LLMs AS RESEARCH PARTICIPANTS: HOUSEHOLDS BUILT FROM SURVEY RECORDS FOR EMPIRICAL ECONOMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Using large language models (LLMs) as research participants in economics requires each simulated person to decide as a real person with the same background would, but their answers have been shown to be biased relative to real people's, and neither the source of this bias nor a reliable correction is known. We propose type sampling, which does not correct the bias but instead unfolds the one answer that has to serve many people sharing the same characteristics into the answers of the people those characteristics could describe. For 500 households from the Panel Study of Income Dynamics and seven LLM families, mean AUC on ten asset-holding questions rises from 0.66 to 0.80, and for average amounts held by groups of similar households, mean log R-squared rises from -0.08 to 0.49 across the three LLMs tested. We then set out a procedure for experiments on simulated households, with a design that keeps the two conditions the same household and criteria that their responses must meet, and apply it to Imbens, Rubin, and Sacerdote's lottery study, which the simulation also extends to prizes of other sizes and payment timings that observed data cannot provide. In the families whose responses meet our three criteria most fully, working hours fall by less than in the original study, and the division of the prize among financial assets and housing is close to the original estimates.

1 INTRODUCTION

Empirical research in the social sciences draws on observations from people with different backgrounds to test hypotheses and estimate causal effects. Can large language models (LLMs) recreate this research setting by simulating a group of participants with different demographic and economic backgrounds? This paper builds simulated participants from the characteristics a survey records by type sampling, a method that writes the people those characteristics could describe and takes one answer from each of them. It then tests whether their answers match those of the real households behind the records, across backgrounds and in aggregate, and uses them in a social experiment.

Building such participants has proved difficult. Argyle et al. showed that LLMs conditioned on real respondents' backgrounds could reproduce meaningful patterns in their answers (Argyle et al., 2023), but later work located where the correspondence breaks down, in plausible averages that coexist with distorted demographic relationships (Bisbee et al., 2024), in response distributions recovered without participant rankings (Cummins, 2025), and in better elicitation that improves agreement yet leaves substantial gaps against empirical benchmarks (Ahnert et al., 2026). What these studies leave unsolved is that each simulated participant may still fail to answer as a real person with the same recorded background would.

A survey records only a few characteristics of each household, such as age, income, and wealth. Many households share the same few characteristics, and because they differ in their histories and preferences, they give different answers to the same question. What a survey reports for those characteristics is therefore an average over many different people. An LLM asked once for such a household does not draw from those people. Its answer comes from the LLM itself, so averaging over households with the same characteristics converges to what the LLM returns for those characteristics, not to what those people would answer, and the difference between the two survives

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.

however many households are asked. This is the systematic error that the literature has documented in several forms, as answers that track some demographic groups far better than others (Santurkar et al., 2023), answers that move with the order and labels of the options (Dominguez-Olmedo et al., 2024), and averages that look right while the relationships among groups are distorted (Bisbee et al., 2024). Which form arises here, and by what rule the LLM settles on its answer, cannot be observed, so rather than correct the error we stop asking for a single answer. The relationship we are after is between recorded characteristics and answers, so we put the missing step back and let many plausible people who fit the characteristics answer, a procedure we propose and call type sampling. One LLM writes descriptions of these people, a second, which can be the same one, answers the survey questions under each description, and we pool the answers, so that a yes-or-no question yields a probability and a numerical question an expected value. The pooled answer stands for the people those characteristics could describe rather than for any one of them, and Section 3 shows that it reduces the error relative to asking once, which is why the paper builds on it.

We evaluate type sampling on 500 households from the 2019 Panel Study of Income Dynamics (Institute for Social Research, 2023), ten holding questions, and seven LLM families, against the survey-style method of asking once per household and question, and we score both methods on whether they rank the households that hold an asset above those that do not and on whether their pooled holding rates match the observed ones. Type sampling improves on asking once on both, with mean AUC rising from 0.66 to 0.80 and holding-rate errors falling on nine of the ten questions, for stocks from 12.8 to 2.4 percentage points in mean absolute error across LLM families and for business ownership from 9.6 to 1.5. We also evaluate numerical answers by comparing simulated and observed asset balances averaged within groups of similar households, and we measure agreement by R-squared on log group means. On this comparison the mean across four assets and three answering LLMs rises from -0.08 to 0.49, all three exceed a supervised learning benchmark for individual retirement accounts, and on savings one exceeds it and another is level with it.

With simulated participants whose answers come closer to those of real households, we can begin to investigate economic questions that would otherwise require costly experiments or data that are difficult to obtain. Agreement with survey answers under observed circumstances does not, however, establish that these participants will respond appropriately to a treatment. We use the effect of lottery winnings on labor supply and asset holdings (Imbens et al., 2001) to demonstrate how to construct and evaluate a social experiment with simulated participants. Before reading any change in their answers as an effect of the prize, we require three things of the simulated households. Their behavior under the control condition should resemble that of comparable real households, the treated households should state that they won and how much they have received and report wealth that reflects that prize, and their other financial answers should agree with that wealth and with one another. The second matters because a household may report unchanged working hours either because it chooses not to reduce them or because its answers fail to reflect the lottery at all. In the LLM families that meet these criteria most fully, working hours fall by less than in the original study, and the division of the prize among financial assets and housing is close to the original estimates. The same simulated households can also receive prizes of other sizes and payment timings, which observed lottery data cannot provide, and Section 4 traces how their responses change with the size and timing of the prize.

This paper contributes to research that uses LLMs as simulated survey respondents and experimental subjects (Argyle et al., 2023; Aher et al., 2023; Xie et al., 2024; Horton et al., 2023; Chen et al., 2023). First, type sampling starts from the characteristics a survey records for one household, writes the people those characteristics could describe, and takes one answer from each of them. Its closest precedents condition LLMs on open-ended backstories (Moon et al., 2024) or request dozens of responses to each survey question (Brand et al., 2024), and in our household-finance application the method improves ranking accuracy and holding-rate accuracy over asking once. Second, LLMs that replace human participants can flatten identity groups (Wang et al., 2025), and we evaluate the answers conditional on one household's recorded characteristics against the real households behind those characteristics, in their rankings, their group holding rates, and their mean balances. This complements benchmarks of distributional alignment (Meister et al., 2025), direct requests for a response distribution (Gong et al., 2026), and fine-tuning on individual responses from 210 experiments whose data cannot recover the distribution conditional on an individual profile (Kolluri et al., 2025). Third, LLMs can predict the results of social science experiments (Ashokkumar et al., 2026), an LLM has been fine-tuned on human choice data (Binz et al., 2025), and simulation has

been proposed as a method of social research (Anthis et al., 2025), yet varying a treatment in a simulation can change variables that should be held fixed (Gui & Toubia, 2023). We therefore set out a procedure for experiments on simulated households, a design in which the two conditions describe the same household and differ only in the intervention, and the criteria that the responses in both conditions must meet before their difference is read as an effect, and we illustrate the procedure with a published lottery study (Imbens et al., 2001). Section 2 describes the method, Section 3 compares simulated with survey answers, Section 4 runs the social experiment, and Section 5 discusses where the method is weak and how far it travels. The appendices give the samples, the prompts, and further results.

