HOUSEHOLDBENCH: EVALUATING LARGE LANGUAGE MODELS AS PREDICTORS OF HOUSEHOLD ECONOMIC BEHAVIOR

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) have the potential to meet a key goal in economics: a quantitative model of household decision making, across a variety of settings. Yet existing evaluations cover few surveys and outcomes, and do not study how households adjust to changing economic conditions. We introduce a new evaluation, HouseholdBench, which unites 6 U.S. household surveys and 32 prediction tasks spanning numeric, categorical and probabilistic outcomes, related to consumption, income, labor, expectations, and housing. Using past behavior, demographics and macroeconomic conditions, the tasks test whether LLMs predict behavior, including how households adjust to changes in various policies. We evaluate 13 proprietary and open-weight LLMs against a no-change baseline and a gradient-boosted tree model. Most LLMs outperform the no-change baseline, including for policy response tasks—with the best model lowering error for numeric outcomes by 12.2%. Across most tasks, gradient-boosted trees rank first; leading proprietary LLMs approach their performance, but open-weight models lag. LLMs exhibit systematic over- and underprediction across different tasks. We identify methods that enable a 4 billion parameter open-weight model to match proprietary models' performance: fine-tuning and aggregating 16 predictions per observation. Improvements generalize to policy-response tasks, which are excluded from fine-tuning.

1 INTRODUCTION

Large language models (LLMs) may be able to meet a key goal in economics—a quantitative model of how economic agents make decisions. To assess this promise, many papers evaluate whether LLMs can predict decision making in various settings, ranging from social science experiments (Filippas et al., 2024; Ashokkumar et al., 2026), to survey response (Argyle et al., 2022; Park et al., 2024), to behavior in economic games (Xie et al., 2025; Huang et al., 2026). A core question in economics is how households make real-world decisions about labor supply and consumption. Household behavior matters because it is a fundamental driver of aggregate outcomes. For instance, how households consume or save after income shocks determines the aggregate effect of fiscal and monetary policies (Kaplan et al., 2018; Auclert, 2019). If LLMs provide a realistic model of household decision making, then one can use them to carry out realistic simulations of the macroeconomy—as the literature in "generative agent-based modelling" has begun to explore (Li et al., 2024; Piao et al., 2025; Karten et al., 2025). In particular, one could simulate how various policies—such as stimulus checks or unemployment insurance changes—affect the economy.

There is not yet a comprehensive evaluation of how well LLMs simulate household economic behavior. Recent work evaluates LLMs on household survey data, for instance predicting a respondent's occupation, employment, or retirement (Athey et al., 2026; Jia et al., 2026; Garzón et al., 2026), and their income or homeownership (Cruz et al., 2024; Gao et al., 2026). However, previous work usually uses few surveys and focuses on few outcomes, meaning the results may be context specific rather than widely applicable. Moreover prior work does not ask whether LLMs can predict how households respond to real-world changes in economic policy—which is critical for realistic policy simulations.

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.

We propose HouseholdBench: a comprehensive benchmark for evaluating whether LLMs can predict household economic behavior across a variety of settings. HouseholdBench uses six public U.S. household surveys with rich longitudinal data on household economic behavior and has a total of 21.7M observations (Figure 1). We construct 32 tasks spanning five topics: consumption and saving, income and resources, labor and retirement, macroeconomic expectations, and housing and location. We cover three types of prediction tasks: numeric prediction, such as total household spending; categorical prediction, such as employment status; and probability prediction, such as households' subjective probability distributions for inflation. To evaluate if LLMs can predict how household behavior adjusts to policy, we include a set of *policy-response tasks*. These tasks ask how households respond to an external, policy-related change such as a stimulus check or a job displacement. Using past behavior, demographics, and macroeconomic conditions, we ask how well each LLM can predict households' future behavior in each task.

We evaluate thirteen LLMs on HouseholdBench. Eight are proprietary models from the GPT and Claude families, and five are open-weight models. We compare them to two statistical references: a *no-change baseline* carries forward the most recent value from the household's history; and a gradient-boosted tree (Chen & Guestrin, 2016) trained task-by-task on the same information that LLMs use.

Given the breadth of HouseholdBench, we can draw widely applicable conclusions about household behavior. Our main findings are as follows. First, we find that most LLMs outperform the no-change baseline, including on policy response tasks, at least for numerical and categorical outcomes. Second, widespread across tasks, there is a clear ranking of models. The gradient-boosted tree predicts best (19% better than the no-change baseline for numeric tasks). Proprietary LLMs from the Claude and GPT families approach this performance, but open-weight models perform worse. The performance of the best proprietary LLMs is notable, given that gradient-boosted trees achieve good performance on tabular prediction tasks (Holzmüller et al., 2024). Third, again widespread across tasks, performance after LLM knowledge cut-offs is equally good, which suggests that the performance of LLMs is not because of data contamination (e.g., public survey microdata may appear in pretraining corpora (Sarkar & Vafa, 2025; Ludwig et al., 2024)). Fourth, we document systematic biases of various kinds, with LLMs systematically overestimating household outcomes on some tasks, and underestimating them on others. Fifth, we uncover conditions under which a small, open-weight model can approach the frontier performance. In particular fine tuning Qwen3.5-4B (Qwen Team, 2026a), and aggregating across multiple predictions, greatly improves performance across a range of

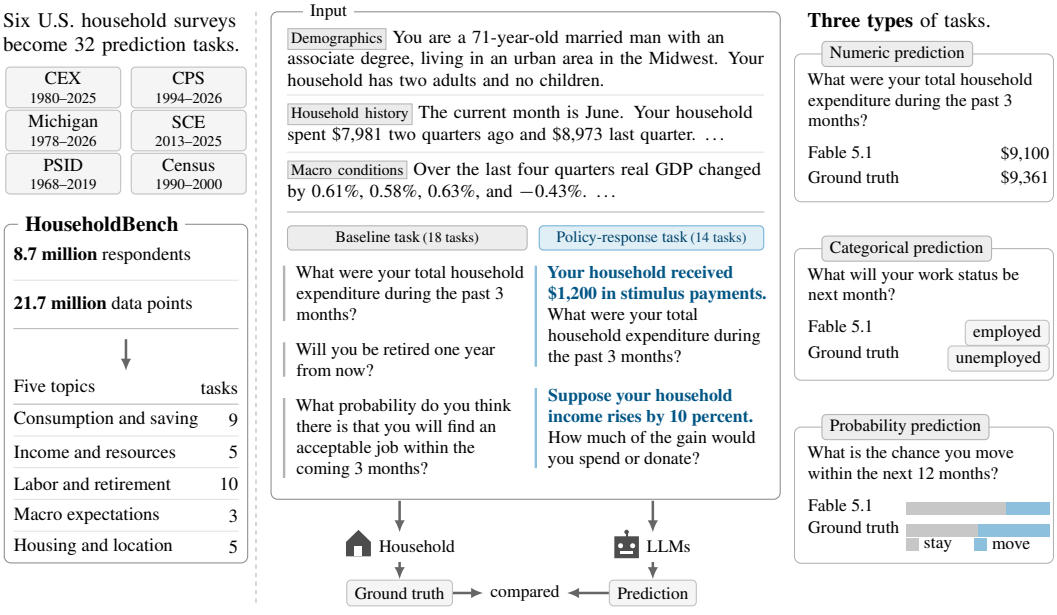


Figure 1: HouseholdBench turns six household surveys into 32 prediction tasks across five topics.

tasks, for instance reducing the numeric error by 24% and outperforming the strongest LLM. Sixth, these forecasting improvements generalise to policy response tasks, which are not used for fine tuning. This step is important because policymakers often contemplate new policies, for which there is no existing data for fine-tuning.

2 HOUSEHOLDBENCH

2.1 DATA

Our main data sources are six leading surveys of U.S. households that offer rich self-reported information on households' socio-demographic background, their economic behavior, and their preferences and beliefs about the future. We choose these surveys because they are publicly available, extremely well-documented and widely used in social science research. Except for the Census, all of them enable longitudinal linking of households or respondents over several waves. Thus, HouseholdBench is able to track households over time and focus on predicting *changes* in behavior, rather than on pure cross-sectional prediction.

Each survey provides high-quality information about a few narrow topics. The **Consumer Expenditure Survey** (CEX) collects detailed information on household consumption expenditure. The **Current Population Survey** (CPS) contains detailed information on employment, unemployment, hours worked and labor earnings. The University of Michigan's **Surveys of Consumers** (Michigan) focus on households' expectations and attitudes about their own economic situation and about the U.S. economy as a whole. The New York Fed's **Survey of Consumer Expectations** (SCE) specializes in consumer beliefs about the economy, including measures of subjective uncertainty, and it also provides a rich set of special modules eliciting consumer preferences directly. The **Panel Study of Income Dynamics** (PSID) has followed a set of U.S. individuals and families continuously since 1968, providing detailed information on income, employment, wealth, and other aspects of economic behavior. Finally, **U.S. Decennial Census** extracts (Census) provide data on demographics and residential mobility. For a detailed discussion of each source, including exact provenance and data cleaning, see Appendix B.

2.2 EVALUATION TASKS

We use micro-data from these surveys to build 32 separate evaluation tasks, grouped into five broad **topics**. They are (1) consumption and saving, (2) income dynamics and household resources, (3) labor supply, job search, and retirement, (4) macroeconomic expectations, and (5) housing and location. These topics cover important aspects of household behavior and expectations and correspond to different *blocks* within a model of household choice. Individual tasks within the topics were designed to zoom in on particular margins or decisions. See Table 1 for a summary of the full suite, and Appendix H for detailed information on each task, including full prompt examples.

Target Variables. Each task features one or several targets (variables to be predicted), which generally are direct, untransformed survey answers. Targets can be *numeric*, like total consumption expenditure in US Dollars, *categorical*, like employment status in a month, or *probabilities*, like beliefs about inflation over the next 12 months.

Baseline and Policy-Response Tasks. We also classify the tasks into two groups based on the information provided in the prompt.

- *Baseline tasks* measure whether models can predict household outcomes or beliefs using information already available about the household. The prompt gives socio-demographic background (sex, race, education, household composition, location), past economic behavior and outcomes, and the macroeconomic environment.[1]
- *Policy-response tasks* ask whether models can predict behavior or beliefs conditional on variation in the economic environment that is plausibly external to the household. That variation is a change in government policy, an observed economic shock, an information treatment, or a hypothetical

[1] All tasks include a common set of core macroeconomic variables. All tasks include the last four quarters of GDP growth, CPI inflation, unemployment and the Federal Funds rate, plus their average values over the previous five years. Some tasks include additional variables if relevant.

Table 1: The 32 HouseholdBench tasks by topic. Baseline tasks condition on the household's history and the macroeconomic environment. Policy-response tasks add a policy, shock, or scenario. Type: N numeric, C categorical, P probabilistic. Appendix H describes every task in full.

Baseline tasks			**Policy-response tasks**		
Task ID	**Prediction target**	**Type**	**Task ID**	**Prediction target**	**Type**
Consumption and saving					
cons_cex_categories	Spending in 12 categories	N	cons_cex_rebate01	Spending after the 2001 tax rebates	N
cons_cex_total	Total, nondurable, and durable spending	N	cons_cex_sspay	Daily spending around Social Security payments	N
cons_psid_wealth	Net worth in the next wave	N	cons_cex_stimulus08	Spending after the 2008 stimulus payments	N
cons_sce_growth	Year-on-year growth in monthly spending	N	cons_psid_jobloss	Food spending the year after a job loss	N
			cons_sce_shock	Spending share of $\pm$10% income changes	N
Income dynamics and household resources					
income_mich_finance	Expected 1-year family-income growth	N	income_cps_displace	Weekly earnings after job displacement	N
income_psid_earnings	Labor earnings 2, 4, and 10 years ahead	N	income_sce_policy	Stated effect of tax and benefit changes	C
income_sce_growth	Expected 1-year household-income growth	N			
Labor supply, job search, and retirement					
labor_cps_jobfind	Next-month status of the unemployed	C	labor_cps_displace	Labor-force status after job displacement	C
labor_cps_retire	Retirement within 12 months, aged 50–75	C	labor_cps_ui	Next-month status under unemployment-insurance durations	C
labor_cps_separation	Next-month status of the employed	C	labor_psid_addedworker	Partner hours and earnings after spousal job loss	N
labor_sce_offer	Acceptance of the first listed job offer	C	labor_sce_reswage	Reservation wage and preferred hours	N
labor_sce_risk	Job-loss, quit, and reemployment chances	P			
labor_sce_search	3- and 12-month job-finding probabilities	P			
Macroeconomic expectations					
macro_mich_outlook	Expected inflation at 1 and 5–10 years	N	macro_sce_revision	Expected inflation after an inflation surprise	N
macro_sce_uncertainty	Inflation and macro-outcome probabilities	P			
Housing and location					
house_census_move	Five-year residential-mobility status	C	house_sce_financing	Home price, down payment, 3 scenarios	N
house_psid_owner	Next-wave homeownership among renters	C	house_sce_lockin	Three-year moving probability with a portable mortgage rate	P
house_sce_move	One-year moving probability	P			

scenario the interviewer provides. Policy-response tasks offer a sharper test of whether models understand how households adjust their behavior when the economic environment changes.

Sample Construction. We filter the raw survey entries using a careful protocol that is harmonised across datasets. For each task we drop records with a missing target or missing recent history, apply survey-specific quality filters, and trim outliers within each period. Each data point is one household in one period. The prompt gives that household's demographics, its own recent history, and current macroeconomic conditions, and policy-response tasks add the policy or shock. The answer is the household's actual survey response. Appendix H gives the filters and a full prompt for each task.

Dataset Splits. After constructing the samples for each task, we divide the dataset into a few splits. We first hold out every data point released after 16th February 2026 as a *post-cutoff test set*. It postdates the knowledge cutoff of most LLMs we evaluate, so it gives valuable information about the contamination effect. Earlier samples are then grouped into calendar quarters, and we randomly split the quarters into training, validation, and pre-cutoff testing in an 80/10/10 ratio. Splitting on quarters means the test set contains macro conditions unseen in the training or validation set. From each test set, we evaluate at most 200 observations per task, sampled across calendar quarters in proportion to the number of eligible observations in each quarter.

2.3 Comparison with Existing Work

Table 2 compares HouseholdBench with related work evaluating LLM predictions of household economic outcomes and beliefs. Existing studies focus on specific topics at a time, and use at most four surveys. HouseholdBench is the only study covering a broad range of five topics (spending, income, jobs and retirement, macro expectations, housing) under a harmonised framework, while using as many as six survey datasets. This breadth is important for drawing robust and widely applicable conclusions about household behavior, which are not specific to a particular survey or outcome.

Table 2: Comparison between HouseholdBench and selected studies.

Study	Data	Years	Outcomes predicted					Design
			Spending	Income	Labor	Macro	Housing	Policy
Athey et al. (2026)	US labor panels (PSID, NLSY)	1979–2021			✓			
Cruz et al. (2024)	American Community Survey (ACS)	2018		✓	✓		✓	
Brynjolfsson et al. (2025)	PSID, UK valuation surveys	2017–2021		✓				
Jia et al. (2026)	Dutch household panel (LISS)	2023–2024			✓		✓	
Wu et al. (2025)	US consumer surveys (SCE, Nielsen)	2018–2023				✓		✓
Zarifhonarvar (2026)	US consumer surveys (SCE, Nielsen)	2019–2025				✓		✓
HouseholdBench (ours)	Six U.S. household surveys	**1968–2026**	✓	✓	✓	✓	✓	✓

Regarding policy response, there are some papers studying how households adjust to changing information or hypothetical policy treatments (Anesti et al., 2025; Wu et al., 2025; Park, 2025; Lin et al., 2026; Liu et al., 2026; Zarifhonarvar, 2026). However our policy-response tasks include not only changing information and hypotheticals, but also real-world episodes with policy changes and other related shocks, including: tax rebates and stimulus payments, unemployment-insurance variation, and household job loss. Asking whether LLMs can match actual household policy responses is critical—the answer tells us whether LLMs can be useful for realistic policy simulations.

3 Experimental Setting

3.1 Model Suite

We benchmark two types of language models, open-weight LLMs and proprietary frontier LLMs. We also include a no-change baseline and XGBoost as statistical reference models.

Open-Weight LLMs. We include open-weight LLMs with various model sizes, including Qwen3.5-4B (Qwen Team, 2026a), Qwen3.6-27B and Qwen3.6-35B-A3B (Qwen Team, 2026b), DeepSeek-V4-Flash and DeepSeek-V4-Pro (DeepSeek-AI, 2026). These span from small dense models to larger mixture-of-experts models.

Proprietary LLMs. We include two families of widely used frontier proprietary models. Within each family we include different capability tiers. For GPT, we include GPT-5.6 luna, GPT-5.6 terra, GPT-5.6 sol (OpenAI, 2026a), and GPT-6 astra (OpenAI, 2026b). For Claude, we include Claude Opus 4.8 (Anthropic, 2026b), Claude Sonnet 5 (Anthropic, 2026d), Claude Opus 5 (Anthropic, 2026c), and Fable 5.1 (Anthropic, 2026a). We run every model under its default inference settings. Details are included in Appendix G.

Statistical and Reference Models. Two statistical models provide reference points for evaluating LLM performance.

- *No-change baseline.* This uses the household's most recent observed value as its prediction. For the policy-response tasks it predicts zero response to the policy or shock. This simple forecast is a standard benchmark in economics (e.g. Meese & Rogoff, 1983; Atkeson & Ohanian, 2001).
- *XGBoost.* XGBoost is a gradient-boosted tree method and achieves good performance on tabular prediction tasks (Holzmüller et al., 2024). XGBoost (Chen & Guestrin, 2016) is trained, task-by-task, on the same household features the LLMs receive as text. We use the pre-tuned parameter settings of Holzmüller et al. (2024), which outperform XGBoost's default parameters. Appendix C gives the features and every parameter value.

3.2 Metrics

We use the following metrics for the three task types.

- *Numeric targets* use relative mean absolute error (RelMAE), the model's absolute error over the absolute error of the no-change baseline on the same samples (Hyndman & Koehler, 2006;

Rank	Model	Average rank within topic						Average score by target type		
		Consumption (# = 9)	Housing (# = 5)	Income (# = 5)	Labor (# = 10)	Macroeconomic (# = 3)	Average rank	RelMAE ↓ (# = 18)	Macro-F1 ↑ (# = 9)	TV ↓ (# = 5)
1	XGBoost	2.3 (0.6)	5.8 (1.0)	2.6 (0.8)	4.4 (0.7)	2.0 (0.8)	3.4 (0.4)	0.811 (0.011)	0.432 (0.014)	0.141 (0.004)
2	Fable 5.1	6.0 (0.5)	**4.4** (0.8)	**2.4** (0.9)	**4.2** (0.6)	**2.0** (0.9)	**3.8** (0.3)	**0.878** (0.013)	**0.413** (0.013)	**0.139** (0.005)
3	GPT-5.6 sol	5.3 (0.5)	5.2 (0.7)	6.2 (1.0)	4.8 (0.5)	6.0 (0.7)	5.5 (0.3)	0.910 (0.013)	0.399 (0.011)	**0.139** (0.005)
4	GPT-6 astra	**4.4** (0.4)	8.6 (0.7)	4.6 (0.7)	6.4 (0.4)	5.0 (0.8)	5.8 (0.3)	0.915 (0.014)	0.369 (0.006)	**0.139** (0.005)
5	Claude Opus 5	4.7 (0.4)	7.2 (0.7)	5.4 (1.0)	5.2 (0.6)	8.7 (1.0)	6.2 (0.3)	0.905 (0.014)	0.406 (0.013)	0.144 (0.005)
6	Claude Opus 4.8	5.4 (0.5)	6.4 (0.8)	6.8 (0.8)	5.4 (0.5)	8.7 (0.9)	6.5 (0.3)	0.919 (0.014)	0.402 (0.009)	0.146 (0.005)
7	GPT-5.6 terra	7.2 (0.6)	7.4 (0.8)	8.4 (0.9)	6.9 (0.5)	6.3 (0.8)	7.3 (0.3)	0.950 (0.013)	0.385 (0.011)	0.147 (0.005)
8	Claude Sonnet 5	8.3 (0.6)	7.4 (0.9)	6.8 (0.8)	8.8 (0.5)	9.0 (0.9)	8.1 (0.3)	0.934 (0.014)	0.361 (0.008)	0.148 (0.005)
9	Qwen3.6-27B	10.4 (0.5)	6.6 (1.0)	8.8 (0.9)	6.6 (0.5)	9.3 (1.0)	8.4 (0.4)	1.007 (0.019)	0.385 (0.011)	0.148 (0.005)
10	GPT-5.6 luna	9.0 (0.6)	5.8 (1.0)	10.6 (0.8)	7.8 (0.6)	9.0 (0.7)	8.4 (0.3)	1.001 (0.018)	0.381 (0.012)	0.143 (0.005)
11	DeepSeek-V4-Pro	9.6 (0.7)	10.8 (1.1)	7.6 (0.9)	8.7 (0.6)	7.3 (0.8)	8.8 (0.4)	1.002 (0.017)	0.375 (0.011)	0.151 (0.005)
12	DeepSeek-V4-Flash	10.4 (0.6)	11.8 (0.9)	10.0 (0.9)	7.8 (0.5)	8.7 (1.0)	9.7 (0.4)	0.995 (0.019)	0.363 (0.010)	0.151 (0.005)
13	No-change baseline	11.4 (0.5)	6.8 (0.8)	12.8 (0.5)	7.9 (0.4)	10.3 (0.4)	9.9 (0.2)	1.000 (0.000)	0.315 (0.003)	0.139 (0.005)
14	Qwen3.6-35B-A3B	10.4 (0.5)	11.2 (1.0)	12.2 (0.8)	9.1 (0.6)	13.0 (0.6)	11.2 (0.3)	0.995 (0.016)	0.365 (0.011)	0.174 (0.006)
15	Qwen3.5-4B	14.9 (0.3)	12.2 (1.0)	14.8 (0.1)	10.7 (0.5)	14.7 (0.2)	13.5 (0.2)	1.304 (0.033)	0.346 (0.012)	0.229 (0.008)

Table 3: Average rank by topic and average score by target type, pre-cutoff test split. Best LLM in bold, second best underlined. Ranks are taken over the models shown. Brackets report bootstrap SEs.

Hewamalage et al., 2023):

$$\text{RelMAE} = \frac{\sum_{i=1}^{n} |\hat{y}_i - y_i|}{\sum_{i=1}^{n} |\tilde{y}_{i,\text{no-change}} - y_i|}.$$

Here y_i denotes the household's answer, $\hat{y}_i$ the model's prediction, and $\tilde{y}_{i,\text{no-change}}$ that of the no-change baseline of Section 3.1, over the n samples of the task. A value below 1.0 beats the no-change baseline. The tasks of the benchmark come on different numerical scales, and RelMAE is a natural way to normalize across them.

- *Categorical targets* use macro-F_1, which computes an F_1 score for each outcome category and averages these scores with equal weight. This matters when some outcomes are much less common than others.
- *Probability targets* use mean total variation (TV) distance between the predicted probability vector $\hat{p}_i$ and the household's own p_i over the target's B bins:

$$\text{TV} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \sum_{b=1}^{B} |\hat{p}_{ib} - p_{ib}|.$$

  It is 0 when the two probability vectors agree perfectly, and 1 when they put their mass on disjoint bins, with no overlap at all.

To obtain a unified score for each model that represents its performance on HouseholdBench, we report the **average rank** across all 32 tasks. Lower average ranks indicate better performance. Since we only evaluate models on relatively small fractions of our full testing samples, we assess sampling variation in all our final results and measures using a standard i.i.d. bootstrapping procedure (10,000 bootstrap samples for each task), paired across models.

4 Results and Discussion

4.1 Main Results

Most LLMs outperform the no-change baseline on numerical and categorical prediction, but not probability prediction. Table 3 reports the pre-cutoff leaderboard, where we rank the models by their average rank over the five topics. Most LLMs are better than the no-change baseline, which

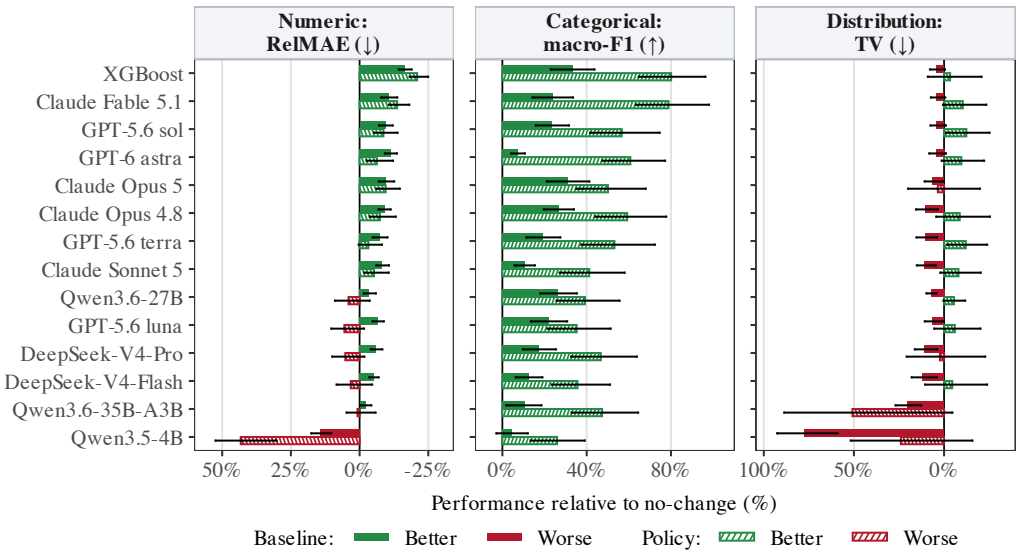


Figure 2: Model performance relative to the no-change benchmark on baseline and policy-response tasks. Solid bars report mean scores for baseline tasks, while hatched bars report mean scores for policy-response tasks. Whiskers are pointwise 95% basic IID bootstrap intervals.

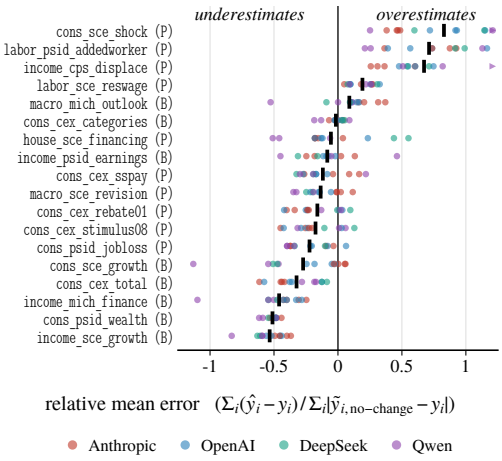
ranks 13th. Among the LLMs, Fable 5.1 achieves the strongest overall LLM performance, with a RelMAE of 0.878 and a macro-F1 of 0.413. The proprietary models also rank above the open-weight models: the best open-weight model (Qwen3.6-27B) only ranks 9th.

For probability tasks, no model improves on the no-change baseline. This is due to models underestimating the persistence of beliefs. On the four baseline probability tasks, 26.9% of target probability vectors exactly repeat the previous household report. Appendix Figure D1 separates performance on changed and unchanged reports. For GPT-5.6 sol and Claude Opus 5, gains on changed reports are more than offset by errors on unchanged reports.

Leading LLMs also improve over the no-change prediction on policy-response tasks. Figure 2 considers performance improvements over the no-change model separately for baseline and policy-response tasks. Fable 5.1, Claude Opus 5, and GPT-5.6 sol reduce numeric error by 13.6%, 9.5%, and 8.7% on policy-response tasks, even without task-specific training. Their baseline-task gains are 10.5%, 9.6%, and 9.4%, respectively. These results illustrate the flexibility of LLMs in predicting household responses to policy changes.

