SCE-BENCH: MULTI-SCALE BEHAVIORAL ALIGNMENT OF LLM AGENTS IN REAL-WORLD MICROECONOMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

The adoption of Large Language Models within Agent-Based Modeling enables the simulation of economic behaviors, yet the empirical validity of these LLM-driven agents remains largely unverified. Current benchmarks often emphasize normative optimization or game-theoretic tasks, omitting the temporal consistency and individual heterogeneity necessary for realistic economic modeling. In response, we introduce **SCE-Bench**, derived from the Survey of Consumer Expectations using over 70,000 nationally representative observations from 2016 to 2023. We establish a hierarchical evaluation protocol that measures alignment at both the aggregate level, through temporal correlation and distribution matching, and the individual level, using a Monte Carlo-based approach to assess individual differences and statistical indistinguishability. Our empirical analysis covers comparisons across model characteristics, prompt variations, and temporal contexts. The results indicate that LLMs often exhibit an "Average Person" bias and lack precise temporal tracking. Meanwhile, cost-effective open-weight models emerge as viable alternatives for large-scale simulations. These findings, along with insights into feature selection, verbalization, role-playing perspective, and temporal contexts, establish practical guidelines for configuring LLM economic agents in realistic simulations. Finally, we further validate the generalization of our framework and findings on cross-domain tasks and downstream simulation systems. Dataset, code, and experimental results are available at `https://anonymous.4open.science/r/SCE-Bench-BBA7`.

1 INTRODUCTION

The integration of Large Language Models (LLMs) into Agent-Based Modeling (ABM) has brought new potential to economic simulations (Gürcan, 2024; Li et al., 2024). By utilizing their aptitude for processing unstructured data and reasoning, LLM-based agents are increasingly applied to model household behaviors. In these simulations, agents observe dynamic economic variables (*e.g.*, inflation, unemployment rate) and internal states (*e.g.*, savings, income) to determine microeconomic actions, such as spending and labor supply (Li et al., 2024; Mi et al., 2025). These decisions shape the macroeconomic environment, creating a closed-loop feedback system that generates both micro-behavioral trajectories and emergent macro-indicators. However, a significant methodological discrepancy exists between traditional ABM and LLM-based approaches. Traditional macro-ABMs (Fagiolo et al., 2019; Axtell & Farmer, 2025) rely on explicit, rule-based micro-foundations, where agent behaviors are governed by defined functions $f_m(x) = y$ with parameters calibrated through historical data or empirical evidence. In contrast, replacing rule-based agents with LLMs results in an opaque micro-foundation $f_{LLM}(\cdot)$. Current research (Li et al., 2024) often relies on the unverified assumption that LLM behavioral patterns approximate human behavior, *i.e.*, $f_{LLM} \approx f_{human}$.

As illustrated in Tab. 1, existing works evaluate LLM behavior through two distinct methodological tracks defined by their specific objectives. The first category focuses on **normative optimization**, assessing whether models can meet theoretical benchmarks within specific economic domains (Mi et al., 2025; Raman et al., 2025; 2024). STEER-ME (Raman et al., 2025), for instance, tests microeconomic reasoning via questions that require an optimal solution derived from economic equations. In contrast, a second stream of research emphasizes **behavioral alignment** to evaluate

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Comparison of economics-oriented benchmarks for evaluating LLM behavior. Temp. /Distr. /Indiv. denote temporal, distributional, and individual alignment, respectively.

Benchmark	Goal	Task	Model Property	Temp.	Distr.	Indiv.	Validation Points
Steer (Raman et al., 2024)	Normative Optimization	MCQA	microeconomic reasoning	/	/	/	/
Steer-me (Raman et al., 2025)	Normative Optimization	Free-text QA and MCQA	microeconomic reasoning	/	/	/	/
Econgym (Mi et al., 2025)	Normative Optimization	Multi-agent Games	Policy optimization	/	/	/	/
Agent-trust (Xie et al., 2024)	Behavioral Alignment	Trust Games	Trust	×	×	×	Human Avg. Only
Rational Players (Fan et al., 2024)	Behavioral Alignment	Game-theoretic Games	Rationality	×	×	×	Human Avg. Only
Turing Experiment (Aher et al., 2023)	Behavioral Alignment	Ultimatum Game	General representativeness	×	×	×	Human Avg. Only
Nicer than Humans (Fontana et al., 2025)	Behavioral Alignment	Prisoner's Dilemma	Cooperative behavior	×	×	×	Human Avg. Only
Take Caution (Gao et al., 2025)	Behavioral Alignment	Money Request Game	General representativeness	×	✓	×	108 Human Samples
Homo Silicus (Horton, 2023)	Behavioral Alignment	Behavioral Econ. games	Social pref., fairness perception	×	×	×	Human Avg. Only
Decision-Making (Jia et al., 2024)	Behavioral Alignment	Lottery Games	Risk aversion, Prob. weight	×	×	×	Human Avg. Only
SCE-Bench (Ours)	**Behavioral Alignment**	**Real-life SCE survey**	**Credit, spending, labor**	✓	✓	✓	**70,094 Human Samples**

how closely model outputs mirror human behavioral patterns (Jia et al., 2024; Horton, 2023; Gao et al., 2025; Fontana et al., 2025; Aher et al., 2023; Fan et al., 2024; Xie et al., 2024). Specifically, Horton (2023) utilizes game-theoretic experiments to measure the congruence between LLM and human trust dynamics by evaluating decisions on monetary allocations such as \$6:\$4 or \$8:\$2. Please refer to Appendix B for additional related work.

While these simplified protocols provide initial insights, they often lack the granularity of real-world microeconomic decisions. Even among studies evaluating behavioral alignment, two methodological facets remain overlooked: (1) **Temporal Dynamics**: Reliable policy prediction requires agents to exhibit temporal consistency and adapt to evolving economic conditions, as opposed to maintaining static preferences (*e.g.*, fixed risk aversion); and (2) **Individual Heterogeneity**: Aggregate metrics often obscure distributional shifts, whereas precise modeling must represent how different individuals (*e.g.*, billionaires *vs.* white-collar workers) respond distinctly to shocks.

To address these limitations, we introduce **SCE-Bench** for evaluating LLM behavioral alignment in real-world microeconomic contexts. It directly isolates and evaluates the LLM behavioral decision kernel before it is coupled with higher-level simulator dynamics. We utilize the Survey of Consumer Expectations (Armantier et al., 2016) (SCE), a nationally representative dataset conducted by the Federal Reserve Bank of New York covering Jan. 2016 to Dec. 2023 with 70,094 observations, which is integrated with relevant macroeconomic indicators. Our evaluation structure consists of aggregate-level and individual-level assessments. At the aggregate level, we measure the alignment of temporal trends (*Spearman*), deviations (*RMSE*), and response distributions (*JS divergence*) between LLM agents and human cohorts across various demographic segments. At the individual level, because a single survey observation represents only one historical realization of human behavior, direct point-to-point comparisons with LLM outputs are naturally imprecise. Therefore, we develop a Monte Carlo approach to evaluate individual alignment. For each profile, we obtain N model-elicited responses generated in one call and use bootstrap resampling to construct a statistically estimated predictive interval from these responses. We then assess whether the ground-truth human response falls within this interval, thereby operationalizing the statistical indistinguishability between simulated and real agents. The proportion of human samples covered by these bootstrap-based intervals is defined as the *Coverage Rate (CR)*. Moreover, we compute the average of the N responses to serve as a representative summary point. Finally, we calculate the *Mean Absolute Error* between this average value and the actual human response.

Our empirical analysis evaluates model architectures based on accessibility, reasoning type, and release date; prompt variations including input features, feature verbalization formats, and role-playing; and temporal contexts across varying historical memories. We also conduct generalization studies. Experimental results provide the following observations: (1) **Behavioral Alignment Biases**: LLMs reveal an "Average Person" bias and temporal tracking failures caused by distributional averaging, which requires calibration for long-tail heterogeneity. (2) **Model Selection Trade-offs**: Newer proprietary models are not always superior to smaller or older models; consequently, more affordable models can serve as cost-effective options for large-scale simulations. (3) **Feature Representation Constraints**: Curating core features (such as housing status) pertinent to specific economic decisions leads to better alignment predictions, as excessive variables often introduce significant bias and noise to LLMs. (4) **Structural Prompt Optimization**: Third-person perspectives surpass standard second-person perspectives, as models inherently perform inference from a third-party standpoint rather than genuinely embodying a first-person persona. (5) **Temporal Consistency**: Incorporating historical response context is critical, as it prevents the model from treating them as

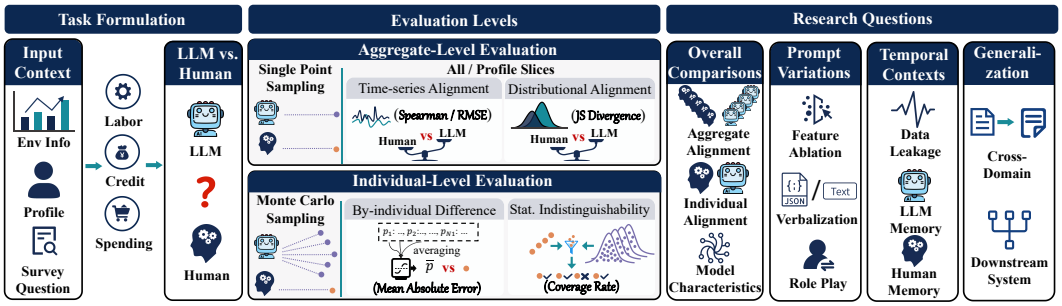


Figure 1: Illustration of the multi-level evaluation framework. Left: Task formulation. Mid: Hierarchical evaluation levels. Right: Four systematic research questions.

isolated events. (6) **Cross-domain generalization**: Evaluations in the political domain reveal similar difficulties in temporal tracking and evidence of pattern collapse. (7) **Downstream simulation system transferability**: Performance on SCE-Bench offers preliminary insight into the aggregated macroeconomic indicators produced by downstream LLM-driven simulation systems. Overall, these findings establish methodological guidelines for improving the realism and cost-efficiency of LLM-based economic agents. Our contributions can be summarized as follows:

- We propose SCE-Bench, which benchmarks LLM agents against empirical microeconomic behaviors rather than stylized laboratory games, with over 70k human observations from the SCE.
- We formulate a multi-scale validation protocol that assesses alignment through aggregate distributions and the statistical indistinguishability of individual decisions.
- We characterize methodological trade-offs among model characteristics, prompt structure, and behavioral realism, establishing evidence-based guidelines for aligning microeconomic behavior. We further demonstrate that our framework and findings generalize to cross-domain tasks and downstream economic simulation systems.

2 Evaluation Framework

The overall evaluation framework of SCE-Bench is illustrated in Fig. 1. This section outlines the task formulation, the evaluation settings, and the research questions.

2.1 Task Formulation

Every four months, the SCE administers questionnaires regarding credit, labor, and spending, providing participants with a $20 incentive. These are distributed to approximately 1,300 U.S. household heads, sampled to reflect national demographic distributions. We use publicly available SCE data covering 24 waves from Jan. 2016 to Dec. 2023, choosing one question from each domain that is primarily applicable to economic simulations. Each question q comprises n samples, with $n = 26,416$ for spending, $16,568$ for labor, and $27,110$ for credit. The selected survey questions are as follows:

Spending (spending forecast): Participants report their expected total spending changes over the coming year. ***Credit (mortgage application probability)***: Participants provide the likelihood of applying for a mortgage within the next year. ***Labor (offer acceptance probability)***: Participants provide the probability of accepting an offer at 0.9-1.0 times their expected wage.

Each sample includes the response from the human participant, and the response generated by model m, where m belongs to: {GPT-3.5-turbo, GPT-4o-mini, GPT-5-mini-nothinking (reasoning_effort = minimal), GPT-5-mini-thinking (reasoning_effort = medium), Gemini-3-Flash-nothinking (think_level = minimal), Gemini-3-Flash-thinking (think_level = medium), Qwen3-30B-A3B (enable_thinking=False), Qwen3-30B-A3B (enable_thinking=True)}. Hereafter, these models are denoted as GPT3.5, GPT4om, GPT5m-nt, GPT5m-t, GEM3f-nt, GEM3f-t, QWE3-nt, and QWE3-t. Each query for the LLM incorporates three components: (1) respondent demographics x; (2) historical macroeconomic variables for the given period (unemployment rate, inflation rate, and federal funds rate from the previous four months) e; and (3) the question q. We formalize **behavioral**

alignment as the similarity between the human conditional decision distribution and the LLM-agent conditional decisions under a similar decision context:

$$P_{\text{human}}(y \mid x, e, q) \quad \text{vs.} \quad P_{\text{LLM}}(y \mid x, e, q),$$

The formulation itself is general and can be applied to other domains with different contexts. Detailed data descriptions and preprocessing details are provided in Appendices D.1, D.2, and D.3.

2.2 Evaluation Levels

Our evaluation comprises two hierarchical levels:

Aggregate-level: Alignment between LLM and human populations is assessed by generating a single prediction per demographic profile, computing mean trajectories, and quantifying turning-point trend and distance alignment using the **Spearman correlation** and **RMSE**. Distributional similarity is measured with **JS divergence**, applied either to the full population or demographic subgroups.

Individual-level: For each profile, $N = 50$ Monte Carlo realizations are generated to estimate the model's predictive distribution. The aggregated mean prediction is compared to the human response using *MAE*. To evaluate statistical consistency, we further construct a bootstrap-based 90% predictive interval from the generated realizations, and define the *CR* as the proportion of human samples falling within this statistically estimated interval. Details of the metrics are provided in Appendix D.4.