2 THE METHOD

Every step below starts from two things a household survey records. The first is the state of a household, what the survey records about it, such as age, education, and wealth, and the second is its decisions, what the survey then asks, such as whether it holds an asset and how much of it. We treat a household's decisions as depending on its state, and what the method has to do is produce, from the state alone, decisions that match the ones the households actually made. The experiment in Section 4 then changes one element of the state, wealth, and observes what happens to the decisions. We start by giving an LLM a household's state and asking it the survey's question. The LLM receives the state and then the question worded as the survey puts it to a person, once per household and question, and the reply is recorded the way the survey records an answer, as a yes or no, a point on a scale, or a dollar amount. This is no more than fielding the survey on the LLM, one household at a time, and we refer to it as the survey-style method and take it as our starting point. Asked once, the LLM has to give a single answer for a state that many different people share, and answers obtained this way carry the systematic error described in the introduction. Since the rule behind the error is not known, the method below does not try to correct it and instead unfolds the single answer by writing out the people a state admits, letting each of them answer, and averaging.

The unfolding is direct. Instead of asking the question, we first give the state to an LLM with one instruction, to write n short descriptions of different people who fit it. The descriptions fill in what the survey does not record, family history, preferences, and the path that brought the household to its present position, and the LLM writing them is not told which survey question will follow. Each description is what we call a type, and the whole procedure type sampling. A LLM, which can be the same one, is then given one type at a time and answers the survey question as that person, once per type, so that a household with n types produces n answers to each question, and the household's prediction is their plain average, which for a holding question is the share of the n who say yes. Writing and answering are separate steps and need not run on the same LLM, and we use fifty descriptions throughout.

The same types also answer the questions about dollar amounts, which the survey asks in two steps, first whether the household holds the asset and then, if it does, how much. Each type answers both steps, so a household's n types give a holding share and, among the types that hold, a set of positive balances, and the household's amount is again an average over its types. Averaging the balances with zeros for the types that do not hold gives the expected amount for a household in this state, which is what Section 3 uses. Averaging the positive balances alone gives the mean amount among holders, with participation taken separately from the holding share, which is what Section 4 uses.

A survey records both a household's state and its decisions, allowing us to compare simulated answers with observed outcomes. Households with similar recorded states can nevertheless differ in their histories, preferences, and circumstances that the survey does not capture, so their answers need not coincide. The observed answers are themselves measured with error, since respondents round amounts and underreport some items (Gideon et al., 2017; Meyer et al., 2015), and about 7 percent of PSID observations report saving larger in absolute value than income (Bosworth & Smart, 2009). We evaluate how well the simulation reproduces the relationship between recorded states and observed decisions. For a holding question, type sampling predicts the share of a household's types who say yes, while the survey-style method returns one or zero. We assess these predictions in two ways. AUC measures how well they rank households that hold the asset above those that do not. Averaging the predictions across households gives a predicted holding rate, which we compare with the observed rate to assess how well the simulation reproduces the level. As an upper-bound

benchmark for how much the recorded state can predict, we also train a supervised learning model, LightGBM (Ke et al., 2017), on the survey's answers for other households and score it on the same households, which shows what a model that has seen thousands of answers gets from the state alone, whereas the simulation sees no answers. Section 3 reports both methods against this benchmark.

Type sampling's numerical prediction for a household is an average over its types, so it estimates what people in that state answer on average, not the answer of the one household the survey recorded. That recorded answer is one draw from those people, whose answers differ widely (Bernheim et al., 2001), since even detailed records leave most of the variation in wealth and in life outcomes unexplained (Venti & Wise, 2001; Salganik et al., 2020). Among the 7,002 PSID households in the cells of net worth, age, and income used in Section 3, about 70 percent of the variance in the ten holding answers lies within the cells on average (Appendix B), so a prediction equal to the cell mean for every household would explain only about 30 percent of the variance household by household. Because the prediction is an estimate of an average, we compare it with the average observed amount among households with similar states, as Sections 3 and 4 describe for each application, and in every comparison both averages either count households that do not hold the asset as zeros or leave them out.

This section has described type sampling and how its answers are compared with survey data, and Section 4 describes in detail how the simulated households are used in a social experiment and what their answers must satisfy before they are read as effects of a treatment.

3 Does it answer like the people?

We study household financial decisions, for which we expect the observed state variables to be particularly informative. We apply the method to the 2019 wave of the Panel Study of Income Dynamics (Institute for Social Research, 2023), the last wave before the pandemic, when household finances were in a comparatively settled state. We deliberately use a small set of eight state variables to describe each household's basic demographic and economic background: age, gender, marital status, children, education, employment, household income, and net worth. The LLM receives only these eight variables, in a fixed wording reproduced in the appendix. We evaluate ten yes-or-no questions about asset holding and four questions requiring numerical answers, which are listed in Tables 1 and 2. We work with 500 of the roughly 9,500 households in the wave, drawn in equal numbers across levels of wealth, education, and age, and the appendix gives the sampling procedure.

We first examine whether the simulated holding decisions reproduce both the ranking of households and the share that holds each asset. Table 1 reports the ten holding questions for seven LLM families, DeepSeek V4 Flash (DS Flash), Kimi K2.5 (Kimi), Grok 4.1 Fast (Grok), GPT-5-mini (GPT-mini), Claude Haiku 4.5 (Haiku), MiniMax-M3 (MiniMax), and qwen3.8-flash (Qwen), each answering the same fifty types per household, written by Kimi. With type sampling the mean AUC across the seven families is 0.80 over the ten questions, highest for home ownership, stocks, and employer plans, and lowest for other debt, other financial assets, and business ownership. The benchmark is the survey-style method of Section 2, the same question put once to the same LLM for the same household. Its ten-question mean is 0.66 against 0.80, and the improvement holds family by family, with each family's mean rising from between 0.65 and 0.68 to between 0.795 and 0.803. On every one of the ten questions its mean AUC is lower as well, by 0.07 for other financial assets to 0.24 for other real estate, and each gap exceeds its sampling uncertainty, judged by a 95% household bootstrap interval (Appendix B). Our upper-bound benchmark, trained on the survey's other households with the same eight state variables, has a mean AUC of 0.81 on the same 500 households, so type sampling, which sees no answers, ranks households almost as well as a model that has seen thousands of them, and the three questions on which type sampling ranks worst are also the three on which the upper bound is lowest, so these holdings are hard to predict from the recorded state by any method. Type sampling also reduces the holding-rate error on nine of the ten questions, with the largest reductions for stocks, IRAs, and savings, and the mean absolute error across the seven families falls from 12.8 to 2.4 percentage points for stocks and from 9.6 to 1.5 for business ownership. The error for home ownership increases from 10.9 to 12.1 percentage points, while the error for other financial assets falls from 24.4 to 16.1 but remains large. The upper bound, by contrast, sits within about one point of the observed rates, so the level rather than the ranking is where the simulation still falls short.

Table 1. Observed holding rates, holding-rate errors, and AUC by question.