Widespread across tasks, LLMs do not outperform XGBoost models trained per task, but Fable 5.1 is close; other proprietary models perform well, and open-weight models lag. XGBoost ranks first on the leaderboard with an average rank of 3.4, against 3.8 for Fable 5.1, and their bootstrap intervals overlap. The two are closest on the categorical and probability tasks, where their intervals overlap: 0.413 against 0.432 in macro-F1, and 0.139 against 0.141 in TV, with Fable 5.1 marginally ahead. XGBoost keeps its lead on the numeric tasks, 0.811 against 0.878. This suggests that the strongest proprietary LLM, without any task-specific training, can predict household behavior nearly as accurately as an XGBoost model trained on each task. Open-weight models perform less well, occupying the bottom ranks of the leaderboard. These results are widespread across tasks.

LLMs' performance on HouseholdBench does not drop after or close to their knowledge cutoff. Panels (a) and (b) of Appendix Figure D2 compare model performance on the matched pre-cutoff and post-cutoff task samples. Across all tasks, we do not see large systematic changes in performance for tasks with numerical targets, suggesting limited look-ahead bias from pre-training. This may reflect the lower likelihood that pre-training data contain individual household outcomes. Appendix Figure D3 performs an additional analysis, looking at average performance over time for a subset of tasks where we have a long sample period. Again, we do not see any systematic degradation of performance as we approach knowledge cutoffs either.



Task (majority class)	Ground truth	Claude Fable 5.1	GPT-6 astra	GPT-5.6 sol	XGBoost
labor_cps_retire (B) (*not retired*)	96%	**100%**	**100%**	96%	**98%**
labor_cps_separation (B) (*still employed*)	94%	**100%**	**100%**	**100%**	**100%**
house_psid_owner (B) (*still renting*)	89%	**98%**	**98%**	**96%**	**97%**
labor_sce_offer (B) (*rejected the offer*)	69%	**76%**	**76%**	**72%**	**80%**
labor_cps_ui (P) (*still unemployed*)	69%	**94%**	**100%**	**100%**	**84%**
house_census_move (B) (*did not move*)	67%	**83%**	**100%**	**88%**	**94%**
labor_cps_displace (P) (*re-employed*)	62%	**80%**	62%	22%	**87%**
income_sce_policy (P) (*average of 4 targets*)	55%	**57%**	**64%**	52%	**72%**
labor_cps_jobfind (B) (*still unemployed*)	50%	**94%**	**100%**	**96%**	**76%**

Share of majority-class answers.

Figure 3: LLMs exhibit systematic bias in predicting household behavior. *Left.* LLMs consistently overestimate household behavior on some tasks and underestimate it on others. *Right.* LLMs overpredict the majority class on categorical tasks. Bold means that a model predicts the majority class more often than the ground truth. (B) and (P) denote baseline and policy-response tasks.

4.2 LLMs' BIAS ON PREDICTING HOUSEHOLD BEHAVIOR

Although leading frontier LLMs predict household behavior well overall, they exhibit systematic biases.

LLMs systematically overestimate household outcomes on some tasks and underestimate them on others. Figure 3 (left) reports each LLM's relative mean error on the 18 numeric tasks, which is RelMAE with the signed error $\hat{y}_i - y_i$ in the numerator. Overestimation is largest on policy-response tasks, where the LLMs predict stronger responses than households report. For example, `cons_sce_shock` asks what share of a hypothetical permanent 10% income gain the household would spend or donate. The median household answers 5%, while the three frontier LLMs (Fable 5.1, GPT-5.6 sol, and GPT-6 astra) predict 30% to 40%.

On categorical prediction tasks, LLMs predict the most common outcome of a household decision more often than households choose it. Figure 3 (right) shows that the three frontier LLMs over-predict the majority class, usually the status quo. The gap is largest on `labor_cps_jobfind` and `labor_cps_ui`, which predict the next-month labor-force status of unemployed workers. On `labor_cps_jobfind`, only half of the workers remain unemployed, but the LLMs predict this for 94% to 100% of them. The LLMs also miss every worker who leaves the labor force on both tasks. Among the three models, GPT-6 astra shows this pattern most strongly: it predicts the majority class for every data point on 5 of the 9 categorical tasks. These biases suggest that more progress is needed before LLMs are fully reliable as simulations of household behavior.

5 IMPROVING LLMs' PREDICTIVE POWER

We have found that open-weight models lag proprietary models for predicting household behavior. This section presents a method that enables a small open-weight model to match leading proprietary LLMs. In particular, we show that combining supervised fine-tuning on household behavior data with aggregation over draws can close the gap.

5.1 FINE-TUNING SETUP

To test the hypothesis that fine-tuning an LLM on household training data could improve its performance on HouseholdBench, we use supervised fine-tuning (SFT), which is commonly used to adapt a general-purpose language model to a specific domain (Wei et al., 2022; Li et al., 2025). We fine-tune models on the training split of the 18 baseline tasks. For each task, we sample at most 6,000 data

	Model	K	Baseline tasks (seen in fine-tuning) RelMAE ↓ (# = 8)	Macro-F1 ↑ (# = 6)	TV ↓ (# = 4)	Policy-response tasks (held out) RelMAE ↓ (# = 10)	Macro-F1 ↑ (# = 3)	TV ↓ (# = 1)
	XGBoost	1	0.84	0.46	0.15	0.79	0.36	0.12
Reference	GPT-6 astra	1	0.89	0.39	**0.15**	0.94	0.32	**0.11**
	Claude Opus 5	1	0.90	**0.47**	**0.15**	0.91	0.28	0.13
	Fable 5.1	1	0.89	0.44	**0.15**	**0.86**	**0.35**	**0.11**
SFT	Qwen3.5-4B	1	1.14	0.39	0.25	1.43	0.25	0.15
	+ SFT	1	1.00 (−0.14)	0.45 (+0.06)	0.16 (−0.09)	1.30 (−0.14)	0.33 (+0.08)	0.17 (+0.01)
	Qwen3.6-27B	1	0.96	0.45	0.16	1.04	0.26	0.12
	+ SFT	1	1.00 (+0.03)	0.46 (+0.01)	0.16 (+0.01)	1.18 (+0.14)	0.30 (+0.04)	0.15 (+0.04)
SFT + aggregation	Qwen3.5-4B	16	1.11	0.38	0.24	1.31	0.24	0.15
	+ SFT	16	**0.87** (−0.24)	0.42 (+0.05)	**0.15** (−0.09)	1.13 (−0.18)	0.31 (+0.06)	0.16 (+0.01)
	Qwen3.6-27B	16	0.95	0.43	**0.15**	1.01	0.25	0.12
	+ SFT	16	**0.87** (−0.08)	0.44 (+0.01)	**0.15** (−0.01)	1.03 (+0.02)	0.27 (+0.02)	0.14 (+0.02)

Table 4: Results of SFT. The K column is the number of aggregated predictions. Best among the language models in bold, second best underlined. XGBoost is refitted on each task it is scored on, so its policy-response scores are not held out.

points from its training split, stratified by quarters. The training set contains 100,702 data points. We fine-tune two widely-used open-weight LLMs, Qwen3.5-4B and Qwen3.6-27B, with LoRA (Hu et al., 2022) for one epoch.

We do not fine tune on the policy response tasks, and instead treat them as a hold out sample, which allows us to test whether fine-tuning on baseline tasks generalizes to policy-response tasks. This step is important because data for policy response tasks is often sparse. Moreover often policymakers contemplate new policies, for which there is no existing data—say, a new form of unemployment insurance. One would like the LLM to make good predictions even for these kinds of policies.

5.2 Fine-Tuning Results

We fine-tune the two models with the data and configuration described above. We also aggregate $K = 16$ draws per model by averaging numerical and probability predictions and taking majority votes for categorical predictions. Table 4 shows the main results.

First, SFT improves Qwen3.5-4B on all three types of prediction tasks, and the gains generalize to the unseen policy-response tasks. On the baseline tasks, SFT improves Qwen3.5-4B's RelMAE by 13%, macro-F_1 by 15%, and TV by 34%. The gains generalize to the unseen policy-response tasks, where RelMAE improves by 10% and macro-F_1 by 30%. Qwen3.6-27B's results are more mixed: SFT improves its macro-F_1 but worsens its RelMAE and TV. This suggests that a model that already performs well on HouseholdBench may not always benefit from additional fine-tuning.

Second, **the fine-tuned models with prediction aggregation can beat the best LLMs on numeric tasks.** We find that the fine-tuned models perform better with prediction aggregation. On the numeric baseline tasks, aggregating $K = 16$ draws lowers the backbone models' RelMAE by only 3% (Qwen3.5-4B) and 1% (Qwen3.6-27B), but the fine-tuned models' by 10% to 13%. With aggregation, the fine-tuned Qwen3.5-4B reaches the best performance on numerical and probability prediction tasks on the baseline tasks, outperforming Fable 5.1. We also note that aggregating for proprietary models does not improve their performance (Appendix F.1). In addition to SFT, we explore fine-tuning with a number-token loss (Zausinger et al., 2025) and report the results in Appendix E.

6 Conclusion

We introduce HouseholdBench, a benchmark that evaluates LLM predictions of household outcomes and beliefs across five areas of economic behavior. Our evaluation shows that LLMs still do not outperform XGBoost overall, although Fable 5.1 comes close. With fine-tuning and aggregation of 16 predictions, small open-weight models outperform leading frontier LLMs evaluated on numeric baseline tasks. By testing these capabilities against household survey data, we provide a comprehensive way to assess LLMs as predictors of household behavior.

Ethics Statement

We build HouseholdBench only from existing public survey micro-data, which the providers release without respondent identities (Appendix B). The study involves no contact with respondents.

AI Use Statement

In this work, we used generative AI tools to support data analysis. We have not used generative AI tools for any other task that requires disclosure. Additionally, we used generative AI tools to aid and polish the writing of this paper. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Reproducibility Statement

Every data point in HouseholdBench comes from a publicly released survey. Appendix B documents how we download and clean each source. For each of the 32 tasks, Appendix H gives the target, predictors, sample construction, data splits, and an example prompt. Appendix G lists the inference settings and fine-tuning hyperparameters of the models, and Appendix C describes the gradient-boosted tree baselines.

References

Felix Aidala, Andrew F. Haughwout, Ben Hyman, Jason Somerville, and Wilbert van der Klaauw. Mortgage rate lock-in and homeowners' moving plans. Liberty Street Economics, Federal Reserve Bank of New York, May 2024. URL https://libertystreeteconomics.newyorkfed.org/2024/05/mortgage-rate-lock-in-and-homeowners-moving-plans/.

Nikoleta Anesti, Edward Hill, and Andreas Joseph. Inflation attitudes of large language models. *arXiv preprint arXiv:2512.14306*, 2025.

Anthropic. Introducing Claude Fable 5.1 and Claude Mythos 5.1. Anthropic Blog, 2026a. URL https://www.anthropic.com/claude-fable-and-mythos-5-1.

Anthropic. Introducing Claude Opus 4.8. Anthropic Blog, 2026b. URL https://www.anthropic.com/news/claude-opus-4-8. May 28, 2026.

Anthropic. Introducing Claude Opus 5. Anthropic Blog, 2026c. URL https://www.anthropic.com/news/claude-opus-5. July 24, 2026.

Anthropic. Introducing Claude Sonnet 5. Anthropic Blog, 2026d. URL https://www.anthropic.com/news/claude-sonnet-5. June 30, 2026.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua Gubler, Christopher Michael Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *CoRR*, abs/2209.06899, 2022. doi: 10.48550/ARXIV.2209.06899. URL https://doi.org/10.48550/arXiv.2209.06899.

Olivier Armantier, Giorgio Topa, Wilbert van der Klaauw, and Basit Zafar. An overview of the survey of consumer expectations. *Economic Policy Review*, 23(2):51–72, 2017. URL https://www.newyorkfed.org/research/epr/2017/epr_2017_overview-of-sce_armantier.

Olivier Armantier, Leo Goldman, Gizem Koşar, Giorgio Topa, Wilbert van der Klaauw, and John C. Williams. What are consumers' inflation expectations telling us today? Liberty Street Economics, Federal Reserve Bank of New York, February 2022. URL https://libertystreeteconomics.newyorkfed.org/2022/02/what-are-consumers-inflation-expectations-telling-us-today/.

Ashwini Ashokkumar, Luke Hewitt, Isaias Ghezae, and Robb Willer. Large language models can predict the results of social science experiments. *Nature*, pp. 1–8, 2026.

Susan Athey, Herman Brunborg, Tianyu Du, Ayush Kanodia, and Keyon Vafa. LABOR-LLM: Language-Based Occupational Representations with Large Language Models, 2026. URL https://arxiv.org/abs/2406.17972v4. arXiv:2406.17972v4.

Andrew Atkeson and Lee E. Ohanian. Are Phillips curves useful for forecasting inflation? *Federal Reserve Bank of Minneapolis Quarterly Review*, 25(1):2–11, 2001. URL https://www.minneapolisfed.org/research/quarterly-review/are-phillips-curves-useful-for-forecasting-inflation.

Adrien Auclert. Monetary policy and the redistribution channel. *American Economic Review*, 109(6):2333–2367, 2019. doi: 10.1257/aer.20160137. URL https://www.aeaweb.org/articles?id=10.1257/aer.20160137.

Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K. Eckstein, Noémi Élteto, Thomas L. Griffiths, Susanne Haridi, Akshay K. Jagadish, Ji-An Li, Alexander Kipnis, Sreejan Kumar, Tobias Ludwig, Marvin Mathony, Marcelo G. Mattar, Alireza Modirshanechi, Surabhi S. Nath, Joshua C. Peterson, Milena Rmus, Evan M. Russek, Tankred Saanum, Johannes A. Schubert, Luca M. Schulze Buschoff, Nishad Singhi, Xin Sui, Mirko Thalmann, Fabian J. Theis, Vuong Truong, Vishaal Udandarao, Konstantinos Voudouris, Robert C. Wilson, Kristin Witte, Shuchen Wu, Dirk U. Wulff, Huadong Xiong, and Eric Schulz. A foundation model to predict and capture human cognition. *Nat.*, 644 (8078):1002–1009, 2025. doi: 10.1038/S41586-025-09215-4. URL https://doi.org/10.1038/s41586-025-09215-4.

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.

James Brand, Ayelet Israeli, and Donald Ngwe. Using LLMs for Market Research. *Available at SSRN 4395751*, 2026. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4395751. April 30, 2026 revision; first circulated in 2023.

Erik Brynjolfsson, José Ramón Enríquez, Sophia Kazinnik, and David Nguyen. Augmenting survey data with generative ai: An application to economic research. *Available at SSRN 6343598*, 2025.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. Association for Computing Machinery, 2016. doi: 10.1145/2939672.2939785.

André F. Cruz, Moritz Hardt, and Celestine Mendler-Dünner. Evaluating language models as risk scores. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/b0a4b3e384b4554e65a47ad1f6b0310a-Abstract-Datasets_and_Benchmarks_Track.html.

Davis Daumler, Esther Friedman, and Fabian T. Pfeffer. PSID-SHELF user guide and codebook, 1968–2021, beta release. Technical Report PSID-SHELF Data Documentation 2025-01, Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI, 2025. URL https://doi.org/10.7302/25205.

DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence. *CoRR*, abs/2606.19348, 2026. doi: 10.48550/ARXIV.2606.19348. URL https://doi.org/10.48550/arXiv.2606.19348.

Henry S. Farber, Jesse Rothstein, and Robert G. Valletta. The effect of extended unemployment insurance benefits: Evidence from the 2012–2013 phase-out. *American Economic Review*, 105 (5):171–176, 2015. doi: 10.1257/aer.p20151088. URL https://doi.org/10.1257/aer.p20151088.

Federal Housing Finance Agency. House price index, 2026. URL https://www.fhfa.gov/data/house-price-index.

Federal Reserve Bank of New York. Center for microeconomic data: Data bank, 2026a. URL https://www.newyorkfed.org/microeconomics/databank.html.

Federal Reserve Bank of New York. Survey of consumer expectations: Frequently asked questions, 2026b. URL https://www.newyorkfed.org/microeconomics/sce/sce-faq.

Federal Reserve Bank of St. Louis. FRED-MD and FRED-QD databases, 2026. URL https://www.stlouisfed.org/research/economists/mccracken/fred-databases. HouseholdBench uses the July 2026 FRED-QD vintage.

Anastassia Fedyk, Ali Kakhbod, Peiyao Li, and Ulrike Malmendier. AI and Perception Biases in Investments: An Experimental Study. *Available at SSRN 4787249*, 2024. doi: 10.2139/ssrn.4787249. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4787249. Revised December 12, 2025.

Apostolos Filippas, John J. Horton, and Benjamin S. Manning. Large language models as simulated economic agents: What can we learn from homo silicus? In Dirk Bergemann, Robert Kleinberg, and Daniela Sabán (eds.), *Proceedings of the 25th ACM Conference on Economics and Computation, EC 2024, New Haven, CT, USA, July 8-11, 2024*, pp. 614–615. ACM, 2024. doi: 10.1145/3670865.3673513. URL https://doi.org/10.1145/3670865.3673513.

Sarah Flood, Miriam King, Renae Rodgers, Steven Ruggles, J. Robert Warren, Daniel Backman, Etienne Breton, Grace Cooper, Julia A. Rivera Drew, Stephanie Richards, David Van Riper, and Kari C. W. Williams. IPUMS CPS: Version 13.0 [dataset], 2025. URL https://doi.org/10.18128/D030.V13.0.

Andreas Fuster and Basit Zafar. The sensitivity of housing demand to financing conditions: Evidence from a survey. *American Economic Journal: Economic Policy*, 13(1):231–265, 2021. doi: 10.1257/pol.20150337. URL https://doi.org/10.1257/pol.20150337.

Andreas Fuster and Basit Zafar. Replication data for: The sensitivity of housing demand to financing conditions, 2022. Version V1, distributed October 15, 2022.

Wayne Gao, Sukjin Han, and Annie Liang. How well do llms predict human behavior? A measure of their pretrained knowledge. *CoRR*, abs/2601.12343, 2026. doi: 10.48550/ARXIV.2601.12343. URL https://doi.org/10.48550/arXiv.2601.12343.

Rubén Garzón, Pauline Baron, Vincent Grari, Jonne Kamphorst, Michael Bernstein, and Marcin Detyniecki. From demographics to survey anchors: Evaluating LLM agents for modeling retirement attitudes. *CoRR*, abs/2605.16303, 2026. doi: 10.48550/ARXIV.2605.16303. URL https://doi.org/10.48550/arXiv.2605.16303.

Hansika Hewamalage, Klaus Ackermann, and Christoph Bergmeir. Forecast evaluation for data scientists: common pitfalls and best practices. *Data Mining and Knowledge Discovery*, 37(2): 788–832, 2023.

David Holzmüller, Léo Grinsztajn, and Ingo Steinwart. Better by default: Strong pre-tuned MLPs and boosted trees on tabular data. In *Advances in Neural Information Processing Systems*, volume 37, pp. 26577–26658, 2024. doi: 10.52202/079017-0837.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Jin Huang, Yutong Xie, Wanli Song, Xingjian Zhang, Walter Yuan, Matthew O. Jackson, and Qiaozhu Mei. Behaviorbench: Benchmarking foundation models for behavioral science tasks. *CoRR*, abs/2606.24162, 2026. doi: 10.48550/ARXIV.2606.24162. URL https://doi.org/10.48550/arXiv.2606.24162.

Rob J Hyndman and Anne B Koehler. Another look at measures of forecast accuracy. *International journal of forecasting*, 22(4):679–688, 2006.

Inter-university Consortium for Political and Social Research. Consumer expenditure survey series, 2026. URL https://www.icpsr.umich.edu/web/ICPSR/series/20.

IPUMS CPS. Displaced worker supplement sample notes, 2026. URL https://cps.ipums.org/cps/dw_sample_notes.shtml.

IPUMS USA. Description of samples, 2026. URL https://usa.ipums.org/usa/sampdesc.shtml.

Mumin Jia, Yilin Chen, Divya Sharma, and Jairo Diaz Rodriguez. When can digital personas reliably approximate human survey findings? *CoRR*, abs/2605.10659, 2026. doi: 10.48550/ARXIV.2605.10659. URL https://doi.org/10.48550/arXiv.2605.10659.

David S. Johnson, Jonathan A. Parker, and Nicholas S. Souleles. Household expenditure and the income tax rebates of 2001. *American Economic Review*, 96(5):1589–1610, 2006. doi: 10.1257/aer.96.5.1589. URL https://doi.org/10.1257/aer.96.5.1589.

Greg Kaplan, Benjamin Moll, and Giovanni L. Violante. Monetary policy according to HANK. *American Economic Review*, 108(3):697–743, 2018. doi: 10.1257/aer.20160042. URL https://www.aeaweb.org/articles?id=10.1257/aer.20160042.

Seth Karten, Wenzhe Li, Zihan Ding, Samuel Kleiner, Yu Bai, and Chi Jin. LLM economist: Large population models and mechanism design in multi-agent generative simulacra. *arXiv preprint arXiv:2507.15815*, 2025. URL https://arxiv.org/abs/2507.15815.

Leonard Kinzinger and Jochen Hartmann. Synthetic personalities: How well can llms mimic individual respondents using socio-economic microdata? *arXiv preprint arXiv:2606.04592*, 2026.

Akaash Kolluri, Shengguang Wu, Joon Sung Park, and Michael S. Bernstein. Finetuning llms for human behavior prediction in social science experiments. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 30096–30111. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.EMNLP-MAIN.1530. URL https://doi.org/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, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.829. URL https://aclanthology.org/2024.acl-long.829/.

Sihang Li, Jin Huang, Jiaxi Zhuang, Yaorui Shi, Xiaochen Cai, Mingjun Xu, Xiang Wang, Linfeng Zhang, Guolin Ke, and Hengxing Cai. Scilitllm: How to adapt llms for scientific literature understanding. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=8dzKkeWUUb.

Yumiao Li, Peixin Liu, Donglin Di, Chen Li, and Runhuan Feng. Can Large Language Models Anticipate Behavioral Responses to Social Policies? A Case of Pension Enrollment Prediction among China's Flexible Workers. *arXiv preprint arXiv:2609.05189*, 2026. URL https://arxiv.org/abs/2609.05189v1.

Jianhao Lin, Lexuan Sun, and Yixin Yan. Simulating Macroeconomic Expectations in Survey Experiments with LLM-based Economic Agents. *arXiv preprint arXiv:2505.17648*, 2026. URL https://arxiv.org/abs/2505.17648v5. June 2026 revision (version 5); first posted May 2025.

Guangya Liu, Cheng Wang, Jiangtong Li, Huafei Wu, and Changjun Jiang. Do LLM Agents Really Mimic Humans? Diagnosing and Aligning Microeconomic Behaviors in Macro-ABMs. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 36102–36122. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1799. URL https://aclanthology.org/2026.findings-acl.1799/.

Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. Large language models: An applied econometric framework. *Annual Review of Economics*, 18, 2024.

Michael W. McCracken and Serena Ng. FRED-QD: A quarterly database for macroeconomic research. *Federal Reserve Bank of St. Louis Review*, 103(1):1–44, 2021. doi: 10.20955/r.103.1-44.

Richard A. Meese and Kenneth Rogoff. Empirical exchange rate models of the seventies: Do they fit out of sample? *Journal of International Economics*, 14(1-2):3–24, 1983. doi: 10.1016/0022-1996(83)90017-X.

Keiichi Namikoshi, Alex Filipowicz, David A. Shamma, Rumen Iliev, Candice L. Hogan, and Nikos Aréchiga. Using LLMs to Model the Beliefs and Preferences of Targeted Populations. *arXiv preprint arXiv:2403.20252*, 2024. URL https://arxiv.org/abs/2403.20252v1.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition. OpenAI Blog, 2026a. URL https://openai.com/index/gpt-5-6/.

OpenAI. GPT-6 Astra: A new generation of intelligence. OpenAI Blog, 2026b. URL https://openai.com/index/gpt-6-astra/.

Joon Sung Park, Carolyn Q Zou, Jonne Kamphorst, Niles Egan, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Percy Liang, Robb Willer, et al. Llm agents grounded in self-reports enable general-purpose simulation of individuals. *arXiv preprint arXiv:2411.10109*, 2024.

Sungho Park. ChatGPT: Augmenting or Replacing Intelligence in Economic Surveys? Discussion Paper 56/2025, Fondazione GRINS, 2025. URL https://grins.it/sites/default/files/2025-12/Chatgpt__augmenting_or_replacing_intelligence_in_economic_surveys_.pdf.

Jonathan A. Parker, Nicholas S. Souleles, David S. Johnson, and Robert McClelland. Consumer spending and the economic stimulus payments of 2008. *American Economic Review*, 103(6):2530–2553, October 2013. doi: 10.1257/aer.103.6.2530.

Tianyi Peng, Melanie Brucks, George Gui, Daniel J Merlau, Grace Jiarui Fan, Malek Ben Sliman, Eric J Johnson, Abdullah Althenayyan, Silvia Bellezza, Dante Donati, et al. Digital twins are funhouse mirrors: Five systematic distortions. *Science Advances*, 12(36):eaeh8260, 2026.

Fabian T. Pfeffer, Davis Daumler, and Esther Friedman. PSID-SHELF, 1968–2021: The PSID's social, health, and economic longitudinal file, beta release [dataset], 2025. URL https://doi.org/10.3886/E194322V2.

Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, Chen Gao, Fengli Xu, Fang Zhang, Ke Rong, Jun Su, and Yong Li. AgentSociety: Large-scale simulation of LLM-driven generative agents advances understanding of human behaviors and society. *arXiv preprint arXiv:2502.08691*, 2025. URL https://arxiv.org/abs/2502.08691.

Qwen Team. Qwen3.5. Qwen Team blog, 2026a. URL https://qwen.ai/blog?id=qwen3.5. February 15, 2026.

Qwen Team. Qwen3.6. Qwen Team blog, 2026b. URL https://qwen.ai/blog?id=qwen3.6-35b-a3b. April 14, 2026.

Steven Ruggles, Sarah Flood, Matthew Sobek, Daniel Backman, Grace Cooper, Julia A. Rivera Drew, Stephanie Richards, Renae Rodgers, Jonathan Schroeder, and Kari C. W. Williams. IPUMS USA: Version 16.0 [dataset], 2025. URL https://doi.org/10.18128/D010.V16.0.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29971–30004. PMLR, 2023. URL https://proceedings.mlr.press/v202/santurkar23a.html.

Suproteem K Sarkar and Keyon Vafa. Lookahead bias in pretrained language models. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*, 2025. URL https://openreview.net/forum?id=s6WkKKBgw3.

Melvin Stephens, Jr. Worker displacement and the added worker effect. *Journal of Labor Economics*, 20(3):504–537, 2002. doi: 10.1086/339615. URL https://doi.org/10.1086/339615.

Melvin Stephens, Jr. "3rd of tha Month": Do Social Security Recipients Smooth Consumption Between Checks? *American Economic Review*, 93(1):406–422, 2003. doi: 10.1257/000282803321455386. URL https://doi.org/10.1257/000282803321455386.

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 Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 21147–21170. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.ACL-LONG.1028. URL https://doi.org/10.18653/v1/2025.acl-long.1028.

University of Michigan, Survey Research Center. Surveys of consumers: Frequently asked questions, 2026a. URL https://data.sca.isr.umich.edu/faq.php.

University of Michigan, Survey Research Center. Surveys of consumers. Public microdata and documentation archive, 2026b. URL https://data.sca.isr.umich.edu/.

U.S. Bureau of Economic Analysis. Personal income by state, 2026. URL https://www.bea.gov/data/income-saving/personal-income-by-state.

U.S. Bureau of Labor Statistics. *Handbook of Methods: Consumer Expenditures and Income*, 2022. URL https://www.bls.gov/opub/hom/cex/.

U.S. Bureau of Labor Statistics. Consumer expenditure surveys public use microdata. Consumer Expenditure Surveys program, 2026a. URL https://www.bls.gov/cex/pumd_data.htm. Data and documentation archive.