2.3 Research Questions

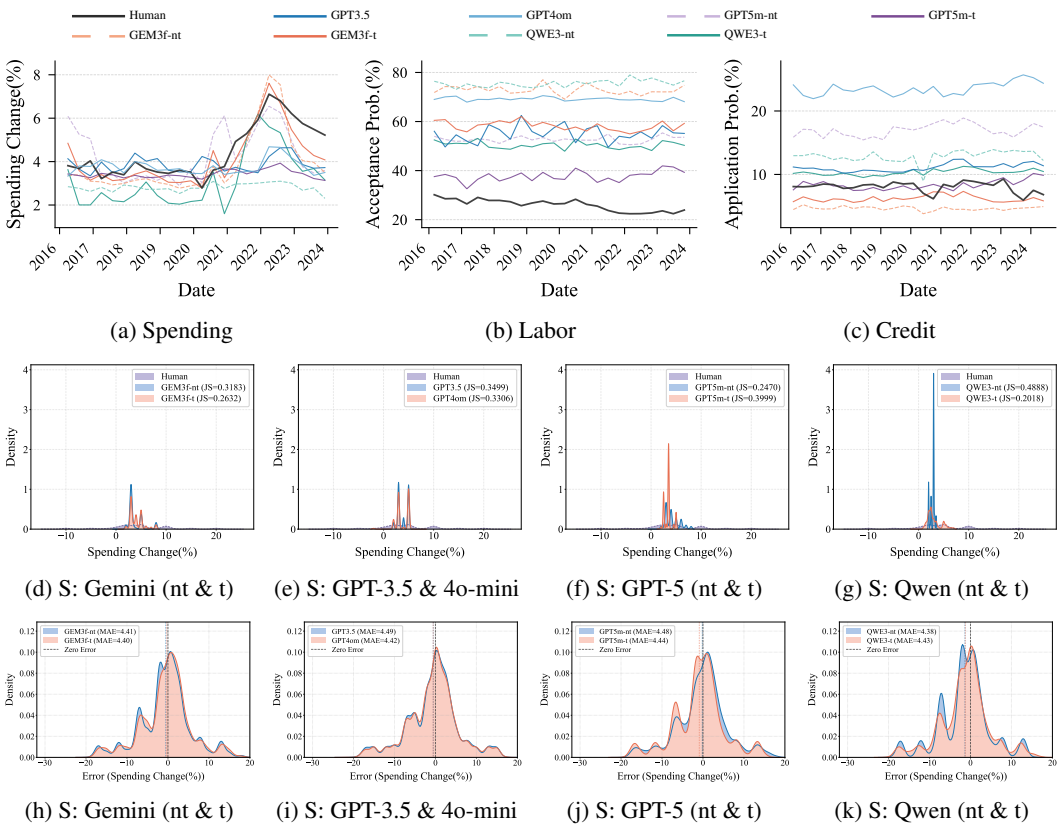
- **RQ 1: Overall comparisons of model alignment.** To what extent do models reproduce human behavior across credit, labor, and spending? We examine shared limitations in aggregate-level (RQ 1.1) and individual-level (RQ 1.2) alignment, and how reasoning capability and architecture generation influence behavioral realism (RQ 1.3).
- **RQ 2: Influence of prompt variations.** We examine how input configuration affects behavioral realism by ablating specific demographic and macroeconomic features (RQ 2.1), comparing structured JSON with natural-language feature descriptions (RQ 2.2), and contrasting second-person prompts with third-person role-playing (RQ 2.3).
- **RQ 3: Effect of temporal contexts.** We assess how environmental states and prior behavioral choices influence alignment. Random perturbations of macroeconomic indicators probe sensitivity to exact contextual values (RQ 3.1). We separately test the effects of providing model-generated (RQ 3.2) and ground-truth human (RQ 3.3) responses from the previous time step.
- **RQ 4: Generalization of the framework and findings.** We examine whether the findings observed on SCE extend to repeated, environment-conditioned political-survey tasks using ANES (American National Election Studies, 2026) (RQ 4.1), and what insights the benchmark provides for downstream, micro-behavior-driven macroeconomic simulations beyond the original survey domains (RQ 4.2).

3 Experimental Results

In our evaluation, the default setup uses the features listed in Appendix Tab. 5, following a second-person role-play format and natural language feature verbalization. For RQ 2 and RQ 3, we adopt Qwen3-30B-A3B (enable_thinking=False), chosen as a strong open-weight model with broad community adoption. Temperature is set to standard 1.0 to encourage diversity, following related works (Guo et al., 2023; Lorè & Heydari, 2023). Specific prompts are provided in Appendix F.

3.1 RQ 1: Overall Comparisons

Fig. 2 presents the time series, response distributions, and individual-level difference distributions. Fig. 3 reports point estimates; 95% paired-bootstrap confidence intervals are provided in Appendix E.4. The main benchmark uses three core ABM decisions, while supplementary experiments on durable purchases, saving adjustment, and labor-force participation reproduce the main alignment patterns (Appendix E.1).



(a) Spending (b) Labor (c) Credit

(d) S: Gemini (nt & t) (e) S: GPT-3.5 & 4o-mini (f) S: GPT-5 (nt & t) (g) S: Qwen (nt & t)

(h) S: Gemini (nt & t) (i) S: GPT-3.5 & 4o-mini (j) S: GPT-5 (nt & t) (k) S: Qwen (nt & t)

Figure 2: The top row shows time series comparisons across three tasks (RQ 1.1), the middle row shows distribution differences for the spending task (RQ 1.1), and the bottom row shows individual-level differences for the spending task (RQ 1.2). For full visualizations, see Fig. 8 and Fig. 9 in the Appendix.

3.1.1 RQ 1.1: Aggregate-level Alignment

As shown in the first row of Fig. 2 and the first column of Fig. 3, LLMs exhibit **difficulties in capturing temporal trends**. For instance, few models accurately tracked the fluctuations resulting from the pandemic shock from 2020 to 2022. Concerning the deviation between model and human trajectories, we find a **systematic overestimation** of offer acceptance probabilities, which may distort downstream predictions regarding employment dynamics. In terms of **distributional alignment**, as illustrated in the middle row of Fig. 2, a marked divergence exists between human and model behaviors. While human responses show broad distributions with long-tailed characteristics, LLMs tend to suffer from **mode collapse**, concentrated on a limited set of values. This failure to represent low-frequency patterns prevents the essential modeling of population heterogeneity (Kirman, 2006; Michelacci & Paciello, 2024).

3.1.2 RQ 1.2: Individual-level Analysis

As shown in the fifth column of Fig. 3, the *Coverage Rate* for the majority of models remains low. For instance, in the *Labor* domain, no model exceeds a *Coverage Rate* of 0.30. Sensitivity analyses over $N \in \{10, 25, 50, 100\}$ and temperature $T \in \{0.5, 1.0, 1.3, 1.6\}$ show that the results stabilize by $N = 50$ and that temperature variation does not explain the main alignment gap (Appendix E.3). Combined with the distributional analysis, these findings suggest that models essentially function as an **"Average Person,"** failing to reflect the heterogeneity and long-tail distribution of real human behaviors. Moreover, as shown in the bottom row of Fig. 2, the bias detected at the aggregate level is not driven by outliers but represents a systematic deviation across most individuals. Besides,

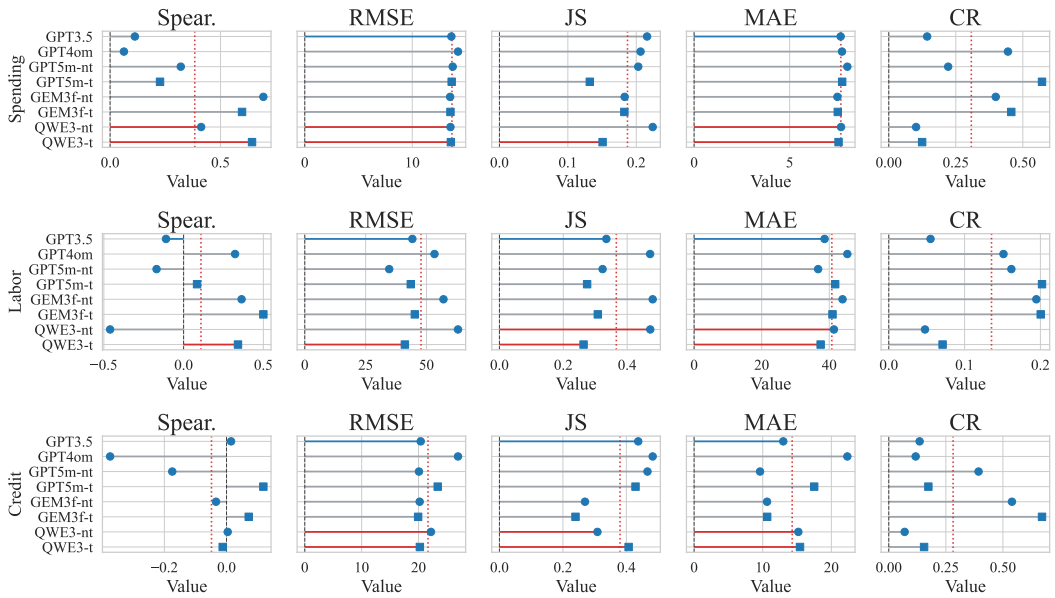


Figure 3: Comparative alignment across LLMs (RQ 1.3). Each marker denotes one model's performance under the domain–metric pair (higher is better for *Spearman* ρ and *Coverage Rate*; lower is better for *RMSE*, *JS divergence*, and *MAE*). Red lines highlight cases where QWE3 (nt/t) outperforms at least one model among GPT5m-nt, GPT5m-t, GEM3f-nt and GEM3f-t, while blue lines highlight analogous wins for GPT3.5. • denotes no-thinking (nt), and ■ denotes thinking (t).

Table 2: Feature removal improvement ratio across tasks and evaluation metrics. Values indicate the percentage of ablations whose removal improves performance relative to the full-feature baseline; higher values indicate that a feature is more often harmful under the corresponding metric.

	Spending					**Labor**					**Credit**				
	Spear.	RMSE	JS	MAE	CR	Spear.	RMSE	JS	MAE	CR	Spear.	RMSE	JS	MAE	CR
Improve ratio (%)	**0**	62	**31**	77	69	69	77	**46**	**0**	**23**	**31**	**23**	69	**0**	**23**

alignment at the aggregate level often results from systematic individual deviations canceling each other out (*e.g.*, positive and negative errors offsetting). Thus, while aggregate means may align, individual fidelity remains low.

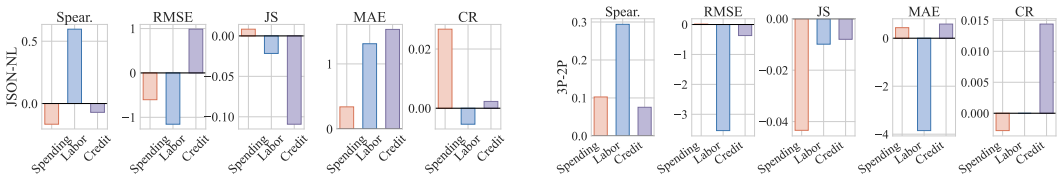
3.1.3 RQ 1.3: Impact of Model Characteristics

As shown in Fig. 3, we evaluate model performance across three domains by comparing the same models with and without *Thinking* mode. **Enabling this mode generally improves alignment in both temporal trends and distributions**, helping models better capture economic fluctuations and encouraging broader exploration, which lowers *JS* divergence and increases *CR*.

We further examine model evolution by comparing `GPT-3.5-Turbo`, `GPT-4o-mini`, and `GPT-5-mini`. The results show a non-linear progression in performance: **newer is not always "better."** As illustrated by the blue lines in Fig. 3, the oldest model in our experiment, `GPT-3.5-Turbo`, occasionally achieves superior alignment in certain metrics compared to newer LLMs, such as `GPT-5-mini` and the `Gemini-3-FLASH` series.

Moreover, the `Qwen` series demonstrates that smaller open-weight models do not inherently lag behind strong proprietary baselines. As illustrated by the red lines in Fig. 3, `Qwen` frequently outperforms proprietary counterparts across multiple scenarios. For instance, in the Labor domain, `Qwen3` achieves the closest distributional alignment and reports the lowest individual differences.

Additional frontier-scale experiments with Claude Opus 4.7 and GPT-5.6 Sol show that reasoning improves alignment within each model family, but frontier capability does not consistently surpass the previous metric-wise best results (Appendix E.2).



(a) Verbalization Formats (JSON − NL) (b) Perspective (third-person − second-person)

Figure 4: **(a) Impact of verbalization formats (RQ 2.2).** The y-axis reports the performance difference between JSON and natural-language prompting (JSON − NL). For *Spearman* and *Coverage Rate*, larger positive values indicate better performance under JSON prompting. For *RMSE*, *JS divergence* and *MAE*, negative values indicate better performance under JSON prompting. **(b) Perspective analysis (RQ 2.3).** The y-axis reports the performance difference between third-person and second-person perspective (third-person − second-person), following the interpretation as (a).

Table 3: Effect of random perturbations (RQ 3.1). We compare performance using original versus randomly perturbed environmental numerical values. Values report the change of random perturbations compared with original environmental values, computed as $Random - Original$.

Spending					**Labor**					**Credit**				
Spear.	RMSE	JS	MAE	CR	Spear.	RMSE	JS	MAE	CR	Spear.	RMSE	JS	MAE	CR
-0.01	+0.07	0.00	+0.03	-0.02	+0.09	-0.65	0.00	-0.64	+0.01	-0.03	+0.12	+0.01	+0.16	+0.01

Overall, increasing the inference-time *Thinking* setting improves alignment compared with the same model configured with minimal or disabled thinking. Given that large-scale simulations require extensive inference, strong open-weight models offer cost-effective alternatives. Results broken down by income segments are presented in Appendix E.6, and cost analysis is provided in Appendix E.7.

3.2 RQ 2: Influence of Prompt Variations

3.2.1 RQ 2.1: Impact of specific profile and environment features

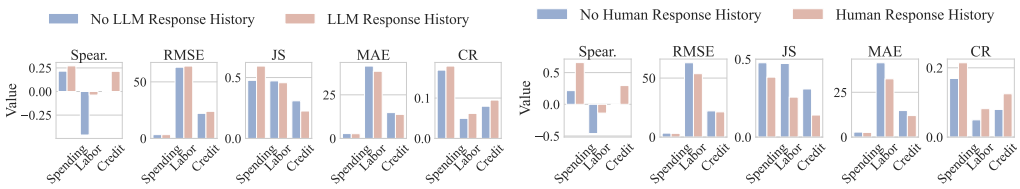
As shown in Tab. 2, our ablation studies show that models have limited capacity to process demographic and environmental features; adding information about the profile and environment does not ensure better alignment. **In many instances, removing a considerable amount of features leads to improved results**. For instance, for the *JS divergence* of the credit question, 69% of features act as noise. Three features are critical: **Housing Status, Age, and Employment Status.** For detailed feature ablation results and economic analysis, please refer to the Appendix E.8.