Question	Observed (%)	MAE (pp)			AUC		
		Survey style	Type sampling	LightGBM	Survey style	Type sampling	LightGBM
Stocks	13.4	12.8	2.4	0.9	0.777	0.861	0.878
Savings	84.6	12.3	3.5	0.6	0.572	0.804	0.801
Employer plan	39.8	10.6	9.1	1.2	0.780	0.858	0.874
IRA	24.8	13.2	4.1	1.5	0.711	0.806	0.827
Business	11.0	9.6	1.5	2.5	0.533	0.748	0.785
Home	61.0	10.9	12.1	1.9	0.788	0.881	0.888
Vehicle	80.4	10.5	5.7	1.8	0.643	0.797	0.819
Other real estate	11.2	5.3	4.1	0.8	0.557	0.800	0.800
Other financial assets	11.0	24.4	16.1	1.9	0.673	0.741	0.644
Other debt	47.2	8.7	5.3	0.3	0.598	0.701	0.758
Mean	**38.4**	**11.8**	**6.4**	**1.3**	**0.663**	**0.800**	**0.807**

Note. Both measures are averaged over the seven families, and LightGBM is trained on the survey's other households with the same eight variables and scored on the same 500 (Appendix B).

Two further points bear on these results. The first asks whether repetition alone would close the gap. Repeating the survey-style question fifty times per household and averaging the answers, which we tried on three of the seven families, raises the mean AUC only from 0.66 to 0.71 and leaves the holding-rate error as high as it was. The second concerns who writes the types. In Table 1 every family answers types written by Kimi, and three of the families also wrote types of their own, and for the two of them that answered both pools, Kimi's types do somewhat better but not by much, so the choice of writer matters little. Appendix B reports these runs, together with the versions and dates of every family.

We next examine how closely the simulation reproduces numerical answers, and compare averages among households with similar states as set out in Section 2. The upper-bound benchmark is again LightGBM, here trained to predict asset balances, with training details in the appendix. We form cells by crossing net worth deciles, age quintiles, and income quintiles, with about 125 populated. Within each cell, we compare the mean predicted amount with the observed mean from the full survey sample assigned to that cell, and include zeros for nonholding in both. The score is the R-squared of log cell means, since amounts are skewed enough that on the raw scale a few of the richest cells would decide it.

Table 2 compares numerical answers from three LLMs. On savings GPT-5.4 exceeds the upper bound and Grok is level with it, and all three exceed it for individual retirement accounts. For stocks, only GPT-5.4 exceeds it. All three fall far short on employer retirement plans, partly because of the wording of the question. The PSID records the balance of a plan held through the respondent's present job, so that someone who is not working has none, whereas our question asked about any employer retirement plan, and the simulated retired households report balances many times those in the data, which the appendix breaks down by employment status. Among working households average amounts are close to the data, and the simulation cannot tell which of them have a plan, since whether an employer offers one is not among the eight state variables. On average across the three LLMs, asking once predicts three of the four amounts less accurately than predicting the same value for every cell, as the negative log R-squared in Table 2 shows, and asking fifty times and averaging raises the mean log R-squared across the four assets from -0.08 to 0.25, still below 0.49 for type sampling.

A last question is what the fifty types stand for. They are descriptions of different people who could all have the household's recorded state, written without any attempt to match them to the groups that make up the population, and the prediction gives each of them one fiftieth of the weight. We do not claim that a type occurs in the population as often as it is generated, and the checks in Appendix B show that the two choices behind this design are not critical. Fifty types are enough, since in the family tested the ranking accuracy has largely stabilized by that number, and equal weights are enough as well, since giving common response patterns more or less weight does not raise the AUC above the plain average, whose holding-rate error is also close to the lowest.

Table 2. Log R-squared of predicted against observed cell-mean dollar amounts.

	Asking once	Asking fifty times	Type sampling			
Asset	Mean of three LLMs	Mean of three LLMs	DS Flash	GPT-5.4	Grok	LightGBM
Stocks	−0.403	0.097	0.486	0.530	0.475	0.526
Savings	0.420	0.722	0.633	0.872	0.810	0.805
Employer plans	−0.211	−0.174	0.087	−0.103	0.000	0.523
IRA	−0.132	0.341	0.675	0.672	0.696	0.639

Note. Cells cross net worth deciles, age quintiles, and income quintiles, 125 for the LLMs and 124 for LightGBM. Nonholders count as zero and holders with a missing balance are left out (Appendix B).

4 Can simulated households be used in a social experiment?

Simulated households created by type sampling, whose survey answers resemble those of real households with the same background, open the way to a wide range of research, because social experiments can be designed and run on them in simulation. By an experiment we mean here an intervention that we assign to the simulated households and whose effect on their decisions we then measure, with the households serving as the laboratory, which extends the reach of experiments to interventions that cannot be run on people, such as a large prize paid out over twenty years. The study of treatment effects, that is, of how an intervention changes household outcomes, is one example and the one we develop in this section. Observed data can rarely answer such questions because each household is observed under only one set of circumstances, whereas a simulated household can be placed under both the treated and the untreated condition, with an intervention that we choose. Answering like real households in the circumstances the survey observed does not, however, make such an experiment credible, since a simulated household can describe its current situation accurately and still respond to a change in ways that no real household would, so the households have to be coherent in both conditions, as the real households they stand for without the intervention and as people who have registered it with the intervention. We therefore set out an approach for using simulated households in an empirical study of treatment effects, which states the criteria that their responses have to meet and how each criterion is tested, and we illustrate it with a lottery experiment for which a published estimate exists.

We revisit Imbens, Rubin, and Sacerdote's study of how lottery winnings affect labor supply and asset holdings (Imbens et al., 2001). They surveyed, in 1995 and 1996, Massachusetts lottery players who had won prizes between 1984 and 1988, and combined the answers with Social Security earnings records. A study of this kind is difficult to carry out. Winners have to be traced years after the prize and asked to recall their circumstances at the time of winning, those who reply are a self-selected part of those asked, and the players they are compared with are drawn from a different population and differ from them in how much they play. A simulated version of the study avoids these difficulties, and the rest of this section describes its design, the criteria that the simulated responses have to meet, and how the estimated effects compare with those of the original study.

We apply type sampling to two hundred households that the PSID interviewed in both 2009 and 2019. Each household is described to the LLM twice, once for the control condition and once for the treated condition, and the two descriptions share everything except the lottery. Fifty types are written for each description as before, and five LLM families answer as those types. The shared part of a description gives the household's 2019 age, sex, marital status, and education, its income and net worth ten years earlier, and the income and net worth it would have reached in 2019 without a lottery. This last pair is a projection computed from the panel, at the level of cells defined by 2009 age, income, and wealth, from what the households in the same cell had reached in 2019. We use this projection rather than the household's own 2019 record for the same reason that answers are averaged over types. A state admits many people and many paths, the record is one draw from them, and the description is meant to carry the state rather than the draw. The treated description adds that the household won a lottery in 2009 paying 150,000 dollars a year for twenty years, 3,000,000 dollars in total, and that ten years of payments remain, and later in this section we also vary the size and timing of the prize. This is the least the LLM needs to answer as a winner, and Appendix E gives the argument and the designs we tried before settling on it. The prompt also states the household's employment status in 2019 without a lottery, set in turn to working, retired, or otherwise not working, so each household is described under three employment statuses, each with a control and a treated

description and its own fifty types, and the probability of each status, projected from the same cells, enters the estimator below. The types are written separately for the two descriptions, so the prize can change who the types are as well as how they answer. The households answer questions on hours worked per week, employment status, asset holdings and balances, and saving, and they also report their current net worth, which the second criterion below uses.