U.S. Bureau of Labor Statistics. Consumer price index data, 2026b. URL https://www.bls.gov/cpi/data.htm.

U.S. Bureau of Labor Statistics. Current population survey: Design, 2026c. URL https://www.bls.gov/opub/hom/cps/design.htm.

U.S. Bureau of Labor Statistics. Current population survey: History, 2026d. URL https://www.bls.gov/opub/hom/cps/history.htm.

U.S. Bureau of Labor Statistics. Local area unemployment statistics, 2026e. URL https://www.bls.gov/lau/.

U.S. Census Bureau. *Current Population Survey, January 2024: Displaced Worker Supplement File Technical Documentation*, 2024a. URL https://www2.census.gov/programs-surveys/cps/techdocs/cpsjan24.pdf.

U.S. Census Bureau. Latest data releases: 2024, 2024b. URL https://www.census.gov/data/what-is-data-census-gov/latest-releases.2024.html.

U.S. Census Bureau. About the 2000 census, 2026a. URL https://www.census.gov/programs-surveys/decennial-census/decade/2000/about-2000.html.

U.S. Census Bureau. Consumer expenditure surveys, 2026b. URL https://www.census.gov/programs-surveys/ce.html.

U.S. Census Bureau. Current population survey: Frequently asked questions, 2026c. URL https://www.census.gov/programs-surveys/cps/about/faqs.html.

U.S. Census Bureau. Decennial census data, 2026d. URL https://www.census.gov/programs-surveys/decennial-census/data.html.

Hexi Wang, Yujia Zhou, Bangde Du, Weihang Su, Xinyuan Cao, Qingyi Pan, Qingyao Ai, Yueyue Wu, Min Zhang, and Yiqun Liu. Mitigating Identity Essentialism in LLM Agents with Longitudinal Life Trajectories. *arXiv preprint arXiv:2608.19621*, 2026. URL https://arxiv.org/abs/2608.19621v2.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=gEZrGCozdqR.

Jing Cynthia Wu, Jin Xi, and Shihan Xie. Llm survey framework: Coverage, reasoning, dynamics, identification. Technical report, National Bureau of Economic Research, 2025.

Yutong Xie, Zhuoheng Li, Xiyuan Wang, Yijun Pan, Qijia Liu, Xingzhi Cui, Kuang-Yu Lo, Ruoyi Gao, Xingjian Zhang, Jin Huang, Walter Yuan, Matthew O. Jackson, and Qiaozhu Mei. Be.fm: Open foundation models for human behavior. *CoRR*, abs/2505.23058, 2025. doi: 10.48550/ARXIV.2505.23058. URL https://doi.org/10.48550/arXiv.2505.23058.

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

Jonas Zausinger, Lars Pennig, Anamarija Kozina, Sean Sdahl, Julian Sikora, Adrian Dendorfer, Timofey Kuznetsov, Mohamad Hagog, Nina Wiedemann, Kacper Chlodny, Vincent Limbach, Anna Ketteler, Thorben Prein, Vishwa Mohan Singh, Michael M. Danziger, and Jannis Born. Regress, don't guess: A regression-like loss on number tokens for language models. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/zausinger25a.html.

A Related Work

In the main text and Table 2, we focus on a few representative studies from the rapidly growing literature that evaluates or fine-tunes LLMs using household survey microdata, but there are others that deserve discussion. Within the topics covered by HouseholdBench, further evaluations use PSID data to predict wages and homeownership (Gao et al., 2026), SOEP data to reconstruct held-out survey answers (Kinzinger & Hartmann, 2026), and SHARE data to predict retirement-related attitudes (Garzón et al., 2026). Other studies fine-tune LLMs to predict pension enrollment in Chinese household surveys (Li et al., 2026), or combine individual model adaptation with longitudinal life histories to predict survey responses (Wang et al., 2026). The expectations literature includes comparisons with UK inflation beliefs (Anesti et al., 2025), Italian household expectations and responses to information (Park, 2025), and human survey responses to hypothetical macroeconomic shocks and housing information (Lin et al., 2026). Liu et al. (2026) compare LLM spending expectations and job-acceptance intentions with Survey of Consumer Expectations responses before using the agents in an economic simulation. Related survey-based evaluations also examine decisions outside our specific tasks, including investment perceptions (Fedyk et al., 2024), product choices and willingness to pay (Brand et al., 2026), and electric-vehicle preferences and responses to information (Namikoshi et al., 2024). These studies provide valuable evidence within particular economic domains, but, as we argued in the main text, HouseholdBench provides a more systematic and comprehensive evaluation of LLMs across a broad set of economic outcomes and settings. Also, policy-response evaluations in the cited studies are limited to stated responses to information or hypothetical scenarios, while HouseholdBench additionally evaluates predictions against realized household outcomes after real economic events.

HouseholdBench also sits within a broader literature simulating human behavior with LLMs. LLMs can simulate results from social science experiments (Filippas et al., 2024; Ashokkumar et al., 2026) and approximate the survey response distributions of demographic subgroups when conditioned on respondent demographics (Argyle et al., 2022). Yet these simulations remain inaccurate: LLM opinions stay misaligned with those of demographic groups even after steering (Santurkar et al., 2023), synthetic responses vary less than real ones (Bisbee et al., 2024), and predictions for specific individuals correlate weakly with their actual responses (Peng et al., 2026). Fine-tuning on human behavioral data can narrow this gap. For example, Binz et al. (2025) fine-tune an LLM on choices from 160 psychology experiments and outperform traditional cognitive models. Kolluri et al. (2025) fine-tune on responses from social science experiments, Xie et al. (2025) on economic games and personality surveys, and Suh et al. (2025) on public opinion surveys to predict response distributions across demographic subgroups. HouseholdBench extends this broader literature to fine-tune LLMs with household economic decisions.

B Data Sources

This appendix describes each survey and the preparation of the common data files used to construct HouseholdBench tasks.

B.1 Consumer Expenditure Survey: Diary Survey

Additional Information. The U.S. Bureau of Labor Statistics designs and publishes the Consumer Expenditure Surveys, which the U.S. Census Bureau administers. The current continuous program began in 1980. In the Diary Survey, each sampled consumer unit records all expenditures for two consecutive weeks. The survey focuses on frequent purchases and also records demographic characteristics, income, and benefit receipt. A recent annual sample selects roughly 18,000 addresses and yields about 6,700 usable pairs of diaries (U.S. Bureau of Labor Statistics, 2022; U.S. Census Bureau, 2026b). BLS publishes the public-use files annually. Exact historical posting dates are unavailable for the years used here. We assign each release to March 31 of the following year.

Downloading and Cleaning the Data. We download the 1982–1989 fixed-width files from ICPSR and the 1990–2011 comma-delimited files from BLS (Inter-university Consortium for Political and Social Research, 2026; U.S. Bureau of Labor Statistics, 2026a). We combine the consumer-unit, individual, diary-week, and expenditure files. We use the accompanying documentation to harmonize

identifiers, demographics, resources, benefit receipt, diary dates, and expenditure categories across releases. We assign each expenditure to its recorded date and sum expenditures within consumer unit, date, and category. We add days with no expenditure so that each valid diary week contains seven rows. We leave the date and daily expenditures missing when the source date cannot be recovered. For the Social Security task, we reproduce the recipient and expenditure definitions in Stephens (2003). The cleaned file contains 2,720,802 consumer-unit-day rows for 205,003 consumer units, with each consumer unit contributing either seven or fourteen rows.

B.2 Consumer Expenditure Survey: Interview Survey

Additional Information. The Interview Survey is the second component of the Consumer Expenditure Surveys. It follows sampled addresses quarterly and generally interviews a consumer unit four times. Interviews record large and regularly incurred expenditures over the preceding three months. They also record household composition, employment, income, assets, and liabilities. The current design contacts roughly 13,000 addresses and obtains about 5,000 usable interviews per quarter (U.S. Bureau of Labor Statistics, 2022; U.S. Census Bureau, 2026b). BLS publishes annual public-use packages. We use official release dates for 2015–2024. Exact posting dates are unavailable for 1980–2014, so we assign those releases to March 31 of the following year.

Downloading and Cleaning the Data. We download the BLS annual public-use packages for 1980, 1981, and 1984–2024 (U.S. Bureau of Labor Statistics, 2026a). We take the FMLI consumer-unit summary files and construct one row per interview. Annual packages overlap, so some interviews appear twice. We identify an interview by NEWID, year, and month and keep the copy from the later package. We harmonize demographic, geographic, income, and expenditure fields across questionnaire changes. We create separate household identifiers for each survey design because raw identifiers can repeat after a redesign. We calculate quarterly expenditure as the sum of its three monthly components. Before-tax income covers the preceding twelve months. The cleaned file contains 1,040,182 consumer-unit interviews. These interviews form 375,509 linked series, each containing one to four quarterly interviews.

B.3 Current Population Survey: Basic Monthly Survey

Additional Information. The U.S. Census Bureau conducts the Current Population Survey for the U.S. Bureau of Labor Statistics. It has run monthly since 1940 and represents the civilian noninstitutional population. The current design includes about 62,000 eligible housing units and 54,000 completed interviews each month. These interviews describe roughly 105,000 people aged 16 or older (U.S. Bureau of Labor Statistics, 2026c;d). The Basic Monthly Survey records household composition, demographics, employment, unemployment, and labor-force participation. Households enter for four months, leave for eight months, and return for four months. Public-use files are usually available 30–45 days after collection ends (U.S. Census Bureau, 2026c). We use exact Census release dates from January 2020 onward. For earlier months, we use 30 days after the Saturday ending the interview week that contains the nineteenth of the month.

Downloading and Cleaning the Data. We download one IPUMS CPS Version 13.0 extract and its codebook (Flood et al., 2025). We interpret each variable as specified in the codebook. We recode documented missing and not-in-universe values as missing and record whether each variable exists in each month. We add a release date for every survey month. For longitudinal tasks, we link records only when CPSIDP is positive and unique in both months. We require the records to occupy the expected positions in the survey rotation and to report consistent sex, race, and age. The cleaned Basic file contains 79,330,448 person-month rows from January 1976 through June 2026. A person can appear in as many as eight monthly interviews under the 4-8-4 design.

B.4 Current Population Survey: Displaced Worker Supplement

Additional Information. The Census Bureau administers the Displaced Worker Supplement for BLS to CPS respondents aged 20 or older. It has generally run every two years, in January or February, since 1984. The supplement records involuntary job loss, the lost job, unemployment, benefits, relocation, subsequent work, current earnings, and health insurance. Its recall window

covered five years in 1984–1992 and three years from 1994 onward. Since 1998, it has excluded self-employed workers and workers expecting recall (IPUMS CPS, 2026; U.S. Census Bureau, 2024a). DWS files are published separately from Basic CPS files. We use the concurrent Basic release date when an exact DWS date is unavailable (U.S. Census Bureau, 2024b).

Downloading and Cleaning the Data. We obtain the DWS variables in the same IPUMS CPS extract used for the Basic survey (Flood et al., 2025). We apply the same definitions of missing values and the same release dates. The cleaned file contains 1,840,305 person-wave records with a DWS status across 21 supplements from 1984 through 2024. Each person identifier appears once in these data. Of these records, 95,904 identify a displaced worker.

B.5 University of Michigan Surveys of Consumers

Additional Information. The University of Michigan Survey Research Center has conducted the Surveys of Consumers since 1946. The monthly survey covers the contiguous United States and the District of Columbia. It records household finances, buying conditions, and expectations about inflation, unemployment, interest rates, and business conditions. The final sample for each month now contains about 1,000 households. Each month combines new respondents with repeat interviews about six and twelve months after the first interview (University of Michigan, Survey Research Center, 2026b). Public respondent files generally appear four weeks after the final monthly release (University of Michigan, Survey Research Center, 2026a). From 1991 onward, we use the final release date plus 28 days. For earlier months, we use the last day of the survey month because the official release table does not cover those years.

Downloading and Cleaning the Data. We download the Cross-Section Archive data file, Stata dictionary, and codebook (University of Michigan, Survey Research Center, 2026b). We use the dictionary and codebook to assign variable labels, value labels, and missing values. We apply the historical revisions supplied with the archive. We link repeat interviews using the identifiers for the interviews about six and twelve months earlier. The cleaned file contains 342,345 respondent-month rows from January 1978 through February 2026. These rows form 217,641 linked respondent panels with one, two, or three interviews.

B.6 Survey of Consumer Expectations

Additional Information. The Federal Reserve Bank of New York has conducted the Survey of Consumer Expectations monthly since June 2013. This internet panel follows roughly 1,200–1,300 adult household heads for as long as twelve months. It records probabilistic expectations about inflation, labor markets, income, spending, credit, housing, and household finances. Topical modules cover related subjects (Armantier et al., 2017; Federal Reserve Bank of New York, 2026b). Microdata is released with a nine-month lag for Core and Credit Access data, and an 18-month lag for Household Spending, Public Policy, Housing, and Labor Market data. The New York Fed does not list separate lags for Household Finance, Informal Work, or Job Search, so we use nine months for those modules. The Public Policy workbook available on May 27, 2026 contained April 2025 samples, despite the stated 18-month lag, so we therefore use nine months for Public Policy as well.

Downloading and Cleaning the Data. We download the Core and topical workbooks from the New York Fed Data Bank (Federal Reserve Bank of New York, 2026a). We import each workbook separately and use common names and formats for respondent identifiers, dates, weights, and variable labels. The Core file contains 186,660 respondent-month rows for 24,592 respondent identifiers. The median respondent appears in nine months. Although the stated panel duration is twelve months, 774 respondent identifiers appear in more than twelve months and the observed maximum is sixteen. The eight topical files contain between 6,809 and 38,105 rows each. We recover the February 2014 home-financing experiment, which is not included in the standard release, from the replication package of Fuster & Zafar (2021; 2022).

B.7 Panel Study of Income Dynamics

Additional Information. The University of Michigan Survey Research Center began the Panel Study of Income Dynamics in 1968. Its initial nationally representative sample included more than 18,000 people in about 5,000 families. The survey follows original sample families, their descendants, and later immigrant refresher samples. One adult generally answers the family interview. The survey covers employment, income, wealth, expenditure, health, education, and family structure. Interviews were annual through 1997 and have been biennial since 1999. PSID does not follow a fixed publication lag, so we use December 31 of the survey year for waves through 2019. The 2021 wave is dated June 30, 2023, the month of the official user guide.

Downloading and Cleaning the Data. We rely on PSID-SHELF (Pfeffer et al., 2025; Daumler et al., 2025), an unofficial harmonization effort that covers 42 waves through 2021 and was published on February 24, 2025. We download the long person-wave PSID-SHELF file and the Complete Main Study file from its V2 release. We standardize variable names, sample status, family roles, survey years, income years, occupations, value labels, and release dates. We also select 306 fields from the raw Complete Main Study file (also provided in the release) to add to PSID-SHELF, covering spouse roles, annual hours, food expenditure, and reasons for job endings. The final file contains 3,533,082 person-wave rows for 84,121 permanent person identifiers and 42 survey waves. It contains one row for every person and wave, including years when the person was not interviewed or was not active in the sample.

B.8 U.S. Decennial Census

Additional Information. The Census Bureau has conducted a national population census every ten years since 1790. Sampling for the detailed long form began in 1940 and ended after 2000, when the American Community Survey replaced it. The 2000 long form contained 52 questions and went to about one in six households. It covered demographics, education, employment, income, migration, and housing (U.S. Census Bureau, 2026d;a). We use two IPUMS USA 1-percent samples: 1990 and 2000 (Ruggles et al., 2025; IPUMS USA, 2026). We use December 31 of each census year as the release date because a consistent set of exact dates is unavailable.

Downloading and Cleaning the Data. We download the six samples in one IPUMS USA Version 16.0 Stata extract with its codebook (Ruggles et al., 2025). We retain the six sample codes and apply the IPUMS definitions of missing values. We keep the two 1970 forms separate because they asked different questions. We harmonize migration variables across years and identify topcoded and censored income values. The cleaned file contains 13,445,846 person rows across six independent cross-sections. Each person appears once in a given cross-section.

B.9 Additional Data Sources

National Macroeconomic and Financial Series. We construct the quarterly national variables from the July 2026 FRED-QD file (McCracken & Ng, 2021; Federal Reserve Bank of St. Louis, 2026). We use real GDP, unemployment, all-items CPI, the federal funds rate, house prices, the S&P 500, and the 30-year mortgage rate. The corresponding FRED-QD columns are GDPC1, UNRATE, CPIAUCSL, FEDFUNDS, USSTHPI, S&P 500, and MORTGAGE30US. For each survey sample, we use only quarters completed before the survey date. We include the preceding four quarterly values and one five-year summary. We calculate quarterly percentage changes and 20-quarter compound growth for quantities and prices. Rate variables enter as quarterly levels and 20-quarter averages. These series include revisions present in the July 2026 file. They are not historical vintages available on each survey date.

Price and State Macroeconomic Series. We download monthly all-items CPI-U and the eight major-group indexes from BLS (U.S. Bureau of Labor Statistics, 2026b). The groups cover food and beverages, housing, apparel, transportation, medical care, recreation, education and communication, and other goods and services. We average the monthly indexes within each quarter and calculate quarterly percentage changes. For the Census migration task, we also use three additional sources: annual state unemployment from BLS Local Area Unemployment Statistics (U.S. Bureau of Labor

Statistics, 2026e), from which we use the unadjusted M13 annual average; state and national per-capita personal income from BEA SAINC1 line 3, where we calculate annual growth (U.S. Bureau of Economic Analysis, 2026); and the FHFA traditional all-transactions quarterly house-price index for each state and the nation (Federal Housing Finance Agency, 2026), which we average within year to calculate annual growth.

C XGBOOST MODELS

Specification. All models are estimated in Python 3.13.5 using XGBoost 3.3.0 (Chen & Guestrin, 2016). For each task, the model receives the same predictor variables represented in the household profile supplied to the language models. We adopt the pre-tuned parameter values proposed by Holzmüller et al. (2024) and select the number of boosting rounds using the relevant HouseholdBench evaluation metric. Training stops after 300 consecutive rounds without improvement. Predictions use the iteration with the best validation criterion; the model is not subsequently refit on the combined training and validation samples. We think this specification balances performance and training cost in terms of time and compute. Experiments with package-default parameters and no early stopping were dominated by this specification, while full random search over hyperparameter values yielded small performance gains, on the order of 1 to 2%, and increased training time by a factor of 8. Results are available on request.

Hyperparameter values. Here are the exact settings for each target type:

Numeric targets. We train one regression for each outcome with the following parameter values: at most 1,000 trees, learning rate 0.05, maximum depth 9, row sampling 0.70, full column sampling, minimum child weight 2, and zero L1, L2, and split-gain penalties. The loss criterion option is `reg:absoluteerror`. Direct early stopping minimizes validation mean absolute error.

Categorical targets. We train one classifier for each outcome, with the following parameter values: at most 1,000 trees, learning rate 0.08, maximum depth 6, row sampling 0.65, column sampling by level 0.90, minimum child weight 0.000005, and the same zero penalties. Binary models use `binary:logistic`; multiclass models use `multi:softprob`. Class predictions select the category with the highest predicted probability. Early stopping in the literature-informed specification minimizes one minus Macro-F1, calculated over the complete set of categories defined for the task.

Probability targets. We train one regression for each entry of the probability vector with the following parameter values: at most 2,000 trees, learning rate 0.05, maximum depth 9, row sampling 0.70, full column sampling, minimum child weight 2, and zero L1, L2, and split-gain penalties. The loss criterion option is `reg:absoluteerror`. A binary distribution is estimated through its event probability, with the complement restored during reconstruction. For distributions with more than two outcomes, XGBoost trains one tree per output probability (`one_output_per_tree`) and then predictions are projected onto the unit simplex. Early stopping minimizes validation total-variation distance after projection.

Training samples. For each task, we use at most 400,000 training rows and 50,000 validation rows. Tasks with less data use all of theirs. Across the 32 tasks, the training split holds 3,544,247 rows and the validation split 447,819.

D Additional Results and Discussion

D.1 Persistence in Household Probability Reports

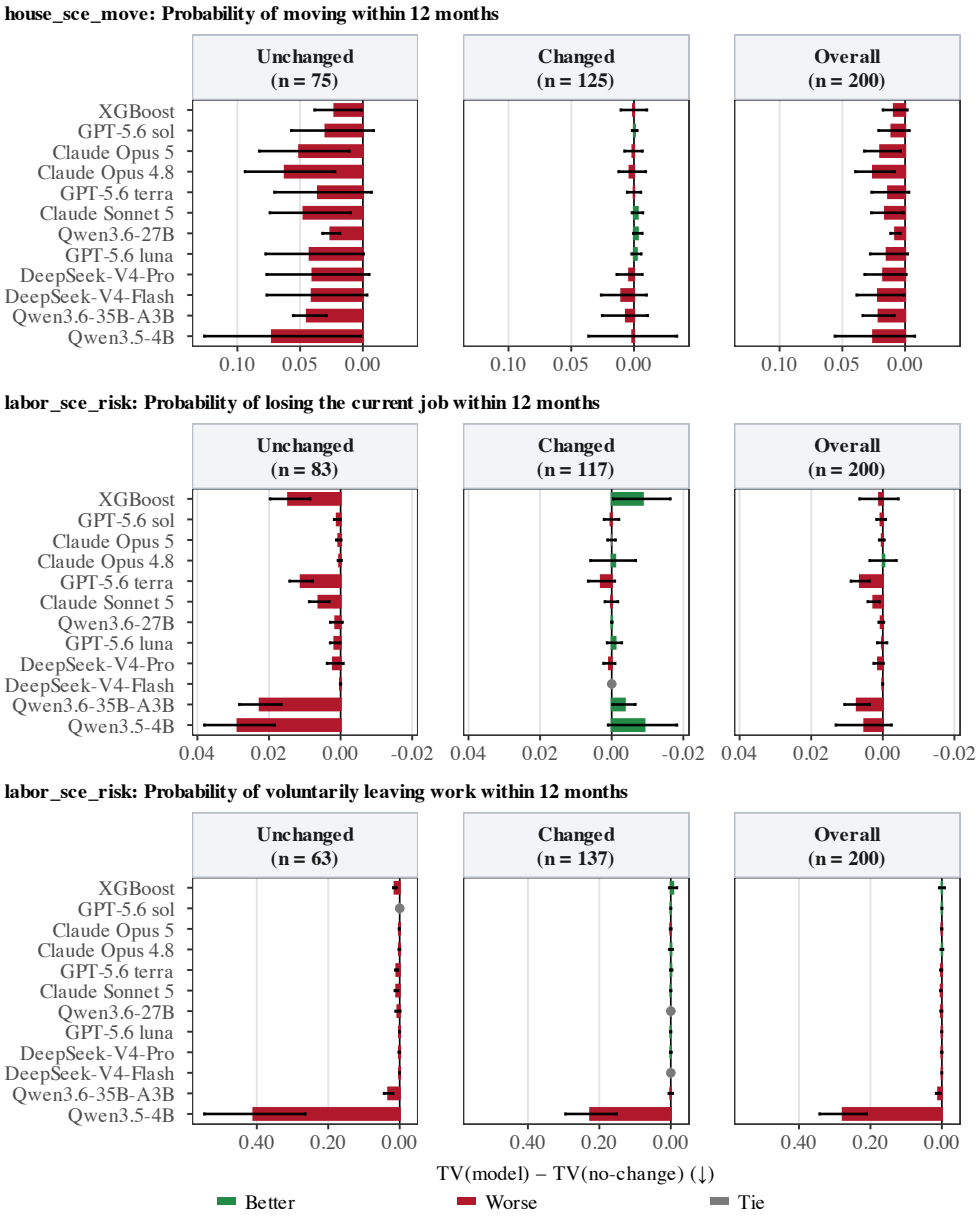


Figure D1: Difference in mean total variation by target, separately for unchanged, changed and all target probability vectors. Bars show $\mathrm{TV}(\mathrm{model}) - \mathrm{TV}(\mathrm{no\text{-}change})$ on the same valid observations, without division. Negative values favor the model; zero denotes equal performance. Each row reports one target; multi-target tasks are not pooled. The figure shows twelve prediction models using pre-cutoff valid responses; the no-change row is omitted. Facet titles count evaluation rows for that target before excluding invalid model responses. Unchanged and Changed partition those rows. Green denotes a negative difference, red a positive difference and grey a tie. Black bars are 95% basic paired IID bootstrap intervals for the difference, resampling rows within task and omitting draws without a valid subgroup observation. Each target uses one scale across its three panels; scales differ across targets. Continued on the next page.

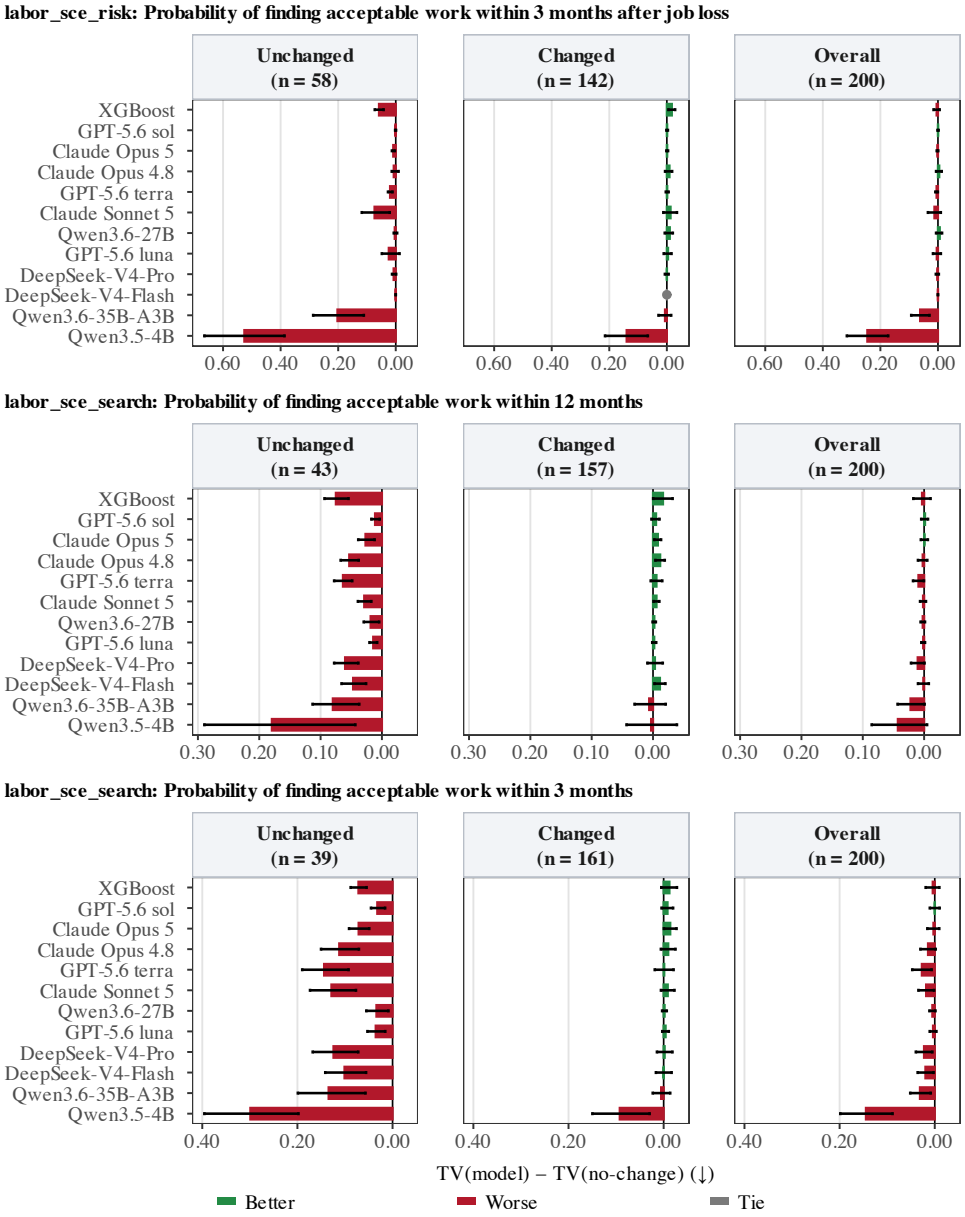


Figure D1: Difference in mean total variation by target (continued). The model set, sample definitions and interval construction are the same as on the first page. Continued on the next page.

