline Breakthrough Pioneer Program of the Ministry of Education of China (Grant No. JYB2025XDXM122).

References

Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International conference on machine learning*, pages 337–371. PMLR.

Jacy Reese Anthis, Ryan Liu, Sean M Richardson, Austin C. Kozlowski, Bernard Koch, Erik Brynjolfsson, James Evans, and Michael S. Bernstein. 2025. Position: LLM social simulations are a promising research method. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 81005–81034.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Anthropic Technical Report*.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Olivier Armantier, Giorgio Topa, Wilbert Van der Klaauw, and Basit Zafar. 2017. An overview of the survey of consumer expectations. *Economic Policy Review*, (23-2):51–72.

Neeraj Arora, Ishita Chakraborty, and Yohei Nishimura. 2025. Ai–human hybrids for marketing research: Leveraging large language models (llms) as collaborators. *Journal of Marketing*, 89(2):43–70.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection.

Rüdiger Bachmann, Tim O Berg, and Eric R Sims. 2015. Inflation expectations and readiness to spend: Cross-sectional evidence. *American Economic Journal: Economic Policy*, 7(1):1–35.

Reuben M Baron and David A Kenny. 1986. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, 51(6):1173.

William A Brock and Cars H Hommes. 1998. Heterogeneous beliefs and routes to chaos in a simple asset pricing model. *Journal of Economic dynamics and Control*, 22(8-9):1235–1274.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, and 1 others. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Caoyun Fan, Jindou Chen, Yaohui Jin, and Hao He. 2024. Can large language models serve as rational players in game theory? a systematic analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17960–17967.

Nicoló Fontana, Francesco Pierri, and Luca Maria Aiello. 2025. Nicer than humans: How do large language models behave in the prisoner's dilemma? In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 522–535.

Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2024a. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications*, 11(1):1–24.

Chen Gao, Fengli Xu, Xu Chen, Xiang Wang, Xiangnan He, and Yong Li. 2024b. Simulating human society with large language model agents: City, social media, and economic system. In *Companion Proceedings of the ACM Web Conference 2024*, pages 1290–1293.

Yuan Gao, Dokyun Lee, Gordon Burtch, and Sina Fazelpour. 2025. Take caution in using llms as human surrogates. *Proceedings of the National Academy of Sciences*, 122(24):e2501660122.

Robert J Gordon. 2010. Okun's law and productivity innovations. *American Economic Review*, 100(2):11–15.

George Gui and Olivier Toubia. 2023. The challenge of using llms to simulate human behavior: A causal inference perspective. *Columbia Business School Research Paper*, (4650172).

Jiaxian Guo, Bo Yang, Paul Yoo, Bill Yuchen Lin, Yusuke Iwasawa, and Yutaka Matsuo. 2023. Suspicion-agent: Playing imperfect information games with theory of mind aware gpt-4. *arXiv preprint arXiv:2309.17277*.

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

Greg Kaplan, Benjamin Moll, and Giovanni L Violante. 2018. Monetary policy according to hank. *American Economic Review*, 108(3):697–743.

Greg Kaplan and Giovanni L Violante. 2014. A model of the consumption response to fiscal stimulus payments. *Econometrica*, 82(4):1199–1239.

Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. 2024. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on machine learning research*.

Nian Li, Chen Gao, Mingyu Li, Yong Li, and Qingmin Liao. 2024. Econagent: large language model-empowered agents for simulating macroeconomic activities. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 15523–15536.

Alejandro Lopez-Lira. 2025. Can large language models trade? testing financial theories with llm agents in market simulations. *arXiv preprint arXiv:2504.10789*.

Nunzio Lorè and Babak Heydari. 2023. Strategic behavior of large language models: Game structure vs. contextual framing. *arXiv preprint arXiv:2309.05898*.

Robert E Lucas and Leonard A Rapping. 1969. Price expectations and the phillips curve. *The American Economic Review*, 59(3):342–350.

Qirui Mi, Qipeng Yang, Zijun Fan, Wentian Fan, Heyang Ma, Chengdong Ma, Siyu Xia, Bo An, Jun Wang, and Haifeng Zhang. 2025. Econgym: A scalable ai testbed with diverse economic tasks. *arXiv preprint arXiv:2506.12110*.

Frederic S Mishkin. 1995. Symposium on the monetary transmission mechanism. *Journal of Economic perspectives*, 9(4):3–10.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Judea Pearl. 2009. *Causality*. Cambridge university press.

Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. 2024. Cooperate or collapse: Emergence of sustainable cooperation in a society of llm agents. *Advances in Neural Information Processing Systems*, 37:111715–111759.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, and 1 others. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

Jiakai Tang, Heyang Gao, Xuchen Pan, Lei Wang, Haoran Tan, Dawei Gao, Yushuo Chen, Xu Chen, Yankai Lin, Yaliang Li, Bolin Ding, Jingren Zhou, Jun Wang, and Jiayao Wen. 2024. Gensim: A general social simulation platform with large language model based agents. *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*.