Employment status enters the estimate in two ways. Which status a household would hold in 2019 without a lottery is uncertain, so the estimate weights the three statuses by the share of the PSID households in its 2009 cell that held each status in 2019. Given a description with one status, the types then choose for themselves, since each reports its own 2019 status, and the share of a household's types that report each status, with and without the prize, is what the prize changes. The effect of the prize on an outcome therefore has two parts. The first compares the answers of the types that report the initial status with and without the prize, the same kind of person answering differently. The second takes the change in the shares of types that report each status once the household has won, from working to retired for instance, and values each shift at the average outcome among types that hold that status without a lottery, pooled across households, so types that report retirement after winning are given the hours and balances of the retired rather than their own answers. Averaging the two parts over the three initial statuses with the data weights gives the household's effect, which therefore separates into a change within employment statuses and a change from movement between them, a separation that observed data on lottery winners cannot provide, and the effect on each outcome is the average of these household effects. Appendix E states the projection, the reference levels, and the estimator formally. The population, period, and prize differ from those of the original study, whose estimates therefore serve as a comparison of direction and magnitude rather than as a target to reproduce.

Three criteria test whether the simulated households are coherent under the control description and under the treated one. The first, control consistency, requires that without the lottery the simulated households behave as comparable real households do. We compare their five asset balances and four holdings with those of the twenty PSID households closest to them in age, wealth, income, and education within their projected 2019 age and wealth cell, with an upper-bound benchmark of the same kind as before alongside. The second, treatment recognition, requires that the circumstances the treated households report be consistent with the treatment. They should state that they won and how much they have received, and the prize should appear in the net worth they report. We measure the latter as the increase of a household's average reported net worth over its projected net worth, divided by the 1,500,000 dollars received, and count the households whose share lies between 5% and 100%, a band calibrated on this sample rather than fixed in advance. The third, internal consistency, requires that a treated household's asset answers and the net worth it reports agree with each other. Two signals serve, in the answers given with the employment status set to working, a stock balance above the reported net worth, which debts the questions do not ask about could produce but which most families' answers rarely show, and saving over the past two years that, taken per year, exceeds projected annual income plus the annual prize payment. The families in which either signal is common are examined further in Appendix E.

All five families meet the first criterion. Without the lottery, each ranks households' holdings with an AUC between 0.85 and 0.98 and fits the four balances at least as well as the upper-bound benchmark (Appendix E). The second criterion is met by three of them. In DS Flash, Kimi, and Grok, the wealth that treated households report rises by between 14 and 27 percent of the cash received at the median, and the rise lies between 5 and 100 percent of the cash for 86 to 100 percent of households, whereas in GPT-mini and Haiku the median rise is 5 to 6 percent and only about half of the households fall in that range, so that the prize leaves little trace in their balance sheets. On the third criterion Grok falls short as well, and Haiku falls short for a second reason. Grok's treated households report saving more over two years than their income and prize payments could have provided in 12 percent of their answers, against at most 1.3 percent in any other family, and Haiku's stock balances exceed its reported net worth in 12 percent of its treated answers, against 10 percent for Grok and at most 4 percent for the other three families. The answers in which stocks exceed net worth, in Haiku and Grok alike, turn out on inspection to have copied the no-lottery net worth from the description as their total while listing assets that include the prize, so the total repeats the prompt rather than reporting the household's position. DS Flash and Kimi therefore meet the three criteria most fully, and we interpret the magnitudes of the effects for these two families only, with the effects of all five and the full figures behind each criterion in Appendix E.

Table 3 sets the effects in DS Flash and Kimi, the two families that meet the criteria most fully, beside the estimates of the original study (Imbens et al., 2001; 1999). The division of the prize among assets is close to the original, with the responses of financial assets, assets outside retirement accounts, and house value inside a standard error of the original estimates except for Kimi's assets outside retirement accounts, and the two families depart from the original in the same two places, since both add to retirement accounts, where the original study finds no response, and both let house values rise far ahead of mortgages. Working hours fall in both families, by less than in the original study once each prize contrast is expressed per dollar of annual prize (Appendix E), and most of the reduction comes from households working less rather than leaving work. Hours fall by similar amounts in all five families, whatever wealth each registers (Appendix E), so the hours response alone would not have shown which families registered the prize.

Table 3. Effects of the lottery in DS Flash and Kimi against the original study.

	Per dollar received				Ratio	Per week	Ratio
	Financial assets	Outside retirement	Retirement accounts	House value	Mortgage share	Hours	Employment share
Original	0.16	0.18	n.s.	0.037	0.72–0.77	−9.3	n.a.
DS Flash	0.204	0.160	0.044	0.047	0.128	−3.58	0.110
Kimi	0.142	0.115	0.028	0.034	0.152	−3.75	0.196

Note. Mortgage share is the mortgage increase over the house-value increase, and Employment share is the part of the hours effect that comes from changes in employment shares.

This comparison uses a single prize, whereas a simulated household can be given several, so the experiment can also show how the same households respond as the prize grows and as its form changes, which observed lottery data cannot. Figure 1 plots, for the four outcomes that the original study also reports, the responses to annuities of 25,000, 75,000, and 150,000 dollars a year and to a lump sum of 3,000,000 dollars paid in 2009, with the figures in Appendix E. The two families move almost together on every outcome. The hours response flattens as the prize grows, from 2.0 to 3.6 hours a week in DS Flash and from 2.2 to 3.7 in Kimi as the annual payment rises from 25,000 to 150,000 dollars, whereas the share of each dollar received that is held as financial assets rises with the prize, from 0.09 to 0.20 and from 0.08 to 0.14, so the larger the prize, the larger the fraction of it that is saved. Paid at once, the same nominal total reduces hours by 0.82 and 0.91 of the annuity's reduction, a ratio computed from our own responses that agrees with the 0.88 reported for Swedish winners paid in installments or at once (Cesarini et al., 2017), and it is held as financial assets several times as much per dollar, so the households treat the annuity as income and the lump sum as wealth. The one outcome on which the two families part company is the mortgage, where the lump sum lowers the probability of holding one by 4 percentage points in DS Flash and by 17 in Kimi.

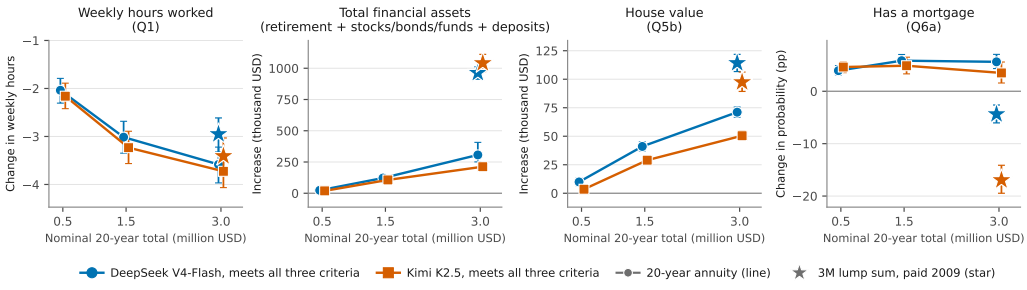

Figure 1. Responses of DS Flash and Kimi to the three annuities and the lump sum, with 95 percent household bootstrap intervals.