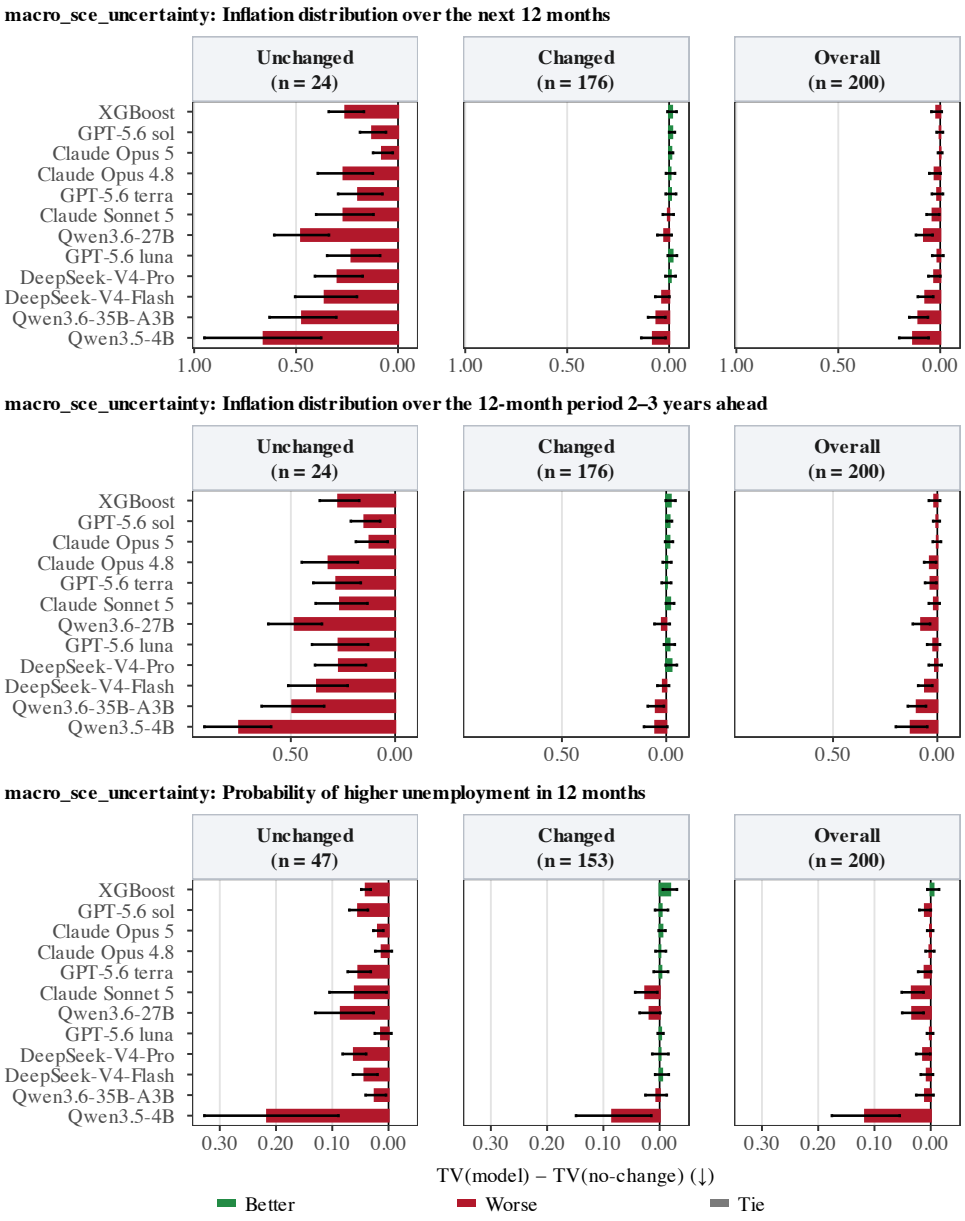


Figure D1: Difference in mean total variation by target (continued). The model set, sample definitions and interval construction are the same as on the first page. Continued on the next page.

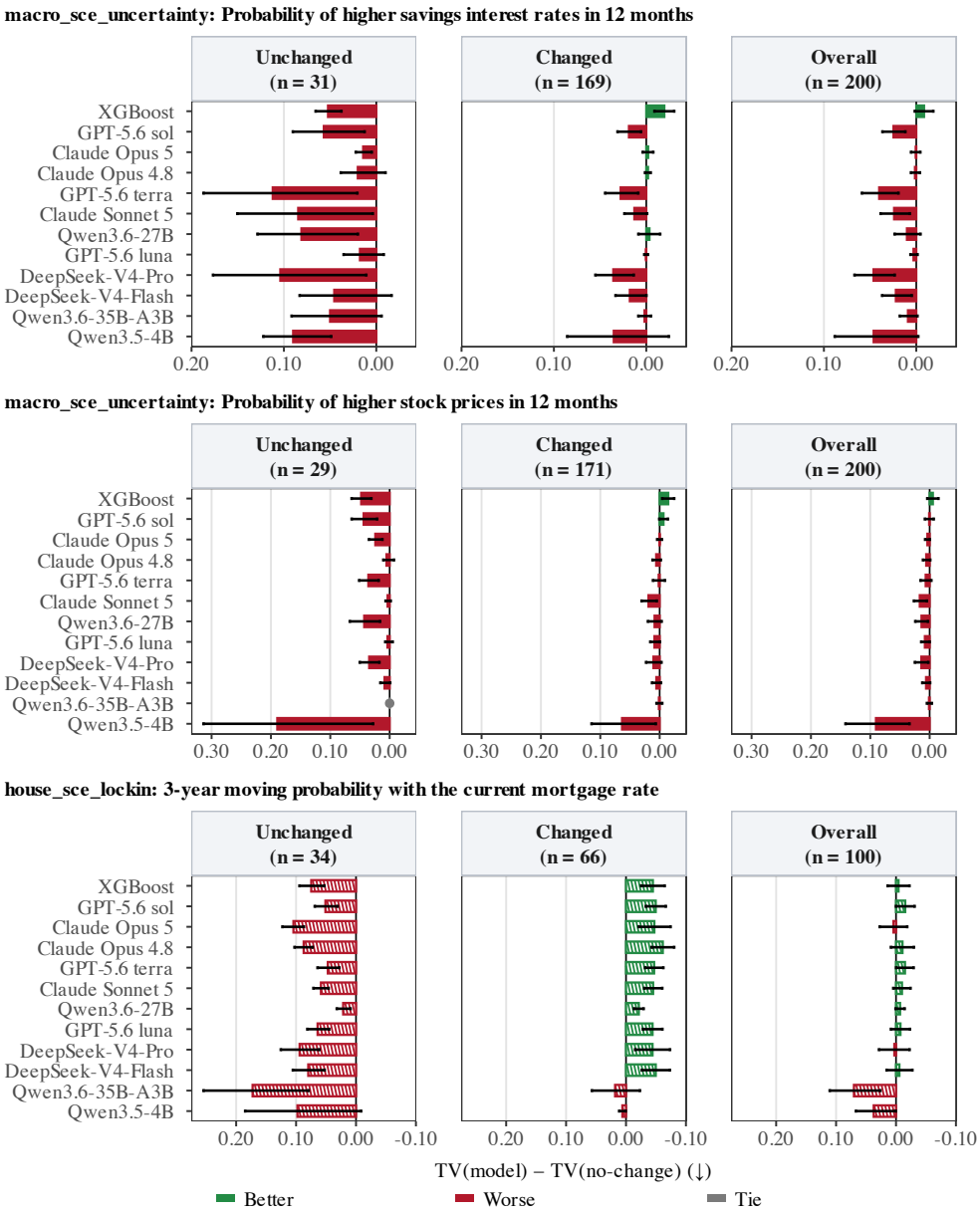


Figure D1: Difference in mean total variation by target (continued). Solid bars denote baseline tasks and hatched bars denote the policy task. Each model and its no-change comparator use the same valid observations in both the point estimate and every bootstrap draw.

D.2 BENCHMARK RESULTS

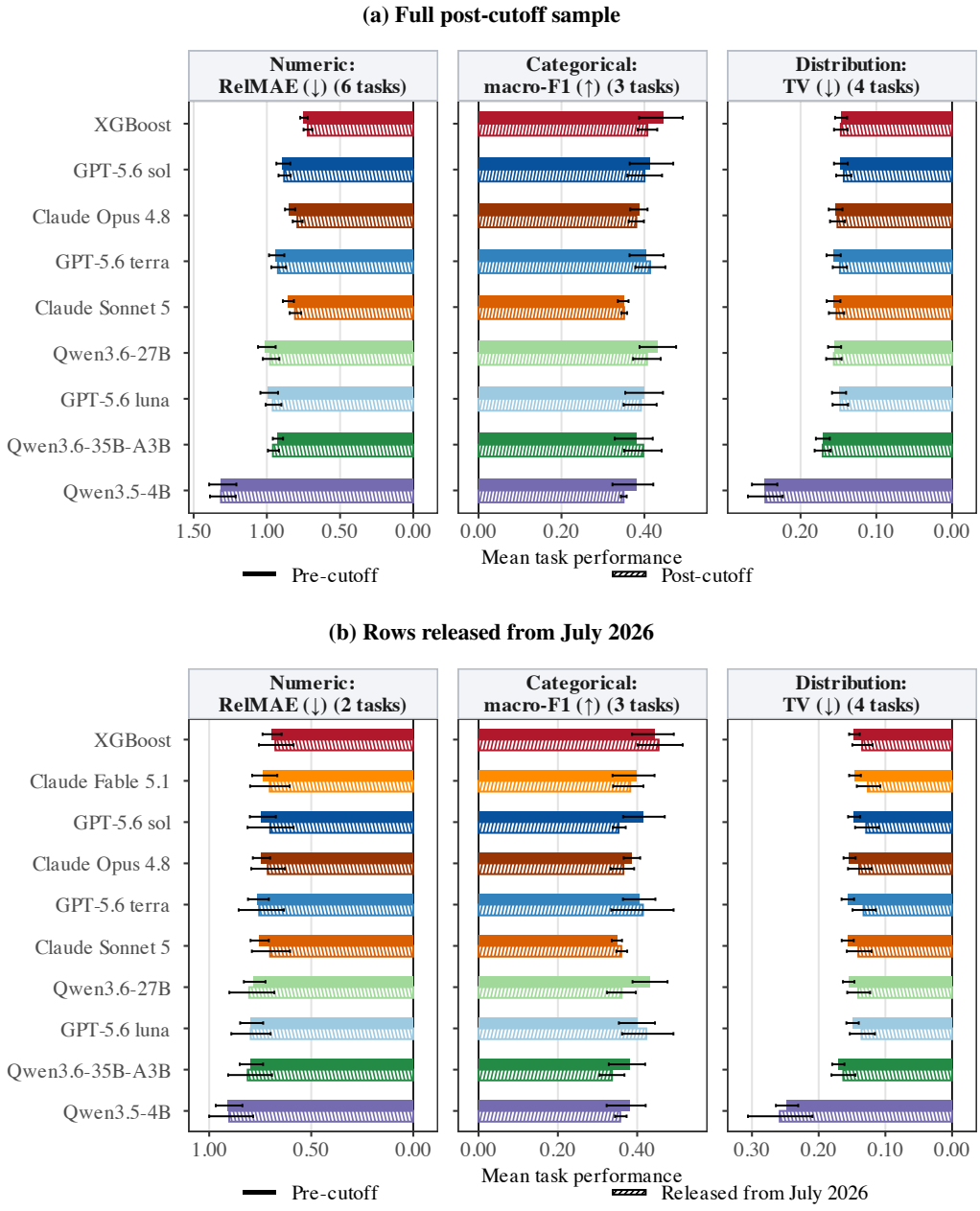


Figure D2: Mean performance before and after the cutoff. Panel (a) compares matched pre-cutoff and post-cutoff task samples for the main-leaderboard models evaluated in both samples, excluding the no-change baseline. Panel (b) compares pre-cutoff performance with performance on the 454 rows across nine tasks released on or after 1 July 2026, the first release wave after Claude Fable 5.1's June 2026 knowledge cutoff; its pre-cutoff sample covers the same nine tasks. Models follow the main-leaderboard order. Solid bars report equal-task means of the raw task metric in each panel's pre-cutoff sample; hatched bars report the corresponding later-sample means. Whiskers are pointwise 95% basic IID bootstrap intervals. The RelMAE and total-variation axes are reversed because lower values indicate better performance.

D.3 PERFORMANCE OVER TIME

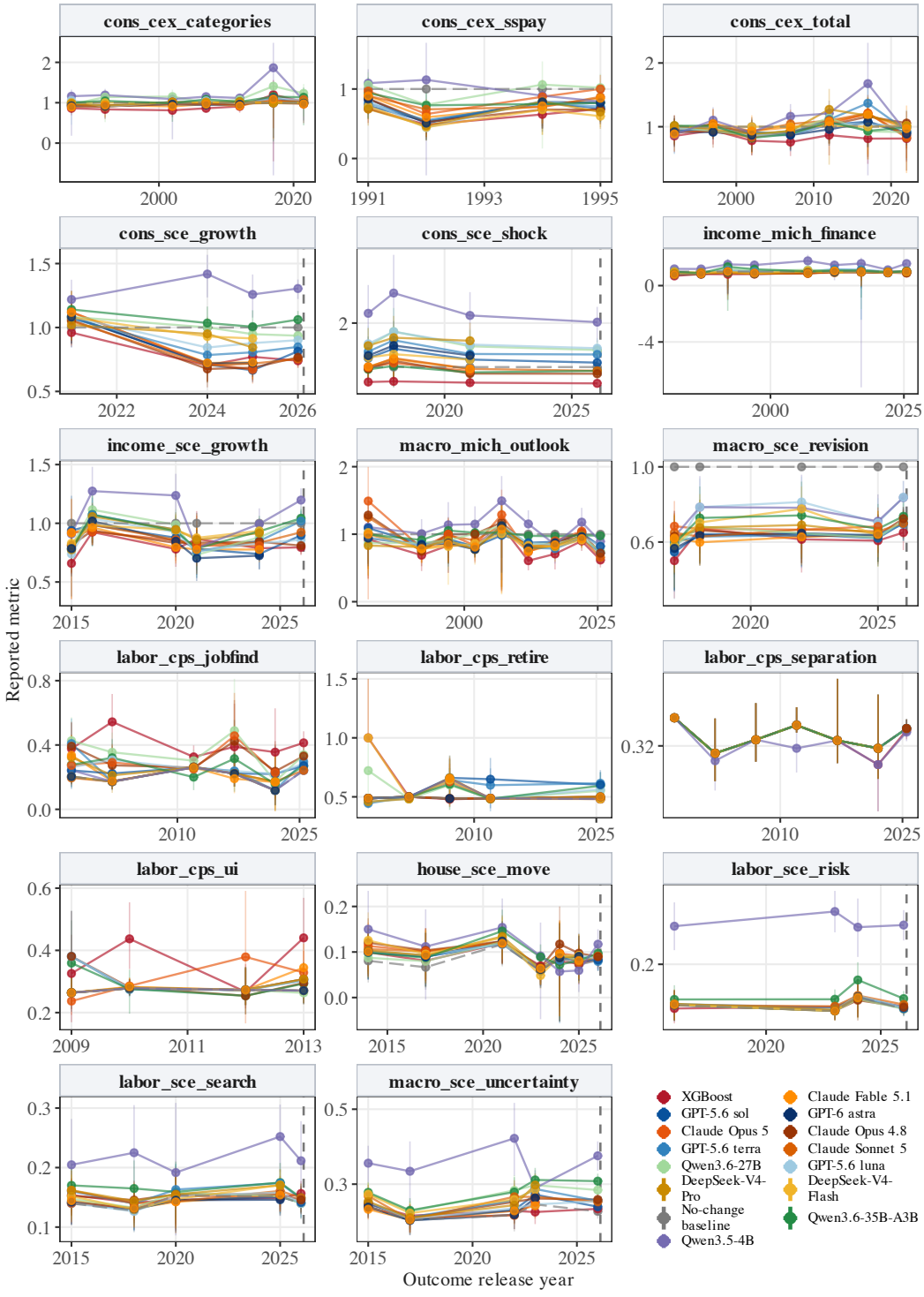


Figure D3: Model performance over outcome-release time for the main-leaderboard models. The legend follows the main-leaderboard order. Vertical bars are pointwise 95% basic IID bootstrap intervals.

E FINE-TUNING WITH A NUMBER-TOKEN LOSS

Besides supervised fine-tuning (Section 5.1), we test SFT with a *number-token loss* (SFT w/ NTL) (Zausinger et al., 2025). Our motivation is that two of our three target types (numeric and probability prediction) ask the model to predict numbers, and that SFT performs suboptimally on numerical prediction. The intuition is that the cross-entropy loss used in SFT treats the digit tokens as unordered labels and cannot express the closeness of predicted numbers. Motivated by this, Zausinger et al. (2025) propose adding a number-token loss, which minimizes the Wasserstein distance between the predicted and ground-truth numerical numbers. We expect that SFT with NTL brings better performance on the numerical and probability prediction tasks compared to SFT. We note that NTL loss only adds about 1% computation overhead.

			Baseline tasks (seen in fine-tuning)			**Policy-response tasks** (held out)		
	Model	K	**RelMAE** ↓ (# = 8)	**Macro-F1** ↑ (# = 6)	**TV** ↓ (# = 4)	**RelMAE** ↓ (# = 10)	**Macro-F1** ↑ (# = 3)	**TV** ↓ (# = 1)
	XGBoost	1	0.84	0.46	0.15	0.79	0.36	0.12
Reference	GPT-6 astra	1	0.89	0.39	0.15	0.94	0.32	**0.11**
	Claude Opus 5	1	0.90	**0.47**	0.15	0.91	0.28	0.13
	Fable 5.1	1	0.89	0.44	0.15	**0.86**	**0.35**	**0.11**
SFT	Qwen3.5-4B	1	1.14	0.39	0.25	1.43	0.25	0.15
	+ SFT	1	1.00 (−0.14)	0.45 (+0.06)	0.16 (−0.09)	1.30 (−0.14)	0.33 (+0.08)	0.17 (+0.01)
	+ SFT w/ NTL	1	0.97 (−0.17)	0.46 (+0.07)	0.15 (−0.10)	1.23 (−0.20)	0.33 (+0.07)	0.14 (−0.02)
	Qwen3.6-27B	1	0.96	0.45	0.16	1.04	0.26	0.12
	+ SFT	1	1.00 (+0.03)	0.46 (+0.01)	0.16 (+0.01)	1.18 (+0.14)	0.30 (+0.04)	0.15 (+0.04)
	+ SFT w/ NTL	1	0.97 (+0.01)	0.44 (0.00)	0.15 (0.00)	1.11 (+0.07)	0.27 (+0.01)	0.13 (+0.02)
SFT + aggregation	Qwen3.5-4B	16	1.11	0.38	0.24	1.31	0.24	0.15
	+ SFT	16	**0.87** (−0.24)	0.42 (+0.05)	0.15 (−0.09)	1.13 (−0.18)	0.31 (+0.06)	0.16 (+0.01)
	+ SFT w/ NTL	16	**0.87** (−0.24)	0.43 (+0.06)	**0.14** (−0.10)	1.12 (−0.19)	0.32 (+0.08)	0.14 (−0.01)
	Qwen3.6-27B	16	0.95	0.43	0.15	1.01	0.25	0.12
	+ SFT	16	**0.87** (−0.08)	0.44 (+0.01)	0.15 (−0.01)	1.03 (+0.02)	0.27 (+0.02)	0.14 (+0.02)
	+ SFT w/ NTL	16	**0.87** (−0.08)	0.45 (+0.02)	**0.14** (−0.01)	1.01 (0.00)	0.26 (+0.01)	0.13 (+0.01)

Table E1: Results of SFT and SFT with the number-token loss. The K column is the number of aggregated predictions. Best among the language models in bold, second best underlined. XGBoost is refitted on each task it is scored on, so its policy-response scores are fitted rather than held out.

SFT with the number-token loss achieves better performance on numeric and probability tasks than SFT alone (Table E1). Compared with SFT alone, adding the NTL lowers RelMAE by a further 3% for Qwen3.5-4B and 2% for Qwen3.6-27B on the baseline tasks. On the policy-response tasks, the additional improvement is 5% and 6%. It also lowers TV by 7% to 19% across the two models. However, the additional gains from NTL diminish after prediction aggregation.

F Model Ensembles

F.1 Aggregating Model Predictions

In Table F1, we show the detailed results of aggregating a model's predictions. We note that aggregating the two predictions of proprietary models (i.e., Fable 5.1, Claude Opus 5, and GPT-5.6 sol) does not change their prediction performance: the relative changes on numeric and probability tasks are minimal (less than 0.5%), and the changes on categorical tasks are mostly negative. This implies that aggregating more predictions from proprietary models would not improve the performance either. On the contrary, on the numeric and probability prediction tasks, the fine-tuned Qwen3.5-4B and Qwen3.6-27B improve from the aggregation with only two predictions ($K = 2$), and the improvement becomes larger at $K = 16$.

	Model	$K = 1$	$K = 2$	Improvement of $K = 2$ from $K = 1$ (%)	$K = 16$	Improvement of $K = 16$ from $K = 1$ (%)
Numeric, RelMAE ↓ (# = 18)						
Proprietary	Fable 5.1	0.878	0.879	−0.2	–	–
	Claude Opus 5	0.905	0.902	0.3	–	–
	GPT-5.6 sol	0.910	0.910	0.0	–	–
Open-weight	Qwen3.5-4B	1.304	1.272	2.5	1.223	6.3
	+ SFT	1.164	1.087	6.6	1.017	12.6
	+ SFT w/ NTL	1.117	1.059	5.2	1.008	9.8
	Qwen3.6-27B	1.007	0.999	0.8	0.984	2.2
	+ SFT	1.101	1.037	5.8	0.956	13.1
	+ SFT w/ NTL	1.050	1.001	4.6	0.950	9.5
Categorical, Macro-F1 ↑ (# = 9)						
Proprietary	Fable 5.1	0.413	0.400	−3.1	–	–
	Claude Opus 5	0.406	0.401	−1.2	–	–
	GPT-5.6 sol	0.399	0.400	0.2	–	–
Open-weight	Qwen3.5-4B	0.346	0.332	−4.1	0.332	−4.0
	+ SFT	0.410	0.409	−0.2	0.384	−6.4
	+ SFT w/ NTL	0.416	0.401	−3.4	0.396	−4.8
	Qwen3.6-27B	0.385	0.381	−1.1	0.374	−2.9
	+ SFT	0.406	0.397	−2.2	0.386	−4.9
	+ SFT w/ NTL	0.388	0.397	2.4	0.388	0.0
Probability, TV ↓ (# = 5)						
Proprietary	Fable 5.1	0.139	0.139	0.0	–	–
	Claude Opus 5	0.144	0.144	0.2	–	–
	GPT-5.6 sol	0.139	0.140	−0.4	–	–
Open-weight	Qwen3.5-4B	0.229	0.233	−1.6	0.221	3.4
	+ SFT	0.164	0.156	4.5	0.148	9.3
	+ SFT w/ NTL	0.148	0.146	1.5	0.142	4.4
	Qwen3.6-27B	0.148	0.148	−0.1	0.146	0.8
	+ SFT	0.161	0.155	4.0	0.147	9.0
	+ SFT w/ NTL	0.147	0.145	1.1	0.140	4.7

Table F1: Aggregating predictions of one model, pre-cutoff test split. $K = 1$ is the score of one draw, and $K = 2$ and $K = 16$ are the scores of the aggregate of two and sixteen draws. Each improvement is the change from $K = 1$ as a percentage of the $K = 1$ score, positive where the metric gets better.

G Inference and Fine-Tuning Settings

G.1 Inference Settings

We prompt every model at its provider's default sampling settings. The Qwen models are run with non-thinking mode and use the recommended non-thinking inference parameters for general tasks (temperature=0.7, top-p=0.8, top-k=20)[2]. The GPT models and Claude models run at their default reasoning effort. The DeepSeek models run at medium reasoning effort.

G.2 Fine-Tuning Settings

We fine-tune Qwen3.5-4B and Qwen3.6-27B with LoRA (Hu et al., 2022) on the 100,702 training data points described in Section 5.1. Both models share the hyperparameters in Table G1. We train each model on four NVIDIA A100 80GB GPUs.

Hyperparameter	**Value**
LoRA rank	8
LoRA alpha	32
LoRA dropout	0.05
LoRA target modules	all linear layers of the language model
Optimizer	AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay 0.1)
Peak learning rate	5×10^{-5}
Learning-rate schedule	linear warmup over the first 10% of steps, then cosine decay
Gradient clipping	1.0
Epochs	1
Effective batch size	128
Maximum sequence length	2,048 tokens
Precision	bfloat16
Random seed	42

Table G1: Fine-tuning hyperparameters, shared by Qwen3.5-4B and Qwen3.6-27B and by both training losses.

H Evaluation Tasks

This appendix describes each of the 32 tasks in HouseholdBench in detail, including information about the target, the predictors, sample construction and splits, and a full prompt example.

H.1 cons_cex_categories

Target. Household expenditure over the last three months, in current U.S. Dollars, for twelve categories: food, alcoholic beverages, housing, apparel and services, transportation, health care, entertainment, personal care, reading, education, tobacco, and miscellaneous. These aggregate to total expenditure, as used in other CEX tasks such as `cons_cex_total`. We take these expenditure variables directly from the raw CEX data, where they are constructed as aggregates of more detailed item-level expenditures reported directly by respondents.

Predictors. Spending history (up to three prior three-month periods of category expenditure), before-tax household income over the past 12 months as recorded on the current interview row, household composition (household size, number of adults, number of children), respondent demographics and geography (age, sex, race, education, marital status, region, urban or rural status), core macro context (real GDP growth, unemployment rate, headline CPI inflation, and the effective federal funds rate over the four most recently completed calendar quarters, plus five-year quarterly references), and disaggregated CPI inflation (major-group inflation over the same four calendar quarters, plus five-year quarterly averages).

[2] `https://huggingface.co/Qwen/Qwen3.5-4B`

No-change baseline. For each expenditure category, the no-change prediction is the household's expenditure in that category reported in the immediately preceding interview, conducted three months earlier.

Sample construction. Starting from our clean CEX Interview panel, we merge core macro variables from FRED-QD and disaggregated CPI from the BLS's CPI database. We then construct up to three lags of income and expenditure variables. These need to form a continuous history: if a household missed an interview, we set deeper lags to missing as well. Next, we drop rows that are missing one or more of the target variables, and also rows with current or first-history predictors missing (further lags are allowed to remain missing). Then, we require current annual before-tax household income to be at least $1,000, current and lagged total expenditure to be at least $250 minimum, and require every displayed expenditure category to be nonnegative. Finally, we trim the top and bottom percentiles of current annual income, current total expenditure and lagged total expenditure strictly, with the trimming bounds computed within each year and calendar quarter.

Prompt example

`cons_cex_categories`

System. You are the head of an American household making decisions and forming expectations about income, spending, and saving.

User. Here is some background information about yourself and your household. You are a 32-year-old white man, are currently married, and have a graduate or professional degree. You currently reside in an urban area in the West. Your household has 2 adults and 1 child, for 3 people in total.