John B Taylor. 1993. Discretion versus policy rules in practice. In *Carnegie-Rochester conference series on public policy*, volume 39, pages 195–214. Elsevier.

Leigh Tesfatsion and Kenneth L. Judd, editors. 2006. *Handbook of computational economics: agent-based computational economics*, volume 2. Elsevier.

Lei Wang, Zheqing Zhang, and Xu Chen. 2025. Investigating and extending homans' social exchange theory with large language model based agents. In *Proceedings of ACL*, pages 9762–9777.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jeffrey M Wooldridge. 2010. *Econometric analysis of cross section and panel data*. MIT press.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Shiyang Lai, Kai Shu, Jindong Gu, Adel Bibi, Ziniu Hu, David Jurgens, and 1 others. 2024. Can large language model agents simulate human trust behavior? *Advances in neural information processing systems*, 37:15674–15729.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

YANG Yuzhe, Yifei Zhang, Minghao Wu, Kaidi Zhang, Yunmiao Zhang, Honghai Yu, Yan Hu, and Benyou Wang. Twinmarket: A scalable behavioral and social simulation for financial markets. In *ICLR 2025 Workshop on World Models: Understanding, Modelling and Scaling*.

Overall Introduction

This appendix provides supplementary information to support our main findings. Section A describes the *SCE* dataset and its preprocessing procedures. Section B details the implementation of our evaluation and the *HSRCT* intervention. Section C presents additional experimental results.

A SCE Data Configuration

A.1 Data Cleaning and Sample Specifications

Prior to analysis, raw data undergoes rigorous preprocessing to filter outliers and incomplete entries. We apply twofold exclusion criteria, discarding: (1) records lacking essential demographic metadata; and (2) instances with missing responses to any sub-components of multi-part questions. Aligning with the SCE official release schedule and data availability, this study utilizes data from specific collection windows. Sample sizes for the analyzed periods are detailed as follows:

- **Labor Domain:** Analysis covers waves from Mar 2016 through Nov 2023, specifically Mar 2016 ($N = 1,086$), Jul 2016 ($N = 1,134$), Nov 2016 ($N = 1,150$), Mar 2017 ($N = 1,180$), Jul 2017 ($N = 1,173$), Nov 2017 ($N = 1,173$), Mar 2018 ($N = 1,140$), Jul 2018 ($N = 1,169$), Nov 2018 ($N = 1,161$), Mar 2019 ($N = 1,189$), Jul 2019 ($N = 1,166$), Nov 2019 ($N = 1,128$), Mar 2020 ($N = 1,130$), Jul 2020 ($N = 1,041$), Nov 2020 ($N = 1,062$), Mar 2021 ($N = 1,069$), Jul 2021 ($N = 1,116$), Nov 2021 ($N = 1,126$), Mar 2022 ($N = 1,120$), Jul 2022 ($N = 1,171$), Nov 2022 ($N = 1,043$), Mar 2023 ($N = 1,082$), Jul 2023 ($N = 1,009$), and Nov 2023 ($N = 985$).

- **Spending Domain:** This segment includes waves from Apr 2016 through Dec 2023, including Apr 2016 ($N = 1,070$), Aug 2016 ($N = 1,082$), Dec 2016 ($N = 1,184$), Apr 2017 ($N = 1,108$), Aug 2017 ($N = 1,161$), Dec 2017 ($N = 1,110$), Apr 2018 ($N = 1,122$), Aug 2018 ($N = 1,139$), Dec 2018 ($N = 1,094$), Apr 2019 ($N = 1,165$), Aug 2019 ($N = 1,120$), Dec 2019 ($N = 1,103$), Apr 2020 ($N = 1,147$), Aug 2020 ($N = 1,022$), Dec 2020 ($N = 1,181$), Apr 2021 ($N = 1,097$), Aug 2021 ($N = 1,120$), Dec 2021 ($N = 1,132$), Apr 2022 ($N = 1,120$), Aug 2022 ($N = 1,164$), Dec 2022 ($N =$

1,036), Apr 2023 ($N = 1,118$), Aug 2023 ($N = 1,006$), and Dec 2023 ($N = 978$).

The complete microdata underlying this analysis can be accessed via the Federal Reserve Bank of New York's digital repository[2].

A.2 Survey Questions

Detailed survey questions are as follows:

• ***Spending (spending forecast)***: What do you expect will happen to the total household spending over the next 12 months? Please assign a probability to each outcome (summing to 100%). (1) Increase by 12% or more, (2) Increase by 8% to 12%, (3) Increase by 4% to 8%, (4) Increase by 2% to 4%, (5) Increase by 0% to 2%, (6) Decrease by 0% to 2%, (7) Decrease by 2% to 4%, (8) Decrease by 4% to 8%, (9) Decrease by 8% to 12%, (10) Decrease by 12% or more.

• ***Labor (offer acceptance probability)***: Think about the job offers that you may receive within the coming four months. What do you think is the percent chance that you will accept a job if it offers an annual salary for the first year that is less than 0.8 times your estimate of the average annual salary for these offers?