5 Discussion

In the setting of this paper type sampling falls short in two places. The first is dispersion. Averaging over types compresses the differences between households, and the variance of the simulated households is a third to a half of the real one (Appendix B), so the simulation reproduces averages

and rankings better than the spread around them. The second is the level of holding rates for particular groups. Grouping the households by each of the eight state variables in turn, with control for multiple testing, we find no group whose ranking or median amounts sit unusually far from the upper-bound benchmark. For holding rates such groups do appear, and every one of them is defined by wealth or income, the richest fifth and the residual categories above all (Appendix D).

We also applied the method to other household surveys. On household-finance surveys from Japan in 2019 and China in 2011, type sampling reaches a mean AUC of 0.76 and 0.91 on the financial assets that both surveys ask about, and on the 2023 wave of the PSID it reaches 0.77 on the ten questions of Table 1, against 0.66 for asking once (Appendix D).

We then extended type sampling from household financial decisions to political attitudes. From the 2022 General Social Survey (Smith et al., 2022) we took 18 questions with two answer options for 500 respondents, among them the 2020 presidential vote, abortion for any reason, the death penalty, and whether same-sex relations are wrong, which ask for political and social attitudes, and whether there is a gun in the home, which asks, like most of the remaining questions, about the respondent's own circumstances. The upper-bound benchmark is again LightGBM trained on the survey's other respondents with the same fields. With the five fields that the survey shares with our household description, age, marital status, education, work status, and income, type sampling reaches a mean AUC of 0.62 over the 18 questions and 0.58 on the four political questions, and the benchmark 0.61 and 0.57, so these fields carry little information about political attitudes and neither method can find what is not there. With the twelve fields that add gender, race, religion, region and place of upbringing, party identification, and political views, the political and social attitude questions rise to 0.80 and the presidential vote to 0.94, the level of the benchmark with the same twelve fields. Of the added fields, party identification and political views carry most of this gain, since with the other ten fields alone the vote falls to 0.73. The questions about the respondent's own circumstances gain little from the added fields under either method. These kinds of answers are more likely related to a life that the fields do not record than to decisions made from the recorded state, as asset holdings are, so a method that starts from these fields has little to work with. Type sampling exceeds asking once on every one of the 18 questions with twelve fields, by 0.08 on average (Table D4 in Appendix D, with DS Flash writing the types and answering). What the simulated respondents add to the benchmark is therefore not accuracy, which the two share, but the possibility of putting the same respondents under an intervention.

We also tried two experiments from the archive of (Ashokkumar et al., 2026), Kennedy's judgments of a neighboring family and Howard's spending forecasts, of the kind used in psychology and behavioral economics in which a manipulation changes what respondents are told (Appendix D, Tables D5 to D7). That archive scores a predicted treatment effect against the observed one after pooling the participants of each condition, whereas our simulated respondents answer one by one, so we can also compare each contrast between the groups the experiment distinguishes, across parties or across instructions. The simulated panel reproduces the direction of 32 of 36 contrasts in Kennedy's experiment and the sign of Howard's main effect, but not their size, which we attribute to the description carrying a household's circumstances while these experiments move an identity or a way of thinking.

Could the agreement with the survey and with the lottery study be recall of training data rather than simulation? The pattern argues against recall as the sole explanation. An LLM reciting the PSID would not answer some questions well and others poorly, asking once would not miss the holding rates by 12 points on average, and the ranking gain also appears in the 2023 wave. In the direct test of Appendix D.5, a fictional crisis with the consequences of 2008 but a name that appears nowhere moves home ownership as much as the real one in DS Flash and by four fifths in Kimi, and a placebo history of a relative's illness moves it by less than one point. The simulation gives each household's answers with and without a treatment and in every employment status, which no experiment on real households observes, and whether these paired answers recover the distribution of gains across real households (Heckman & Smith, 1995) remains a conjecture.

The paper thus provides a panel of simulated households that can be placed under an intervention, and the criteria that test whether the households remain coherent in both conditions. The experiments this opens are those that observed data cannot provide, and the size and timing of a lottery prize are only the first.

Reproducibility statement

The household sample is drawn from the public-use 2019 wave of the Panel Study of Income Dynamics, and Appendix A describes the sampling. Appendix B gives the background template, the ten holding questions, the type-generation and answering prompts, and the version of every LLM used, and Appendix E gives the prompts and the estimator of the lottery experiment. The PSID conditions of use do not allow us to distribute data extracts or household identifiers, so the supplementary material, available at https://anonymous.4open.science/r/llm-households-light-safe-2026/, contains the sampling and description code, the prompts, the version of every LLM, the parsed answers of the LLMs keyed by a study index, and checksums with which a registered PSID user can confirm that a sample rebuilt from the public-use files matches ours. Tables 1 and 2 are recomputed from these without any call to a model, and Table 3, Table E9, and Figure 1 are recomputed from the parsed lottery answers once the lottery sample is rebuilt, whose selection the package documents and whose identifiers can be shared only through the PSID Public Data Extract Repository.

Ethics statement

The study uses public-use survey data and generates answers from language models. No human participants were recruited, and no individual PSID household is identified. The simulated participants are not a substitute for consent-based research with people, and Section 5 discusses the questions on which their answers do not match real respondents.

AI use statement

Large language models are the object of this study, and every answer of the simulated households analyzed in Sections 3 to 5 was generated by the LLMs listed in Appendix B as part of the method. Beyond that, we used generative AI tools to write and test code, to clean and reformat data, to draft and edit text, and to check numbers and citations against their sources. The authors set the questions, the design, the arguments, and the interpretation of every result, reviewed all AI-assisted work, and take full responsibility for the content of this paper.

References

Gati Aher, Rosa I. Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 337–371. PMLR, 2023. arXiv:2208.10264.

Georg Ahnert, Anna-Carolina Haensch, Barbara Plank, and Markus Strohmaier. Survey response generation: Generating closed-ended survey responses in-silico with large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 41554–41577, 2026. doi: 10.18653/v1/2026.acl-long.1927. URL https://aclanthology.org/2026.acl-long.1927/.

Jacy Reese Anthis, Ryan Liu, Sean M. Richardson, Austin C. Kozlowski, Bernard Koch, James Evans, Erik Brynjolfsson, and Michael Bernstein. Position: LLM social simulations are a promising research method. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 81005–81034. PMLR, 2025. arXiv:2504.02234.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023. doi: 10.1017/pan.2023.2.

Ashwini Ashokkumar, Luke Hewitt, Isaias Ghezae, and Robb Willer. Large language models can predict the results of social science experiments. *Nature*, 656(8126):115–122, 2026. doi: 10.1038/s41586-026-10742-x.

B. Douglas Bernheim, Jonathan Skinner, and Steven Weinberg. What accounts for the variation in retirement wealth among u.s. households? *American Economic Review*, 91(4):832–857, 2001. doi: 10.1257/aer.91.4.832.

Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K. Eckstein, et al. A foundation model to predict and capture human cognition. *Nature*, 644(8078):1002–1009, 2025. doi: 10.1038/s41586-025-09215-4. arXiv:2410.20268.

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: 401–416, 2024. doi: 10.1017/pan.2024.5.

Barry P. Bosworth and Rosanna Smart. Evaluating micro-survey estimates of wealth and saving. Working Paper CRR WP 2009-4, Center for Retirement Research at Boston College, 1 2009. URL https://www.brookings.edu/wp-content/uploads/2016/06/01_micro_survey_bosworth.pdf.