Here are your household's recent spending and income. Note that the current month is February, and that $ denotes amounts in U.S. dollars. Between nine and twelve months ago, your household spent $1,170 on food; $15 on alcoholic beverages; $4,369 on housing; $424 on apparel and services; $2,739 on transportation; $1,172 on health care; $496 on entertainment; $248 on personal care; $64 on reading; $0 on education; $0 on tobacco; and $750 on miscellaneous. Between six and nine months ago, your household spent $1,352 on food; $155 on alcoholic beverages; $2,494 on housing; $369 on apparel and services; $3,403 on transportation; $644 on health care; $600 on entertainment; $108 on personal care; $20 on reading; $0 on education; $0 on tobacco; and $0 on miscellaneous. Between three and six months ago, your household spent $1,830 on food; $0 on alcoholic beverages; $2,436 on housing; $1,171 on apparel and services; $1,886 on transportation; $1,449 on health care; $572 on entertainment; $90 on personal care; $63 on reading; $0 on education; $0 on tobacco; and $30 on miscellaneous. Your before-tax household income over the past 12 months was $47,050.

You are also aware of recent national economic conditions. Over the four most recently completed calendar quarters, listed from oldest to most recent, U.S. real GDP changed by 0.94%, 0.83%, 1.33%, and 1.64% quarter over quarter; the unemployment rate was 4.30%, 4.27%, 4.23%, and 4.07%; overall consumer prices changed by 0.37%, 0.75%, 0.74%, and 0.74% quarter over quarter; and the effective federal funds rate was 4.73%, 4.75%, 5.09%, and 5.31%. Over the five years ending in the most recently completed quarter, compound average quarterly real GDP growth was 1.02%, the average unemployment rate was 4.93%, compound average quarterly CPI inflation was 0.59%, and the average effective federal funds rate was 5.38%.

You are also aware of changes in prices for specific spending categories. Over the same four quarters, listed from oldest to most recent, prices changed by 0.61%, 0.26%, 0.41%, and 0.63% quarter over quarter for food and beverages; 0.35%, 0.55%, 0.65%, and 0.63% quarter over quarter for housing; -1.18%, 0.61%, -0.38%, and 0.03% quarter over quarter for apparel; -0.50%, 2.02%, 1.74%, and 1.46% quarter over quarter for transportation; 0.77%, 0.85%, 0.95%, and 0.90% quarter over quarter for medical care; 0.33%, 0.26%, 0.03%, and 0.07% quarter over quarter for recreation; 0.20%, 0.26%, 0.33%, and 0.46% quarter over quarter for education and communication; and 4.29%, 0.48%, 1.50%, and 1.39% quarter over quarter for other goods and services. Over the same five-year period, compound average quarterly inflation was 0.62% for food and beverages, 0.63% for housing, -0.05% for apparel, 0.40% for transportation, 0.84% for medical care, 0.46% for recreation, 0.63% for education and communication, and 1.34% for other goods and services.

Given this information, how much did your household spend in each of the categories during the past three months, in current U.S. dollars?

Return only a valid JSON array with 12 numeric elements, in this order: food, alcoholic beverages, housing, apparel and services, transportation, health care, entertainment, personal care, reading, education, tobacco, and miscellaneous. Your output must follow this exact array structure: [v_1, v_2, ..., v_12]. Replace every v placeholder with a single number. Do not include keys or explanatory text.

Assistant. [2750, 130, 3184, 535, 3655, 572, 1171, 72, 66, 0, 0, 14]

H.2 CONS_CEX_TOTAL

Target. Total, nondurable, and durable household expenditure over the last three months, in current U.S. dollars. We construct these measures from the expenditure summaries in the CEX, following the definitions in Parker et al. (2013). Total expenditure equals the survey's total expenditure measure less cash contributions, life and other personal insurance, and contributions to retirement plans, pensions, and Social Security. Nondurable expenditure is the sum of spending on food and alcoholic beverages,

utilities, household operations, public transportation, gasoline and motor oil, personal care, tobacco, miscellaneous items, apparel, health care, and reading. Durable expenditure equals total expenditure less nondurable expenditure.

Predictors. Spending history (up to three prior three-month periods of total, nondurable, and durable expenditure), before-tax household income over the past 12 months, household composition (household size, number of adults, number of children), respondent demographics and geography (age, sex, race, education, marital status, region, urban or rural status), and core macro context (real GDP growth, unemployment rate, headline CPI inflation, and the effective federal funds rate over the four most recently completed calendar quarters, plus five-year quarterly references).

No-change baseline. For each expenditure measure, the no-change prediction is the household's expenditure reported in the immediately preceding interview, conducted three months earlier.

Sample construction. Starting from our clean CEX Interview panel, we merge core macro variables from FRED-QD and construct up to three lags of income and expenditure variables. These need to form a continuous history: if a household missed an interview, we set deeper lags to missing as well. Next, we drop rows that are missing one or more of the target variables, and also rows with current or first-history predictors missing (further lags are allowed to remain missing). Next, we require current annual before-tax household income to be at least $1,000, and current and available lagged total expenditure to be at least $250. We also require current and available lagged values of total, nondurable, and durable expenditure to be nonnegative. Finally, within each year and calendar quarter, we trim the top and bottom percentiles of current annual income and current and available lagged total expenditure.

Prompt example

`cons_cex_total`

System. You are the head of an American household making decisions and forming expectations about income, spending, and saving.

User. Here is some background information about yourself and your household. You are a 72-year-old white woman, are currently divorced, and have a high school diploma. You currently reside in an urban area in the West. Your household has 1 adult and no children.

Here are your household's recent spending and income. Note that the current month is March, and that $ denotes amounts in U.S. dollars. Between nine and twelve months ago, your total household expenditure was $2,135, non-durable expenditure was $1,290, and durable expenditure was $846. Between six and nine months ago, your total household expenditure was $22,364, non-durable expenditure was $1,334, and durable expenditure was $21,030. Between three and six months ago, your total household expenditure was $4,527, non-durable expenditure was $1,360, and durable expenditure was $3,166. Your before-tax household income over the past 12 months was $5,826.

You are also aware of recent national economic conditions. Over the four most recently completed calendar quarters, listed from oldest to most recent, U.S. real GDP changed by 0.94%, 0.83%, 1.33%, and 1.64% quarter over quarter; the unemployment rate was 4.30%, 4.27%, 4.23%, and 4.07%; overall consumer prices changed by 0.37%, 0.75%, 0.74%, and 0.74% quarter over quarter; and the effective federal funds rate was 4.73%, 4.75%, 5.09%, and 5.31%. Over the five years ending in the most recently completed quarter, compound average quarterly real GDP growth was 1.02%, the average unemployment rate was 4.93%, compound average quarterly CPI inflation was 0.59%, and the average effective federal funds rate was 5.38%.

Given this information, what were your total household expenditure, non-durable expenditure, and durable expenditure during the past three months, in current U.S. dollars?

Return only a valid JSON array with exactly 3 numeric elements, in this order: total_expenditure, non_durable_expenditure, durable_expenditure. Your output must follow this exact array structure: [v_1, v_2, v_3]. Replace every v placeholder with a single number. Do not include keys or explanatory text.

Assistant. [4824, 1379, 3446]

H.3 CONS_PSID_WEALTH

Target. Household net worth in the next PSID interview, usually two years after the current interview and sometimes one year later in the early part of the sample. We use the PSID's summary measure of net worth, which combines the household's reported assets, including savings, financial investments, home equity, and other assets, and subtracts its reported debts.

Predictors. Household resources (income, net worth, savings, financial investments, home equity, other debt), household history (up to three prior samples of those resource measures and housing tenure), housing tenure (own, rent, or another arrangement), household composition (presence of a spouse or partner, marital status, household size, number of children), respondent demographics and geography (age, sex, race and ethnicity, region, metropolitan status when available), core macro context (real GDP growth, unemployment rate, headline CPI inflation, and the effective federal funds rate over the four most recently completed quarters, plus five-year quarterly references), and financial and housing-market context (house-price growth, S&P 500 growth, and the 30-year mortgage rate over the same periods).

No-change baseline. The no-change prediction is the household's net worth in the current PSID interview.

Sample construction. Starting from our clean PSID household panel, we merge national economic, financial-market, and housing-market variables. We link each household to its next consecutive PSID interview. Next, we drop rows that are missing the next-interview net-worth target or any required current predictor: age, sex, race or ethnicity, region, housing tenure, spouse or partner status, marital status, family income, net worth, or the economic, financial-market, and housing-market variables. Current household size, number of children, metropolitan status, the individual wealth components, and all prior household histories are allowed to remain missing. Finally, within the relevant survey year, we trim the top and bottom percentiles of current and available lagged family income and of current, available lagged, and next-interview net worth.

Prompt example

`cons_psid_wealth`

System. You are the head of an American household making decisions about income, saving, housing, and household wealth.

User. Here is some background information about yourself and your household. You are a 57-year-old White woman. You live in the South, in a metropolitan area. You do not have a spouse or partner in your household and are not legally married. Your household has 1 person, no children.

Here is your household's financial situation now and in past years. Note that $ denotes amounts in U.S. dollars. Your current household income is $60,704, and your total net worth is $1,000. You have $0 in savings, $0 in financial investments, $0 in home equity, and $0 in other debt. You currently rent your home. Six years ago, your household income was $52,656, your total net worth was $17,000, your savings were $4,000, your financial investments were $0, your home equity was $0, your other debt was $0, and your household rented its home. Four years ago, your household income was $18,428, your total net worth was $10,000, your savings were $0, your financial investments were $0, your home equity was $0, your other debt was $0, and your household rented its home. Two years ago, your household income was $55,080, your total net worth was $4,000, your savings were $4,000, your financial investments were $0, your home equity was $0, your other debt was $3,000, and your household rented its home.

You are also aware of recent national economic conditions. Over the four most recently completed calendar quarters, listed from oldest to most recent, U.S. real GDP changed by -0.35%, 1.29%, 1.22%, and 0.51% quarter over quarter; the unemployment rate was 6.67%, 6.20%, 6.07%, and 5.70%; overall consumer prices changed by 0.62%, 0.53%, 0.26%, and -0.25% quarter over quarter; and the effective federal funds rate was 0.07%, 0.09%, 0.09%, and 0.10%. Over the five years ending in the most recently completed quarter, compound average quarterly real GDP growth was 0.57%, the average unemployment rate was 8.03%, compound average quarterly CPI inflation was 0.44%, and the average effective federal funds rate was 0.12%. You are also aware of recent financial and housing-market conditions. Over the same four completed quarters, listed from oldest to most recent, U.S. house prices changed by 0.49%, 1.50%, 1.06%, and 0.54% quarter over quarter; the S&P 500 changed by 3.61%, 3.60%, 3.98%, and 1.83% quarter over quarter; and the average 30-year mortgage rate was 4.36%, 4.23%, 4.14%, and 3.97%. Over the same five-year period, compound average quarterly house-price growth was -0.15%, compound average quarterly S&P 500 growth was 3.12%, and the average 30-year mortgage rate was 4.19%.

Given this information, in two years, what will your household's total net worth be, in current U.S. dollars?

Return only a valid JSON array containing exactly one integer, with no explanatory text. Your output must follow this exact array structure: [v_1]. Replace every v placeholder with a single number.

Assistant. [2000]

H.4 CONS_SCE_GROWTH

Target. The percentage change in the household's current monthly spending relative to 12 months earlier. The SCE first asks whether monthly spending is higher, lower, or unchanged and then asks by what percentage it changed. We combine the direction and size of the reported change so that increases are positive and decreases are negative.

Predictors. Household finances and expectations (current financial situation, expected financial situation one year ahead, one-year inflation expectation, longer-run inflation expectation, spending relative to budget when available), work situation (current work-status indicators), respondent demographics, resources, and geography (age, numeracy, region, education, household before-tax income bracket), and core macro context (real GDP growth, unemployment rate, headline CPI inflation, and the effective federal funds rate over the four most recently completed quarters, plus five-year quarterly references).

No-change baseline. The no-change prediction is no change in the household's monthly spending relative to 12 months earlier.

Sample construction. We merge the SCE Household Spending module with respondent information from the SCE Core survey and national economic variables from FRED-QD. Next, we drop rows that are missing either part of the spending-growth target or any current predictor. We then require respondents to be between 18 and 100 years old. Finally, within each survey month, we trim the top and bottom percentiles of the reported spending change.

Prompt example

`cons_sce_growth`

System. You are the head of an American household making decisions and forming expectations about income, spending, and saving.

User. Here is some background information about yourself and your household. You are a 36-year-old, have a college education or more, live in the West, and have a high numeracy score.

Here are your household finances, work situation, and expectations. Note that the current month is August, and that $ denotes amounts in U.S. dollars. Your household's total pre-tax income over the past 12 months was between $30,000 and $39,999. Your household is somewhat better off than it was 12 months ago, and you expect your household finances to be about the same 12 months from now. You expect inflation to be 1.1% over the next 12 months and 1.0% over the next three years. You are working full-time, and your household currently uses a budget.

You are also aware of recent national economic conditions. Over the four most recently completed calendar quarters, listed from oldest to most recent, U.S. real GDP changed by 1.22%, 0.51%, 0.90%, and 0.62% quarter over quarter; the unemployment rate was 6.07%, 5.70%, 5.53%, and 5.43%; overall consumer prices changed by 0.26%, -0.25%, -0.65%, and 0.68% quarter over quarter; and the effective federal funds rate was 0.09%, 0.10%, 0.11%, and 0.12%. Over the five years ending in the most recently completed quarter, compound average quarterly real GDP growth was 0.58%, the average unemployment rate was 7.60%, compound average quarterly CPI inflation was 0.43%, and the average effective federal funds rate was 0.12%.

Compared with 12 months ago, by what percentage has your household's current monthly spending changed? Use a positive number if spending is higher now and a negative number if spending is lower now.

Return only a valid JSON array containing exactly one number. Your output must follow this exact array structure: [v_1]. Replace every v placeholder with a single number. Do not include explanatory text.

Assistant. [-2.0]

H.5 CONS_CEX_REBATE01

This task is inspired by Johnson et al. (2006).

Target. Total, durable, and nondurable household expenditure over the last three months, in current U.S. dollars. We construct these measures from the expenditure summaries in the CEX, following the definitions in Parker et al. (2013). Total expenditure equals the survey's total expenditure measure less cash contributions, life and other personal insurance, and contributions to retirement plans, pensions, and Social Security. Nondurable expenditure sums food and alcoholic beverages, utilities, household operations, public transportation, gasoline and motor oil, personal care, tobacco, miscellaneous items, apparel, health care, and reading. Durable expenditure equals total expenditure less nondurable expenditure.

Predictors. Spending history (up to three prior three-month periods of total, nondurable, and durable expenditure), before-tax household income over the past 12 months, household composition (household size, number of adults, number of children), respondent demographics and geography (age, sex, race, education, marital status, region, urban or rural status), core macro context (real GDP growth, unemployment rate, headline CPI inflation, and the effective federal funds rate over the four

most recently completed quarters, plus five-year quarterly references), and amount received under the 2001 federal income-tax rebates (the amount received during the current three-month period and during prior three-month periods, including zero amounts).

No-change baseline. For each expenditure measure, the no-change prediction is the household's expenditure in the immediately preceding three-month period.

Sample construction. Starting from our clean CEX Interview panel, we merge the 2001 tax-rebate supplement and national economic variables from FRED-QD. We keep interviews conducted from January 2001 through March 2002 while the rebate questions were fielded, and construct up to three continuous lags of income, expenditure, and assigned rebate amounts. Next, we drop rows that are missing one or more target variables, any current demographic, geographic, income, or macroeconomic predictor, any first-history expenditure, or the first-history assigned rebate. The current assigned rebate and the second and third histories of expenditures and rebates are allowed to remain missing. We then require current annual before-tax household income to be at least $1,000, current and available lagged total expenditure to be at least $250, and current and available lagged values of total, nondurable, and durable expenditure to be nonnegative. Finally, within each year and calendar quarter, we trim the top and bottom percentiles of current annual income and current and available lagged total expenditure.

Prompt example

`cons_cex_rebate01`

System. You are the head of an American household making decisions and forming expectations about income, spending, and saving.

User. Here is some background information about yourself and your household. You are a 48-year-old white woman, are currently widowed, and have a bachelor's degree. You currently reside in an urban area in the South. Your household has 1 adult and no children.

Here are your household's recent spending and income. Note that the current month is December, and that $ denotes amounts in U.S. dollars. Between nine and twelve months ago, your total household expenditure was $4,797, non-durable expenditure was $2,407, and durable expenditure was $2,390. Between six and nine months ago, your total household expenditure was $3,966, non-durable expenditure was $2,151, and durable expenditure was $1,815. Between three and six months ago, your total household expenditure was $4,046, non-durable expenditure was $2,229, and durable expenditure was $1,817. Your before-tax household income over the past 12 months was $28,500.

You are also aware of recent national economic conditions. Over the four most recently completed calendar quarters, listed from oldest to most recent, U.S. real GDP changed by 0.60%, -0.33%, 0.62%, and -0.40% quarter over quarter; the unemployment rate was 3.90%, 4.23%, 4.40%, and 4.83%; overall consumer prices changed by 0.71%, 0.96%, 0.70%, and 0.28% quarter over quarter; and the effective federal funds rate was 6.47%, 5.59%, 4.33%, and 3.50%. Over the five years ending in the most recently completed quarter, compound average quarterly real GDP growth was 0.89%, the average unemployment rate was 4.46%, compound average quarterly CPI inflation was 0.61%, and the average effective federal funds rate was 5.34%.

The federal government has recently implemented a policy of sending tax rebates to households who satisfy certain criteria. Over the past three months, your household received a total of $300 as part of the program. Before this, the tax-rebate amounts recorded for your household were $300 between three and six months ago, $0 between six and nine months ago, and $0 between nine and twelve months ago.

Given this information, what were your total household expenditure, non-durable expenditure, and durable expenditure during the past three months, in current U.S. dollars?

Return only a valid JSON array with exactly 3 numeric elements, in this order: total_expenditure, non_durable_expenditure, durable_expenditure. Your output must follow this exact array structure: [v_1, v_2, v_3]. Replace every v placeholder with a single number. Do not include keys or explanatory text.

Assistant. [4643, 2768, 1875]

H.6 CONS_CEX_SSPAY

This task is inspired by Stephens (2003).

Target. Total, food-at-home, and food-away-from-home household expenditure on a given day, in current U.S. dollars. We construct these measures by assigning each purchase recorded in the CEX Diary to its reported date and summing purchases within the three categories, as defined in Stephens (2003), for that day. A day with no recorded spending in a category is assigned zero.

Predictors. Spending history (total, food-at-home, and food-away-from-home expenditure on each of the previous seven calendar days), household resources (annual household income and Social Security or Railroad Retirement income), household composition and respondent demographics, calendar timing (weekday, day of month, month, and the household's position relative to its scheduled Social Security payment date), and core macro context (national economic conditions over the most recently completed quarters).

No-change baseline. For each expenditure measure, the no-change prediction is the household's expenditure on the immediately preceding calendar day.

Sample construction. Starting from our clean daily CEX Diary panel for 1986–1996, we add scheduled Social Security payment dates and national economic variables, and construct spending histories for the previous seven calendar days. We keep non-January days in a seven-day window around a scheduled payment for households in which the reference person or spouse reports Social Security retirement or disability income. We require both weeks of the household's two-week diary to be complete, with valid nonoverlapping dates and usable dated purchases. Next, we drop rows that are missing one or more target variables, any current predictor, or any of the seven daily spending histories. No target or predictor shown to the model is allowed to remain missing. We then require positive household income and Social Security income, Social Security income to account for between 0 and 100 percent of household income, all displayed spending to be nonnegative, and food-at-home plus food-away-from-home spending not to exceed total spending on any displayed day. Finally, within each calendar year, we remove samples above the 99th percentile of annual household income or of total daily spending on the current or any of the previous seven days.

Prompt example

`cons_cex_sspay`

System. You are the head of an American household making decisions and forming expectations about income, spending, and saving.

User. Here is some background information about yourself and your household. You are an 82-year-old white man, are widowed, have an elementary school education, and are not currently working. You live in an urban area in the South. Your household consists of 1 adult and no children. Your household owns its home without a mortgage.

Your household reported about $9,886 in income before tax over the past 12 months. Of this, about $6,186, or 62.6%, came from Social Security and Railroad Retirement. These are annual figures and do not tell you the size of any particular monthly payment.

Here is your household's spending over the last seven days, listed from oldest to most recent. Note that today is Monday, 27th January, and that $ denotes amounts in U.S. dollars.

- 7 days ago: $7.41 in total, including $7.41 on food at home and $0.00 on food away from home.

- 6 days ago: $194.20 in total, including $20.45 on food at home and $0.00 on food away from home.

- 5 days ago: $0.00 in total, including $0.00 on food at home and $0.00 on food away from home.

- 4 days ago: $5.85 in total, including $4.39 on food at home and $0.00 on food away from home.

- 3 days ago: $9.42 in total, including $9.42 on food at home and $0.00 on food away from home.

- 2 days ago: $0.00 in total, including $0.00 on food at home and $0.00 on food away from home.

- 1 day ago: $54.74 in total, including $10.74 on food at home and $0.00 on food away from home.

You are also aware of recent national economic conditions. Over the four most recently completed calendar quarters, listed from oldest to most recent, U.S. real GDP changed by 0.97%, 0.88%, 1.53%, and 0.74% quarter over quarter; the unemployment rate was 7.23%, 7.30%, 7.20%, and 7.03%; overall consumer prices changed by 0.92%, 0.91%, 0.62%, and 1.02% quarter over quarter; and the effective federal funds rate was 8.48%, 7.92%, 7.90%, and 8.10%. Over the five years ending in the most recently completed quarter, compound average quarterly real GDP growth was 0.85%, the average unemployment rate was 8.32%, compound average quarterly CPI inflation was 1.22%, and the average effective federal funds rate was 11.21%.

Social Security benefits are normally paid on the third day of each month. If the third falls on a weekend or federal holiday, payment is made on the preceding business day.

Given this information, how much will your household spend today in total, on food at home, and on food away from home?

Return only a valid JSON array with exactly 3 numeric elements, in this order: total_expenditure, food_at_home, food_away_from_home. Use 0 when there is no spending in a category. Your output must follow this exact array structure: [v_1, v_2, v_3]. Replace every v placeholder with a single number. Do not include keys or explanatory text.

Assistant. [0.0, 0.0, 0.0]

H.7 CONS_CEX_STIMULUS08

This task is inspired by Parker et al. (2013).

Target. Total, durable, and nondurable household expenditure over the last three months, in current U.S. dollars. We construct these measures from the expenditure summaries in the CEX, following the definitions in Parker et al. (2013). Total expenditure equals the survey's total expenditure measure less cash contributions, life and other personal insurance, and contributions to retirement plans, pensions, and Social Security. Nondurable expenditure sums food and alcoholic beverages, utilities, household operations, public transportation, gasoline and motor oil, personal care, tobacco, miscellaneous items, apparel, health care, and reading. Durable expenditure equals total expenditure less nondurable expenditure.

Predictors. Spending history (up to three prior three-month periods of total, nondurable, and durable expenditure), before-tax household income over the past 12 months, household composition (household size, number of adults, number of children), respondent demographics and geography (age, sex, race, education, marital status, region, urban or rural status), core macro context (real GDP growth, unemployment rate, headline CPI inflation, and the effective federal funds rate over the four most recently completed quarters, plus five-year quarterly references), and amount received under the 2008 economic stimulus payment program (the amount received during the current three-month period and during prior periods, including zero amounts).

No-change baseline. For each expenditure measure, the no-change prediction is the household's expenditure in the immediately preceding three-month period.

Sample construction. Starting from our clean CEX Interview panel, we merge the 2008 stimulus-payment supplement and national economic variables from FRED-QD. We keep interviews conducted from September 2007 through March 2009 and require the household to have been interviewed while the stimulus questions were fielded, from June 2008 through March 2009. We construct up to three continuous lags of income, expenditure, and assigned payment amounts. Next, we drop rows that are missing one or more target variables, any current demographic, geographic, income, or macroeconomic predictor, or the first lag of the expenditure fields. The current and first lag of the assigned payment amounts are set to zero when no payment is assigned. Second and third lags of expenditures and assigned payments are allowed to remain missing. We then require current annual before-tax household income to be at least $1,000, current and available lagged total expenditure to be at least $250, and current and available lagged values of total, nondurable, and durable expenditure to be nonnegative. Finally, within each year and calendar quarter, we trim the top and bottom percentiles of current annual income and current and available lagged total expenditure.

Prompt example

`cons_cex_stimulus08`

System. You are the head of an American household making decisions and forming expectations about income, spending, and saving.

User. Here is some background information about yourself and your household. You are a 66-year-old white woman, are currently divorced, and have a graduate or professional degree. You currently reside in a rural area in the Midwest. Your household has 1 adult and no children.

Here are your household's recent spending and income. Note that the current month is September, and that $ denotes amounts in U.S. dollars. Between nine and twelve months ago, your total household expenditure was $3,896, non-durable expenditure was $1,616, and durable expenditure was $2,280. Between six and nine months ago, your total household expenditure was $5,464, non-durable expenditure was $3,404, and durable expenditure was $2,060. Between three and six months ago, your total household expenditure was $4,994, non-durable expenditure was $2,326, and durable expenditure was $2,668. Your before-tax household income over the past 12 months was $41,273.

You are also aware of recent national economic conditions. Over the four most recently completed calendar quarters, listed from oldest to most recent, U.S. real GDP changed by 0.58%, 0.63%, -0.43%, and 0.60% quarter over quarter; the unemployment rate was 4.67%, 4.80%, 5.00%, and 5.33%; overall consumer prices changed by 0.63%, 1.23%, 1.08%, and 1.30% quarter over quarter; and the effective federal funds rate was 5.07%, 4.50%, 3.18%, and 2.09%. Over the five years ending in the most recently completed quarter, compound average quarterly real GDP growth was 0.70%, the average unemployment rate was 5.08%, compound average quarterly CPI inflation was 0.82%, and the average effective federal funds rate was 3.27%.

The federal government has recently implemented a policy of sending stimulus payments to households who satisfy certain criteria. Over the past three months, your household received a total of $600 as part of the program. Before this, the stimulus-payment

amounts recorded for your household were $600 between three and six months ago, $0 between six and nine months ago, and $0 between nine and twelve months ago.

Given this information, what were your total household expenditure, non-durable expenditure, and durable expenditure during the past three months, in current U.S. dollars?

Return only a valid JSON array with exactly 3 numeric elements, in this order: total_expenditure, non_durable_expenditure, durable_expenditure. Your output must follow this exact array structure: [v_1, v_2, v_3]. Replace every v placeholder with a single number. Do not include keys or explanatory text.

Assistant. [3802, 2193, 1609]

H.8 CONS_PSID_JOBLOSS

Target. The household's annual cash expenditure on food in the calendar year after a job loss, in current U.S. dollars. We construct this measure from the PSID questions on food purchased for use at home and food purchased away from home, excluding food paid for with public assistance.

Predictors. Household food spending history (up to three prior annual samples), household resources (family income and each adult's labor earnings, annual hours, and employment status), job-loss circumstances (which adult lost a job and whether the loss was caused by a layoff, firing, plant closure, or employer move), household composition and respondent demographics (age, sex, education, race, marital status, presence of a spouse or partner, household size, number of children), and core macro context (national economic conditions before and during the job-loss year).

No-change baseline. The no-change prediction is the household's most recently observed annual cash food expenditure before the job-loss year.

Sample construction. Starting from PSID household records for 1968–1992, we combine the reports of the two adults in each household, construct cash food expenditure, and link each household to its next interview and to as many as three earlier interviews. A job-loss sample requires a layoff, firing, plant closure, or employer move after the affected adult had been working; multiple losses reported by the same household for the same period are combined. A control sample requires every current adult to have answered the job-loss questions and no adult to report a current or earlier qualifying loss, and we retain at most one control sample per household. Next, we drop rows that are missing the positive cash-food target, current predictors or the first lag of spending before the job loss event. Current education, race or ethnicity, region, housing tenure, prior job-loss details, and history records beyond the mandatory one are allowed to remain missing. We keep households whose adults are between 25 and 65 years old and whose spouse or partner structure is unchanged across the relevant interviews. Finally, within each survey year, we trim the top and bottom percentiles of every family-income value shown in the household's history.