A.3 Estimation of Expected Change

For the *Spending* domain, respondents provide density forecasts rather than point estimates. To parameterize these subjective expectations, we adopt the methodology established in the original SCE study (Armantier et al., 2017). Respondents assign probabilities $p_{i,j}$ to a set of pre-defined intervals with boundaries $E = [-30, -12, -8, -4, -2, 0, 2, 4, 8, 12, 30]$. We first normalize these probabilities to sum to unity. Interval midpoints are then mapped onto a standard $[0, 1]$ support. We estimate the sample mean $\hat{\mu}_i$ and variance $\hat{\sigma}_i^2$ via the method of moments:

$$\hat{\mu}_i = \sum_j p_{i,j} \cdot c_j, \quad \hat{\sigma}_i^2 = \sum_j p_{i,j}(c_j - \hat{\mu}_i)^2 \quad (5)$$

where c_j denotes the normalized midpoint of the j-th interval. The shape parameters α and β of the Beta distribution are computed as:

$$\alpha = \hat{\mu}_i \left(\frac{\hat{\mu}_i(1-\hat{\mu}_i)}{\hat{\sigma}_i^2} - 1 \right) \quad (6)$$

$$\beta = (1-\hat{\mu}_i) \left(\frac{\hat{\mu}_i(1-\hat{\mu}_i)}{\hat{\sigma}_i^2} - 1 \right) \quad (7)$$

[2] https://www.newyorkfed.org/microeconomics/sce

Model	Overall	Demographic Attributes			
		Age	Edu.	Home.	Income
GPT-3.5 Turbo	0.22	-0.05	**0.12**	-0.08	0.04
GPT-5	**0.68**	**0.15**	-0.02	0.00	**0.09**
Claude 3.5 Haiku	0.14	-0.13	0.05	-0.04	**-0.11**
Gemini 2.5 Flash	0.36	0.03	-0.15	0.11	-0.05

Table 5: Spearman correlation coefficients (ρ) of Offer Acceptance task.

The mean of the fitted Beta distribution, $\frac{\alpha}{\alpha+\beta}$, is rescaled to the original domain to derive the final expected spending change for respondent i. This parametric strategy yields a reliable point estimate by incorporating the distributional characteristics of the reported density.

B Implementation Details

B.1 Resource Consumption

The total cost for conducting all experiments, including both analysis and intervention phases, amounted to approximately $500 in API usage fees across various providers and around 11 hours.

B.2 Detailed Settings of LLM Calls

For both evaluation and intervention, we set the temperature to 1.0 and *max_tokens* to 4,000 for all LLMs. When interfacing with *GPT-5*, the *reasoning_effort* parameter is configured to "medium," while all other settings remain at their default values. To optimize inference efficiency, we employ a multi-threaded parallel calling strategy with 18 concurrent workers. To control for generation stochasticity, we conduct five independent runs for all LLM-generated economic decisions across both the diagnosis and intervention phases, reporting the averaged results. For experimental results of the Intervention, markers denoted with * indicate statistical significance ($p < 0.05$) compared to the corresponding baseline, as verified by a paired t-test.

B.3 Prompts and Generated Graph Nodes

We present all the prompts we used in Figs. 11–17. We present all the graph nodes we created for four tasks (SCE-Spending, SCE-Labor, EconAgent-Spending, and EconAgent-Labor) in Figs. 18–21.

C Additional Experiment Results

C.1 Labor Analysis Results

As shown in Tab. 5, consistent with the spending domain, standard LLMs fail to capture temporal

Model	Prompt	Overall	Time Period	
			Pre-Cutoff	Post-Cutoff
GPT-3.5 Turbo	w/ Date	0.53	0.71	0.25
	w/o Date	0.22	0.15	0.29
GPT-5	w/ Date	**0.69**	**0.76**	**0.62**
	w/o Date	0.68	0.75	0.61

Table 6: Analysis of the impact of data leakage on alignment performance.

Model	Overall	Demographic Attributes			
		Age	Edu.	Home.	Income
GPT-3.5 Turbo	0.22	-0.05	0.12	-0.08	0.04
GPT-5	0.68	0.15	-0.02	0.00	0.09
GPT-3.5 Turbo + **HSRCT**	**0.82***	**0.34***	**0.38***	**0.29***	**0.41***

Table 7: Spearman correlation coefficients (ρ) on the Offer Acceptance task.

shifts in labor supply behaviors. While humans dynamically adjust labor supply intentions in response to environmental changes , LLM agents exhibit weakly correlated trends. This reinforces our hypothesis that LLMs lack the intrinsic capability to translate raw macroeconomic signals into domain-specific micro-decisions. Moreover, regarding alignment between distinct demographic groups and their corresponding LLM agents, correlations are negligible. In Tab. 6, we further corroborate the risk of data leakage observed in the spending domain.