James Brand, Ayelet Israeli, and Donald Ngwe. Using GPT for market research. In *Proceedings of the 25th ACM Conference on Economics and Computation (EC '24)*, pp. 613, 2024. doi: 10.1145/3670865.3673479. Full paper: Harvard Business School Marketing Unit Working Paper 23-062, SSRN 4395751, doi 10.2139/ssrn.4395751.

David Cesarini, Erik Lindqvist, Matthew J. Notowidigdo, and Robert Östling. The effect of wealth on individual and household labor supply: Evidence from swedish lotteries. *American Economic Review*, 107(12):3917–3946, 2017. doi: 10.1257/aer.20151589.

Yiting Chen, Tracy Xiao Liu, You Shan, and Songfa Zhong. The emergence of economic rationality of GPT. *Proceedings of the National Academy of Sciences*, 120(51):e2316205120, 2023. doi: 10.1073/pnas.2316205120.

Jamie Cummins. The threat of analytic flexibility in using large language models to simulate human data. *arXiv preprint arXiv:2509.13397*, 2025.

Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. Questioning the survey responses of large language models. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, pp. 45850–45878, 2024. doi: 10.52202/079017-1458. Oral; arXiv:2306.07951.

Michael Gideon, Brooke Helppie-McFall, and Joanne W. Hsu. Heaping at round numbers on financial questions: The role of satisficing. *Survey Research Methods*, 11(2):189–214, 2017. doi: 10.18148/srm/2017.v11i2.6782. URL https://ojs.ub.uni-konstanz.de/srm/article/view/6782.

Eric Gong, Nathan E. Sanders, and Bruce Schneier. Characterizing the ability of llms to recapitulate americans' distributional responses to public opinion polling questions across political issues, 2026.

George Gui and Olivier Toubia. The challenge of using LLMs to simulate human behavior: A causal inference perspective. Working Paper 4650172, Social Science Research Network, 2023. arXiv:2312.15524.

James J. Heckman and Jeffrey A. Smith. Assessing the case for social experiments. *Journal of Economic Perspectives*, 9(2):85–110, 1995. doi: 10.1257/jep.9.2.85.

John J. Horton, Apostolos Filippas, and Benjamin S. Manning. Large language models as simulated economic agents: What can we learn from homo silicus? Working Paper 31122, National Bureau of Economic Research, 2023. Issued April 2023, revised February 2026; all authors contributed equally.

Guido W. Imbens, Donald B. Rubin, and Bruce Sacerdote. Estimating the effect of unearned income on labor supply, earnings, savings, and consumption: Evidence from a survey of lottery players. Working Paper 7001, National Bureau of Economic Research, 1999.

Guido W. Imbens, Donald B. Rubin, and Bruce I. Sacerdote. Estimating the effect of unearned income on labor earnings, savings, and consumption: Evidence from a survey of lottery players. *American Economic Review*, 91(4):778–794, 2001. doi: 10.1257/aer.91.4.778.

Institute for Social Research. Panel study of income dynamics, public use dataset, 2023.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *NeurIPS*, 2017.

Akaash Kolluri, Shengguang Wu, Joon Sung Park, and Michael S. Bernstein. Finetuning LLMs for human behavior prediction in social science experiments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 30084–30099, 2025. doi: 10.18653/v1/2025.emnlp-main.1530. arXiv:2509.05830.

Nicole Meister, Carlos Guestrin, and Tatsunori Hashimoto. Benchmarking distributional alignment of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 24–49, 2025. doi: 10.18653/v1/2025.naacl-long.2. arXiv:2411.05403.

Bruce D. Meyer, Wallace K. C. Mok, and James X. Sullivan. Household surveys in crisis. *Journal of Economic Perspectives*, 29(4):199–226, 2015. doi: 10.1257/jep.29.4.199. Working paper version: NBER Working Paper 21399.

Suhong Moon, Marwa Abdulhai, Minwoo Kang, Joseph Suh, Widyadewi Soedarmadji, Eran Kohen Behar, David M. Chan, and John Canny. Virtual personas for language models via an anthology of backstories. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 19864–19897, 2024. doi: 10.18653/v1/2024.emnlp-main.1110. arXiv:2407.06576.

Matthew J. Salganik, Ian Lundberg, Alexander T. Kindel, et al. Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*, 117(15):8398–8403, 2020. doi: 10.1073/pnas.1915006117.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29971–30004. PMLR, 2023. arXiv:2303.17548.

Tom W. Smith, Michael Davern, Jeremy Freese, and Stephen L. Morgan. General social surveys, 1972–2022, 2022.

Steven F. Venti and David A. Wise. Choice, chance, and wealth dispersion at retirement. In *Aging Issues in the United States and Japan*, pp. 25–64. University of Chicago Press, 2001. doi: 10.7208/chicago/9780226620831.003.0002.

Angelina Wang, Jamie Morgenstern, and John P. Dickerson. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, 7(3):400–411, 2025. doi: 10.1038/s42256-025-00986-z. arXiv:2402.01908.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Shiyang Lai, Kai Shu, Jindong Gu, Adel Bibi, Ziniu Hu, David Jurgens, James Evans, Philip Torr, Bernard Ghanem, and Guohao Li. Can large language model agents simulate human trust behavior? In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, pp. 15674–15729, 2024. doi: 10.52202/079017-0501. arXiv:2402.04559.

APPENDIX

A HOUSEHOLD SAMPLE

The sampling frame is the 2019 PSID extract. Records are eligible if the reference person's age is observed and at least 18, education is recorded, and the longitudinal household weight is positive. We stratify the eligible records by five net-worth groups, five education groups, and four age groups, giving 100 cells. The education groups are less than high school, high school diploma, some college, bachelor's degree, and graduate or professional degree. The age groups are 18–34, 35–49, 50–64, and 65 or older.

Net-worth cutoffs are weighted quintiles using the longitudinal household weights. The sampling implementation takes the logarithm of positive net worth and maps nonpositive values to zero before calculating the cutoffs. The stored cutoffs correspond to approximately \$1,330, \$33,200, \$139,000, and \$438,800. Five records are drawn from each cell, with selection probabilities proportional to the longitudinal household weights. Sampling is without replacement unless a cell contains fewer than five eligible records, in which case it is with replacement. The random seed is 42. The saved evaluation sample contains 500 records, with five in each of the 100 cells.

B PROMPTS, MODEL CONFIGURATIONS, AND DETAILS OF SECTION 3

B.1 BACKGROUND TEMPLATE

The survey information supplied about each household consists of age, gender, marital status, children, education, employment, household income, and net worth. The following template is shared by the inspected holding-question and numerical-answer implementations. Instructions and questions are added separately. Under type sampling, the answering prompt also includes a generated type description; this description is not additional observed survey information.

```
Here is your background:

You are {a_an} {age}-year-old {gender},
{marital_status}{children_text}.  Your highest level of
education is {education}.  You are currently
{employment_status}.
Your annual household income is approximately {income}.
Your household's total net worth (assets minus debts) is
approximately {net_worth}.
```

B.2 SURVEY-STYLE HOLDING QUESTIONS

The following system and user messages reproduce the survey-style implementation used by the rerun pipeline. The user message includes all ten holding questions in one call. Each household therefore produces one answer per question, although the questions are bundled rather than issued in ten separate calls.