3.2.2 RQ 2.2: Impact of Different Feature Verbalization Formats

As illustrated in Fig. 4a, distinct patterns emerge across domains. Overall, **neither the JSON nor the natural language format demonstrates consistent superiority.** Specifically, JSON minimizes semantic ambiguity by explicitly separating keys from values, whereas natural language descriptions tend to introduce syntactic complexity and greater variability.

3.2.3 RQ 2.3: Impact of Role-playing Perspectives

As shown in Fig. 4b, **third-person framing consistently results in closer trend alignment and distribution alignment**. This advantage likely arises from the reduction of "self-bias." Second-person prompts ("You are") often activate safety constraints or an idealized assistant persona. Conversely, the third-person approach presents the task as an objective simulation, leading to the generation of more realistic and varied human behaviors conditioned on the environment changes.



(a) LLM response history. (b) Human response history.

Figure 5: **(a) Effect of LLM response history (RQ 3.2).** We compare performance with and without appending the model's past answers. **(b) Effect of ground-truth human response history (RQ 3.3).** We compare model performance with and without access to real historical behavior data.

3.3 RQ 3: Effect of Temporal Contexts

3.3.1 RQ 3.1: Data Leakage

As shown in Tab. 3, perturbing numerical values (with 1% random noise) within the environmental context resulted in minimal changes to model performance. Notably, this holds true for Qwen3-30B-A3B (enable_thinking=False), even though the model was released 1.5 years after our survey cutoff. This suggests that the model **does not rely on memorizing the exact environment indicators in the training data, but instead conducts inference using the provided logic and context**. Moreover, many of the findings are negative, such as mode collapse. Therefore, any potential data leakage is unlikely to substantially affect these conclusions, since these negative outcomes arise even when the model may have encountered the relevant data during training. Post-cutoff Spending results are similarly stable (Appendix E.5).

3.3.2 RQ 3.2: Effect of Incorporating Model-generated History

As illustrated in Fig. 5a, including the model's prior responses (Self-Memory) increases the *Spearman* coefficient, decreases the *MAE*, and improves the *Coverage Rate*. This improvement demonstrates that Self-Memory reinforces path dependency and consistency. Economic behaviors are rarely independent events; they display smoothing and habit formation. **Even with hallucinated initial history, maintaining consistency enables the model to simulate a coherent economic narrative**, which reduces random fluctuations and mirrors the inertia of human financial decisions.

3.3.3 RQ 3.3: Effect of Incorporating Ground-truth Human History

As shown in Fig. 5b, incorporating ground-truth human history leads to large improvements across all metrics. This result confirms the high autocorrelation within microeconomic data, where past financial behavior serves as a strong predictor of future actions. By anchoring predictions to actual historical data, we markedly reduce the model's search space, thereby improving its alignment with human behaviors. For real-world simulation systems, **the timely integration of historical agent data can significantly benefit realism and accuracy** in LLM-based simulations.

3.4 RQ 4: Generalization of the Framework and Findings

3.4.1 RQ 4.1: Cross-domain Generalization

We evaluate cross-domain generalization using the American National Election Studies (ANES) dataset, focusing on the political party "Feeling Thermometer" task. Each LLM is provided with a profile including age, gender, race/ethnicity, education, and family income. Environmental information includes election type, incumbent president's party, party control of Congress, unemployment rate, inflation rate, a one-sentence summary of Democratic Party positions, and a one-sentence summary of Republican Party positions. The task asks, for example, "How would you rate the Democratic Party on a 0–100 feeling thermometer?" For the 18 waves between 1978 and 2024 in the original dataset, we randomly sample 100 respondents per wave. The prompt is presented in Fig. 29.

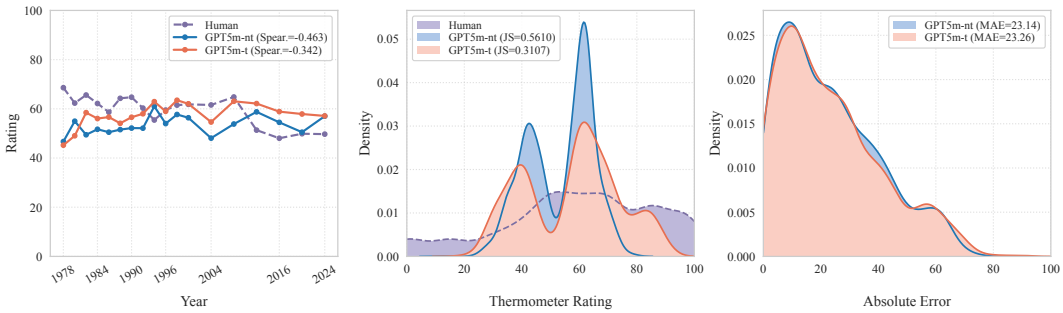


Figure 6: Political party thermometer rating task.

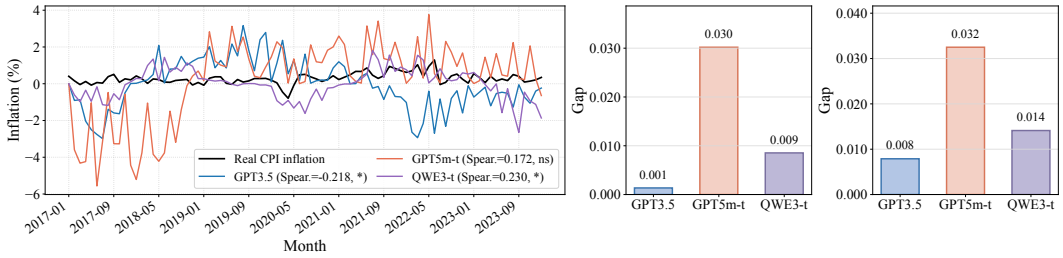


Figure 7: EconAgent simulation results. Left: real and simulated monthly inflation. Middle and right: consumption gaps across income groups and employment gaps across age groups.

Unlike prior static ANES evaluations (Argyle et al., 2023), we formulate the task as a repeated, environment-conditioned evaluation. Figure 6 shows that even strong LLMs exhibit poor temporal tracking and mode collapse, although stronger Thinking partially mitigates the latter.

3.4.2 RQ 4.2: Generalization in Downstream Macroeconomic Simulation Systems

We conduct a monthly historical re-enactment in *EconAgent* (Li et al., 2024) by replacing its endogenous environmental inputs with observed unemployment, interest-rate, and inflation data. Agents generate work and consumption propensities, from which the simulator derives aggregate inflation and subgroup heterogeneity; we retain the original prompts and replace only the environmental data. We assess temporal transfer by comparing simulated aggregate inflation with historical CPI. Because EconAgent does not provide respondent-level human outcomes, subgroup validation focuses on consumption gaps across income groups and employment gaps across age cohorts.

Figure 7 shows that all models deviate substantially from historical inflation. Nevertheless, the temporal ranking, Qwen3-30B-A3B $>$ GPT-5-mini $>$ GPT-3.5-turbo, is consistent with their aggregate trend performance on SCE-Bench. Lower-income groups are expected to exhibit higher consumption propensities, while younger cohorts exhibit higher employment propensities. GPT-5-mini most closely reproduces these patterns, followed by Qwen3-30B-A3B and GPT-3.5-turbo, consistent with their relative individual-level performance on SCE-Bench. This agreement suggests that benchmark performance transfers partially to downstream economic simulation.

4 Conclusion

We introduce SCE-Bench to bridge the gap between LLM capabilities and the empirical requirements of economic Agent-Based Modeling. Across 70,000+ observations, LLMs exhibit "Average Person" bias and limited temporal precision, while greater general capability does not guarantee simulation fidelity. Our analyses provide practical guidelines for configuring LLM agents and demonstrate partial transfer to cross-domain tasks and downstream simulation systems. We hope SCE-Bench supports more accurate and distinctively human-like economic simulations.

AI USE STATEMENT

We use generative language models as experimental agents to generate simulated economic and political-survey responses and, in designated conditions, model-generated response histories. These outputs are evaluated against human survey observations rather than reported as human data. The evaluated models and inference settings are described in Section 3 and Appendix E.7, and prompt examples are provided in Appendix F. We additionally used AI tools to check for grammatical errors and refine the wording of the manuscript. The authors remain responsible for the final text, methods, results, and AI-assisted artifacts.

ETHICS STATEMENT

This study analyzes existing, publicly available, de-identified SCE and ANES survey data; it does not recruit participants or collect new responses through interaction with human subjects. Data-use and institutional-review considerations are discussed in Appendix C. Economic, demographic, and political attributes warrant care even in de-identified data: combinations of attributes may create re-identification risks, and model outputs may reproduce stereotypes or underrepresent particular groups. Reuse and redistribution of survey data should respect the source providers' terms and privacy protections.

REPRODUCIBILITY STATEMENT

The evaluation protocol is described in Section 2, with data sources and preprocessing in Appendix D and metric definitions in Appendix D.4. Section 3 reports the default experimental settings; Appendix F provides prompt examples, and Appendix E.7 reports inference backends, hardware, output-token limits, and computational costs. The anonymous repository for the accompanying data, code, and experimental results is linked in the abstract. These materials document the evaluation procedure, rather than guaranteeing identical stochastic model outputs across reruns. The absence of repeated full-benchmark runs is noted in Appendix E.7.

REFERENCES

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

American National Election Studies. ANES Time Series Cumulative Data File, 2026. URL `https://www.electionstudies.org`. Dataset and documentation, February 5, 2026 version.

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

Olivier Armantier, Giorgio Topa, Wilbert Van der Klaauw, and Basit Zafar. An overview of the survey of consumer expectations. 2016.

Robert L Axtell and J Doyne Farmer. Agent-based modeling in economics and finance: Past, present, and future. *Journal of Economic Literature*, 63(1):197–287, 2025.

James Bisbee, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 32 (4):401–416, 2024. doi: 10.1017/pan.2024.5.

Yong Cao, Haijiang Liu, Arnav Arora, Isabelle Augenstein, Paul Röttger, and Daniel Hershcovich. Specializing large language models to simulate survey response distributions for global populations. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3141–3154. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.162.

Christopher D Carroll. Buffer-stock saving and the life cycle/permanent income hypothesis. *The Quarterly journal of economics*, 112(1):1–55, 1997.

Ami Dalloul and Moritz Pfeifer. Can llms mimic household surveys?: From representative agents to population distributions, 2026. SSRN working paper.

Angus Deaton. Saving and liquidity constraints, 1989.

Giorgio Fagiolo, Mattia Guerini, Francesco Lamperti, Alessio Moneta, and Andrea Roventini. Validation of agent-based models in economics and finance. In *Computer simulation validation: fundamental concepts, methodological frameworks, and philosophical perspectives*, pp. 763–787. Springer, 2019.

Caoyun Fan, Jindou Chen, Yaohui Jin, and Hao He. 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, pp. 17960–17967, 2024.

Nicoló Fontana, Francesco Pierri, and Luca Maria Aiello. 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, pp. 522–535, 2025.

Milton Friedman. Theory of the consumption function. 2018.

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

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

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

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

Önder Gürcan. Llm-augmented agent-based modelling for social simulations: Challenges and opportunities. *HHAI 2024: Hybrid Human AI Systems for the Social Good*, pp. 134–144, 2024.

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

Matteo Iacoviello. House prices, borrowing constraints, and monetary policy in the business cycle. *American economic review*, 95(3):739–764, 2005.

Jingru Jessica Jia, Zehua Yuan, Junhao Pan, Paul McNamara, and Deming Chen. Decision-making behavior evaluation framework for llms under uncertain context. *Advances in Neural Information Processing Systems*, 37:113360–113382, 2024.

Alan Kirman. Heterogeneity in economics. *Journal of Economic Interaction and Coordination*, 1(1): 89–117, 2006.

Akaash Kolluri, Shengguang Wu, Joon Sung Park, and Michael S. Bernstein. Finetuning llms for human behavior prediction in social science experiments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 30096–30111. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1530.

Nian Li, Chen Gao, Mingyu Li, Yong Li, and Qingmin Liao. 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)*, pp. 15523–15536, 2024.

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

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

John Joseph McCall. Economics of information and job search. *The Quarterly Journal of Economics*, 84(1):113–126, 1970.

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

Claudio Michelacci and Luigi Paciello. Ambiguity aversion and heterogeneity in households' beliefs. *American Economic Journal: Macroeconomics*, 16(2):95–126, 2024.

Franco Modigliani and Richard Brumberg. Utility analysis and the consumption function: an interpretation of cross-section data 1. In *Post-keynesian economics*, pp. 388–436. Routledge, 2013.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Narun Raman, Taylor Lundy, Thiago Amin, Jesse Perla, and Kevin Leyton-Brown. Steer-me: Assessing the microeconomic reasoning of large language models. *arXiv preprint arXiv:2502.13119*, 2025.

Narun Krishnamurthi Raman, Taylor Lundy, Samuel Joseph Amouyal, Yoav Levine, Kevin Leyton-Brown, and Moshe Tennenholtz. Steer: Assessing the economic rationality of large language models. In *International Conference on Machine Learning*, pp. 42026–42047. PMLR, 2024.

Joseph Suh, Erfan Jahanparast, Suhong Moon, Minwoo Kang, and Serina Chang. Language model fine-tuning on scaled survey data for predicting distributions of public opinions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21147–21170. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1028.

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

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*.

Ali Zarifhonarvar. Generating inflation expectations with large language models. *Journal of Monetary Economics*, 157:103859, 2026. doi: 10.1016/j.jmoneco.2025.103859.

A LIMITATIONS

First, our study relies on the 2016–2023 Survey of Consumer Expectations (SCE) dataset, which is among the most comprehensive and representative public microeconomic panels available for our evaluation. We acknowledge that earlier historical periods and more recent time steps are not included in our analysis. In particular, the public release of SCE microdata typically involves a substantial time lag, and extending the dataset beyond the current window was either not publicly available or not fully compatible with our benchmark design at the time of the study. Within the official release scope, we make full use of the high-quality SCE data to ensure rigor and representativeness.