C.2 Labor Intervention Results

As illustrated in Tab. 7, the proposed framework significantly enhances the alignment of behavioral trends between LLM agents and human labor data. Notably, applying our intervention to the GPT-3.5 Turbo backbone yields performance surpassing the larger-scale GPT-5 model (0.82 vs. 0.68), demonstrating the efficacy of structured causal reasoning over model scale in complex economic simulations. Beyond aggregate trends, our method effectively recovers population heterogeneity, a dimension where standard LLMs exhibit severe mode collapse. As shown in Tab. 7, while GPT-5 attains moderate overall correlation, its alignment with distinct demographic groups drops to negligible levels (e.g., $\rho \approx 0.0$ for Age and Homeownership). In contrast, our framework restores these nuances, maintaining consistent positive correlations across all attributes. This confirms that our framework effectively translates profile attributes into heterogeneous labor supply decisions, avoiding the generic "average" behavior typical of baseline models.

Setting	Overall	Demographic Attributes			
		Age	Edu.	Home.	Income
w/o Media.	0.55	0.28	0.31	0.22	0.33
w/o Moder.	0.64	0.18	0.22	0.15	0.25
w/o Search	0.52	0.12	0.09	0.08	0.14
HSRCT Framework	**0.82***	**0.34***	**0.38***	**0.29***	**0.41***

Table 8: **Ablation study on the Offer Acceptance task.** The results are reported in Spearman's rank correlation coefficient (ρ).

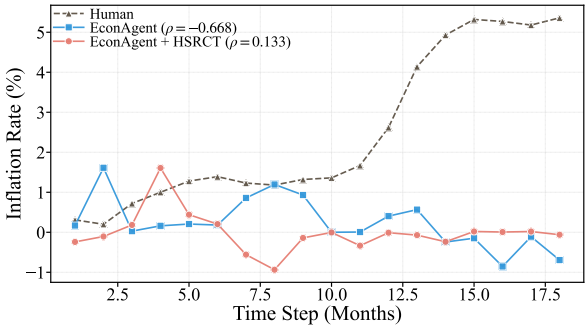


Figure 9: **Macro-level Simulation.** Comparison of simulated inflation rates against empirical data.

To validate the necessity of each module within the labor context, we conduct an ablation study (Tab. 8) with three variants. Results demonstrate that eliminating any component leads to consistent performance degradation.

C.3 Inflation Results of Macro-level Validity

Fig. 9 illustrates the trajectory of the inflation rate within the *EconAgent* system under the HSRCT intervention. In contrast to unemployment rates, which derive directly from LLM responses, inflation is modulated by a combination of LLM outputs and intrinsic system dynamics, including tax regulations and price adjustment mechanisms. Attributable to these structural constraints, the improvement in trend alignment for inflation is comparatively moderate.

C.4 Cumulative Net Causal Strength Analysis

To quantify the aggregate impact of identified causal pathways on economic trends, we introduce *Cumulative Net Causal Strength* (CNCS).

Let $\mathcal{P}_t^+$ and $\mathcal{P}_t^-$ denote sets of selected paths exhibiting positive (spending-increasing) and negative (spending-decreasing) dominant signs at time step t, respectively. Let $w(p)$ represent the strength of a specific path p. We calculate the instantaneous net strength $S_{net}(t)$ as the difference between total

Variable	Encoding
Age	1: 0–10, 2: 11–20, 3: 21–30, 4: 31–40, 5: 41–50, 6: 51–60, 7: 61–70, 8: 71–80, 9: 81–90, 10: 90+
Gender	1: Female, 2: Male
Education	1: <High school, 2: High school, 3: Some college, 4: Associate, 5: Bachelor, 6: Master, 7: Doctor, 8: Professional, 9: Other
Marital status	1: Married/Partnered, 2: Single
Home ownership	1: Own, 2: Rent, 3: Other
Own other home	1: Yes, 2: No
Health	1: Excellent, 2: Very good, 3: Good, 4: Fair, 5: Poor
Income	1: <10k, 2: 10–20k, 3: 20–30k, 4: 30–40k, 5: 40–50k, 6: 50–60k, 7: 60–75k, 8: 75–100k, 9: 100–150k, 10: 150–200k, 11: >200k
Employment status	1: Full-time, 2: Part-time, 3: Unemployed (seeking), 4: Laid off, 5: On leave, 6: Disabled, 7: Retired, 8: Student, 9: Homemaker, 10: Other

Table 9: Encoding schema for profile variables.