System message:

```
You are participating in a household financial decision
study.  You will be given a description of a person's
demographic and financial background.  Based on this
background, please answer ALL questions as this person would
realistically answer.  For each question, respond with ONLY
A (Yes) or B (No).  Use the format:  Q1:  A  Q2:  B  etc.
```

User message template:

```
Here is your background:
```

```
You are {a_an} {age}-year-old {gender},
{marital_status}{children_text}.  Your highest level of
education is {education}.  You are currently
{employment_status}.
Your annual household income is approximately {income}.
Your household's total net worth (assets minus debts) is
approximately {net_worth}.

--

Please answer each of the following questions as this person
would.  For each question, respond with A (Yes) or B (No).

Q1:  Do you directly own any stocks or shares in publicly
traded companies through a brokerage account?  A. Yes  B. No
Q2:  Do you have any money in checking or savings accounts?
A. Yes  B. No
Q3:  Does your employer offer a retirement plan (like a
401k) that you participate in?  A. Yes  B. No
Q4:  Do you have a private retirement account (like an IRA)
that you set up yourself?  A. Yes  B. No
Q5:  Do you own a business or have an ownership interest in
a business?  A. Yes  B. No
Q6:  Do you own your home (as opposed to renting)?  A. Yes
B. No
Q7:  Do you own a vehicle (car, truck, or motorcycle)?  A.
Yes  B. No
Q8:  Do you own any real estate other than your primary
home?  A. Yes  B. No
Q9:  Do you have any other financial assets (bonds, life
insurance cash value, etc.)?  A. Yes  B. No
Q10:  Do you have any debts other than your mortgage,
vehicle loans, and student loans?  A. Yes  B. No

Respond in this exact format (one line per question):
Q1:  A
Q2:  B
Q3:  A
...and so on for all 10 questions.
```

B.3 TYPE GENERATION

The ten-type and fifty-type implementations use different instructions. The latter also requests additional dimensions of diversity and unusual types; its wording cannot be obtained simply by replacing 10 with 50. Both variants are reproduced here. The instructions are experimental inputs, not claims that the resulting descriptions sample the population distribution.

Ten-type system message:

```
You are a social scientist studying the diversity of
lifestyles within demographic groups.  Given a demographic
profile, generate exactly 10 distinct types of people who
could match this profile.  Each type should differ in
personality, daily routine, social circle, financial habits,
and life priorities.  Be realistic -- include both
financially engaged and disengaged types.  Format each type
as a short paragraph (2-3 sentences).
```

Ten-type user message template:

```
Here is your background:

You are {a_an} {age}-year-old {gender},
{marital_status}{children_text}.  Your highest level of
```

```
education is {education}.  You are currently
{employment_status}.
Your annual household income is approximately {income}.
Your household's total net worth (assets minus debts) is
approximately {net_worth}.
--

Generate exactly 10 distinct types of real people who could
match this demographic profile.  They should vary in:
- Daily routine and time availability
- Financial knowledge and engagement
- Social circle and influences
- Risk attitudes and life priorities
- Access to and awareness of financial services

Format your response as:
Type 1:  [description]
Type 2:  [description]
...
Type 10:  [description]
```

Fifty-type system message:

```
You are a social scientist studying the diversity of
lifestyles within demographic groups.  Given a demographic
profile, generate exactly 50 distinct types of people who
could match this profile.  Each type should differ from all
others in personality, daily routine, social circle,
financial habits, and life priorities.  Be realistic --
include both financially engaged and disengaged types.
Cover the full spectrum:  savers vs spenders, planners vs
procrastinators, risk-takers vs risk-averse, socially
connected vs isolated, employed vs underemployed, etc.
Format each type as a short paragraph (2-3 sentences).  Do
not repeat any type.
```

Fifty-type user message template:

```
Here is your background:

You are {a_an} {age}-year-old {gender},
{marital_status}{children_text}.  Your highest level of
education is {education}.  You are currently
{employment_status}.
Your annual household income is approximately {income}.
Your household's total net worth (assets minus debts) is
approximately {net_worth}.

--

Generate exactly 50 distinct types of real people who could
match this demographic profile.  They should span maximum
diversity across:
- Daily routine and time availability
- Financial knowledge and engagement
- Social circle and influences
- Risk attitudes and life priorities
- Access to and awareness of financial services
- Health, family obligations, hobbies
- Cultural background and values

Cover the full distribution -- not just plausible "average"
types but also unusual, edge, and surprising types within
this demographic.  Each of the 50 must be clearly distinct
from all others.
```

```
Format your response as:
Type 1:  [description]
Type 2:  [description]
...
Type 50:  [description]
```

B.4 HOLDING ANSWERS UNDER TYPE SAMPLING

The generated description replaces `{type_description}`. The background variables and the ten questions are retained. The inspected implementations have minor differences in the response-format example, so both templates are recorded.

Ten-type implementation, system message:

```
You are participating in a household financial decision
study.  You will be given a description of a person's
demographic and financial background, along with additional
details about their lifestyle.  Based on this background,
please answer ALL questions as this person would
realistically answer.  For each question, respond with ONLY
A (Yes) or B (No).  Use the format:  Q1:  A  Q2:  B  etc.
```

User message template:

```
Here is your background:

You are {a_an} {age}-year-old {gender},
{marital_status}{children_text}.  Your highest level of
education is {education}.  You are currently
{employment_status}.
Your annual household income is approximately {income}.
Your household's total net worth (assets minus debts) is
approximately {net_worth}.

Additional context about this specific person:
{type_description}

--

Please answer each of the following questions as this person
would.  For each question, respond with A (Yes) or B (No).

Q1:  Do you directly own any stocks or shares in publicly
traded companies through a brokerage account?  A. Yes  B. No
Q2:  Do you have any money in checking or savings accounts?
A. Yes  B. No
Q3:  Does your employer offer a retirement plan (like a
401k) that you participate in?  A. Yes  B. No
Q4:  Do you have a private retirement account (like an IRA)
that you set up yourself?  A. Yes  B. No
Q5:  Do you own a business or have an ownership interest in
a business?  A. Yes  B. No
Q6:  Do you own your home (as opposed to renting)?  A. Yes
B. No
Q7:  Do you own a vehicle (car, truck, or motorcycle)?  A.
Yes  B. No
Q8:  Do you own any real estate other than your primary
home?  A. Yes  B. No
Q9:  Do you have any other financial assets (bonds, life
insurance cash value, etc.)?  A. Yes  B. No
Q10:  Do you have any debts other than your mortgage,
vehicle loans, and student loans?  A. Yes  B. No

Respond in this exact format (one line per question):
```

```
Q1:  A
Q2:  B
Q3:  A
...and so on for all 10 questions.
```

Fifty-type implementation, system message:

```
You are participating in a household financial decision
study.  You will be given a description of a person's
demographic and financial background, along with additional
details about their lifestyle.  Based on this background,
please answer ALL questions as this person would
realistically answer.  For each question, respond with ONLY
A (Yes) or B (No).  Use the format:  Q1:  A  Q2:  B  etc.
```