Second, we focus on three representative microeconomic behaviors (spending forecast, mortgage application, labor offer acceptance) that are widely adopted in macroeconomic simulation systems. As demonstrated in prior work such as EconAgent, only a small set of core behaviors (consumption propensity, labor supply propensity) is sufficient to drive a complete economic simulation. We do not exhaustively cover all possible household decisions, as this would introduce unnecessary complexity; moreover, these three representative behaviors are adequate to support the core contributions of this paper.

B Related Work

B.1 LLM-Based Economic Simulation

The emergence of LLMs has encouraged the adoption of these models as economic agents (Gao et al., 2024a) to model complex aggregate behaviors with increased accuracy. Initial studies, such as Li et al. (2024), simulate individual labor supply and spending choices to calculate macroeconomic indicators like inflation and unemployment rates. To enable broader applications, Mi et al. (2025) introduce a dedicated platform for diverse simulation tasks, while research (Lopez-Lira, 2025; Yuzhe et al.) further extend the scope of LLM-driven market environments. However, the reliability of this approach rests on a critical, yet often implicit, premise: that LLM agents accurately mirror human economic behaviors. To test this under-examined assumption, our work empirically validates LLM decision-making against real-world survey data rather than relying on stylized facts.

B.2 General LLM Agents

On the one hand, SCE-Bench is not a benchmark for sandbox-style generative agents in the sense of fully open-ended Generative Agents (Park et al., 2023), where the emphasis is on autonomous social interaction, memory, and open-ended planning within an interactive world. On the other hand, it directly evaluates the behavioral decision kernel that is embedded inside simulation systems such as EconAgent (Li et al., 2024), where the LLM is placed inside a richer macroeconomic simulation system but still ultimately needs to map profile and environmental context into labor and consumption decisions. In this sense, our benchmark isolates and evaluates this core behavioral component before it is coupled with higher-level simulator dynamics.

B.3 Behavioral Alignment Evaluation for LLMs

In contrast to benchmarks assessing normative optimization (Raman et al., 2025; 2024), behavioral alignment focuses on how accurately models simulate human actions (Gao et al., 2024b; Aher et al., 2023). Xie et al. (2024) characterize this alignment by measuring the resemblance between LLM and human decisions in comparable contexts (Argyle et al., 2023), building upon the Turing Experiment proposed by Aher et al. (2023). Existing evaluations rely primarily on controlled, game-theoretic settings analyzed at the aggregate level. For example, studies have employed the Iterated Prisoner's Dilemma to evaluate cooperative strategies (Fontana et al., 2025) or repeated Rock-Paper-Scissors to test rationality (Fan et al., 2024). These studies commonly derive an average response through repeated LLM sampling to compare against human means under identical conditions—such as contrasting the mean loss aversion of LLMs with human subjects (Jia et al., 2024).

Survey-based studies provide closer comparisons but address different settings. Bisbee et al. (2024) evaluate static political-opinion responses using ANES feeling thermometers and show that synthetic responses can approximate human means while exhibiting substantially lower variance. In economics, Zarifhonarvar (2026) studies LLM-generated inflation expectations and monetary-policy information treatments, comparing their aggregate patterns with the survey responses, whereas Dalloul & Pfeifer (2026) focus on household inflation expectations and whether LLM populations recover the disagreement and dispersion observed in survey data. These studies do not jointly evaluate respondent-matched decisions across multiple microeconomic action domains under changing macroeconomic environments.

Complementary work improves human simulation through model specialization. Suh et al. (2025) fine-tune models to predict public-opinion distributions for demographic subpopulations, while Cao et al. (2025) specialize models for country-level survey-response distributions. Kolluri et al. (2025) instead fine-tune on individual-level responses from a broad collection of social-science experiments. These approaches study group- or individual-level simulation and demonstrate the value of fine-tuning, whereas SCE-Bench evaluates the off-the-shelf behavioral decision kernels commonly placed inside LLM-based economic simulators.

The contribution of SCE-Bench is the combination of four validation dimensions: conditioning on changing historical economic environments, comparison with matched respondent-level observations, coverage of multiple microeconomic behaviors (spending, labor, and credit), and downstream transfer to a closed-loop ABM. This combination tests whether behavioral discrepancies persist beyond

static survey reproduction and whether benchmark performance predicts the quality of downstream economic simulation.

C Ethical Use of Data and Informed Consent

This study uses the SCE dataset and the ANES dataset, which are publicly available and de-identified resources. In accordance with institutional regulations, formal approval from an Institutional Review Board (IRB) was not required for this work. This determination is based on the fact that the research relies solely on a pre-existing, anonymized dataset that is accessible to the public, meaning no new data was collected through direct interaction with human subjects.

D Data Source

D.1 Representativeness and Fielding Protocol

Post-stratification weights based on American Community Survey (ACS) benchmarks (income, education, region, and age) are applied to ensure national representativeness of the SCE. The demographic profile of the SCE closely mirrors that of the U.S. household head population. The SCE adopts a rotating panel design: each respondent remains in the panel for up to 12 months, with approximately equal numbers of new and departing participants each month. This ensures the sample remains demographically representative over time (Armantier et al., 2016). Each monthly survey wave is distributed in three batches, invited on the 2nd, 11th, and 20th of the month. Panelists are dropped if they miss three consecutive surveys. Approximately 90% of participants in a given month return the following month, ensuring sample stability over time (Armantier et al., 2016). The Survey of Consumer Expectations (SCE) data are provided under the "License for Survey of Consumer Expectations Data and Questions" by the Federal Reserve Bank of New York, granting a worldwide, royalty-free, non-exclusive right to use and distribute the data for research purposes. All microdata are available on the SCE official website.[1]

D.2 Data Preprocessing Details

The **profile data**, **question**, and **ground-truth human decisions** all come from the SCE dataset. For the environmental variables—unemployment rate, inflation rate, and interest rate—we download monthly data from the website of the **Federal Reserve Bank of St. Louis.**[2] . Following standard economic practice, the inflation rate is computed as the year-over-year change in CPI. For each human data point, we include the unemployment rate, inflation rate, and interest rate from the **four months preceding the survey date**.

The preprocessing and validation pipeline follows several steps:

- **Step 1: Original survey size** – All raw survey observations before any preprocessing. This forms the starting benchmark size for each module.
- **Step 2: Human response null** – Rows where the human response is null are dropped, because no valid comparison with the LLM is possible.
- **Step 3: LLM response non-numeric/NaN** – Rows where the LLM output cannot be parsed as a numeric value are removed, since the model failed to produce a valid numeric answer.
- **Step 4: Out-of-range filter [0,100]** – For **Labor** and **Credit**, responses outside [0,100] are dropped, because these questions are explicitly bounded probabilities; out-of-range values indicate invalid responses.
- **Step 5: IQR outlier filter ($\times 15$)** – For **Spending**, rows where the human or LLM value exceeds Q3 + 15 $\times$ IQR are removed. Spending is unbounded, so this conservative filter removes only extreme outliers while preserving the heavy-tailed distribution.
- **Step 6: Final valid observations** – Remaining rows after all preprocessing filters constitute the final benchmark size used for analysis.

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

[2] https://www.stlouisfed.org/

Tab. 4 summarizes the numeric impact of each preprocessing step for **Spending**, **Labor**, and **Credit**. Only the numeric changes and final counts are reported. The first column shows the corresponding step number for easy reference.

Table 4: Numeric impact of preprocessing steps.

Step	Spending	Labor	Credit	Total
1	26,579	26,803	27,251	80,633
2	-40	-10,227	-126	-10,393
3	-101	-8	-15	-124
4	0	0	0	0
5	-22	0	0	-22
6	26,416	16,568	27,110	70,094

D.3 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?
• ***Credit (mortgage application probability)***: Over the next 12 months, what do you think is the percent chance that you will apply for a mortgage or home-based loan?
• ***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 equals 0.9 - 1.0 times your estimate of the average annual salary for these offers?

D.4 Details of Metrics

D.4.1 Aggregate-level

Let T be the total number of time steps. We denote the mean values of the human and model populations at time step t as $\bar{r}_{t,\text{human}}$ and $\bar{r}_{t,m}$, respectively.

The ***Spearman correlation coefficient*** (ρ) is calculated based on the rank order of these mean trajectories:

$$\text{Spearman} = 1 - \frac{6\sum_{t=1}^{T} d_t^2}{T(T^2-1)} \tag{1}$$

where $d_t = \text{rank}(\bar{r}_{t,\text{human}}) - \text{rank}(\bar{r}_{t,m})$ represents the difference between the ranks of the human and model means at time step t.

To justify our choice of *Spearman correlation* (ρ) instead of Pearson correlation as the primary trend-oriented metric, we provide a simple example illustrating temporal ordering and turning-point alignment.

Suppose we observe the aggregate borrowing amount over four seasons. The human trajectory is [100, 120, 110, 130], which rises from Season 1 to 2, falls in Season 3 after a tightening policy shock, and rises again in Season 4. Consider two model predictions: Model A predicts [95, 125, 123, 126], capturing the correct turning-point pattern (up $\rightarrow$ down $\rightarrow$ up), while Model B predicts [100, 118, 121, 130], which is numerically closer overall but misses the key turning point (up $\rightarrow$ up $\rightarrow$ up). The resulting correlations are: for Model A, Spearman = 1.00 and Pearson = 0.82; for Model B, Spearman = 0.80 and Pearson = 0.89. This example shows that even when Pearson correlation is higher, Spearman better captures the temporal behavioral structure and the policy-induced turning points.

We emphasize that Spearman alone is not sufficient, which is why we also report RMSE. The ***Root Mean Squared Error (RMSE)*** quantifies the average deviation between the model output and the

Table 5: Profile variables and macroeconomic environment indicators.

Variable	Encoding
Age	continuous variable
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
State	U.S. state
Inflation rate	continuous variable
Unemployment rate	continuous variable
Federal funds rate	continuous variable

human ground truth:

$$\text{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\bar{r}_{t,\text{human}} - \bar{r}_{t,m})^2} \tag{2}$$

Let P and Q_m denote the empirical response distributions of the aggregated human responses and model m responses, respectively. In our implementation, these distributions are estimated by binning the cleaned response values into 50 histogram bins and normalizing the resulting counts to sum to one. We then define the mixture distribution $M = \frac{1}{2}(P + Q_m)$. The Jensen-Shannon divergence is calculated as:

$$\text{JS}(P \parallel Q_m) = \frac{1}{2}D_{\text{KL}}(P \parallel M) + \frac{1}{2}D_{\text{KL}}(Q_m \parallel M) \tag{3}$$

where D_{KL} denotes the Kullback-Leibler divergence,

$$D_{\text{KL}}(P \parallel M) = \sum_x P(x) \log\left(\frac{P(x)}{M(x)}\right), \tag{4}$$

and x indexes the histogram bins of the response space.

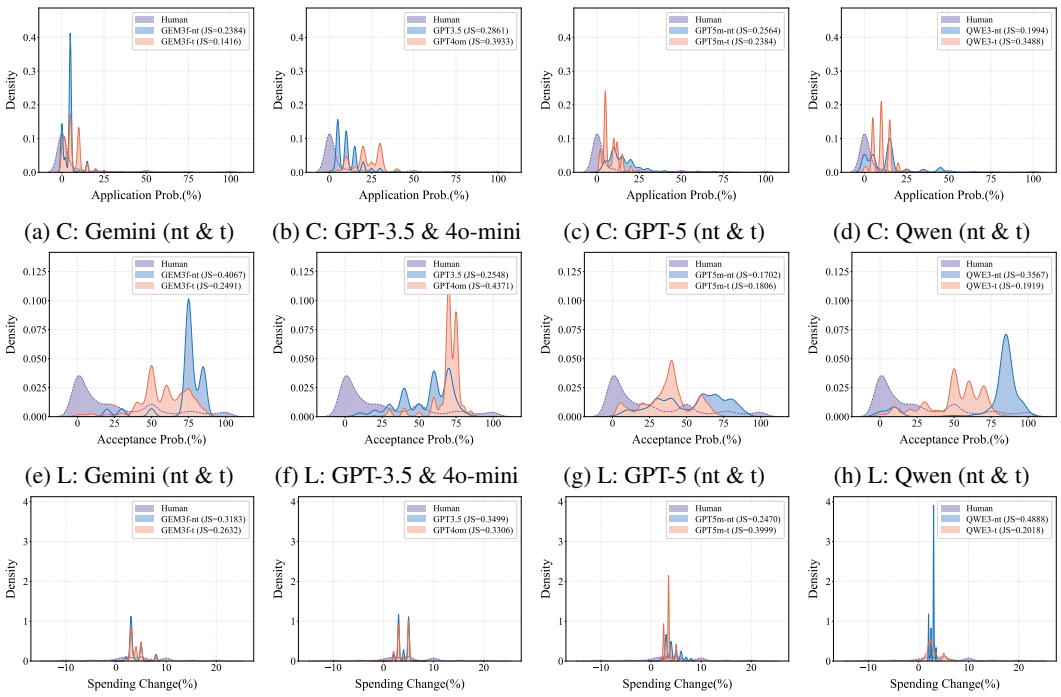
D.4.2 Individual-level

For a specific profile, we use N model-elicited responses generated in one call, denoted as $\mathcal{P}_i = \{p_i^{(1)}, p_i^{(2)}, \ldots, p_i^{(N)}\}$, to characterize the model's elicited response distribution. First, we compute the average of the N responses as the representative model prediction, denoted as $\bar{p}_i = \frac{1}{N}\sum_{k=1}^{N} p_i^{(k)}$. We use the ***Mean Absolute Error (MAE)*** to measure the **individual-level difference** between this aggregated prediction and the human ground truth:

$$\text{MAE} = \frac{1}{|\mathcal{D}|}\sum_{i\in\mathcal{D}} |\bar{p}_i - p_{i,\text{human}}| \tag{5}$$