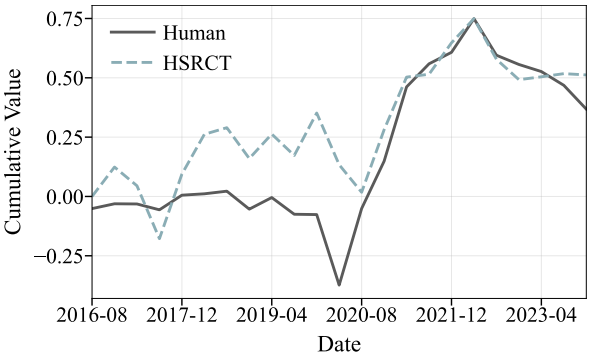


Figure 10: **Temporal Alignment.** Comparison between the Cumulative Net Causal Strength (CNCS) and ground-truth trends in spending forecasts

positive and negative strengths:

$$S_{net}(t) = \sum_{p \in \mathcal{P}_t^+} w(p) - \sum_{p \in \mathcal{P}_t^-} w(p) \tag{8}$$

To capture the temporal accumulation of economic trends, CNCS at time T is defined as the summation of these instantaneous variations:

$$\text{CNCS}_T = \sum_{t=t_0}^{T} S_{net}(t) \tag{9}$$

We visualize the trajectory of the derived CNCS_T against SCE spending data. As presented in Fig. 10, **the CNCS trajectory exhibits strong alignment with ground-truth fluctuations in human consumption decisions.** This strong correlation provides further evidence of HSRCT's effectiveness.

Prompt Template: Mediator Generation

System Instruction:

Role Definition
You are a theoretical economist specializing in **Causal Transmission Dynamics**. Your task is to discover and structure the hidden intermediate mechanisms (Mediators) that explain how macroscopic economic shocks transmit to specific microeconomic household decisions.

Task Description
You are provided with a set of Macroeconomy Environmental Shocks (L_0) and a target Micro-Decision (L_2). You must identify a diverse set of Mediator Nodes (L_1) that bridge L_0 and L_2.

Generation Process
- **Brainstorm Transmission Paths:** Analyze how each shock (Unemployment, Inflation, Interest Rate) physically and psychologically impacts the decision-maker.
- **Identify Latent Variables:** Extract the specific intermediate variables (nouns) acting as bridges in these paths.
- **Cluster & Label:** Group these variables into distinct, categories that represent different dimensions of influence.

Format: Output strictly in JSON. Key names should be the category names you identified.

User Instruction:

Input Data
Shocks (L_0):
- Unemployment Rate (Labor Market Shock)
- Inflation Rate (Price Level Shock)
- Interest Rate (Monetary Policy Shock)

Target Decision (L_2):
- {{TARGET_DECISION}}
(e.g., "Total Household Spending Forecast over the next 12 months")

Execution
Generate the Mediator Nodes (L_1) and organize them into self-identified categories.

Output Schema (JSON)
```
{
  "Category_Name_1": ["Node A", "Node B", ...],
  "Category_Name_2": ["Node C", "Node D", ...],
  ...
}
```

Figure 11: **Prompt template for the Mediator Generation module.** The agent is tasked with identifying intermediate mediators (L_1) that bridge macroeconomic shocks (L_0) and microeconomic decisions (L_2).

Prompt Template: Causal Influence Specification

System Instruction:

Role Definition
You are an expert economic agent specializing in **Causal Reasoning**. Your task is to analyze the direct causal relationship between a specific Source Node (Cause) and a Target Node (Effect).

Task Description
For the given pair of economic variables, you must estimate four key dimensions of their causal link:
- **Directional Sign ($+/-$):** Determine if an increase in the cause leads to an increase ($+$) or decrease ($-$) in the effect.
- **Mediation Effect Strength** $\in [0, 1]$**:** Provide a numerical score representing the magnitude of the impact.
- **Mechanism Explanation:** Provide a concise textual explanation of the economic channel driving this effect.

Format: Return ONLY valid JSON.

User Instruction:

Input Pair
Source (Cause): {{SRC_NODE}} (e.g., "Interest Rate")
Target (Effect): {{DST_NODE}} (e.g., "Borrowing Cost")

Execution
Analyze the causal link and fill the schema.

Output Schema (JSON)
```
{
  "src": "{{SRC_NODE}}",
  "dst": "{{DST_NODE}}",
  "sign": "+/-",
  "strength": 0.85,      // float between 0 and 1
  "mechanism": "Brief explanation of the economic channel..."
}
```

Figure 12: **Prompt template for the Causal Edge Specification module.** The agent evaluates the directional sign, strength, plausibility, and mechanism of potential causal edges between economic variables.

Prompt Template: Heterogeneous Treatment Specification

System Instruction:

Role Definition
You are an expert in household heterogeneity and behavioral economics. Your task is to analyze how demographic and contextual variables moderate a specific causal effect.

Task Description
Given a causal edge (**Cause** $\rightarrow$ **Effect**) and a list of demographic variables (e.g., Age, Income, Education), you must:

- Identify the **Top-2 most influential moderator variables** that govern the heterogeneity of this specific pathway.
- For each identified moderator, provide a **Moderation Score** (scalar $\in [-1.0, 1.0]$) for every specific category (e.g., "Income $<$ \$10k").

Scoring Logic
Positive Score (> 0): Indicates the category amplifies the causal effect magnitude.
Negative Score (< 0): Indicates the category attenuates (weakens) the causal effect.
Near Zero: Indicates a neutral impact.

Format: Return ONLY valid JSON.