Prompt example

`cons_psid_jobloss`

System. You are the head of an American household making decisions about food spending, work, income, and household finances.

User. Here is some background information about yourself and your household. You are a 52-year-old Black woman. You live in the South. You do not have a spouse or partner in your household. Your household has 1 person, no children. You completed 14 years of schooling. You currently own your home.

Here is your household's work, income, and food-spending history. Note that $ denotes amounts in U.S. dollars. Three years ago, you worked 540 hours in total and had $650 in annual labor earnings. Your total family income was $1,250. Two years ago, you worked 233 hours in total and had $350 in annual labor earnings. Your total family income was $1,350. Your household's annual food spending was $520, including food at home and food away from home.

You are also aware of recent national economic conditions. Over the four most recently completed calendar quarters, listed from oldest to most recent, U.S. real GDP changed by -0.15%, 0.14%, 0.92%, and -1.07% quarter over quarter; the unemployment rate was 4.17%, 4.77%, 5.17%, and 5.83%; overall consumer prices changed by 1.60%, 1.40%, 1.04%, and 1.45% quarter over quarter; and the effective federal funds rate was 8.57%, 7.89%, 6.71%, and 5.57%. Over the five years ending in the most recently completed quarter, compound average quarterly real GDP growth was 0.69%, the average unemployment rate was 3.93%, compound average quarterly CPI inflation was 1.11%, and the average effective federal funds rate was 6.08%.

You did not lose your job during the current calendar year.

Given this information, in one year, what will your household's annual food spending be, including food at home and food away from home?

Return only a valid JSON array with exactly one number in current U.S. dollars. Your output must follow this exact array structure: [v_1]. Replace v_1 with one number. Do not include a dollar sign, commas, keys, or explanatory text.

Assistant. [312.0]

H.9 CONS_SCE_SHOCK

Target. How the household says it would adjust spending under two hypothetical income changes: a permanent 10 percent increase in household income and a permanent 10 percent decrease. For the increase, the SCE asks the respondent to allocate the gain among spending or donating, saving or investing, and paying down debt. For the decrease, it asks the respondent to allocate the loss among reducing spending, reducing saving, and increasing borrowing. The task predicts the share of the gain allocated to spending or donating and the share of the loss covered by reducing spending.

Predictors. Spending and budgeting (current spending growth relative to 12 months earlier, spending relative to budget when available), household finances and expectations (current financial situation, expected financial situation one year ahead, one-year and longer-run inflation expectations), work situation (current work-status indicators), respondent demographics, resources, and geography (age, education, household before-tax income bracket, region, numeracy), and core macro context (real GDP growth, unemployment rate, headline CPI inflation, and the effective federal funds rate over the four most recently completed quarters, plus five-year quarterly references). Every respondent considers the same permanent 10 percent income increase and decrease.

No-change baseline. The no-change prediction is that the household assigns none of either income change to spending.

Sample construction. We merge the SCE Household Spending module with respondent information from the SCE Core survey and national economic variables from FRED-QD. Next, we drop rows that are missing one or more of the six allocation answers or any required demographic, income, or macroeconomic predictor. Current spending and budgeting, household-finance, inflation-expectation, and work-status predictors are allowed to remain missing. We then require each allocation to lie between 0 and 100 percent, the three allocations within each scenario to sum to between 99 and 101 percent, and the respondent to be between 18 and 100 years old.

Prompt example

`cons_sce_shock`

System. You are the head of an American household making decisions and forming expectations about income, spending, and saving.

User. Here is some background information about yourself and your household. You are a 36-year-old, have a college education or more, live in the West, and have a high numeracy score.

Here are your household finances, work situation, and expectations. Note that the current month is August, and that $ denotes amounts in U.S. dollars. Your household's total pre-tax income over the past 12 months was between $30,000 and $39,999. Your household is somewhat better off than it was 12 months ago, and you expect your household finances to be about the same 12 months from now. You expect inflation to be 1.1% over the next 12 months and 1.0% over the next three years. Your monthly household spending is 2.0 percent lower than 12 months ago, your household currently uses a budget, and you are working full-time.

You are also aware of recent national economic conditions. Over the four most recently completed calendar quarters, listed from oldest to most recent, U.S. real GDP changed by 1.22%, 0.51%, 0.90%, and 0.62% quarter over quarter; the unemployment rate was 6.07%, 5.70%, 5.53%, and 5.43%; overall consumer prices changed by 0.26%, -0.25%, -0.65%, and 0.68% quarter over quarter; and the effective federal funds rate was 0.09%, 0.10%, 0.11%, and 0.12%. Over the five years ending in the most recently completed quarter, compound average quarterly real GDP growth was 0.58%, the average unemployment rate was 7.60%, compound average quarterly CPI inflation was 0.43%, and the average effective federal funds rate was 0.12%.

For this question, suppose your household income changes unexpectedly next year. In one case, your household income rises by 10 percent. In the other case, your household income falls by 10 percent.

How much of the income gain would you spend or donate, and how much of the income loss would you cover by reducing spending? Give each answer as a number from 0 to 100, not as a fraction.

Return only a valid JSON array with exactly two numeric elements, in this order: percent of the income gain spent or donated; percent of the income loss covered by reducing spending. Your output must follow this exact array structure: [v_1, v_2]. Replace every v placeholder with a single number. Do not include keys or explanatory text.

Assistant. [1.0, 70.0]

H.10 INCOME_MICH_FINANCE

Target. The expected percentage change in the household's income over the next 12 months. The Michigan Survey first asks whether family income is expected to increase, decrease, or remain unchanged and then asks by what percentage it is expected to change. We combine the direction and size so that expected increases are positive and expected decreases are negative.

Predictors. Household and national assessments (personal finances compared with one year earlier, national business conditions compared with one year earlier), household resources (current annual income), household composition and respondent demographics (age, sex, education, marital or partner status, number of adults, number of children), geography (region), and core macro context (real GDP growth, unemployment rate, headline CPI inflation, and the effective federal funds rate over the four most recently completed quarters, plus five-year quarterly references).

No-change baseline. The no-change prediction is no change in household income over the next 12 months.

Sample construction. Starting from the Michigan Survey, we merge national economic variables from FRED-QD. Next, we drop rows that are missing either part of the income-growth target or any current predictor. No target or predictor shown to the model is allowed to remain missing. Finally, within each survey month, we trim the top and bottom percentiles of expected income growth and annual household income.

Prompt example

`income_mich_finance`

System. You are an American adult forming expectations about economic conditions and your own finances.

User. Here is some background information about yourself and your household. You are an 18-year-old man living in the North Central United States. You have never been married, and your highest completed education is a high school diploma. Your household has 2 adults and no children.

Here are your current views on your finances and the economy. Note that the current month is February, and that $ denotes amounts in U.S. dollars. Your current household income is $22,500 per year. You report that your personal finances are better than they were a year ago, while national business conditions are worse than they were a year ago.

You are also aware of recent national economic conditions. Over the four most recently completed calendar quarters, listed from oldest to most recent, U.S. real GDP changed by 1.19%, 1.94%, 1.80%, and 0.00% quarter over quarter; the unemployment rate was 7.50%, 7.13%, 6.90%, and 6.67%; overall consumer prices changed by 1.83%, 1.75%, 1.38%, and 1.47% quarter over quarter; and the effective federal funds rate was 4.66%, 5.16%, 5.82%, and 6.51%. Over the five years ending in the most recently completed quarter, compound average quarterly real GDP growth was 0.68%, the average unemployment rate was 6.74%, compound average quarterly CPI inflation was 1.92%, and the average effective federal funds rate was 7.13%.

By what percentage do you expect your income to change over the next 12 months? Use a negative number if you expect your income to be lower, a positive number if you expect it to be higher, and 0 if you expect it to be about the same.

Return only a valid JSON array containing exactly one number. Your output must follow this exact array structure: [v_1]. Replace every v placeholder with a single number. Do not include explanatory text.

Assistant. [3]

H.11 INCOME_PSID_EARNINGS

Target. The respondent's annual labor earnings two, four, and ten years after the current PSID interview, in current U.S. dollars.

Predictors. Current work and earnings (annual labor earnings, employment status, occupation when available), household resources and composition (household income, presence of a spouse or partner, household size, number of children), respondent demographics (age, sex, race and ethnicity,

education), geography (region), and core macro context (real GDP growth, unemployment rate, headline CPI inflation, and the effective federal funds rate over the four quarters preceding the survey year, plus five-year quarterly references).

No-change baseline. At each horizon, the no-change prediction is the respondent's annual labor earnings in the current interview.

Sample construction. Starting from our clean PSID panel, we keep respondents between 20 and 60 years old, harmonize occupation information, and merge national economic variables. We link each current interview to the same respondent's record exactly two, four, and ten years later. Next, we drop rows that are missing one or more of the three earnings targets or any required current predictor: age, sex, race or ethnicity, region, education, employment status, labor earnings, family income, spouse or partner status, or the macroeconomic variables. Current household size, number of children, metropolitan status, and occupation are allowed to remain missing. Finally, within each survey year, we trim the top and bottom percentiles of current, future, and available prior labor earnings and of current and available prior family income.

Prompt example

`income_psid_earnings`

System. You are an American individual or family making education, work, housing, and long-run financial decisions.

User. Here is some background information about yourself and your household. You are a 25-year-old White man. You live in the South. You completed 11 years of schooling and did not complete high school. You have a spouse or partner in your household. Your household has 5 people, including 3 children.

Here is your current work situation. Note that $ denotes amounts in U.S. dollars. You are currently working. Your labor earnings are $5,800, your household income is $7,900, and your broad occupation is production, transport, or labor.

You are also aware of recent national economic conditions. Over the four most recently completed calendar quarters, listed from oldest to most recent, U.S. real GDP changed by 0.89%, 0.06%, 0.95%, and 0.75% quarter over quarter; the unemployment rate was 3.83%, 3.83%, 3.80%, and 3.90%; overall consumer prices changed by 0.25%, 0.61%, 1.00%, and 1.09% quarter over quarter; and the effective federal funds rate was 4.82%, 3.99%, 3.89%, and 4.17%. Over the five years ending in the most recently completed quarter, compound average quarterly real GDP growth was 1.27%, the average unemployment rate was 4.59%, compound average quarterly CPI inflation was 0.54%, and the average effective federal funds rate was 4.02%.

What will your labor earnings be in current dollars two years from now, four years from now, and ten years from now?

Return only a valid JSON array with exactly 3 numeric elements, in this order: earn_2y, earn_4y, earn_10y. Your output must follow this exact array structure: [v_1, v_2, v_3]. Replace every v placeholder with a single number. Do not include keys or explanatory text.

Assistant. [6100, 5700, 13000]

H.12 INCOME_SCE_GROWTH

Target. The expected percentage change in the household's total income over the next 12 months. The SCE first asks whether total household income is expected to increase, decrease, or remain unchanged and then asks by what percentage it is expected to change. We combine the direction and size so that expected increases are positive and expected decreases are negative.

Predictors. Household finances and health (finances compared with one year earlier, expected finances one year ahead, health when available), work and job-search situation (current work-status indicators, number of jobs, type of employment arrangement, whether the respondent is looking for work, and duration of unemployment or time out of work when available), respondent demographics, resources, and geography (age, education, household before-tax income bracket, region, numeracy), and core macro context (real GDP growth, unemployment rate, headline CPI inflation, and the effective federal funds rate over the four most recently completed quarters, plus five-year quarterly references).

No-change baseline. The no-change prediction is no change in total household income over the next 12 months.

Sample construction. We merge national economic variables from FRED-QD with the SCE Core survey. Next, we drop rows that are missing either part of the income-growth target or any required demographic, income, or macroeconomic predictor. Current household-finance, health, work, and

job-search predictors are allowed to remain missing. We then require respondents to be between 18 and 100 years old. Finally, within each survey month, we trim the top and bottom percentiles of expected income growth.

Prompt example

`income_sce_growth`

System. You are an American adult making household decisions and forming expectations about your finances, work, and the economy.

User. Here is some background information about yourself and your household. You are a 54-year-old, have a high-school education, live in the Midwest, and have a high numeracy score.

Here are your household finances, work situation, and expectations. Note that the current month is June, and that $ denotes amounts in U.S. dollars. Your household's total pre-tax income over the past 12 months was less than $10,000. Your household is much worse off than it was 12 months ago, but you expect your household finances to be about the same 12 months from now.

You are also aware of recent national economic conditions. Over the four most recently completed calendar quarters, listed from oldest to most recent, U.S. real GDP changed by 0.45%, 0.14%, 0.12%, and 0.99% quarter over quarter; the unemployment rate was 8.20%, 8.03%, 7.80%, and 7.73%; overall consumer prices changed by 0.21%, 0.45%, 0.66%, and 0.40% quarter over quarter; and the effective federal funds rate was 0.15%, 0.14%, 0.16%, and 0.14%. Over the five years ending in the most recently completed quarter, compound average quarterly real GDP growth was 0.24%, the average unemployment rate was 8.48%, compound average quarterly CPI inflation was 0.44%, and the average effective federal funds rate was 0.35%.

By what percentage do you expect total household income to change over the next 12 months? Use a negative number for a decline, a positive number for an increase, and 0 for no change.

Return only a valid JSON array containing one number. Your output must follow this exact array structure: [v_1]. Replace every v placeholder with a single number. Do not include explanatory text.

Assistant. [0.0]

H.13 INCOME_CPS_DISPLACE

Target. Current weekly earnings among workers who lost a job and subsequently found another one. We use the Displaced Worker Supplement's current weekly earnings measure, which records weekly pay directly or standardizes pay reported at another frequency to a weekly amount.

Predictors. Respondent demographics and background (age, sex, race and ethnicity, education, marital status, Hispanic origin, nativity, citizenship, veteran status), geography (state, region, metropolitan status), household composition (household role, household size, own children), displacement circumstances (reason and timing of job loss, occupation and industry of the lost job), lost-job history (weekly earnings, tenure), reemployment history (number of jobs held since displacement, weeks without work before the next job), and core macro context (national economic conditions over the most recently completed quarters). The task predicts current earnings conditional on the respondent having lost a job and found another one.

No-change baseline. The no-change prediction is the worker's weekly earnings on the job that was lost.

Sample construction. Starting from the Displaced Worker Supplement, we keep respondents who report a qualifying job loss and are employed at the time of the supplement. We convert current and lost-job pay to weekly earnings. Next, we drop rows that are missing the current earnings target; any required current demographic, household, geographic, or macroeconomic predictor; lost-job weekly earnings or tenure; or either reemployment-history measure. The reported reason and timing of displacement, occupation and industry of the lost job, advance notice, full-time status, union status, employer type, and health-insurance provision are allowed to remain missing. Finally, within each survey wave, we trim the top and bottom percentiles of current and lost-job weekly earnings.

Prompt example

`income_cps_displace`

System. You are an American person answering questions about your income, work, and labor-force situation.

User. Here is some background information about yourself and your household. You are a 40-year-old White woman with a bachelor's degree. You are divorced and live in Alabama, in a metropolitan area in the South. You live alone and have no own children in your household. You are not Hispanic, were born in the United States, are a U.S. citizen, and are not a veteran.

You are also aware of recent national economic conditions. Over the four most recently completed calendar quarters, listed from oldest to most recent, U.S. real GDP changed by 1.40%, 1.70%, 0.83%, and 1.72% quarter over quarter; the unemployment rate was 6.23%, 5.93%, 5.07%, and 4.17%; overall consumer prices changed by 1.02%, 1.82%, 1.59%, and 2.18% quarter over quarter; and the effective federal funds rate was 0.08%, 0.07%, 0.09%, and 0.08%. Over the five years ending in the most recently completed quarter, compound average quarterly real GDP growth was 0.65%, the average unemployment rate was 5.08%, compound average quarterly CPI inflation was 0.71%, and the average effective federal funds rate was 1.09%.

Note that the current month is January, and that $ denotes amounts in U.S. dollars. During the past three years, you lost or left a full-time job at a private for-profit employer because there was insufficient work. You last worked at that job two years ago and received less than 1 month of notice before the job loss. You were not a labor-union member at that job, and the job provided health insurance. You worked at the lost job for about 1.5 years and earned approximately $1,096.16 per week.

Since losing or leaving that job, you have held 1 job. You went 48 weeks without work between the end of the lost job and the start of your next job. You are currently employed.

Approximately how much do you earn per week at your current job?

Return exactly one valid JSON array containing one number in current U.S. dollars per week. Your output must follow this exact array structure: [v_1]. Replace every v placeholder with a single number. Do not include a dollar sign, commas, keys, or explanatory text.

Assistant. [1307.7]

H.14 INCOME_SCE_POLICY

Target. The respondent's stated effect on their own household under four hypothetical policy changes over the next 12 months: changes in federal welfare benefits, unemployment benefits, the payroll tax rate, and the average income tax rate. For each policy, the survey asks whether the effect would be very negative, somewhat negative, neutral, somewhat positive, or very positive. The task predicts one of these five answers for each policy.

Predictors. Household finances and expectations (current financial situation, expected financial situation one year ahead, one-year and longer-run inflation expectations), work situation (current work-status indicators), respondent demographics, resources, and geography (age, education, household before-tax income bracket, region, numeracy), and core macro context (real GDP growth, unemployment rate, headline CPI inflation, and the effective federal funds rate over the four most recently completed quarters, plus five-year quarterly references). The prompt states whether federal welfare benefits, unemployment benefits, the payroll tax rate, and the average income tax rate increase or decrease.

No-change baseline. The no-change prediction is that each of the four policy changes has no effect on the household.

Sample construction. We merge the SCE Public Policy module with respondent information from the SCE Core survey and national economic variables from FRED-QD. Next, we drop rows that are missing one or more of the four household-impact targets, any of the four corresponding policy directions, or any required demographic, income, or macroeconomic predictor. Current household-finance, inflation-expectation, and work-status predictors are allowed to remain missing. We keep answers that map to the five target categories and require respondents to be between 18 and 100 years old.

Prompt example

`income_sce_policy`

System. You are an American adult making household decisions and forming expectations about spending, work, policy, and the economy.

User. Here is some background information about yourself and your household. You are a 54-year-old, have some college education, live in the South, and have a low numeracy score.

Here are your household finances, work situation, and expectations. Note that the current month is April, and that $ denotes amounts in U.S. dollars. Your household's total pre-tax income over the past 12 months was between $75,000 and $99,999. Your household is somewhat better off than it was 12 months ago, and you expect it to be somewhat better off 12 months from now. You expect inflation to be -1.0% over the next 12 months and 6.0% over the next three years. You work full-time.

You are also aware of recent national economic conditions. Over the four most recently completed calendar quarters, listed from oldest to most recent, U.S. real GDP changed by 0.62%, 0.40%, 0.18%, and 0.58% quarter over quarter; the unemployment rate was 5.43%, 5.10%, 5.03%, and 4.90%; overall consumer prices changed by 0.68%, 0.38%, -0.01%, and -0.06% quarter over quarter; and the effective federal funds rate was 0.12%, 0.14%, 0.16%, and 0.36%. Over the five years ending in the most recently completed quarter, compound average quarterly real GDP growth was 0.58%, the average unemployment rate was 6.95%, compound average quarterly CPI inflation was 0.34%, and the average effective federal funds rate was 0.12%.

For this question, suppose the following policy changes happened. Federal welfare benefits decrease, unemployment benefits increase, the payroll tax rate increases, and the average income tax rate increases.

For each policy change, what impact would it have on your own household? Use one of "very_negative", "somewhat_negative", "no_impact", "somewhat_positive", or "very_positive" for each item.

Return only a valid JSON array with exactly four string elements, in this order: welfare benefits; unemployment benefits; payroll tax rate; average income tax rate. Each element must be one of "very_negative", "somewhat_negative", "no_impact", "somewhat_positive", or "very_positive". Your output must follow this exact array structure: ["s_1", "s_2", "s_3", "s_4"]. Replace every s placeholder with one of the permitted strings. Do not include keys or explanatory text.

Assistant. ["somewhat_negative", "somewhat_negative", "somewhat_negative", "somewhat_negative"]

H.15 LABOR_CPS_JOBFIND

Target. Whether a person who is unemployed in the current CPS interview is employed, unemployed, or out of the labor force one month later. We obtain the outcome by matching the person to the next monthly CPS interview and use the survey's standard classification based on reported work, job search, availability, and reasons for not working.

Predictors. Current unemployment (duration, search or temporary-layoff status, timing of last full-time work, and worker class on the most recent job when available), respondent demographics (age, sex, race, education, marital status, Hispanic origin, nativity, citizenship, veteran status), geography (state, region, metropolitan status), household composition (household role, household size, own children), and core macro context (national economic conditions over the most recently completed quarters).

No-change baseline. The no-change prediction is that the person remains unemployed one month later.

Sample construction. Starting from the monthly CPS, we keep people aged 16 or older who are unemployed and scheduled to be interviewed again in the following month. We match each person uniquely to the expected next interview. Next, we drop rows that are missing the next-month labor-force-status target; any current demographic, household, geographic, or macroeconomic predictor; unemployment duration; search or layoff status; or time since last full-time work. Worker class on the most recent job is allowed to remain missing.

Prompt example

`labor_cps_jobfind`

System. You are an American person answering questions about your work, job search, and labor-force situation.

User. Here is some background information about yourself and your household. You are a 31-year-old multiracial woman with some college or an associate degree. You have never been married and live in New Jersey, in a metropolitan area in the Northeast. There are 5 people in your household, and you do not live with any of your own children. You are not Hispanic. You were born outside the United States and are a naturalized U.S. citizen. You are not a veteran.

Here is your current and recent labor-market situation. Note that the current month is December. You are currently unemployed and are seeking full-time work. You have been unemployed for 29 weeks. The last time you worked full-time for at least 2 consecutive weeks was more than 12 months ago. In your most recent job, you worked in the private sector.

You are also aware of recent national economic conditions. Over the four most recently completed calendar quarters, listed from oldest to most recent, U.S. real GDP changed by 0.46%, -0.16%, 0.95%, and 1.08% quarter over quarter; the unemployment rate was 4.13%, 4.13%, 4.20%, and 4.33%; overall consumer prices changed by 0.79%, 0.91%, 0.41%, and 0.76% quarter over quarter; and the effective federal funds rate was 4.65%, 4.33%, 4.33%, and 4.29%. Over the five years ending in the most recently completed quarter, compound average quarterly real GDP growth was 0.78%, the average unemployment rate was 4.30%, compound average quarterly CPI inflation was 1.11%, and the average effective federal funds rate was 3.04%.

Next month, what will be your work status? For this question, "employed" means working for pay or profit or being temporarily absent from a job, "unemployed" means not employed but available for work and actively searching or on temporary layoff, and "not_in_labor_force" means neither employed nor unemployed.

Return exactly one valid JSON array containing one string, using one of: "employed", "unemployed", "not_in_labor_force". Your output must follow this exact array structure: ["s_1"]. Replace every s placeholder with one of the permitted strings. Do not include explanatory text.

Assistant. ["not_in_labor_force"]

H.16 LABOR_CPS_RETIRE

Target. Whether a person who is employed in the current CPS interview is retired 12 months later. We match the person to the CPS interview one year later and classify the person as retired when they are out of the labor force and report retirement as the reason they are not working.

Predictors. Current work situation (actual and usual hours, worker class, multiple-job status, union coverage, weekly earnings, occupation, industry), household circumstances (family-income range, spouse or partner status and, when present, that person's age and labor-force status), health (functional difficulty), respondent demographics and geography (age, sex, race and ethnicity, education, marital status, state, region, metropolitan status, nativity), and core macro context (national economic conditions over the most recently completed quarters).

No-change baseline. The no-change prediction is that the person is not retired 12 months later.

Sample construction. Starting from CPS interviews from 1995 onward, we keep employed people between 50 and 75 years old who are scheduled to be interviewed again 12 months later. We match each person uniquely to the expected later interview. Next, we drop rows that are missing the later labor-force-status target, any required current demographic, household, geographic, or macroeconomic predictor, usual weekly hours, or hours worked in the previous week. Spouse or partner information, worker class, multiple-job and union status, weekly earnings, work-limiting difficulty, family income, occupation, and industry are allowed to remain missing. We also require sex and race to agree across the two interviews, age to advance consistently, and the person identifier to agree when it is available in both records. Finally, within each survey month, we trim the top and bottom percentiles of current weekly earnings when earnings are observed.

Prompt example

labor_cps_retire

System. You are an American person answering questions about your work, job search, and labor-force situation.

User. Here is some background information about yourself and your household. You are a 62-year-old Black woman with a high school diploma or GED. You are widowed and live in Alabama, in a metropolitan area in the South. There are 2 people in your household, including 1 of your own children. You are not Hispanic, were born in the United States, are a U.S. citizen, and are not a veteran. You do not live with a spouse or partner.

Here are your current work, resources, and health. Note that the current month is September, and that $ denotes amounts in U.S. dollars. You are currently employed by a private for-profit employer. You worked 40 hours last week, usually work 40 hours per week, and do not hold multiple jobs. Your union coverage and weekly earnings are not observed. Whether you have a functional difficulty is not observed. Your family-income category is not observed. Your occupation is Building and Grounds Cleaning and Maintenance, and your industry is Private households.

You are also aware of recent national economic conditions. Over the four most recently completed calendar quarters, listed from oldest to most recent, U.S. real GDP changed by 0.58%, 1.15%, 0.35%, and 0.30% quarter over quarter; the unemployment rate was 6.00%, 5.63%, 5.47%, and 5.67%; overall consumer prices changed by 0.93%, 0.58%, 0.73%, and 0.82% quarter over quarter; and the effective federal funds rate was 4.49%, 5.17%, 5.81%, and 6.02%. Over the five years ending in the most recently completed quarter, compound average quarterly real GDP growth was 0.59%, the average unemployment rate was 6.62%, compound average quarterly CPI inflation was 0.82%, and the average effective federal funds rate was 4.67%.

Will you be retired one year from now? Use "retired" if you will be retired and "not_retired" if you will not be retired.

Return exactly one valid JSON array containing one string: "retired" or "not_retired". Your output must follow this exact array structure: ["s_1"]. Replace every s placeholder with one of the permitted strings. Do not include explanatory text.

Assistant. ["not_retired"]

H.17 LABOR_CPS_SEPARATION

Target. Whether a person who is employed in the current CPS interview is employed, unemployed, or out of the labor force one month later. We obtain the outcome by matching the person to the next monthly CPS interview and use the survey's standard classification based on reported work, job search, availability, and reasons for not working.

Predictors. Current work situation (employment status, actual and usual hours, worker class, multiple-job status), respondent demographics (age, sex, race, education, marital status, Hispanic origin, nativity, citizenship, veteran status), geography (state, region, metropolitan status), household composition (household role, household size, own children), and core macro context (national economic conditions over the most recently completed quarters).

No-change baseline. The no-change prediction is that the person remains employed one month later.

Sample construction. Starting from the monthly CPS, we keep people aged 16 or older who are employed and scheduled to be interviewed again in the following month. We match each person uniquely to the expected next interview and require a valid identifier that appears only once in the baseline month. Next, we drop rows that are missing the next-month labor-force-status target; any current demographic, household, geographic, or macroeconomic predictor; actual weekly hours; or usual weekly hours. Worker class and multiple-job status are allowed to remain missing.

Prompt example

labor_cps_separation

System. You are an American person answering questions about your work, job search, and labor-force situation.

User. Here is some background information about yourself and your household. You are a 61-year-old White man with some college or an associate degree. You are divorced and live in Texas, in a metropolitan area in the South. There is 1 person in your household, and you do not live with any of your own children. You are not Hispanic, were born in the United States, are a U.S. citizen, and are not a veteran.

Here is your current job. Note that the current month is January. You are currently employed as self-employed in an incorporated business. You worked 30 hours last week, usually work 30 hours per week, and do not hold multiple jobs.