The ***Coverage Rate (CR)*** is the percentage of human samples falling in the statistically estimated interval: for each human sample $p_{i,\text{human}}$, given N model predictions $\mathcal{P}_i = \{p_{i,1}, \ldots, p_{i,N}\}$, we perform B bootstrap resamples of $\mathcal{P}_i$ and compute the 5th and 95th percentiles for each resample. Aggregating over all bootstrap samples yields a robust $(1-\alpha)$ predictive interval $[l_i, u_i]$, and CR is defined as

$$\text{CR} = \frac{1}{|\mathcal{D}|}\sum_{i\in\mathcal{D}} \mathbb{I}\left(l_i \le p_{i,\text{human}} \le u_i\right), \tag{6}$$

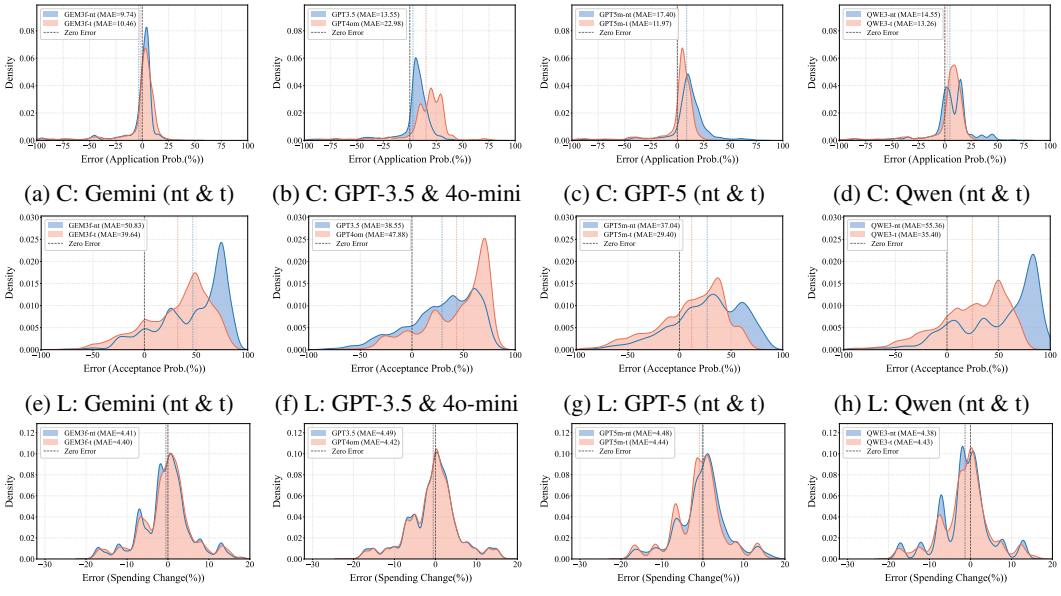


(a) C: Gemini (nt & t) (b) C: GPT-3.5 & 4o-mini (c) C: GPT-5 (nt & t) (d) C: Qwen (nt & t)

(e) L: Gemini (nt & t) (f) L: GPT-3.5 & 4o-mini (g) L: GPT-5 (nt & t) (h) L: Qwen (nt & t)

(i) S: Gemini (nt & t) (j) S: GPT-3.5 & 4o-mini (k) S: GPT-5 (nt & t) (l) S: Qwen (nt & t)

Figure 8: Distribution differences (RQ 1.1 at the aggregation level).



(a) C: Gemini (nt & t) (b) C: GPT-3.5 & 4o-mini (c) C: GPT-5 (nt & t) (d) C: Qwen (nt & t)

(e) L: Gemini (nt & t) (f) L: GPT-3.5 & 4o-mini (g) L: GPT-5 (nt & t) (h) L: Qwen (nt & t)

(i) S: Gemini (nt & t) (j) S: GPT-3.5 & 4o-mini (k) S: GPT-5 (nt & t) (l) S: Qwen (nt & t)

Figure 9: Individual-level differences (RQ 1.2).

where $\mathbb{I}(\cdot)$ is the indicator function. CR jointly reflects location mismatch and dispersion mismatch: low coverage may arise because the elicited response distribution is incorrectly centered, too narrow, or both. Consequently, CR must be interpreted jointly with MAE and JS divergence, which respectively characterize individual-level location error and population-level distributional mismatch.

E Additional Experiment Results

E.1 Additional Behavioral Tasks

To test whether the main findings extend beyond one question per domain, we select three complementary SCE tasks: the probability of purchasing a car or another vehicle within the next four months (26,487 valid instances), the share of a one-time 10% income increase that would be saved or invested (16,318 instances), and the probability of being employed four months from now (26,702 instances). These tasks respectively capture durable consumption, saving adjustment, and labor-force participation. We evaluate four representative GPT-5-mini and Qwen3-30B configurations using the same profile conditioning and evaluation pipeline as the main benchmark.

Table 6: Alignment results on three additional behavioral tasks. Thinking variants are denoted by "t" and non-thinking variants by "nt."

Task	Model	ρ	RMSE	JS	MAE	CR
Vehicle purchase	GPT5m-nt	0.12	9.72	0.40	17.11	0.08
	GPT5m-t	0.30	8.41	0.33	15.44	0.14
	QWE3-nt	0.25	9.21	0.37	16.48	0.09
	QWE3-t	0.29	8.63	0.30	15.71	0.15
Saving adjustment	GPT5m-nt	-0.04	8.31	0.36	31.62	0.07
	GPT5m-t	0.21	7.34	0.28	28.96	0.12
	QWE3-nt	0.13	7.98	0.34	30.91	0.08
	QWE3-t	0.24	7.61	0.27	29.27	0.13
Future employment	GPT5m-nt	-0.11	14.62	0.41	30.18	0.09
	GPT5m-t	0.24	13.76	0.30	26.84	0.16
	QWE3-nt	0.09	15.76	0.38	29.10	0.10
	QWE3-t	0.29	14.09	0.30	26.21	0.16

Across all three tasks, enabling thinking consistently improves distributional and individual alignment relative to the corresponding non-thinking configuration: JS divergence and MAE decrease, while CR increases. Nevertheless, temporal correlations remain modest and CR does not exceed 0.16, reproducing the main finding that reasoning can improve alignment without resolving limited individual-level fidelity.

E.2 Frontier-Scale Model Evaluation

We additionally evaluate Claude Opus 4.7 and GPT-5.6 Sol using the same prompts, SCE instances, preprocessing, and metrics as the main experiments. For Claude Opus 4.7, the no-thinking configuration omits reasoning, while the thinking configuration uses adaptive thinking with high effort. For GPT-5.6 Sol, the no-thinking and thinking configurations use none and medium reasoning effort, respectively.

Table 7: Frontier-scale model results across the three main tasks. "Previous best" reports the best earlier result for each metric and may combine different models rather than representing one configuration.

Task	Model	ρ	RMSE	JS	MAE	CR
Spending	Previous best	0.69	13.52	0.13	7.50	0.60
	Claude Opus 4.7-nt	0.49	13.62	0.18	7.72	0.20
	Claude Opus 4.7-t	0.61	13.56	0.14	7.61	0.28
	GPT-5.6 Sol-nt	0.55	13.59	0.16	7.65	0.24
	GPT-5.6 Sol-t	0.66	13.54	0.14	7.58	0.30
Labor	Previous best	0.50	34.69	0.26	36.70	0.23
	Claude Opus 4.7-nt	0.28	42.10	0.34	39.20	0.10
	Claude Opus 4.7-t	0.45	38.24	0.26	36.92	0.15
	GPT-5.6 Sol-nt	0.33	40.38	0.32	38.11	0.12
	GPT-5.6 Sol-t	0.48	36.97	0.24	36.74	0.16
Credit	Previous best	0.12	19.93	0.24	9.61	0.83
	Claude Opus 4.7-nt	0.04	20.31	0.30	14.28	0.16
	Claude Opus 4.7-t	0.09	20.02	0.25	13.62	0.19
	GPT-5.6 Sol-nt	0.06	20.14	0.28	13.91	0.17
	GPT-5.6 Sol-t	0.11	19.97	0.24	13.37	0.21

Reasoning improves alignment within both frontier families across the three domains, generally reducing JS divergence and MAE while increasing CR. However, the frontier configurations do not consistently exceed the previous metric-wise best results, particularly for individual-level coverage. These results reinforce the main conclusion that greater general model capability and additional reasoning do not by themselves guarantee reliable microeconomic behavioral alignment.

E.3 Elicitation Sensitivity

We test whether the individual-level conclusions depend on the number of model-elicited responses or decoding temperature. Using Qwen3-30B-A3B with thinking disabled and enabled, we vary $N \in \{10, 25, 50, 100\}$ and $T \in \{0.5, 1.0, 1.3, 1.6\}$, and macro-average the results across Spending, Labor, and Credit. The largest changes occur between $N = 10$ and $N = 25$, while the metrics stabilize by $N = 50$.

Table 8: Sensitivity of macro-averaged individual-level results to elicitation size N and decoding temperature T. Ranges summarize all tested temperature values.

Factor	Tested values	QWE3-nt	QWE3-t
Response count	$N = 10, 25, 50, 100$	$N = 50 \to 100$: MAE 21.42 $\to$ 21.39; $\lvert\Delta\text{CR}\rvert < 0.01$	$N = 50 \to 100$: MAE 20.15 $\to$ 20.12; $\lvert\Delta\text{CR}\rvert < 0.01$
Temperature	$T = 0.5, 1.0, 1.3, 1.6$	JS 0.33–0.34; MAE 21.39–21.49; CR 0.08–0.09	JS 0.27–0.28; MAE 20.13–20.22; CR 0.14–0.15

The $N = 50$ to $N = 100$ changes are negligible, supporting the default choice of $N = 50$. Across the tested temperatures, JS, MAE, and CR remain within narrow ranges for both configurations, indicating that greater decoding randomness alone cannot explain the main alignment gap.

E.4 Bootstrap Confidence Intervals

We quantify uncertainty by resampling the existing paired human–model observations with replacement and recomputing the complete aggregation and evaluation pipeline. The tables report point estimates with 95% percentile bootstrap confidence intervals in brackets. These intervals are conditional on the fixed model generations and therefore quantify sampling uncertainty in the evaluated observations, rather than variation across repeated end-to-end LLM inference runs. The intervals for RMSE, JS divergence, MAE, and CR are generally narrow, whereas the intervals for Spearman correlation are wider because the temporal statistic is computed from 24 monthly waves.

Table 9: Point estimates and 95% paired-bootstrap confidence intervals for the Spending task.

Model	ρ	RMSE	JS	MAE	CR
GPT3.5	$0.11\,[-0.05, 0.26]$	$13.64\,[13.30, 13.96]$	$0.22\,[0.18, 0.24]$	$7.67\,[7.25, 8.13]$	$0.16\,[0.15, 0.17]$
GPT4om	$0.06\,[-0.09, 0.22]$	$14.26\,[13.95, 14.59]$	$0.21\,[0.17, 0.23]$	$7.74\,[7.32, 8.17]$	$0.48\,[0.46, 0.50]$
GPT5m-nt	$0.32\,[0.21, 0.45]$	$13.76\,[13.49, 14.08]$	$0.20\,[0.18, 0.22]$	$8.02\,[7.49, 8.62]$	$0.24\,[0.22, 0.25]$
GPT5m-t	$0.23\,[0.08, 0.39]$	$13.67\,[13.35, 14.00]$	$0.13\,[0.11, 0.15]$	$7.75\,[7.23, 8.38]$	$0.60\,[0.58, 0.62]$
GEM3f-nt	$0.69\,[0.61, 0.79]$	$13.52\,[13.21, 13.85]$	$0.18\,[0.14, 0.21]$	$7.50\,[6.94, 8.11]$	$0.44\,[0.43, 0.46]$
GEM3f-t	$0.60\,[0.51, 0.70]$	$13.55\,[13.24, 13.89]$	$0.18\,[0.14, 0.20]$	$7.52\,[6.97, 8.09]$	$0.50\,[0.48, 0.52]$
QWE3-nt	$0.41\,[0.29, 0.53]$	$13.56\,[13.24, 13.87]$	$0.22\,[0.21, 0.24]$	$7.69\,[7.28, 8.13]$	$0.12\,[0.10, 0.13]$
QWE3-t	$0.64\,[0.54, 0.75]$	$13.59\,[13.28, 13.90]$	$0.15\,[0.14, 0.18]$	$7.56\,[7.12, 8.02]$	$0.14\,[0.12, 0.15]$

Table 10: Point estimates and 95% paired-bootstrap confidence intervals for the Labor task.

Model	ρ	RMSE	JS	MAE	CR
GPT3.5	$-0.11\,[-0.31, 0.08]$	$44.20\,[43.88, 44.51]$	$0.34\,[0.33, 0.34]$	$38.61\,[37.47, 39.71]$	$0.06\,[0.05, 0.07]$
GPT4om	$0.32\,[0.07, 0.56]$	$53.29\,[53.01, 53.59]$	$0.47\,[0.47, 0.48]$	$45.35\,[44.14, 46.46]$	$0.17\,[0.15, 0.19]$
GPT5m-nt	$-0.17\,[-0.37, 0.04]$	$34.69\,[34.42, 34.99]$	$0.32\,[0.32, 0.33]$	$36.70\,[35.57, 37.80]$	$0.18\,[0.16, 0.19]$
GPT5m-t	$0.08\,[-0.17, 0.32]$	$43.55\,[43.22, 43.87]$	$0.28\,[0.27, 0.28]$	$41.75\,[40.71, 42.83]$	$0.23\,[0.21, 0.25]$
GEM3f-nt	$0.36\,[0.12, 0.60]$	$56.99\,[56.65, 57.32]$	$0.48\,[0.47, 0.49]$	$43.93\,[42.88, 45.04]$	$0.22\,[0.20, 0.24]$
GEM3f-t	$0.50\,[0.30, 0.68]$	$45.22\,[44.93, 45.52]$	$0.31\,[0.30, 0.32]$	$40.96\,[39.97, 41.94]$	$0.22\,[0.20, 0.24]$
QWE3-nt	$-0.46\,[-0.70, -0.18]$	$62.96\,[62.61, 63.28]$	$0.47\,[0.47, 0.48]$	$41.38\,[40.21, 42.46]$	$0.05\,[0.04, 0.06]$
QWE3-t	$0.34\,[0.11, 0.55]$	$41.13\,[40.85, 41.41]$	$0.26\,[0.26, 0.27]$	$37.45\,[36.43, 38.58]$	$0.09\,[0.07, 0.10]$

Table 11: Point estimates and 95% paired-bootstrap confidence intervals for the Credit task.