User Instruction:

Input Causal Edge
Cause: {{SRC}} $\rightarrow$ **Effect:** {{DST}}
Base Strength: {{STRENGTH}} **Mechanism:** {{MECHANISM}}

Demographic Schema (Candidate Moderators)
(List of variables provided: Age, Gender, Education, Income, Employment Status, etc...)

Execution
Identify the Top-2 moderators and assign per-category moderation scores.

Output Schema (JSON)

```
{
  "src": "{{SRC}}",
  "dst": "{{DST}}",
  "moderators": [
    {
       "variable": "Income", // Example

       "mechanism": "Liquidity constraints differ by income...",
       "scores": {
          "<10k": 0.8,   // Strong amplification
          ">200k": -0.5, // Attenuation
          ...
       }
    },
    { ... } // Second moderator
  ]
}
```

Figure 13: **Prompt template for the Heterogeneous Treatment Specification module.** The agent identifies key demographic moderators and assigns quantitative amplification/attenuation scores to model behavioral diversity.

Prompt Template: Shock Activation and Quantification

System Instruction:

Role Definition
You are an economics expert agent responsible for evaluating macro-environment conditions.

Task Description
Given a natural language description of the macroeconomic environment, evaluate the activation status for three specific nodes: 1. Unemployment rate, 2. Inflation rate, 3. Interest rate.

For each node, you must determine two attributes:
Direction (+/-): "+" indicates the variable is increasing; "-" indicates it is decreasing.
Strength (0.0 to 1.0): A numerical value representing the intensity of the change. 0 implies extremely weak or negligible activation, while 1 implies very strong activation.

Format: Return ONLY valid JSON.

User Instruction:

Input Environment Description
{{ENVIRONMENT_TEXT}}
(e.g., "The inflation rate over the past 4 months is [3.2, 3.4, 3.5, 3.8]. The federal funds rate has increased...")

Execution
Evaluate the Direction and Strength for the three macro nodes based on the description.

Output Schema (JSON)
```
{
  "Unemployment rate": {"direction": "+", "strength": 0.45},
  "Inflation rate":    {"direction": "-", "strength": 0.10},
  "Interest rate":     {"direction": "+", "strength": 0.85}
}
```

Figure 14: **Prompt template for the Shock Activation module.** The agent analyzes natural language macroeconomic descriptions to quantify the direction and strength of specific economic shocks.

Prompt Template: Evidence-Based Causal Verification

System Instruction:

You are an expert Economic Researcher and Causal Analyst Agent. Your goal is to rigorously verify candidate causal edges in an economic causal graph using established academic literature.

You have access to a **Google Scholar Search Tool** (via SerpApi).
You must strictly follow this "Evidence-Based Causal Verification" process:

1. **Query Generation**: Analyze the candidate causal edge (Cause → Effect). Formulate a precise, academic search query targeting the specific economic mechanism. (e.g., "impact of [Cause] on [Effect]", "causal relationship between [Cause] and [Effect]").
2. **Evidence Retrieval**: Use the search tool to find the top-5 most relevant academic papers.
3. **Support Assessment**: Read the abstracts of the retrieved papers. Analyze them to determine if there is a **consensus** or **substantial evidence** supporting the directional causality.
   - Look for keywords like "significantly increases/decreases," "causes," "leads to," "impacts."
   - Be wary of spurious correlations; look for causal language.
4. **Final Decision**:
   - **VERIFIED**: If strong evidence supports the edge.
   - **PRUNED**: If the literature is contradictory, irrelevant, or shows no causal link.

Output Format:
You must conclude with a JSON object:
```
{
  "search_query": "The query you used",
  "key_evidence": "A brief summary of 1-2 key papers...",
  "decision": "VERIFIED" or "PRUNED",
  "confidence_score": 0.0 to 1.0
}
```

User Instruction:

Please verify the following candidate causal edge for validity:
Cause:
Effect:
""link direction""
""link strength""
""link mechanism""
Execute the verification process and provide your final decision.

Figure 15: **Prompt template for the Evidence-Based Causal Verification module.** The agent utilizes external search tools to validate hypothetical causal edges against academic consensus.

Prompt Template: Evaluation of the Spending Forecast Task

System Instruction:

User Profile
{{PROFILE_SPENDING_NODATE}}

Macroeconomic Environment
{{ENVIRONMENT_SPENDING_NODATE}}

Memory Context (Historical Responses)
{{MEMORY_TEXT}}
(Includes up to one prior period's environment and response history: 4 months ago)

User Instruction:

Task Description
You are asked to answer a spending survey. Please answer clearly as instructed.

Total Spending Forecast (QSP7dens)
What do you expect will happen to the total spending of all members of your household (including you) over the next 12 months?
Please provide the percent chance for each option (they should sum to 100%).