You are also aware of recent national economic conditions. Over the four most recently completed calendar quarters, listed from oldest to most recent, U.S. real GDP changed by 0.17%, 0.58%, 0.48%, and 1.36% quarter over quarter; the unemployment rate was 7.13%, 7.07%, 6.80%, and 6.63%; overall consumer prices changed by 0.73%, 0.72%, 0.46%, and 0.83% quarter over quarter; and the effective federal funds rate was 3.04%, 3.00%, 3.06%, and 2.99%. Over the five years ending in the most recently completed quarter, compound average quarterly real GDP growth was 0.57%, the average unemployment rate was 6.42%, compound average quarterly CPI inflation was 0.97%, and the average effective federal funds rate was 5.91%.

One month from now, what will be your work status? For this question, "employed" means working for pay or profit or being temporarily absent from a job, "unemployed" means not employed but available for work and actively searching or on temporary layoff, and "not_in_labor_force" means neither employed nor unemployed.

Return exactly one valid JSON array containing one string, using one of: "employed", "unemployed", "not_in_labor_force". Your output must follow this exact array structure: ["s_1"]. Replace every s placeholder with one of the permitted strings. Do not include explanatory text.

Assistant. ["employed"]

H.18 LABOR_SCE_OFFER

Target. Whether the respondent accepted or rejected a job offer. Respondents list as many as three of their best offers from the previous four months; we take the first listed, which is not necessarily the first one received. We treat the offer as accepted whether or not the respondent still works in that job, and omit respondents who were still deciding.

Predictors. Job-offer context (number of recent offers, annual salary of the first listed offer, whether that offer is full time or part time), reservation wage (the annual full-time wage the respondent would require), household finances and expectations (current and expected finances, one-year and longer-run inflation expectations), work situation (current work-status indicators), respondent demographics, resources, and geography (age, education, household income, region, numeracy), and core macro context (national economic conditions over the most recently completed quarters).

No-change baseline. The no-change prediction is that the respondent rejected the first listed job offer.

Sample construction. We merge the SCE Labor Market supplement with respondent information from the SCE Core survey and national economic variables from FRED-QD. We keep respondents with at least one job offer. Next, we drop rows that are missing the first-offer accept-or-reject target or any required demographic, income, or macroeconomic predictor. The offered salary, full-time or part-time status, reservation wage, household-finance and inflation-expectation measures, and work-status indicators are allowed to remain missing. We then require respondents to be between 18 and 100 years old and require offered annual pay and reservation wage to be at least $500 when those amounts are reported. Finally, within each survey month, we trim the top and bottom percentiles of offered pay and reservation wage when observed.

Prompt example

`labor_sce_offer`

System. You are an American person answering questions about your work, job search, and labor-force situation.

User. Here is some background information about yourself and your household. You are a 66-year-old, have a college education or more, live in the Midwest, and have a high numeracy score.

Here are your household finances, work situation, and expectations. Note that the current month is November, and that $ denotes amounts in U.S. dollars. Your household's total pre-tax income over the past 12 months was between $75,000 and $99,999. Your household is somewhat better off than it was 12 months ago, and you expect it to be somewhat better off 12 months from now. You expect inflation to be 2.1% over the next 12 months and 2.2% over the next three years. You work part-time.

You are also aware of recent national economic conditions. Over the four most recently completed calendar quarters, listed from oldest to most recent, U.S. real GDP changed by 0.87%, -0.35%, 1.29%, and 1.22% quarter over quarter; the unemployment rate was 6.93%, 6.67%, 6.20%, and 6.07%; overall consumer prices changed by 0.37%, 0.62%, 0.53%, and 0.26% quarter over quarter; and the effective federal funds rate was 0.09%, 0.07%, 0.09%, and 0.09%. Over the five years ending in the most recently completed quarter, compound average quarterly real GDP growth was 0.60%, the average unemployment rate was 8.24%, compound average quarterly CPI inflation was 0.49%, and the average effective federal funds rate was 0.12%.

For this question, consider the following job offer. You received at least one job offer in the last four months. The first offer was part-time and had an annual salary of $2,200. Your reservation wage for a full-time job was approximately $45,760 per year.

Thinking about the first job offer you received in the last four months, did you accept or reject that offer? Use "accepted" if you accepted the offer, whether or not you still work at that job. Use "rejected" if you did not accept it.

Return exactly one valid JSON array containing one string, using one of: "accepted", "rejected". Your output must follow this exact array structure: ["s_1"]. Replace every s placeholder with one of the permitted strings. Do not include explanatory text.

Assistant. ["rejected"]

H.19 LABOR_SCE_RISK

Target. The respondent's stated chances of three labor-market events: losing their current job over the next 12 months, leaving it voluntarily over the next 12 months, and finding a new job within three months if they lost their current job. Each answer comes directly from a survey question asking for a probability from 0 to 100 percent. For evaluation, we express each answer as the predicted probabilities that the event does and does not occur.

Predictors. Previous expectations (the same three stated chances one month earlier), current work situation (work-status indicators, number of paid jobs, and type of employment arrangement when available), household finances and health (current finances, expected finances, health when available), respondent demographics, resources, and geography (age, education, household income, region, numeracy), and core macro context (national economic conditions over the most recently completed quarters).

No-change baseline. For each event, the no-change prediction is the respondent's stated probability one month earlier.

Sample construction. We link each respondent's SCE Core survey record to their interview in the immediately preceding month and merge national economic variables from FRED-QD. Next, we drop rows that are missing one or more of the three probability targets, any of the corresponding first-history probabilities, or any required demographic, income, or macroeconomic predictor. Current work, household-finance, and health predictors are allowed to remain missing. We then require every probability to lie between 0 and 100 percent and respondents to be between 18 and 100 years old.

Prompt example

`labor_sce_risk`

System. You are an American person answering questions about your work, job search, and labor-force situation.

User. Here is some background information about yourself and your household. You are a 40-year-old, have a high-school education, live in the West, and have a low numeracy score.

Here are your household finances, work situation, and expectations. Note that the current month is August, and that $ denotes amounts in U.S. dollars. Your household's total pre-tax income over the past 12 months was between $10,000 and $19,999. Your household is somewhat worse off than it was 12 months ago, but you expect your household finances to be about the same 12 months from now. You work full-time and currently have 1 paid job, and you work for someone else. One month ago, you assigned probability 0.500 to losing your current or main job in the following 12 months, probability 0.490 to voluntarily leaving that job in the following 12 months, and probability 0.490 to finding an acceptable job within 3 months if you lost the job that month.

You are also aware of recent national economic conditions. Over the four most recently completed calendar quarters, listed from oldest to most recent, U.S. real GDP changed by 0.14%, 0.12%, 0.99%, and 0.27% quarter over quarter; the unemployment rate was 8.03%, 7.80%, 7.73%, and 7.53%; overall consumer prices changed by 0.45%, 0.66%, 0.40%, and -0.11% quarter over quarter; and the effective federal funds rate was 0.14%, 0.16%, 0.14%, and 0.12%. Over the five years ending in the most recently completed quarter, compound average quarterly real GDP growth was 0.22%, the average unemployment rate was 8.59%, compound average quarterly CPI inflation was 0.37%, and the average effective federal funds rate was 0.25%.

For your current or main job, what probability do you think there is that each of the following will happen: you lose the job in the next 12 months; you voluntarily leave the job in the next 12 months; and you find an acceptable job within 3 months if you lose the job this month?

Return only a valid JSON array containing exactly three two-element probability distributions, in this order: job loss in the next 12 months; voluntary job leaving in the next 12 months; acceptable job finding within 3 months conditional on job loss. Within each distribution, use this order: event does not occur, event occurs. Express each probability as a number from 0 to 1. Each distribution must sum to 1. Your output must follow this exact array structure: [[p_1_1, p_1_2], [p_2_1, p_2_2], [p_3_1, p_3_2]]. Replace every p placeholder with a number from 0 to 1. Do not include keys or explanatory text.

Assistant. [[0.5, 0.5], [0.49, 0.51], [0.5, 0.5]]

H.20 LABOR_SCE_SEARCH

Target. The respondent's stated chance of finding an acceptable job within three months and within 12 months. Both answers come directly from survey questions asked of people who are looking for work and are reported as probabilities from 0 to 100 percent. For evaluation, we express each answer as the predicted probabilities that the person does and does not find an acceptable job within the stated period.

Predictors. Previous expectations (the same two stated chances one month earlier), work and job-search situation (current work-status indicators, active job-search status, unemployment duration), household finances and health (current finances, expected finances, health when available), respondent demographics, resources, and geography (age, education, household income, region, numeracy), and core macro context (national economic conditions over the most recently completed quarters).

No-change baseline. For each horizon, the no-change prediction is the respondent's stated probability one month earlier.

Sample construction. We link each respondent's SCE Core survey record to their interview in the immediately preceding month and merge national economic variables from FRED-QD. We keep respondents who are currently looking for work. Next, we drop rows that are missing either probability target, either corresponding first-history probability, or any required demographic, income, or macroeconomic predictor. Current work and job-search details, household-finance measures, and health are allowed to remain missing. We then require every probability to lie between 0 and 100 percent and respondents to be between 18 and 100 years old.

Prompt example

`labor_sce_search`

System. You are an American person answering questions about your work, job search, and labor-force situation.

User. Here is some background information about yourself and your household. You are an 83-year-old, have a college education or more, live in the West, and have a low numeracy score.

Here are your household finances, work situation, and expectations. Your household's total pre-tax income over the past 12 months was between $30,000 and $39,999. Your household is somewhat worse off than it was 12 months ago, and you expect it to remain somewhat worse off 12 months from now.

Here is your current work and job-search situation. Note that the current month is July, and that $ denotes amounts in U.S. dollars. You are not working but would like to work. You are currently looking for a job and have been unemployed for 37 months. One month ago, you assigned probability 0.450 to finding an acceptable job within the following 12 months and probability 0.250 to finding one within the following 3 months.

You are also aware of recent national economic conditions. Over the four most recently completed calendar quarters, listed from oldest to most recent, U.S. real GDP changed by 0.14%, 0.12%, 0.99%, and 0.27% quarter over quarter; the unemployment rate was 8.03%, 7.80%, 7.73%, and 7.53%; overall consumer prices changed by 0.45%, 0.66%, 0.40%, and -0.11% quarter over quarter; and the effective federal funds rate was 0.14%, 0.16%, 0.14%, and 0.12%. Over the five years ending in the most recently completed quarter, compound average quarterly real GDP growth was 0.22%, the average unemployment rate was 8.59%, compound average quarterly CPI inflation was 0.37%, and the average effective federal funds rate was 0.25%.

What probability do you think there is that you will find an acceptable job within the coming 12 months? What probability do you think there is that you will find an acceptable job within the coming 3 months?

Return only a valid JSON array containing exactly two two-element probability distributions, in this order: acceptable job finding within the coming 12 months; acceptable job finding within the coming 3 months. Within each distribution, use this order: event does not occur, event occurs. Express each probability as a number from 0 to 1. Each distribution must sum to 1. Your output must follow this exact array structure: [[p_1_1, p_1_2], [p_2_1, p_2_2]]. Replace every p placeholder with a number from 0 to 1. Do not include keys or explanatory text.

Assistant. [[0.65, 0.35], [0.8, 0.2]]

H.21 LABOR_CPS_DISPLACE

Target. Whether a worker who reports a past job displacement is employed, unemployed, or out of the labor force at the time of the Displaced Worker Supplement. We take this outcome from the CPS's standard labor-force classification based on reported work, job search, availability, and reasons for not working.

Predictors. Displacement history (lookback period, reason and timing of job loss, advance notice, full-time status, union status when available, employer type, health-insurance provision, and other characteristics of the lost job), respondent demographics and background (age, sex, race and ethnicity, education, marital status, Hispanic origin, nativity, citizenship, veteran status), geography (state, region, metropolitan status), household composition (household role, household size, own children), and core macro context (national economic conditions over the most recently completed quarters).

No-change baseline. The no-change prediction is that the displaced worker is employed at the time of the supplement.

Sample construction. Starting from the Displaced Worker Supplement, we keep respondents who completed the displacement questions and report a qualifying job loss. Next, we drop rows that are missing the current labor-force-status target or any required current demographic, household, geographic, or macroeconomic predictor. The detailed reason and timing of displacement, advance notice, full-time status, union status, employer type, health-insurance provision, and other lost-job characteristics are allowed to remain missing.

Prompt example

`labor_cps_displace`

System. You are an American person answering questions about your work, job search, and labor-force situation.

User. Here is some background information about yourself and your household. You are a 40-year-old White woman with a bachelor's degree. You are divorced and live in Alabama, in a metropolitan area in the South. There is 1 person in your household, and you do not live with any of your own children. You are not Hispanic, were born in the United States, are a U.S. citizen, and are not a veteran.

You are also aware of recent national economic conditions. Over the four most recently completed calendar quarters, listed from oldest to most recent, U.S. real GDP changed by 1.40%, 1.70%, 0.83%, and 1.72% quarter over quarter; the unemployment rate was 6.23%, 5.93%, 5.07%, and 4.17%; overall consumer prices changed by 1.02%, 1.82%, 1.59%, and 2.18% quarter over quarter; and the effective federal funds rate was 0.08%, 0.07%, 0.09%, and 0.08%. Over the five years ending in the most recently completed quarter, compound average quarterly real GDP growth was 0.65%, the average unemployment rate was 5.08%, compound average quarterly CPI inflation was 0.71%, and the average effective federal funds rate was 1.09%.

Note that the current month is January. During the past three years, you lost or left a full-time job at a private for-profit employer because there was insufficient work. You last worked at that job two years ago and received less than 1 month of notice before the job loss. You were not a labor-union member at that job, and the job provided health insurance.

What is your current employment status?

Return exactly one valid JSON array containing one string, using one of: "employed", "unemployed", "not_in_labor_force". Your output must follow this exact array structure: ["s_1"]. Replace every s placeholder with one of the permitted strings. Do not include explanatory text.

Assistant. ["employed"]

H.22 LABOR_CPS_UI

This task is inspired by Farber et al. (2015).

Target. Whether an unemployed person in the current CPS interview is employed, unemployed, or out of the labor force one month later. We obtain the outcome by matching the person to the next monthly CPS interview and use the survey's standard classification based on reported work, job search, availability, and reasons for not working.

Predictors. Current unemployment (duration and whether the person is searching or on temporary layoff), recent work (worker class, occupation, and industry on the most recent job when available), respondent demographics and geography (age, sex, race and ethnicity, education, marital status, state, region, nativity, citizenship, veteran status), household composition (household role, household size, own children), calendar month, core macro context (national economic conditions over the most recently completed quarters), and unemployment-insurance policy (the number of weeks of regular state benefits, temporary federal extensions, and total available benefits in the person's state and month). Policy values are measured on the fifth day of each month, matching the timing of the CPS interview week.

No-change baseline. The no-change prediction is that the person remains unemployed one month later.

Sample construction. Starting from monthly CPS interviews from January 2008 through August 2014, we keep unemployed people aged 18–69 who report an eligible reason for unemployment and are scheduled to be interviewed again in the following month. We merge the unemployment-insurance durations in force in each state and month and national and state economic variables, then match each person uniquely to the next interview. Next, we drop rows that are missing the next-month labor-force-status target; any required current demographic, household, geographic, macroeconomic, or policy predictor; current unemployment duration and search status; or the occupation and industry of the most recent job. Time since last full-time work and worker class are allowed to remain missing. We also require consistent identifiers and demographic characteristics across interviews, a valid survey weight, and internally consistent nonnegative measures of regular, temporary, and total benefit duration.

Prompt example

`labor_cps_ui`

System. You are an American person answering questions about your work, job search, and labor-force situation.

User. Here is some background information about yourself and your household. You are a 41-year-old White man with a bachelor's degree. You have never been married and live in New York, in a metropolitan area in the Northeast. There is 1 person in your household, and you do not live with any of your own children. You are not Hispanic, were born in the United States, are a U.S. citizen, and are not a veteran.

Here is your current and recent labor-market situation. Note that the current month is January. You are currently unemployed and are seeking full-time work. You have been unemployed for 2 weeks. The last time you worked full-time for at least 2 consecutive weeks was within the last 12 months. In your most recent job, you were self-employed in an unincorporated business.

Here is some additional information about your unemployment spell and most recent job. Your unemployment began because a temporary job ended. You are currently actively looking for work. In your most recent job, your industry was construction, and your occupation was management occupations.

You are also aware of recent national economic conditions. Over the four most recently completed calendar quarters, listed from oldest to most recent, U.S. real GDP changed by 0.30%, 0.61%, 0.58%, and 0.63% quarter over quarter; the unemployment rate was 4.50%, 4.50%, 4.67%, and 4.80%; overall consumer prices changed by 0.98%, 1.13%, 0.63%, and 1.23% quarter over quarter; and the effective federal funds rate was 5.26%, 5.25%, 5.07%, and 4.50%. Over the five years ending in the most recently completed quarter, compound average quarterly real GDP growth was 0.76%, the average unemployment rate was 5.17%, compound average quarterly CPI inflation was 0.74%, and the average effective federal funds rate was 3.13%.

Under the rules in your state in the current month, a job loser who met the program requirements could potentially receive unemployment benefits for up to 26 weeks. This total consists of 26 weeks of regular state benefits and 0 weeks of temporary additional benefits. This maximum does not establish that you personally qualified, received benefits continuously, or had that many weeks remaining.

Next month, what will be your work status? For this question, "employed" means working for pay or profit or being temporarily absent from a job, "unemployed" means not employed but available for work and actively searching or on temporary layoff, and "not_in_labor_force" means neither employed nor unemployed.

Return exactly one valid JSON array containing one string, using one of: "employed", "unemployed", "not_in_labor_force". Your output must follow this exact array structure: ["s_1"]. Replace every s placeholder with one of the permitted strings. Do not include explanatory text.

Assistant. ["employed"]

H.23 LABOR_PSID_ADDEDWORKER

This task is inspired by Stephens (2002).

Target. Spouse or partner's annual work hours and labor earnings in the calendar year of the other adult's job loss and in the following calendar year. We take both measures from the PSID's annual questions about hours worked and labor earnings during the preceding year.

Predictors. Work and earnings history for both adults (work status, annual hours, and annual labor earnings in up to three earlier records), household resources (family income in those earlier records), household composition and demographics (each adult's age, sex, education, housing tenure, region, family size, number of children), job-loss circumstances (which adult lost the job, whether the loss

followed a layoff, firing, employer closure, or employer move, and earlier losses when observed), and core macro context (national economic conditions before and during the job-loss period).

No-change baseline. For both target years, the no-change prediction is the nondisplaced adult's annual work hours and labor earnings before the job loss.

Sample construction. Starting from PSID couples observed from 1968–1992, we link the household's interviews before the job loss, during the job-loss period, and in the following period. We keep households in which exactly one adult reports a layoff, firing, plant closure, or employer move after having worked, both adults are observed throughout the relevant interviews, and neither adult reports another job loss in the adjacent periods. Next, we drop rows that are missing one or more of the four labor-supply targets; either adult's current age, sex, or family-composition predictors; either adult's first-history work status, hours, or earnings; first-history family income; or the macroeconomic variables. Current education, housing tenure, and region, prior job-loss details, and the second and third histories are allowed to remain missing. We also require both adults to be between 25 and 65 years old in all three periods, the couple to remain together, and both adults to have answered the job-loss questions. Finally, within each survey year, we trim the top and bottom percentiles of every family-income value shown in the household's history.

Prompt example

`labor_psid_addedworker`

System. You are an American adult making decisions about work, income, and household finances.

User. Here is some background information about yourself and your household. You are a 31-year-old woman. Your spouse or partner is a 30-year-old man. Your household has 5 people, including 3 children. You completed 12 years of schooling. Your spouse or partner completed 14 years of schooling. You currently rent your home. You live in the South.

Here is your household's work and income history. Note that $ denotes amounts in U.S. dollars. Three years ago, you worked 2,080 hours in total and had $6,000 in annual labor earnings. Your spouse or partner worked 2,000 hours in total and had $7,800 in annual labor earnings. Your total family income was $13,800. Two years ago, you worked 2,080 hours in total and had $6,400 in annual labor earnings. Your spouse or partner worked 2,000 hours in total and had $7,000 in annual labor earnings. Your total family income was $13,400. One year ago, you worked 2,000 hours in total and had $6,300 in annual labor earnings. Your spouse or partner worked 2,600 hours in total and had $9,000 in annual labor earnings. Your total family income was $15,300.

You are also aware of recent national economic conditions. Over the four most recently completed calendar quarters, listed from oldest to most recent, U.S. real GDP changed by -1.22%, 0.71%, 1.71%, and 1.35% quarter over quarter; the unemployment rate was 8.27%, 8.87%, 8.47%, and 8.30%; overall consumer prices changed by 2.14%, 1.20%, 2.01%, and 1.84% quarter over quarter; and the effective federal funds rate was 6.30%, 5.42%, 6.16%, and 5.41%. Over the five years ending in the most recently completed quarter, compound average quarterly real GDP growth was 0.78%, the average unemployment rate was 6.11%, compound average quarterly CPI inflation was 1.68%, and the average effective federal funds rate was 6.83%.

During the current calendar year, your spouse or partner lost their job because of a layoff or firing.

The first two outcomes below cover the entire current calendar year. They can therefore include work before as well as after your spouse or partner lost their job. The last two outcomes cover one year from now.

Given this information, what are your annual work hours and labor earnings in the current calendar year, and what will they be in one year?

Return only a valid JSON array with exactly four numbers, in this order: your annual work hours in the current calendar year; your annual labor earnings in the current calendar year; your annual work hours in one year; your annual labor earnings in one year. Earnings must be in current U.S. dollars. Your output must follow this exact array structure: [v_1, v_2, v_3, v_4]. Replace every v placeholder with a single number. Do not include dollar signs, commas, keys, or explanatory text.

Assistant. [1760.0,8300.0,1984.0,8200.0]

H.24 LABOR_SCE_RESWAGE

Target. The lowest hourly wage the respondent would accept for a new job and the number of hours per week they would prefer to work at that wage. Both targets come directly from the SCE Job Search supplement. When the respondent reports an acceptable wage by week or year instead of by hour, we convert it to an hourly amount using the reported preferred hours.

Predictors. Work and job-search situation (current work-status indicators, number of current jobs when available, duration of job search, months since last paid work, whether the respondent has ever had paid work), pay benchmark (hourly-equivalent pay and usual weekly hours in the current job or,

when unavailable, the most recent job), household finances and expectations (current and expected finances, one-year and longer-run inflation expectations), respondent demographics, resources, and household context (age, education, household income, region, numeracy, household size, children under 18), and core macro context (national economic conditions over the most recently completed quarters).

No-change baseline. The no-change prediction uses hourly pay and weekly hours in the respondent's current job, or in the last job if the respondent is not currently working.

Sample construction. We merge the SCE Job Search supplement with respondent information from the SCE Core survey and national economic variables from FRED-QD, and convert all reported pay amounts to hourly values. Next, we drop rows that are missing either target, any required demographic, income, or macroeconomic predictor, or a complete pay-and-hours benchmark from either the current job or the most recent job. Household size, number of children, household-finance and inflation-expectation measures, and other work and job-search details are allowed to remain missing. We then require the acceptable wage to be positive, preferred weekly hours to be between 1 and 168, and respondents to be between 18 and 100 years old. Finally, within each survey month, we trim the top and bottom percentiles of the acceptable wage and current-job or last-job hourly pay.

Prompt example

`labor_sce_reswage`

System. You are an American person answering questions about your work, job search, and labor-force situation.

User. Here is some background information about yourself and your household. You are a 40-year-old, have a high-school education, live in the West, and have a low numeracy score. Your household has 4 people, including 2 children under 18.

Here are your household finances, work situation, and expectations. Note that the current month is October, and that $ denotes amounts in U.S. dollars. Your household's total pre-tax income over the past 12 months was between $10,000 and $19,999. Your household is somewhat worse off than it was 12 months ago, and you expect it to be about the same 12 months from now. You expect inflation to be 2.2% over the next 12 months and 6.5% over the next three years. You work full-time. You recently spent 14 days searching for work. Your current or main job pays approximately $9.00 per hour, and you usually work 40 hours per week.

You are also aware of recent national economic conditions. Over the four most recently completed calendar quarters, listed from oldest to most recent, U.S. real GDP changed by 0.12%, 0.99%, 0.27%, and 0.85% quarter over quarter; the unemployment rate was 7.80%, 7.73%, 7.53%, and 7.23%; overall consumer prices changed by 0.66%, 0.40%, -0.11%, and 0.54% quarter over quarter; and the effective federal funds rate was 0.16%, 0.14%, 0.12%, and 0.08%. Over the five years ending in the most recently completed quarter, compound average quarterly real GDP growth was 0.29%, the average unemployment rate was 8.65%, compound average quarterly CPI inflation was 0.32%, and the average effective federal funds rate was 0.16%.

For this question, consider the following job preferences.

What is the lowest hourly wage, before taxes and other deductions, that you would accept for a job in a line of work you would consider? Assuming you could find suitable work, how many hours per week would you prefer to work on a new job?

Return only a valid JSON array with exactly two numeric elements, in this order: hourly reservation wage; preferred weekly work hours. Your output must follow this exact array structure: [v_1, v_2]. Replace every v placeholder with a single number. Do not include keys or explanatory text.

Assistant. [9.81, 40.0]

H.25 MACRO_MICH_OUTLOOK

Target. The respondent's expected inflation rate over the next 12 months and their expected average annual inflation rate over the next five to ten years. For each horizon, the Michigan Survey first asks whether prices are expected to rise, fall, or remain unchanged and then asks by what percentage. We combine the direction and magnitude into a signed inflation expectation.

Predictors. Previous expectations (the same two inflation expectations from the respondent's earlier interview six to eight months before), household and national assessments (personal finances and national business conditions compared with one year earlier, assessment of government economic policy), buying conditions (major household items, homes, and vehicles when available), household resources (annual income), household composition and respondent demographics (age, sex, education, marital or partner status, number of adults, number of children), geography (region), and core macro context (national economic conditions over the most recently completed quarters).

No-change baseline. For each horizon, the no-change prediction is the respondent's inflation expectation from the earlier interview.

Sample construction. We link each respondent to a unique earlier Michigan Survey interview conducted six, seven, or eight months before and merge national economic variables from FRED-QD. We keep pairs in the correct chronological order whose earlier interview matches the survey's recorded link and was available by the date of the current response. Next, we drop rows that are missing either current inflation-expectation target, either corresponding first-lag expectation, or any required current predictor. Assessments of whether it is a good time to buy a home, sell a house, or buy a vehicle are allowed to remain missing. We also require the categorical responses shown to the model to have valid meanings. Finally, within each survey month, we trim the top and bottom percentiles of both current inflation expectations and annual household income, and apply the same trimming rule to each earlier expectation within its earlier survey month.

Prompt example

`macro_mich_outlook`

System. You are the head of an American household forming expectations about economic conditions, both for the country in general and for your household in particular.

User. Here is some background information about yourself and your household. You are a 62-year-old man living in the West. You are married or partnered, and your highest completed education is some college without a college degree. Your household has 2 adults and no children.

Here are your current views on your finances and the economy. Note that the current month is November, and that $ denotes amounts in U.S. dollars. Your current household income is $29,000 per year. You report that your personal finances are better than they were a year ago, while national business conditions are worse than they were a year ago. You rate the government's handling of economic policy as only fair. You think now is a good time to buy major household items, buy a home, sell a house, and buy a vehicle.

6 months ago, your expected price change over the next 12 months was 2.0%. Your expected average annual price change over the next 5 to 10 years was 2.0%.

You are also aware of recent national economic conditions. Over the four most recently completed calendar quarters, listed from oldest to most recent, U.S. real GDP changed by 0.35%, 1.20%, 1.08%, and 0.99% quarter over quarter; the unemployment rate was 7.10%, 7.37%, 7.60%, and 7.63%; overall consumer prices changed by 0.83%, 0.68%, 0.77%, and 0.76% quarter over quarter; and the effective federal funds rate was 4.82%, 4.02%, 3.77%, and 3.26%. Over the five years ending in the most recently completed quarter, compound average quarterly real GDP growth was 0.66%, the average unemployment rate was 6.06%, compound average quarterly CPI inflation was 1.05%, and the average effective federal funds rate was 7.01%.