Model	ρ	RMSE	JS	MAE	CR
GPT3.5	$0.02\,[-0.25, 0.29]$	$20.38\,[20.02, 20.72]$	$0.44\,[0.43, 0.44]$	$12.99\,[12.40, 13.62]$	$0.21\,[0.19, 0.23]$
GPT4om	$-0.38\,[-0.64, -0.11]$	$26.94\,[26.71, 27.16]$	$0.48\,[0.48, 0.49]$	$22.33\,[21.78, 22.90]$	$0.14\,[0.13, 0.16]$
GPT5m-nt	$-0.18\,[-0.43, 0.09]$	$20.08\,[19.70, 20.47]$	$0.47\,[0.46, 0.47]$	$9.61\,[8.89, 10.32]$	$0.50\,[0.48, 0.52]$
GPT5m-t	$0.12\,[-0.17, 0.38]$	$23.37\,[23.08, 23.68]$	$0.43\,[0.43, 0.43]$	$17.50\,[16.93, 18.08]$	$0.32\,[0.30, 0.34]$
GEM3f-nt	$-0.03\,[-0.32, 0.26]$	$20.20\,[19.77, 20.62]$	$0.27\,[0.27, 0.27]$	$10.63\,[9.92, 11.34]$	$0.72\,[0.71, 0.74]$
GEM3f-t	$0.07\,[-0.21, 0.34]$	$19.93\,[19.54, 20.35]$	$0.24\,[0.24, 0.24]$	$10.65\,[9.95, 11.35]$	$0.83\,[0.81, 0.84]$
QWE3-nt	$0.00\,[-0.27, 0.30]$	$22.18\,[21.87, 22.49]$	$0.31\,[0.30, 0.31]$	$15.19\,[14.62, 15.76]$	$0.08\,[0.07, 0.09]$
QWE3-t	$-0.01\,[-0.29, 0.27]$	$20.23\,[19.86, 20.59]$	$0.41\,[0.40, 0.41]$	$15.43\,[14.83, 16.02]$	$0.21\,[0.19, 0.23]$

E.5 Post-Cutoff Robustness

To complement the numerical perturbation test in RQ 3.1, we construct a Spending-only post-cutoff slice covering June 2024 to April 2025 (11 waves; 11,638 observations). We evaluate GPT-3.5-turbo (September 2021 cutoff), GPT-4o-mini (October 2023 cutoff), and the minimal- and medium-reasoning GPT-5-mini configurations (May 2024 cutoff). Table 12 reports the post-cutoff result minus the original full-period result, so values close to zero indicate that the evaluation conclusion is stable across the two periods.

Table 12: Post-cutoff Spending results relative to the original full-period results. Each entry is $\mathrm{PostCutoff} - \mathrm{Original}$.

Model	$\Delta\rho$	ΔJS	ΔRMSE	ΔMAE	ΔCR
GPT-3.5-turbo	-0.03	-0.04	-0.09	-0.20	$+0.03$
GPT-4o-mini	-0.07	$+0.02$	$+0.02$	-0.11	-0.04
GPT-5-mini-nt	-0.03	$+0.05$	-0.06	-0.15	$+0.02$
GPT-5-mini-t	-0.04	-0.03	-0.25	-0.25	$+0.08$

Across all four configurations, the changes are small relative to the scale of the original metrics. The post-cutoff slice therefore provides a cleaner robustness check than perturbing already observed macroeconomic values and supports the conclusion that the main findings are not driven by memorization of the original survey period. This test does not rule out all possible forms of contamination, but it reduces concern that the reported alignment patterns depend on exposure to the 2016–2023 SCE period.

Table 13: Per-100-request serving cost and configuration for the Exp1 N1 and N50 experiments. Parallel elapsed time and token consumption are normalised to a common base of 100 requests to allow direct comparison across settings. Serving settings report the execution backend, GPU memory, maximum output tokens, and request pool size.

Block	Model	Setting	Parallel elapsed / 100 req (s)	Mean latency / req (s)	Tokens / 100 req (K)
N1	GPT-3.5-turbo	API, max tokens 4096, pool 150	0.67	0.89	45.4
N1	GPT-4o-mini	API, max tokens 4096, pool 150	2.41	3.12	45.3
N1	GPT-5-mini (minimal)	API, max tokens 40960, pool 150	2.26	2.92	46.4
N1	GPT-5-mini (medium)	API, max tokens 40960, pool 150	5.08	6.60	89.6
N1	Gemini-3-Flash (no thinking)	API, max tokens 40960, pool 150	9.24	12.03	47.7
N1	Gemini-3-Flash (thinking)	API, max tokens 40960, pool 150	12.84	17.03	149.9
N1	Qwen3-30B Instruct (vLLM)	local vLLM, 1×A800 80GB, max tokens 4096, pool 150	2.26	2.82	47.5
N1	Qwen3-30B thinking (vLLM)	local vLLM, 1×A800 80GB, max tokens 4096, pool 150	11.50	55.67	181.9
N50	GPT-3.5-turbo	API, max tokens 4096, pool 150	11.81	11.65	69.9
N50	GPT-4o-mini	API, max tokens 4096, pool 150	3.08	4.09	70.8
N50	GPT-5-mini (minimal)	API, max tokens 40960, pool 150	18.49	6.41	69.9
N50	GPT-5-mini (medium)	API, max tokens 40960, pool 150	11.81	14.05	180.3
N50	Gemini-3-Flash (no thinking)	API, max tokens 40960, pool 150	8.22	7.63	76.6
N50	Gemini-3-Flash (thinking)	API, max tokens 40960, pool 150	20.55	22.91	326.6
N50	Qwen3-30B Instruct (vLLM)	local vLLM, 1×A800 80GB, max tokens 4096, pool 150	2.05	14.37	77.2
N50	Qwen3-30B thinking (vLLM)	local vLLM, 1×A800 80GB, max tokens 4096, pool 150	64.20	67.15	209.3

E.6 Income Segment Analysis

Building on the overall comparison in Fig. 3, we further stratify the evaluation by income level to examine how LLM–human alignment varies across socioeconomic groups (Figs. 10, 11 and 12). A clear and consistent pattern emerges: LLMs exhibit the strongest alignment with human behavior in the middle-income group, while alignment deteriorates for both low-income and high-income individuals. Specifically, alignment degrades toward the tails of the income distribution, where economic behavior becomes either more constrained (low-income) or more idiosyncratic and strategic (high-income). This highlights an important limitation for LLM-based economic agents: while they can serve as reasonable simulators for average agents, caution is required when deploying them to model behavior at socioeconomic extremes.

E.7 Cost Analysis

We report the inference time and token consumption for the main experiments. As shown in Tab. 13, due to the large scale of our benchmark and the substantial token usage, we did not perform repeated end-to-end inference runs; paired-bootstrap confidence intervals conditional on the fixed model generations are reported in Appendix E.4.

E.8 Economic Analysis

The findings can be understood through established economic theories:

- **Age:** Based on *Life-Cycle Theory* (Modigliani & Brumberg, 2013; Friedman, 2018), households consider expected lifetime resources rather than current income alone. Younger households typically have higher expected future income and stronger borrowing motives. Removing age reduces the model's ability to capture intertemporal incentives.
- **Housing status:** Following *Collateral Constraint Theory* (Iacoviello, 2005), housing is a pledgeable asset that relaxes borrowing constraints and changes credit capacity. Asset ownership affects how much a household can borrow and the sensitivity of borrowing to macro-financial conditions.
- **Employment status:** According to *Job Search Theory* and *Reservation Wage Theory* (McCall, 1970), job acceptance depends on whether the offered wage exceeds the worker's reservation threshold, which in turn depends on employment status, outside options, and expected future offers. Removing employment status reduces the model's ability to capture key determinants of labor supply and household borrowing.
- **Macro variables by domain:** Guided by *Intertemporal Choice Theory* and *Precautionary Saving Theory* (Deaton, 1989; Carroll, 1997), interest rates, inflation, and unemployment affect household uncertainty, consumption affordability, and labor-offer evaluation. Their effects vary across domains.

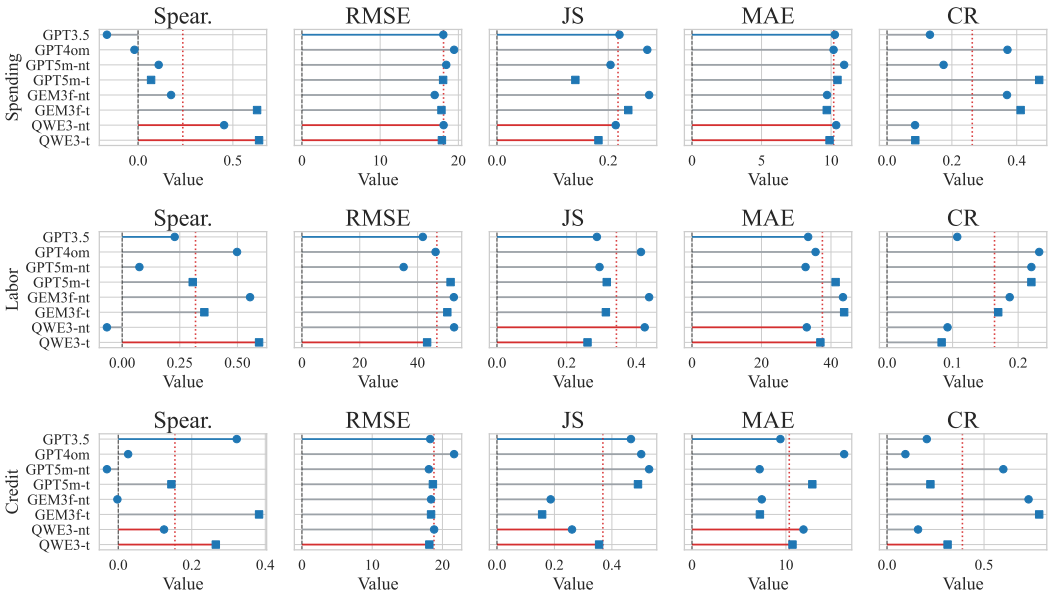


Figure 10: Comparative alignment across LLMs for low-income individuals.

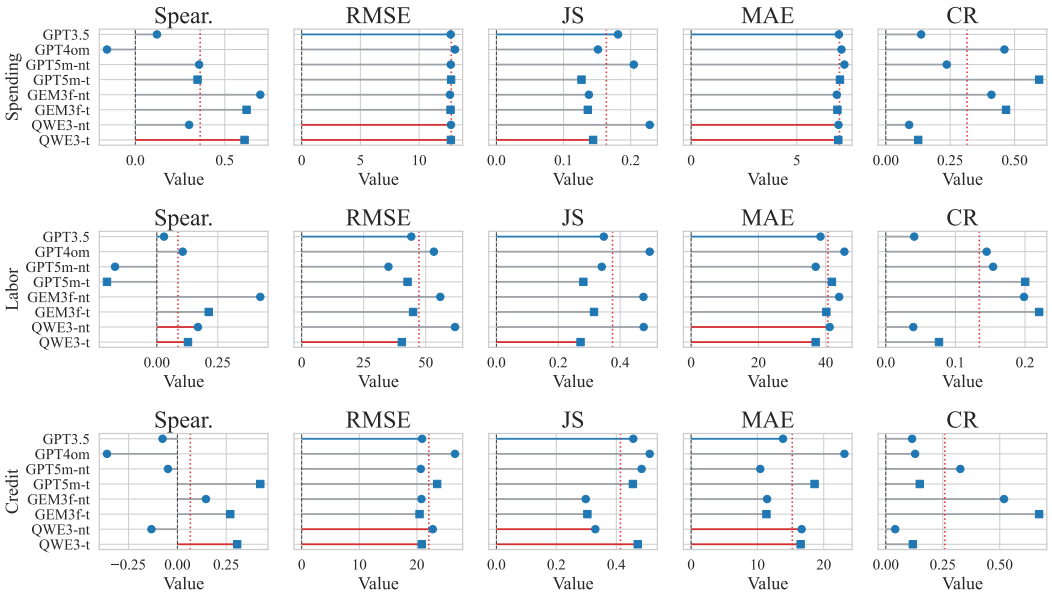


Figure 11: Comparative alignment across LLMs for middle-income individuals.

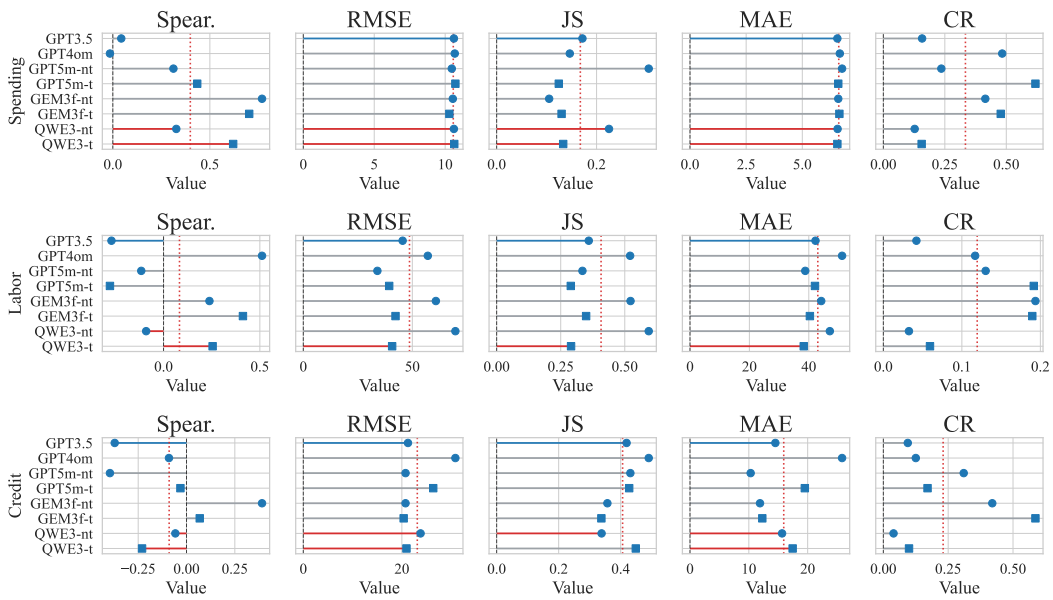


Figure 12: Comparative alignment across LLMs for high-income individuals.