* (1) Increase by 12% or more: ___ %
* (2) Increase by 8% to 12%: ___ %
* (3) Increase by 4% to 8%: ___ %
* (4) Increase by 2% to 4%: ___ %
* (5) Increase by 0% to 2%: ___ %
* (6) Decrease by 0% to 2%: ___ %
* (7) Decrease by 2% to 4%: ___ %
* (8) Decrease by 4% to 8%: ___ %
* (9) Decrease by 8% to 12%: ___ %
* (10) Decrease by 12% or more: ___ %

Output Format Requirement
Please output **only** a JSON object in this format, without any commentary or extra text:

```
{
  "QSP7dens": {
    1: x, 2: x, 3: x, 4: x, 5: x, 6: x, 7: x, 8: x, 9: x, 10: x
  }
}
```

Figure 16: **Template of the prompt used for the evaluation Household Spending Forecast Survey.** The prompt integrates user profile, macroeconomic conditions, and historical memory context, requiring the agent to output a structured JSON distribution.

Instance Prompt: Spending Forecast (with the HSRCT method)

System Instruction:

User Profile
Your age is 32. Your gender is Male. Your highest level of education is: Bachelor's degree. Your marital status is: Married or partnered. You live in Nevada. Your housing status is: Own. You don't own another home. Your health status is: Very good. Your current employment status includes: working full-time. Your household's annual income is in the range: $75,000 to $99,999.

Macroeconomic Environment
The inflation rate over the past 4 months is: [0.31, 0.2, 0.72, 1.0]. The federal funds rate over the past 4 months is: [0.05, 0.05, 0.08, 0.09]. The unemployment rate over the past 4 months is: [14.8, 13.2, 11.0, 10.2].

Memory from Previous Survey
Previous environment:
The inflation rate over the past 4 months is: [2.32, 2.6, 2.34, 1.49]. The federal funds rate over the past 4 months is: [1.55, 1.55, 1.58, 0.65]. The unemployment rate over the past 4 months is: [3.6, 3.6, 3.5, 4.4].
Previous response:
{'QSP7dens': {1: 0.0, 2: 0.0, 3: 80.0, 4: 20.0, 5: 0.0, 6: 0.0, 7: 0.0, 8: 0.0, 9: 0.0, 10: 0.0}}

Causal Transmission Insight
The increase in Inflation rate can lead to an increase in Total Household Spending Forecast over the next 12 months through the channel "Wage growth expectation". The effect strength is expressed in a normalized scale from 0 to 1 (higher means a stronger effect): 0.8220.

Edge Directions
- Inflation rate → Wage growth expectation: positive (+)
- Wage growth expectation → Total Household Spending Forecast over the next 12 months: positive (+)

Moderation Effects (User-specific)
Moderation on edge Inflation rate → Wage growth expectation:
- Your category **Employment status = Full-time** strengthens (enhances) this edge
- Your category **Income = 75-100k** strengthens (enhances) this edge

Moderation on edge Wage growth expectation → Total Household Spending Forecast over the next 12 months:
- Your category **Employment status = Full-time** strengthens (enhances) this edge
- Your category **Income = 75-100k** strengthens (enhances) this edge

User Instruction:

You are asked to answer a spending survey. Please answer clearly as instructed.
–
Total Spending Forecast (QSP7dens)
What do you expect will happen to the total spending of all members of your household (including you) over the next 12 months?
Please provide the percent chance for each option (they should sum to 100%).

* (1) Increase by 12% or more: ___ %
* (2) Increase by 8% to 12%: ___ %
* (3) Increase by 4% to 8%: ___ %
* (4) Increase by 2% to 4%: ___ %
* (5) Increase by 0% to 2%: ___ %
* (6) Decrease by 0% to 2%: ___ %
* (7) Decrease by 2% to 4%: ___ %
* (8) Decrease by 4% to 8%: ___ %
* (9) Decrease by 8% to 12%: ___ %
* (10) Decrease by 12% or more: ___ %

IMPORTANT:
You must output ONLY the FINAL updated distribution.
Your output must be a JSON with the following structure:

```
{
  "QSP7dens": {"1": y1, "2": y2, "3": y3, "4": y4, "5": y5, "6": y6, "7": y7, "8": y8,"9": y9,"10": y10
   }
}

Values must sum to 100.
Do NOT output text outside the JSON.
Do NOT use "'json blocks.
```

Figure 17: **Instance prompt structure for the decision-making with the HSRCT intervention.** The prompt incorporates historical memory, causal transmission insights, and user-specific moderation effects to enhance the agent's reasoning process.