Over the next 12 months, by what percent do you expect prices in general to change? Over the next 5 to 10 years, by what percent per year on average do you expect prices in general to change? Use a negative number if you expect prices to fall.

Return only a valid JSON array with exactly 2 elements, in this order: expected_inflation_1y, expected_inflation_5_to_10y_annual. Both elements must be numbers in percent. Your output must follow this exact array structure: [v_1, v_2]. Replace every v placeholder with a single number. Do not include keys or explanatory text.

Assistant. [3.0, 2.0]

H.26 MACRO_SCE_UNCERTAINTY

Target. Five probability vectors reported by the respondent: inflation over the next year, inflation over the 12-month period two to three years ahead, and whether U.S. unemployment, the average savings-account interest rate, and U.S. stock prices will be higher one year ahead. For each inflation horizon, the survey asks the respondent to divide 100 percentage points across ten possible inflation ranges. For each of the other three outcomes, it asks directly for the chance that the outcome will be higher.

Predictors. Previous expectations (all five probability vectors from one month earlier), household finances and work situation (current household finances, current work-status indicators), respondent demographics, resources, and geography (age, education, household before-tax income bracket, region, numeracy), and core macro context (real GDP growth, unemployment rate, headline CPI inflation, and the effective federal funds rate over the four most recently completed quarters, plus five-year quarterly references).

No-change baseline. The no-change prediction carries forward all five probability vectors from one month earlier.

Sample construction. We link each respondent to their SCE interview in the immediately preceding month and merge national economic variables from FRED-QD. Next, we drop rows that are missing one or more target answers, any of the five first-history probability vectors, or any required demographic, income, or macroeconomic predictor. Current household-finance and work-status predictors are allowed to remain missing. We require every probability to lie between 0 and 100 percent, the probabilities over the inflation ranges to sum to 100 percent at each horizon apart from rounding, and respondents to be between 18 and 100 years old. Finally, within each survey month, we trim the top and bottom percentiles of the interquartile range implied by each respondent's inflation probabilities.

Prompt example

`macro_sce_uncertainty`

System. You are an American adult making household decisions and forming expectations about your finances, work, housing, and the economy.

User. Here is some background information about yourself and your household. You are a 66-year-old, have a college education or more, live in the Northeast, and have a high numeracy score.

Here are your household finances, work situation, and expectations. Note that the current month is July, and that $ denotes amounts in U.S. dollars. Your household's total pre-tax income over the past 12 months was between $100,000 and $149,999. Your household is somewhat better off than it was 12 months ago, but you expect your household finances to be somewhat better off 12 months from now. One month ago, your probabilities for inflation over the following 12 months, in the same ten-bin order used below, were [0.1, 0.15, 0.25, 0.5, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0]. Your probabilities for the 12-month period two to three years ahead were [0.15, 0.2, 0.5, 0.15, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0]. You assigned probability 0.400 to the U.S. unemployment rate being higher 12 months later, probability 0.700 to the average interest rate on savings accounts being higher, and probability 0.500 to U.S. stock prices being higher.

You are also aware of recent national economic conditions. Over the four most recently completed calendar quarters, listed from oldest to most recent, U.S. real GDP changed by 0.14%, 0.12%, 0.99%, and 0.27% quarter over quarter; the unemployment rate was 8.03%, 7.80%, 7.73%, and 7.53%; overall consumer prices changed by 0.45%, 0.66%, 0.40%, and -0.11% quarter over quarter; and the effective federal funds rate was 0.14%, 0.16%, 0.14%, and 0.12%. Over the five years ending in the most recently completed quarter, compound average quarterly real GDP growth was 0.22%, the average unemployment rate was 8.59%, compound average quarterly CPI inflation was 0.37%, and the average effective federal funds rate was 0.25%.

Now think about the different things that may happen to inflation. For each of the following ten ranges, what probability do you think there is that inflation over the next 12 months will fall in that range? Use this order: inflation at least 12 percent, inflation 8 to 12 percent, inflation 4 to 8 percent, inflation 2 to 4 percent, inflation 0 to 2 percent, deflation 0 to 2 percent, deflation 2 to 4 percent, deflation 4 to 8 percent, deflation 8 to 12 percent, deflation at least 12 percent. Then, for the 12-month period two to three years from now, what probability do you think there is that inflation will fall in each of the same ranges? Finally, what probability do you think there is that 12 months from now the U.S. unemployment rate will be higher than it is now, that the average interest rate on savings accounts will be higher than it is now, and that U.S. stock prices will be higher than they are now?

Return only a valid JSON array containing exactly five probability distributions, in this order: the ten-element one-year inflation distribution; the ten-element two-to-three-year inflation distribution; the two-element higher-unemployment distribution; the two-element higher-savings-rate distribution; the two-element higher-stock-price distribution. Preserve the inflation-bin order stated above. Within each two-element distribution, use this order: event does not occur, event occurs. Express each probability as a number from 0 to 1. Each distribution must sum to 1. Your output must follow this exact array structure: [[p_1_1, ..., p_1_10], [p_2_1, ..., p_2_10], [p_3_1, p_3_2], [p_4_1, p_4_2], [p_5_1, p_5_2]]. Replace every p placeholder with a number from 0 to 1. Do not include keys or explanatory text.

Assistant. [[0.0, 0.0, 0.75, 0.2, 0.05, 0.0, 0.0, 0.0, 0.0, 0.0], [0.0, 0.0, 0.05, 0.65, 0.3, 0.0, 0.0, 0.0, 0.0, 0.0], [0.84, 0.16], [0.25, 0.75], [0.44, 0.56]]

H.27 MACRO_SCE_REVISION

This task is inspired by Armantier et al. (2022).

Target. The respondent's inflation expectations in their twelfth monthly SCE interview: expected inflation over the next 12 months and expected average annual inflation over the 12-month period two to three years ahead. For both horizons, we use the mean implied by the probabilities the respondent assigns to different possible inflation ranges.

Predictors. Earlier expectations (both inflation expectations from the respondent's first monthly interview), respondent demographics and resources at the final interview (age, education, household income, region, numeracy, household finances, work situation), core macro context (national economic conditions over the most recently completed quarters), and inflation since the first interview (the observed change in the seasonally adjusted U.S. consumer price index over the intervening 11 months, the earlier one-year expectation converted to the same 11-month period, and the difference

between the two). The task therefore asks how the respondent revises expectations after observing inflation during the panel.

No-change baseline. For each horizon, the no-change prediction is the respondent's inflation expectation in their first monthly interview.

Sample construction. We link each respondent's first and twelfth SCE interviews, require the two interviews to be exactly 11 months apart, and merge national economic variables from FRED-QD and the seasonally adjusted U.S. consumer price index. We calculate inflation over the 11 completed months between the interviews, convert the earlier 12-month expectation to an 11-month rate using constant monthly growth, and subtract the converted expectation from observed inflation. Next, we drop pairs that are missing either final-expectation target, either first-history expectation, any required final-interview demographic, income, or macroeconomic predictor, either price-index value, or any of the three inflation-comparison measures. Final-interview household-finance and work-status predictors are allowed to remain missing. We also require the earlier one-year expectation to exceed -100 percent and respondents to be between 18 and 100 years old. Finally, within each survey month, we trim the top and bottom percentiles of both final inflation expectations.

Prompt example

`macro_sce_revision`

System. You are an American adult making household decisions and forming expectations about your finances, work, housing, and the economy.

User. Here is some background information about yourself and your household. You are a 28-year-old, have a college education or more, live in the Midwest, and have a low numeracy score.

Here are your household finances, work situation, and expectations. Your household's total pre-tax income over the past 12 months was between $10,000 and $19,999. Your household is about as well off as it was 12 months ago, and you expect your household finances to remain about the same 12 months from now.

You are also aware of recent national economic conditions. Over the four most recently completed calendar quarters, listed from oldest to most recent, U.S. real GDP changed by 0.27%, 0.85%, 0.87%, and -0.35% quarter over quarter; the unemployment rate was 7.53%, 7.23%, 6.93%, and 6.67%; overall consumer prices changed by -0.11%, 0.54%, 0.37%, and 0.62% quarter over quarter; and the effective federal funds rate was 0.12%, 0.08%, 0.09%, and 0.07%. Over the five years ending in the most recently completed quarter, compound average quarterly real GDP growth was 0.48%, the average unemployment rate was 8.57%, compound average quarterly CPI inflation was 0.52%, and the average effective federal funds rate was 0.13%.

Now consider how actual consumer prices have evolved compared with your expectations over recent months. Note that the current month is May, and that $ denotes amounts in U.S. dollars. Eleven months ago, you expected consumer prices to increase by 12.4% over the following 12 months, and to increase by 10.3% over the 12-month period two to three years ahead. Assuming a constant monthly expected inflation rate over the following 12 months, your first expectation corresponds to an expected increase of 11.3% over the first 11 months of that period. Actual consumer prices increased by 2.0% over those 11 months. This was 9.3 percentage points below the implied 11-month expectation.

Given this information, by what percentage do you now expect consumer prices to change over the next 12 months? By what percentage do you now expect them to change over the 12-month period starting two years from now?

Return only a valid JSON array with exactly two numbers, in this order: expected consumer-price change over the next 12 months; expected consumer-price change over the 12-month period starting two years from now. Use a negative number if you expect consumer prices to fall. Your output must follow this exact array structure: [v_1, v_2]. Replace every v placeholder with a single number. Do not include keys or explanatory text.

Assistant. [5.071, 3.569]

H.28 HOUSE_CENSUS_MOVE

Target. Whether an adult lives in the same home as five years earlier, a different home in the same state, or a different state. We construct the three categories from the Census questions on residence five years ago and current and previous state of residence.

Predictors. Respondent characteristics five years earlier (age, sex, race and ethnicity, education), location five years earlier (state of residence), economic conditions in that state (annual unemployment rate, per-capita personal income growth, house-price growth), and core macro context before the five-year period (real GDP growth, unemployment rate, headline CPI inflation, and the effective federal funds rate over the preceding four quarters, plus five-year quarterly references).

No-change baseline. The no-change prediction is that the respondent remains in the same home.

Sample construction. Starting from the 1990 and 2000 Census samples, we use each person's reported residence five years earlier to identify the origin state, then merge state unemployment, income, and house-price data and national economic variables. We keep people who were at least 30 years old at the start of the five-year period. Next, we drop rows that are missing the migration target or any current predictor. No target or predictor shown to the model is allowed to remain missing.

Prompt example

`house_census_move`

System. You are an American adult making decisions about housing and residential location.

User. Here is some background information about yourself and where you lived five years ago. Note that the current month is April. Five years ago, you were a 64-year-old mexican, Hispanic woman with grade 5, 6, 7, or 8. Your state of residence five years ago was Texas.

You are also aware of national economic conditions before that five-year period began. Over the four calendar quarters immediately before that period, listed from oldest to most recent, U.S. real GDP changed by 1.95%, 1.73%, 0.96%, and 0.82% quarter over quarter; the unemployment rate was 7.87%, 7.43%, 7.43%, and 7.30%; overall consumer prices changed by 1.42%, 0.94%, 0.87%, and 0.86% quarter over quarter; and the effective federal funds rate was 9.69%, 10.56%, 11.39%, and 9.27%. Over the five years ending in the last of those four quarters, compound average quarterly real GDP growth was 0.64%, the average unemployment rate was 8.32%, compound average quarterly CPI inflation was 1.64%, and the average effective federal funds rate was 12.26%.

You are also aware of annual state and national conditions in the completed calendar year immediately before that five-year period began. Unemployment in your state was 6.10%, compared with 7.51% nationally; per-capita personal income was growing at 7.60% in your state and 9.35% nationally; and house prices were growing at 0.98% in your state and 4.47% nationally.

Compared with where you lived five years ago, do you now live at the same residence, at a different residence in the same state, or in a different state?

Return only a valid JSON array containing exactly one string, using one of: "stay", "move_within_state", "move_to_different_state". Your output must follow this exact array structure: ["s_1"]. Replace every s placeholder with one of the permitted strings. Do not include explanatory text.

Assistant. ["stay"]

H.29 HOUSE_PSID_OWNER

Target. Whether a household that currently rents owns a home in its next PSID interview, usually two years later and sometimes one year later in the early part of the sample. We use the housing-tenure answer in the matched later interview and classify the household as an owner or nonowner.

Predictors. Household resources (income and available wealth components), household history (up to three prior samples of income, wealth, and housing tenure), household composition (presence of a spouse or partner, marital status, household size, number of children), respondent demographics and geography (age, sex, race and ethnicity, region, metropolitan status when available), core macro context (real GDP growth, unemployment rate, headline CPI inflation, and the effective federal funds rate over the most recently completed quarters, plus five-year quarterly references), and financial and housing-market context (house-price growth, S&P 500 growth, and the 30-year mortgage rate over the same periods).

No-change baseline. The no-change prediction is that the household remains a renter in the next PSID interview.

Sample construction. Starting from our clean PSID household panel, we keep current renters between 25 and 75 years old and link each household to its next consecutive interview. We merge national economic, financial-market, and housing-market variables. Next, we drop rows that are missing the next-interview homeownership target or any required current predictor: age, sex, race or ethnicity, region, housing tenure, spouse or partner status, marital status, family size, number of children, family income, or the core macroeconomic variables. Metropolitan status, current wealth fields, changes since the previous interview, all prior household histories, and the financial- and housing-market variables are allowed to remain missing. Finally, within each survey year, we trim the top and bottom percentiles of current and available lagged family income and of current net worth when observed.

Prompt example

`house_psid_owner`

System. You are an American adult making decisions about housing and residential location.

User. Here is some background information about yourself and your household. You are a 57-year-old White woman living in the South, in a metropolitan area. You do not have a spouse or partner in your household and are not legally married. Your household has 1 person and no children. Compared with two years ago, household size has fallen by 1, the number of children is unchanged, partner status is unchanged, and marital status is unchanged.

Here are your household's housing, finances, and past household records. Note that $ denotes amounts in U.S. dollars. You currently rent your home. Your current household income is $77,187, and your total net worth is $1,269. You have $0 in savings, $0 in financial investments, and $0 in home equity. Six years ago, your household rented its home, household income was $72,924, total net worth was $23,610, savings were $5,555, and financial investments were $0. Four years ago, your household rented its home, household income was $25,142, total net worth was $13,307, savings were $0, and financial investments were $0. Two years ago, your household rented its home, household income was $71,953, total net worth was $5,157, savings were $5,157, and financial investments were $0.

You are also aware of recent national economic conditions. Over the four most recently completed calendar quarters, listed from oldest to most recent, U.S. real GDP changed by -0.35%, 1.29%, 1.22%, and 0.51% quarter over quarter; the unemployment rate was 6.67%, 6.20%, 6.07%, and 5.70%; overall consumer prices changed by 0.62%, 0.53%, 0.26%, and -0.25% quarter over quarter; and the effective federal funds rate was 0.07%, 0.09%, 0.09%, and 0.10%. Over the five years ending in the most recently completed quarter, compound average quarterly real GDP growth was 0.57%, the average unemployment rate was 8.03%, compound average quarterly CPI inflation was 0.44%, and the average effective federal funds rate was 0.12%. You are also aware of recent financial and housing-market conditions. Over the same four completed quarters, listed from oldest to most recent, U.S. house prices changed by 0.49%, 1.50%, 1.06%, and 0.54% quarter over quarter; the S&P 500 changed by 3.61%, 3.60%, 3.98%, and 1.83% quarter over quarter; and the average 30-year mortgage rate was 4.36%, 4.23%, 4.14%, and 3.97%. Over the same five-year period, compound average quarterly house-price growth was -0.15%, compound average quarterly S&P 500 growth was 3.12%, and the average 30-year mortgage rate was 4.19%.

In two years, will your household own its home?

Return only a valid JSON array containing exactly one string, using one of: "yes", "no". Your output must follow this exact array structure: ["s_1"]. Replace every s placeholder with one of the permitted strings. Do not include explanatory text.

Assistant. ["no"]

H.30 HOUSE_SCE_MOVE

Target. The respondent's stated chance of moving to a different primary residence over the next 12 months. The answer comes directly from an SCE question asking for a probability from 0 to 100 percent. For evaluation, we express it as the predicted probabilities that the respondent moves and does not move.

Predictors. Previous expectation (the same stated moving probability one month earlier), residence and moving circumstances (current residence and moving-related answers when available), household finances and expectations (current and expected finances), work and job-search situation (current work-status indicators and search information when available), respondent demographics, resources, and geography (age, education, household before-tax income bracket, region, numeracy), and core macro context (real GDP growth, unemployment rate, headline CPI inflation, and the effective federal funds rate over the four most recently completed quarters, plus five-year quarterly references).

No-change baseline. The no-change prediction is the respondent's stated moving probability one month earlier.

Sample construction. We link each SCE Core survey record to the respondent's interview in the immediately preceding month and merge national economic variables from FRED-QD. Next, we drop rows that are missing the current moving-probability target, the corresponding first-history probability, or any required demographic, income, or macroeconomic predictor. Current residence and moving circumstances, household-finance measures, work-status indicators, and job-search information are allowed to remain missing. We then require both probabilities to lie between 0 and 100 percent and respondents to be between 18 and 100 years old.

Prompt example

house_sce_move

System. You are an American adult making decisions about housing and residential location.

User. Here is some background information about yourself and your household. You are a 58-year-old, have some college education, live in the West, and have a low numeracy score.

Here are your household finances, work situation, expectations, and housing. Note that the current month is July, and that $ denotes amounts in U.S. dollars. Your household's total pre-tax income over the past 12 months was between $30,000 and $39,999. Your household is about the same than it was 12 months ago, and you expect your household finances to be somewhat better off 12 months from now. You have not moved recently. One month ago, you thought the probability that you would move to a different primary residence over the following 12 months was 0.300.

You are also aware of recent national economic conditions. Over the four most recently completed calendar quarters, listed from oldest to most recent, U.S. real GDP changed by 0.14%, 0.12%, 0.99%, and 0.27% quarter over quarter; the unemployment rate was 8.03%, 7.80%, 7.73%, and 7.53%; overall consumer prices changed by 0.45%, 0.66%, 0.40%, and -0.11% quarter over quarter; and the effective federal funds rate was 0.14%, 0.16%, 0.14%, and 0.12%. Over the five years ending in the most recently completed quarter, compound average quarterly real GDP growth was 0.22%, the average unemployment rate was 8.59%, compound average quarterly CPI inflation was 0.37%, and the average effective federal funds rate was 0.25%.

What probability do you think there is that over the next 12 months you will move to a different primary residence?

Return only a valid JSON array containing exactly one two-element probability distribution. Use this order: probability of not moving, probability of moving. Express each probability as a number from 0 to 1. The distribution must sum to 1. Your output must follow this exact array structure: [[p_1_1, p_1_2]]. Replace every p placeholder with a number from 0 to 1. Do not include keys or explanatory text.

Assistant. [[1.0, 0.0]]

H.31 HOUSE_SCE_FINANCING

This task is inspired by Fuster & Zafar (2021).

Target. The respondent's maximum home purchase price and down payment under three financing scenarios: freedom to choose any down payment at the original mortgage rate, a mortgage rate two percentage points above or below the original rate, and receipt of a cash inheritance. The six targets are the direct dollar answers to the price and down-payment questions under these three scenarios.

Predictors. Initial housing choice (maximum purchase price, down payment, mortgage rate, monthly payment, comparable-home value), household housing and financial circumstances (tenure, savings, debt, credit-score range, moving expectations, views about property investment), respondent preferences and characteristics (risk tolerance, numeracy, demographics, income, location), and core macro context (national economic conditions over the most recently completed quarters). The experiment randomly assigns an initial mortgage rate of 4.5 or 6.5 percent, then allows a flexible down payment, changes the rate by two percentage points, and adds an inheritance. The model sees the initial choice and predicts all six later answers together.

No-change baseline. The no-change prediction repeats the respondent's initial maximum price and down payment under each of the three later scenarios.

Sample construction. Starting from the February 2014 home-financing experiment data from Fuster & Zafar (2021)'s replication package, we add the survey release date and national economic variables available before the survey. Next, we drop rows that are missing one or more of the six targets or any current predictor. Current home value and housing debt are required for homeowners but are allowed to remain missing for renters; no other predictor shown to the model is allowed to remain missing. We require every stated purchase price to be positive, each down payment to be at least the experiment's rounded 5 percent minimum and no greater than the stated price, and all answers to lie within the ranges allowed by the questionnaire. Finally, we trim the top and bottom percentiles of all the monetary variables.

Prompt example

`house_sce_financing`

System. You are an American adult making decisions about housing, borrowing, and household finances.

User. Here is some background information about yourself and your household. You are a 40-year-old woman, are married, and your highest completed education is less than a college degree. You live in the West and have 2 children under 18 and 1 adult child living with you. You answered 2 of 5 numeracy questions correctly.

Here are your household finances, housing situation, preferences, and expectations. Note that $ denotes amounts in U.S. dollars. Your household income is between $10,000 and $20,000. You rent your current home. Your liquid savings are less than $5,000, your non-housing debt is between $5,000 and $30,000, and your credit-score band is below 620. A typical comparable home costs about $230,000. On a scale from 1 to 10, your willingness to take risks is 2. You give a 50.0% chance of moving within the next three years. You do not regard property in your ZIP code as a good investment.

You are also aware of recent national economic conditions. Over the four most recently completed calendar quarters, listed from oldest to most recent, U.S. real GDP changed by 0.99%, 0.27%, 0.85%, and 0.87% quarter over quarter; the unemployment rate was 7.73%, 7.53%, 7.23%, and 6.93%; overall consumer prices changed by 0.40%, -0.11%, 0.54%, and 0.37% quarter over quarter; and the effective federal funds rate was 0.14%, 0.12%, 0.08%, and 0.09%. Over the five years ending in the most recently completed quarter, compound average quarterly real GDP growth was 0.44%, the average unemployment rate was 8.65%, compound average quarterly CPI inflation was 0.45%, and the average effective federal funds rate was 0.14%.

Imagine that you move to a comparable city and are considering a comparable home where you plan to remain for a long time. You do not need to sell a current home. With a 20% down payment and a 30-year mortgage rate of 4.5%, the most you would pay for the home is $60,000, implying a down payment of $12,000. This is equivalent to an effective total monthly payment of about $827 for a 30-year fixed-rate loan, including principal and interest, estimated mortgage-interest tax savings assuming that 95% of interest is deductible at a 34% income tax rate, and maintenance, property taxes, and homeowners insurance equal in total to 3.35% of the comparable-home value per year. For reference, the principal-and-interest payment alone on each $100,000 borrowed is about $507 per month at 4.5% and $632 per month at 6.5%.

Now consider three further situations, in this order.

Situation 1: The mortgage rate remains 4.5%, but you may choose any down payment of at least 5% of the home price. You receive no extra cash before purchase.

Situation 2: The mortgage rate is 6.5%, and you may again choose any down payment of at least 5% of the home price. You receive no extra cash before purchase.

Situation 3: The mortgage rate remains 6.5%, you may choose any down payment of at least 5%, and you unexpectedly receive $100,000 in cash before buying the home. You may use all, some, or none of this cash for the down payment.

For each situation, state the maximum home price you would be willing to pay and the dollar down payment you would choose.

Return only a valid JSON array with exactly 6 numbers in this order: situation_1_maximum_price, situation_1_down_payment, situation_2_maximum_price, situation_2_down_payment, situation_3_maximum_price, situation_3_down_payment. All six numbers must be nominal U.S. dollars. Your output must follow this exact array structure: [v_1, v_2, v_3, v_4, v_5, v_6]. Replace every v placeholder with a single number. Do not include keys or explanatory text.

Assistant. [65000, 5000, 65000, 7000, 85000, 35000]

H.32 HOUSE_SCE_LOCKIN

This task is inspired by Aidala et al. (2024).

Target. The respondent's stated chance of moving to a different primary residence over the next three years if they could keep their current mortgage interest rate after moving. The answer comes directly from an SCE question asking for a probability from 0 to 100 percent. For evaluation, we express it as the predicted probabilities that the respondent moves and does not move.

Predictors. Ordinary moving expectation (the respondent's stated three-year moving probability without mortgage portability), mortgage and housing circumstances (current mortgage rate, home value, mortgage balance, homeownership, the latest national 30-year mortgage rate, and the difference between the current and national rates), household finances and expectations (current and expected finances), work situation (current work-status indicators), household composition and respondent demographics, resources, and geography (age, education, household before-tax income bracket, region, numeracy, marital or partner status), and core macro context (national economic conditions over the most recently completed quarters). The hypothetical scenario allows the household to carry its current mortgage rate to a new home.

No-change baseline. The no-change prediction is the respondent's ordinary three-year moving probability reported in the same survey.

Sample construction. We merge the SCE Housing module with respondent information from the SCE Core survey, national economic variables from FRED-QD, and the national 30-year mortgage rate. We keep respondents with an active mortgage. Next, we drop rows that are missing the moving-probability target under mortgage portability; the ordinary moving probability; current mortgage status or rate; or any required demographic, income, macroeconomic, or national-mortgage-rate predictor. Current home values, mortgage balance, spouse or partner status, household-finance measures, and work-status indicators are allowed to remain missing. We then require both the ordinary and mortgage-portability moving probabilities to lie between 0 and 100 percent, the current mortgage rate to be positive, and respondents to be between 18 and 100 years old. Finally, within each survey month, we trim the top and bottom percentiles of the current mortgage rate and observed housing values.

Prompt example

`house_sce_lockin`

System. You are an American adult making decisions about housing and residential location.

User. Here is some background information about yourself and your household. You are a 38-year-old, have a high-school education, live in the Midwest, have a high numeracy score, and are married or living with a partner.

Here are your household finances, work situation, expectations, housing, and mortgage situation. Note that the current month is February, and that $ denotes amounts in U.S. dollars. Your household's total pre-tax income over the past 12 months was between $150,000 and $199,999. Your household is about as well off as it was 12 months ago, and you expect your household finances to remain about the same 12 months from now. You expect inflation to be -0.5% over the next 12 months and -2.6% over the next three years. You work full-time. A typical local home is worth about $200,000, while your primary residence was purchased for about $520,000 and is now worth about $600,000. Your home loan consists of mortgage debt only, with a balance of about $400,000 and an interest rate of 3.0%. The national average 30-year mortgage rate is 6.7%, which is 3.7 percentage points above your current mortgage rate. Given your current situation, you think the probability that you would move to a different primary residence over the next three years is 0.010.

You are also aware of recent national economic conditions. Over the four most recently completed calendar quarters, listed from oldest to most recent, U.S. real GDP changed by -0.25%, 0.16%, 0.72%, and 0.69% quarter over quarter; the unemployment rate was 3.87%, 3.63%, 3.53%, and 3.57%; overall consumer prices changed by 2.20%, 2.35%, 1.32%, and 1.05% quarter over quarter; and the effective federal funds rate was 0.12%, 0.77%, 2.19%, and 3.65%. Over the five years ending in the most recently completed quarter, compound average quarterly real GDP growth was 0.57%, the average unemployment rate was 4.93%, compound average quarterly CPI inflation was 0.95%, and the average effective federal funds rate was 1.23%.

For this question, consider the following mortgage-rate scenario. Suppose that, if you moved and bought a different home, you could keep the same interest rate as your current mortgage.

Under that scenario, what probability do you think there is that you would move to a different primary residence over the next three years?

Return only a valid JSON array containing exactly one two-element probability distribution. Use this order: probability of not moving under the scenario, probability of moving under the scenario. Express each probability as a number from 0 to 1. The distribution must sum to 1. Your output must follow this exact array structure: [[p_1_1, p_1_2]]. Replace every p placeholder with a number from 0 to 1. Do not include keys or explanatory text.

Assistant. [[0.99, 0.01]]