Table 14: **Ablation study: feature importance analysis (demographics vs. macro).** This table reports raw metric changes after removing features (Δ = Ablated − Full). Arrows (↑/↓) denote the favorable direction; for higher-is-better metrics, positive values indicate improvement, while for lower-is-better metrics, negative values indicate improvement. **Bold** values indicate the strongest impact within each group.

Feature	Spending					Labor					Credit				
	$\rho\uparrow$	RMSE↓	JS↓	MAE↓	CR↑	$\rho\uparrow$	RMSE↓	JS↓	MAE↓	CR↑	$\rho\uparrow$	RMSE↓	JS↓	MAE↓	CR↑
Demographics															
Age	-0.181	+0.088	-0.001	-0.230	+0.004	+0.065	**+1.513**	**+0.070**	+4.990	-0.020	+0.072	+0.851	+0.013	+17.740	-0.057
Education	-0.470	-0.981	+0.060	-0.210	+0.002	-0.141	**-1.575**	-0.003	+4.480	-0.013	+0.050	+0.575	**-0.067**	+17.600	-0.061
Employment	-0.294	-0.148	+0.056	-0.230	-0.005	-0.009	-0.086	+0.041	+5.090	-0.021	-0.135	+0.268	+0.040	+17.950	-0.062
Gender	-0.089	+0.061	-0.003	-0.250	**+0.008**	+0.170	+0.259	+0.016	+4.820	-0.013	**-0.203**	+0.337	-0.015	+17.770	-0.054
Health	**-0.538**	-1.095	+0.063	-0.210	-0.001	+0.174	-0.864	-0.018	+5.070	-0.022	-0.116	-0.059	-0.007	+17.890	-0.058
Housing	-0.152	-0.411	+0.068	-0.210	-0.007	-0.033	-0.346	+0.022	**+5.170**	**-0.029**	-0.181	**+9.873**	+0.025	+17.710	-0.059
Income	-0.155	+0.011	-0.014	**-0.260**	-0.003	+0.208	-0.581	-0.007	+4.770	-0.021	-0.150	+0.415	-0.042	+17.720	-0.060
Marital	-0.337	**-1.155**	+0.069	-0.220	+0.002	-0.008	-1.060	-0.010	+4.780	-0.021	-0.054	+0.346	-0.010	+17.460	**-0.063**
Own Other Home	-0.110	-0.771	**+0.075**	-0.230	+0.002	+0.039	+0.253	+0.014	+4.960	-0.023	-0.101	+0.470	-0.026	+17.570	-0.059
State	-0.430	-0.961	+0.067	-0.230	+0.002	**+0.269**	-0.625	+0.001	+4.870	-0.021	-0.005	+0.559	-0.011	**+17.970**	-0.061
Macro-Environment															
Inflation	-0.142	**+0.648**	**-0.018**	+0.370	**+0.075**	+0.032	**-2.144**	**-0.040**	+3.230	**+0.028**	**+0.102**	**-0.326**	-0.013	+1.230	**+0.016**
Interest Rate	-0.223	+0.344	+0.006	+0.350	+0.072	+0.191	-1.280	-0.003	**+3.630**	+0.014	+0.020	-0.082	-0.005	**+1.400**	+0.012
Unemployment	**-0.454**	**-0.838**	**+0.074**	**+0.380**	+0.071	**+0.326**	-1.794	**+0.016**	+3.610	+0.012	-0.033	+0.084	**+0.056**	+1.260	+0.010

F Experimental Details

We present prompt examples for the different settings in this section; see Figs. 13–29.

Instance Prompt: Credit Access Survey (Pointwise Estimation)

System Instruction: Context & Persona

Roleplay Instruction
You are a real person with the following demographic profile. Answer the survey below, giving your best reasonable estimates.
User Profile
Your age is 78. Your gender is Male. Your highest level of education is: Master's degree. Your marital status is: Married or partnered. You live in North Dakota. Your housing status is: Own. You don't own another home. Your health status is: Excellent. Your current employment status includes: retired or early retired. Your household's annual income is in the range: $50,000 to $59,999.
Macroeconomic Environment
The inflation rate over the past 4 months is: [0.85, 0.89, 1.17, 1.08].
The federal funds rate over the past 4 months is: [0.38, 0.36, 0.37, 0.37].
The unemployment rate over the past 4 months is: [4.9, 5.0, 5.1, 4.8].

User Instruction: Task & Constraints

You are asked to answer a credit survey. Please answer clearly as instructed.
-- Question --
Over the next 12 months, what do you think is the percent chance that you will **Apply for a mortgage or home-based loan**?
-- Output Requirement --
Please output only a single numerical value representing the probability (0-100).

- Format: A number between 0 and 100.
- Example: If you think there is a X% chance, output 'X'.
- Constraint: Please provide the specific number that reflects your situation.
- **Strictly NO % symbol, words, or JSON. Just the number.**

Figure 13: **Instance prompt for the credit access survey (pointwise estimation).**

Instance Prompt: Credit Access Survey (Sampling Distribution)

System Instruction: Context & Persona

Roleplay Instruction
Now you are a real person with the following demographic profile. Answer the survey below, giving your best reasonable estimates.
User Profile
Your age is 68. Your gender is Male. Your highest level of education is: Some college but no degree. Your marital status is: Married or partnered. You live in Colorado. Your housing status is: Own. You don't own another home. Your health status is: Very good. Your current employment status includes: retired or early retired. 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: [1.67, 1.83, 1.74, 1.68].
The federal funds rate over the past 4 months is: [2.38, 2.4, 2.13, 2.04].
The unemployment rate over the past 4 months is: [3.6, 3.7, 3.6, 3.5].

User Instruction: Task & Constraints

-- Question --
Over the next 12 months, what do you think is the percent chance that you will **Apply for a mortgage or home-based loan**?
-- Output Requirement --
To simulate variability, generate 50 independent samples of probability estimates. Each sample should represent one possible answer from a person with your profile.

- Format: Output a list of 50 numbers, each between 0 and 100.
- Example: [X1, X2, ...]
- Constraint: Output ONLY the list. No explanation, no text, no markdown.

Figure 14: **Instance prompt for the credit access survey (sampling distribution).**

Instance Prompt: Labor Supply Survey (Pointwise Estimation)

System Instruction: Context & Persona

Roleplay Instruction
Now you are a real person with the following demographic profile. Answer the survey below, giving your best reasonable estimates.
User Profile
Your age is 76. Your gender is Female. Your highest level of education is: High school diploma. Your marital status is: Single. You live in Florida. Your housing status is: Own. You don't own another home. Your health status is: Fair. Your current employment status includes: working part-time. Your household's annual income is in the range: $20,000 to $29,999.
Macroeconomic Environment
The inflation rate over the past 4 months is: [0.44, 0.64, 1.24, 0.85].
The federal funds rate over the past 4 months is: [0.12, 0.24, 0.34, 0.38].
The unemployment rate over the past 4 months is: [5.1, 5.0, 4.8, 4.9].

User Instruction: Task & Constraints

You are asked to answer a labor survey. Please answer clearly as instructed.
-- Question --
Think about the job offers that you may receive within the coming four months. If the job with the best offer has an annual salary for the first year that equals **0.9-1.0 times your expected annual salary**, what do you think is the percent chance that you will accept the job?
-- Output Requirement --
Please output only a single numerical value representing the percentage chance (probability) based on your best estimate.

- Range: The value must be between 0 and 100 (inclusive).
- Examples of format: If you think there is a X% chance, output 'X'.
- Constraint: Please provide the specific number that reflects your situation.
- **Strictly NO % symbol, words, or JSON. Just the number.**

Figure 15: **Instance prompt for the labor supply survey (pointwise estimation).**

Instance Prompt: Labor Supply Survey (Sampling Distribution)

System Instruction: Context & Persona

Roleplay Instruction
Now you are a real person with the following demographic profile. Answer the survey below, giving your best reasonable estimates.
User Profile
Your age is 43. Your gender is Female. Your highest level of education is: Some college but no degree. Your marital status is: Single. You live in Michigan. Your housing status is: Rent. You don't own another home. Your health status is: Good. Your current employment status includes: homemaker. Your household's annual income is in the range: Less than $10,000.
Macroeconomic Environment
The inflation rate over the past 4 months is: [1.88, 2.0, 1.8, 1.67].
The federal funds rate over the past 4 months is: [2.41, 2.42, 2.39, 2.38].
The unemployment rate over the past 4 months is: [3.8, 3.7, 3.6, 3.6].

User Instruction: Task & Constraints

-- Question --
Think about the job offers that you may receive within the coming four months. If the job with the best offer has an annual salary for the first year that equals **0.9–1.0 times your expected annual salary**, what do you think is the percent chance that you will accept the job?
-- Output Requirement --
To simulate variability, generate 50 independent samples of probability estimates. Each sample should represent one possible answer from a person with your profile.

- Format: Output a list of 50 numbers, each between 0 and 100.
- Example: [X1, X2, ...]
- Constraint: Output ONLY the list. No explanation, no text, no markdown.

Figure 16: **Instance prompt for the labor supply survey (sampling distribution).**

Instance Prompt: Household Spending Survey (Pointwise Estimation)

System Instruction: Context & Persona

Roleplay Instruction
Now you are a real person with the following demographic profile. Answer the survey below, giving your best reasonable estimates.
User Profile
Your age is 59. Your gender is Female. Your highest level of education is: Bachelor's degree. Your marital status is: Married or partnered. You live in Maine. Your housing status is: Own. You own another home. Your health status is: Excellent. Your current employment status includes: working full-time. Your household's annual income is in the range: $100,000 to $149,999.
Macroeconomic Environment
The inflation rate over the past 4 months is: [0.64, 1.24, 0.85, 0.89].
The federal funds rate over the past 4 months is: [0.24, 0.34, 0.38, 0.36].
The unemployment rate over the past 4 months is: [5.0, 4.8, 4.9, 5.0].

User Instruction: Task & Constraints

You are asked to answer a spending survey. Please answer clearly as instructed.
-- Question --
Now think about your total household spending (groceries, housing, medical, transportation, large items, etc.).
What do you expect will happen to the total spending of all members of your household (including you) over the next 12 months?
-- Output Requirement --
Please output only a single numerical value representing the percentage change based on your best estimate.

- Format: Positive number for increase, negative number for decrease.
- Examples: If you expect a X% increase, output 'X'. If decrease, output '-X'.
- Constraint: Please provide the specific number that reflects your situation.
- **Strictly NO % symbol, words, or JSON. Just the number.**

Figure 17: **Instance prompt for the household spending survey (pointwise estimation).**

Instance Prompt: Household Spending Survey (Sampling Distribution)

System Instruction: Context & Persona

Roleplay Instruction
Now you are a real person with the following demographic profile. Answer the survey below, giving your best reasonable estimates.
User Profile
Your age is 80. Your gender is Male. Your highest level of education is: Bachelor's degree. Your marital status is: Single. You live in Florida. Your housing status is: Own. You don't own another home. Your health status is: Very good. Your current employment status includes: working part-time. Your household's annual income is in the range: $20,000 to $29,999.
Macroeconomic Environment
The inflation rate over the past 4 months is: [2.47, 2.78, 2.81, 2.85].
The federal funds rate over the past 4 months is: [1.69, 1.7, 1.82, 1.91].
The unemployment rate over the past 4 months is: [4.0, 3.8, 4.0, 3.8].

User Instruction: Task & Constraints

-- Question --
What do you expect will happen to the total spending of all members of your household (including you) over the next 12 months? (Think about groceries, housing, medical, transportation, etc.)
-- Output Requirement --
To simulate variability, generate 50 independent samples of percentage change estimates. Each sample should represent one possible answer from a person with your profile.

- Format: Output a list of 50 numbers. Positive for increase, negative for decrease.
- Example: [X1, X2, ...]
- Constraint: Output ONLY the list. No explanation, no text, no markdown.

Figure 18: **Instance prompt for the household spending survey (sampling distribution).**

Instance Prompt: Credit Access Survey (Ablation: Age Removed, Pointwise Estimation)

System Instruction: Context & Persona

Roleplay Instruction
Now you are a real person with the following demographic profile. Answer the survey below, giving your best reasonable estimates.
User Profile
Your gender is Female. Your highest level of education is: Some college but no degree. Your marital status is: Single. You live in Minnesota. Your housing status is: Own. You don't own another home. Your health status is: Good. Your current employment status includes: retired or early retired. Your household's annual income is in the range: $20,000 to $29,999.
Macroeconomic Environment
The inflation rate over the past 4 months is: [0.13, 0.44, 0.64, 1.24].
The federal funds rate over the past 4 months is: [0.12, 0.12, 0.24, 0.34].
The unemployment rate over the past 4 months is: [5.0, 5.1, 5.0, 4.8].

User Instruction: Task & Constraints

You are asked to answer a credit survey. Please answer clearly as instructed.
-- Question --
Over the next 12 months, what do you think is the percent chance that you will **Apply for a mortgage or home-based loan**?
-- Output Requirement --
Please output only a single numerical value representing the probability (0-100).

- Format: A number between 0 and 100.
- Example: If you think there is a X% chance, output 'X'.
- Constraint: Please provide the specific number that reflects your situation.
- **Strictly NO % symbol, words, or JSON. Just the number.**

Figure 19: **Instance prompt for the credit access survey without age (pointwise estimation).**

Instance Prompt: Credit Access Survey (Ablation: Age Removed, Sampling Distribution)

System Instruction: Context & Persona

Roleplay Instruction
Now you are a real person with the following demographic profile. Answer the survey below, giving your best reasonable estimates.
User Profile
Your gender is Female. Your highest level of education is: Some college but no degree. Your marital status is: Single. You live in Minnesota. Your housing status is: Own. You don't own another home. Your health status is: Good. Your current employment status includes: retired or early retired. Your household's annual income is in the range: $20,000 to $29,999.
Macroeconomic Environment
The inflation rate over the past 4 months is: [2.49, 2.15, 2.0, 1.49].
The federal funds rate over the past 4 months is: [2.19, 2.2, 2.27, 2.4].
The unemployment rate over the past 4 months is: [3.8, 3.8, 3.9, 4.0].