Graph Nodes (SCE-spending)

System Definition: Valid Node Space

```
"A" (Macro Shocks): {
   "Interest rate", "Inflation rate", "Unemployment rate"
}
"B" (Latent Mediators): {
  "B1_CreditFinancial": { "Credit availability", "Borrowing cost", "Lending standards", "Household debt burden", "Mortgage availability", "Financial liquidity", "Market liquidity", "Household leverage", "Asset valuation", "Wealth effect" },
  "B2_IncomeExpectations": { "Disposable income expectation", "Future labor income expectation", "Job stability perception", "Working hours expectation", "Wage growth expectation", "Employer hiring expectation", "Perceived unemployment risk", "Government transfer expectation" },
  "B3_ExpectationsConfidence": { "Consumer confidence", "Household sentiment", "Economic outlook expectation", "Inflation expectation", "Real interest rate expectation", "Cost-of-living expectation", "Recession expectation", "Uncertainty perception", "Risk appetite" },
  "B4_RealEconomyOutlook": { "Labor demand expectation", "Business investment expectation", "Hiring intention perception", "Productivity expectation", "Economic activity expectation", "Capacity utilization expectation" },
  "B5_BehavioralConstraints": { "Liquidity constraints", "Precautionary saving motive", "Consumption smoothing behavior", "Intertemporal substitution", "Credit constraints", "Borrowing constraints", "Risk aversion", "Loss aversion", "Durable-nondurable substitution" }
}
"C" (Target Decision): {
   "Total Household Spending Forecast over the next 12 months"
}
```

Figure 18: **Defined node space for the Household Spending (SCE) causal graph.** The taxonomy is compressed into a dictionary format containing Macro Shocks (A), five categories of Latent Mediators (B1-B5), and the Target Decision (C).

Graph Nodes (SCE-Labor)

System Definition: Valid Node Space

```
"A" (Macro Shocks): {
   "Interest rate", "Inflation rate", "Unemployment rate"
}
"B" (Latent Mediators): {
  "B1_FinancialBuffer_Liquidity": { "Savings buffer", "Household debt burden", "Liquidity constraints", "Unemployment benefits valuation", "Non-labor income expectation", "Financial distress", "Asset valuation" },
  "B2_LaborMarket_Leverage": { "Job offer arrival rate expectation", "Perceived labor market tightness", "Bargaining power", "Skill transferability", "Alternative employment options", "Long-term unemployment stigma", "Employer hiring intention perception" },
  "B3_RealWage_CostOfLiving": { "Inflation expectation", "Real wage expectation", "Cost-of-living expectation", "Purchasing power perception", "Wage growth expectation" },
  "B4_SearchBehavior_Psychology": { "Value of leisure", "Risk aversion", "Economic outlook expectation", "Loss aversion", "Search friction perception", "Career path expectation" }
}
"C" (Target Decision): {
   "Probability of accepting an offer below expected salary"
}
```

Figure 19: **Defined node space for the Labor Market (SCE) causal graph.** The taxonomy includes Macro Shocks (A), Latent Mediators (B) categorized by financial buffer, market leverage, real wage, and search psychology, and the Target Decision (C).

Graph Nodes (EconAgent-Spending)

System Definition: Valid Node Space

```
"A" (Macro Shocks): {
   "Interest rate", "Inflation rate", "Unemployment rate"
}
"B" (Latent Mediators): {
  "B1_IncomeSecurity": { "Household disposable income expectation", "Income stability perception",
"Unemployment risk expectation", "Government transfer income expectation" },
  "B2_EssentialPriceExpectations": { "Food price expectation next month", "Energy price expectation
next month", "Housing cost expectation next month", "Overall essential goods inflation expectation"
},
  "B3_SavingsConstraints": { "Precautionary saving motive strength", "Household savings buffer
level", "Liquidity constraint perception", "Debt repayment burden" },
  "B4_ConsumptionConfidence": { "Consumer confidence in short-term economy", "Cost-of-living
pressure perception", "Risk aversion to consumption cutbacks", "Future financial security perception"
}
}
"C" (Target Decision): {
   "the proportion of all your savings and income you intend to spend on essential goods"
}
```

Figure 20: **Defined node space for the EconAgent Spending causal graph.** The taxonomy encompasses Macro Shocks (A), Latent Mediators (B) covering income security, price expectations, savings constraints, and confidence, and the Target Decision (C).

Graph Nodes (EconAgent-Labor)

System Definition: Valid Node Space

```
"A" (Macro Shocks): {
   "Interest rate", "Inflation rate", "Unemployment rate"
}
"B" (Latent Mediators): {
  "B1_JobMarketSignals": { "Employer hiring intention next month", "Short-term labor demand
expectation", "Perceived job scarcity next month", "Temporary layoff risk perception", "Seasonal
labor demand fluctuation" },
  "B2_ShortTermIncomeExpectations": { "Expected monthly wage for next job", "Income stability of
current job next month", "Overtime pay expectation next month", "Minimum wage adjustment expectation"
},
  "B3_WorkDecisionConstraints": { "Cost of job search next month", "Skill mismatch constraint
short-term", "Financial buffer for job switching", "Time constraint for job search next month" },
  "B4_SubjectiveConfidence": { "Confidence in job market next month", "Inflation expectation impact
on real wage", "Risk appetite for job change", "Short-term economic uncertainty perception" }
}
"C" (Target Decision): {
   "Willingness to work next month"
}
```

Figure 21: **Defined node space for the EconAgent Labor Supply causal graph.** The taxonomy includes Macro Shocks (A), Mediators (B) focusing on job signals, income expectations, constraints, and confidence, and the Target Decision (C).