User Instruction: Task & Constraints

-- Question --
Over the next 12 months, what do you think is the percent chance that you will **Apply for a mortgage or home-based loan**?
-- Output Requirement --
To simulate variability, generate 50 independent samples of probability estimates. Each sample should represent one possible answer from a person with your profile.

- Format: Output a list of 50 numbers, each between 0 and 100.
- Example: [X1, X2, ...]
- Constraint: Output ONLY the list. No explanation, no text, no markdown.

Figure 20: **Instance prompt for the credit access survey without age (sampling distribution).**

Instance Prompt: Credit Access Survey (Third-Person Perspective, Pointwise Estimation)

System Instruction: Context & Persona (Third-Person)

Prediction Instruction
The following is a profile of a real participant. Based on the profile, predict how the participant with such a profile would answer the survey below, giving the best reasonable estimates.

Participant Profile
The participant's age is 64. The participant's gender is Female. The participant's highest level of education is: Master's degree. The participant's marital status is: Single. The participant lives in California. The participant's housing status is: Own. The participant does not own another home. The participant's health status is: Very good. The participant's current employment status includes: working full-time. The participant's household's annual income is in the range: $50,000 to $59,999.

Macroeconomic Environment
The inflation rate over the past 4 months is: [0.13, 0.44, 0.64, 1.24].
The federal funds rate over the past 4 months is: [0.12, 0.12, 0.24, 0.34].
The unemployment rate over the past 4 months is: [5.0, 5.1, 5.0, 4.8].

User Instruction: Task & Constraints

The participant is asked to answer a credit survey. The participant should answer clearly as instructed.

-- Question --
Over the next 12 months, what does the participant think is the percent chance that they will **Apply for a mortgage or home-based loan**?

-- Output Requirement --
Please output only a single numerical value representing the probability (0-100).

- Format: A number between 0 and 100.
- Example: If the participant thinks there is a X% chance, output 'X'.
- Constraint: Please provide the specific number that reflects their situation.
- **Strictly NO % symbol, words, or JSON. Just the number.**

Figure 21: **Instance prompt for third-person credit access prediction (pointwise estimation).**

Instance Prompt: Credit Access Survey (Third-Person Perspective, Sampling Distribution)

System Instruction: Context & Persona (Third-Person)

Prediction Instruction
The following is a profile of a real participant. Based on the profile, predict how the participant would answer the survey below, giving the best reasonable estimates.

Participant Profile
The participant's age is 73. The participant's gender is Male. The participant's highest level of education is: Master's degree. The participant's marital status is: Single. The participant lives in Wisconsin. The participant's housing status is: Own. The participant does not own another home. The participant's health status is: Good. The participant's current employment status includes: retired or early retired. The participant's household's annual income is in the range: $40,000 to $49,999.

Macroeconomic Environment
The inflation rate over the past 4 months is: [1.64, 1.73, 1.93, 2.18].
The federal funds rate over the past 4 months is: [1.04, 1.15, 1.16, 1.15].
The unemployment rate over the past 4 months is: [4.3, 4.3, 4.4, 4.3].

User Instruction: Task & Constraints

The participant is asked to answer a credit survey. The participant should answer clearly as instructed.

-- Question --
Over the next 12 months, what does the participant think is the percent chance that they will **Apply for a mortgage or home-based loan**?

-- Output Requirement --
To simulate variability, generate 50 independent samples of probability estimates. Each sample should represent one possible answer from a person with the participant's profile.

- Format: Output a list of 50 numbers, each between 0 and 100.
- Example: [X1, X2, ...]
- Constraint: Output ONLY the list. No explanation, no text, no markdown.

Figure 22: **Instance prompt for third-person credit access prediction (sampling distribution).**

Instance Prompt: Credit Access Survey (JSON Feature Injection, Pointwise Estimation)

System Instruction: Context & Persona (JSON Format)

Roleplay Instruction
Now you are a real person with the following demographic profile. Answer the survey below, giving your best reasonable estimates.
User Profile (JSON)
{"age": 66, "gender": "Female", "education": "Master's degree", "marital_status": "Single", "state_residence": "Illinois", "housing_status": "Own", "own_other_home": true, "health_status": "Excellent", "employment_status": ["retired or early retired", "other"]}
Macroeconomic Environment (JSON)
{"inflation_rate": [0.13, 0.44, 0.64, 1.24], "federal_funds_rate": [0.12, 0.12, 0.24, 0.34], "unemployment_rate": [5.0, 5.1, 5.0, 4.8]}

User Instruction: Task & Constraints

You are asked to answer a credit survey. Please answer clearly as instructed.
-- Question --
Over the next 12 months, what do you think is the percent chance that you will **Apply for a mortgage or home-based loan**?
-- Output Requirement --
Please output only a single numerical value representing the probability (0-100).

- Format: A number between 0 and 100.
- Example: If you think there is a X% chance, output 'X'.
- Constraint: Please provide the specific number that reflects your situation.
- **Strictly NO % symbol, words, or JSON. Just the number.**

Figure 23: **Instance prompt for JSON-based credit access prediction (pointwise estimation).**

Instance Prompt: Credit Access Survey (JSON Feature Injection, Sampling Distribution)

System Instruction: Context & Persona (JSON Format)

Roleplay Instruction
Now you are a real person with the following demographic profile. Answer the survey below, giving your best reasonable estimates.
User Profile (JSON)
{"age": 66, "gender": "Female", "education": "Master's degree", "marital_status": "Single", "state_residence": "Illinois", "housing_status": "Own", "own_other_home": true, "health_status": "Excellent", "employment_status": ["retired or early retired", "other"]}
Macroeconomic Environment (JSON)
{"inflation_rate": [0.13, 0.44, 0.64, 1.24], "federal_funds_rate": [0.12, 0.12, 0.24, 0.34], "unemployment_rate": [5.0, 5.1, 5.0, 4.8]}

User Instruction: Task & Constraints

You are asked to answer a credit survey. Please answer clearly as instructed.
-- Question --
Over the next 12 months, **what does the participant think is the percent chance** that they will Apply for a mortgage or home-based loan?
-- Output Requirement --
To simulate variability, generate 50 independent samples of probability estimates. Each sample should represent one possible answer from a person with the participant's profile.

- Format: Output a list of 50 numbers, each between 0 and 100.
- Example: [X1, X2, ...]
- Constraint: Output ONLY the list. No explanation, no text, no markdown.

Figure 24: **Instance prompt for JSON-based credit access prediction (sampling distribution).**

Instance Prompt: Credit Access Survey (With Historical Memory, Pointwise Estimation)

System Instruction: Context & Persona

Roleplay Instruction
Now you are a real person with the following demographic profile. Answer the survey below, giving your best reasonable estimates.
User Profile
Your age is 61. Your gender is Male. Your highest level of education is: Bachelor's degree. Your marital status is: Single. You live in New York. Your housing status is: Own. You own another home. Your health status is: Very good. Your current employment status includes: retired or early retired. 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: [0.13, 0.44, 0.64, 1.24].
The federal funds rate over the past 4 months is: [0.12, 0.12, 0.24, 0.34].
The unemployment rate over the past 4 months is: [5.0, 5.1, 5.0, 4.8].

User Instruction: Task & Constraints

You are asked to answer a credit survey. Please answer clearly as instructed.
-- Historical Context (Memory) --
When making the previous decision four months ago, the responses of you are as follows. Consider the environmental changes over the past four months to determine your decision outcome this time:
Previous Decision: 0
-- Question --
Over the next 12 months, what do you think is the percent chance that you will **Apply for a mortgage or home-based loan**?
-- Output Requirement --
Please output only a single numerical value representing the probability (0-100).

- Format: A number between 0 and 100.
- Example: If you think there is a X% chance, output 'X'.
- Constraint: Please provide the specific number that reflects your situation.
- **Strictly NO % symbol, words, or JSON. Just the number.**

Figure 25: **Instance prompt for memory-based credit access prediction (pointwise estimation).**

Instance Prompt: Credit Access Survey (Longitudinal Simulation, Sampling Distribution)

System Instruction: Context & Persona

Roleplay Instruction
Now you are a real person with the following demographic profile. Answer the survey below, giving your best reasonable estimates.
User Profile
Your age is 61. Your gender is Male. Your highest level of education is: Bachelor's degree. Your marital status is: Single. You live in New York. Your housing status is: Own. You own another home. Your health status is: Very good. Your current employment status includes: retired or early retired. 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: [0.13, 0.44, 0.64, 1.24].
The federal funds rate over the past 4 months is: [0.12, 0.12, 0.24, 0.34].
The unemployment rate over the past 4 months is: [5.0, 5.1, 5.0, 4.8].

User Instruction: Task & Constraints

You are asked to answer a credit survey. Please answer clearly as instructed.
-- Historical Context (Memory) --
When making the previous decision **four months ago**, the responses of you are as follows. Consider the environmental changes over the past four months to determine your decision outcome this time:
Previous Responses: [0, 0, 0,

Instance Prompt: Credit Access Survey (Pointwise Estimation)

System Instruction: Context & Persona

Roleplay Instruction
Now you are a real person with the following demographic profile. Answer the survey below, giving your best reasonable estimates.
User Profile
Your age is 72. Your gender is Male. Your highest level of education is: Associate degree. Your marital status is: Single. You live in Arkansas. Your housing status is: Rent. You don't own another home. Your health status is: Very good. Your current employment status includes: working part-time. Your household's annual income is in the range: $10,000 to $19,999.
Macroeconomic Environment
The inflation rate over the past 4 months is: [0.85, 0.89, 1.17, 1.08].
The federal funds rate over the past 4 months is: [0.38, 0.36, 0.37, 0.37].
The unemployment rate over the past 4 months is: [4.9, 5.0, 5.1, 4.8].

User Instruction: Task & Constraints

You are asked to answer a credit survey. Please answer clearly as instructed.
-- Social Context (Peer Reference) --
When making the previous decision four months ago, the responses from **real individuals with the same profile background as yours** were as follows. You should carefully review and learn from these responses, then based on these references, consider the environmental changes over the past four months to determine your decision outcome this time:
Peer Reference Value: 0.0
-- Question --
Over the next 12 months, what do you think is the percent chance that you will **Apply for a mortgage or home-based loan**?
-- Output Requirement --
Please output only a single numerical value representing the probability (0-100).

- Format: A number between 0 and 100.
- Example: If you think there is a X% chance, output 'X'.
- Constraint: Please provide the specific number that reflects your situation.
- **Strictly NO % symbol, words, or JSON. Just the number.**

Figure 27: **Instance prompt for peer-history credit access prediction (pointwise estimation).**

Instance Prompt: Credit Access Survey (Sampling Distribution)

System Instruction: Context & Persona

Roleplay Instruction
Now you are a real person with the following demographic profile. Answer the survey below, giving your best reasonable estimates.
User Profile
Your age is 69. Your gender is Male. Your highest level of education is: Some college but no degree. Your marital status is: Single. You live in Massachusetts. Your housing status is: Own. You own another home. Your health status is: Good. Your current employment status includes: retired or early retired. 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: [1.67, 1.83, 1.74, 1.68].
The federal funds rate over the past 4 months is: [2.38, 2.4, 2.13, 2.04].
The unemployment rate over the past 4 months is: [3.6, 3.7, 3.6, 3.5].

User Instruction: Task & Constraints

You are asked to answer a credit survey. Please answer clearly as instructed.
-- Social Context (Peer Reference) --
When making the previous decision four months ago, the responses from **real individuals with the same profile background as yours** were as follows. You should carefully review and learn from these responses, then based on these references, consider the environmental changes over the past four months to determine your decision outcome this time:
Peer Reference Value: 0.0
-- Question --
Over the next 12 months, what do you think is the percent chance that you will **Apply for a mortgage or home-based loan**?
-- Output Requirement --
To simulate variability, generate 50 independent samples of probability estimates. Each sample should represent one possible answer from a person with your profile.

- Format: Output a list of 50 numbers, each between 0 and 100.
- Example: [X1, X2, ...]
- Constraint: Output ONLY the list. No explanation, no text, no markdown.

Figure 28: **Instance prompt for peer-history credit access prediction (sampling distribution).**

Instance Prompt: ANES Democratic Party Feeling Thermometer

System Instruction:

```
Now you are a real person with the following demographic
profile.
```

Your profile:

```
- age: 49
- gender: male
- race/ethnicity: White non-Hispanic
- education: high school
- family income: 68th to 95th income percentile
```

Your survey environment:

```
- election type: midterm
- incumbent president's party: Democratic
- party control of Congress: Democratic
- monthly unemployment rates over the 12 months before your
survey: [6.4, 6.3, 6.3, 6.1, 6.0, 5.9, 6.2, 5.9, 6.0, 5.8,
5.9, 6.0]
- monthly year-over-year inflation rates over the 12 months
before your survey: [6.81, 6.24, 6.38, 6.5, 7.14, 7.44, 7.73,
7.86, 8.48, 8.93, 8.87, 8.99]
- Democratic Party positions: The Democratic Party generally
supported an active federal role in economic management,
social programs, labor protections, and civil rights.
- Republican Party positions: The Republican Party generally
emphasized lower taxes, reduced federal regulation, fiscal
restraint, and a stronger anti-communist foreign policy.
```

Question:

```
How would you rate the Democratic Party on a 0-100 feeling
thermometer, where 0 means very cold/unfavorable, 50 means
neutral, and 100 means very warm/favorable?
```

Output Instruction:

```
Return exactly one numeric answer. The answer must be between
0 and 100. Output only the number, with no prose, no JSON,
and no keys.
```

Figure 29: **Instance prompt for the ANES Democratic Party feeling thermometer task.**