User message template:

```
Here is your background:

You are {a_an} {age}-year-old {gender},
{marital_status}{children_text}.  Your highest level of
education is {education}.  You are currently
{employment_status}.
Your annual household income is approximately {income}.
Your household's total net worth (assets minus debts) is
approximately {net_worth}.

Additional context about this specific person:
{type_description}

--

Please answer each of the following questions as this person
would.  For each question, respond with A (Yes) or B (No).

Q1:  Do you directly own any stocks or shares in publicly
traded companies through a brokerage account?  A. Yes  B. No
Q2:  Do you have any money in checking or savings accounts?
A. Yes  B. No
Q3:  Does your employer offer a retirement plan (like a
401k) that you participate in?  A. Yes  B. No
Q4:  Do you have a private retirement account (like an IRA)
that you set up yourself?  A. Yes  B. No
Q5:  Do you own a business or have an ownership interest in
a business?  A. Yes  B. No
Q6:  Do you own your home (as opposed to renting)?  A. Yes
B. No
Q7:  Do you own a vehicle (car, truck, or motorcycle)?  A.
Yes  B. No
Q8:  Do you own any real estate other than your primary
home?  A. Yes  B. No
Q9:  Do you have any other financial assets (bonds, life
insurance cash value, etc.)?  A. Yes  B. No
Q10:  Do you have any debts other than your mortgage,
vehicle loans, and student loans?  A. Yes  B. No

Respond in this exact format (one line per question):
Q1:  A
Q2:  B
...and so on for all 10 questions.
```

B.5 NUMERICAL ANSWERS

Table 2 reports balances for directly held stocks, savings, employer retirement plans, and individual retirement accounts. The bundled numerical prompt also asks weekly and annual working hours. We

reproduce the complete six-item block, including those hours questions, to distinguish the questions actually asked from the four financial outcomes reported in Table 2.

Each financial question asks first about holding and then about the amount, with a zero specified for nonholding. The employer-retirement wording refers to an employer retirement plan without restricting it to the current employer; this differs from the current-employer participation question in B.2 and B.4. These original wordings are retained here.

System message:

```
You are participating in a household financial study.  You
will be given a description of a person's demographic and
financial background.  Based on this background, please
estimate the financial quantities asked, as this person
would realistically have.  Give concrete answers -- do not
refuse, hedge, or say 'it depends'.  If a quantity is zero
(e.g., this person has no stocks), answer $0 (point) or A
(bracket).
```

Type-sampling user message template:

```
Here is your background:

You are {a_an} {age}-year-old {gender},
{marital_status}{children_text}.  Your highest level of
education is {education}.  You are currently
{employment_status}.
Your annual household income is approximately {income}.
Your household's total net worth (assets minus debts) is
approximately {net_worth}.

Additional context about this specific person:
{type_description}

--

Please answer the following 6 items about this person.  For
each item, answer first whether this person has it (A=Yes /
B=No), then if A, the amount.  If B, the amount is
automatically 0.

Q1a (Stocks):  Does this person directly own any stocks in
publicly traded companies through a brokerage account (NOT
through retirement plans)?
A. Yes
B. No
Q1b (Stocks - amount):  If Q1a=A, dollar amount?  If B, $0.

Q2a (Savings):  Does this person have any money in savings
accounts, CDs, or money market accounts?
A. Yes
B. No
Q2b (Savings - amount):  If Q2a=A, dollar amount?  If B, $0.

Q3a (Employer retirement):  Does this person have an
employer retirement plan (401k, 403b, pension, Keogh)?
A. Yes
B. No
Q3b (Employer retirement - amount):  If Q3a=A, dollar
amount?  If B, $0.

Q4a (IRA): Does this person have a private retirement
account (IRA)?
A. Yes
B. No
```

```
Q4b (IRA - amount):  If Q4a=A, dollar amount?  If B, $0.

Q5a (Currently working?):  Does this person currently work
at a main job?
A. Yes
B. No
Q5b (Hours/week):  If Q5a=A, hours per week?  If B, 0.

Q6a (Worked last year?):  Did this person work at all last
year?
A. Yes
B. No
Q6b (Hours last year):  If Q6a=A, total hours last year?  If
B, 0.

Format strictly:
Q1a:  ___
Q1b:  $___
Q2a:  ___
Q2b:  $___
Q3a:  ___
Q3b:  $___
Q4a:  ___
Q4b:  $___
Q5a:  ___
Q5b:  ___
Q6a:  ___
Q6b:  ___
```

For the corresponding survey-style numerical prompt, the generated-description block is omitted. The background is followed directly by the separator and the same six-item question block.

B.6 Model configurations for the main holding comparison

Table 1 uses seven answering LLM families. Under type sampling, all seven answer the same pool of descriptions generated by Kimi K2.5, with fifty descriptions for each household. The table below reports the answering versions and means over the ten holding questions. Qwen and MiniMax answer with thinking disabled in the request parameters. GPT-5-mini uses reasoning. The DeepSeek Flash results use V4.

	AUC		MAE (pp)	
Answering LLM	Survey style	Type sampling	Survey style	Type sampling
Claude Haiku 4.5	0.676	0.800	8.6	4.4
Kimi K2.5	0.674	0.802	13.8	7.6
GPT-5-mini	0.652	0.801	14.7	4.9
MiniMax-M3	0.660	0.796	13.8	7.3
qwen3.8-flash	0.655	0.800	8.3	7.8
Grok 4.1 Fast	0.667	0.803	13.6	6.5
DeepSeek V4 Flash	0.660	0.795	10.0	6.2

For each question, holding-rate MAE is the mean across families of the absolute difference between predicted and observed holding rates, calculated on that family's valid households. Errors are taken before averaging across families. Thus, over the ten questions, MAE falls from 11.8 to 6.4 percentage points; averaging predicted rates across families before taking absolute differences instead gives 10.8 and 5.3. Table 1 uses the former measure. Its observed-rate column describes the fixed 500-household sample.

Missing answers are excluded for the affected question. In the MiniMax-M3 survey-style run, three responses omit five answers in total. Under type sampling, three responses have no parseable answers (two refusals and one format failure), and 27 have some unparseable answers. Two responses contain visible reasoning despite the request to disable it. The Qwen runs have no such exceptions.

Each model–method combination is one run, and B.16 reports bootstrap intervals over households for the seven-family means of Table 1.

These configurations apply to Table 1. The numerical comparisons in Table 2, the type analyses, and the lottery study use their own batches and model pairings. They are not reruns with the seven-family configuration above.

B.7 Supplementary comparison with DeepSeek V4 Pro

DeepSeek V4 Pro generates and answers its own types. Reasoning is enabled under both methods. We report this comparison separately because it does not use the common Kimi type pool in Table 1. The table gives AUC and absolute holding-rate errors for this single family. The ten-question mean AUC rises from 0.661 to 0.799, and the mean absolute holding-rate error falls from 9.9 to 7.4 percentage points. This comparison does not isolate the effect of reasoning.

	AUC		MAE (pp)	
Question	Survey style	Type sampling	Survey style	Type sampling
Stocks	0.785	0.853	13.4	5.3
Savings	0.543	0.836	13.6	5.2
Employer plan	0.772	0.863	4.8	14.7
IRA	0.727	0.814	15.2	0.4
Business	0.534	0.723	9.8	0.2
Home	0.794	0.888	4.6	14.7
Vehicle	0.621	0.781	13.6	4.1
Other real estate	0.585	0.780	4.8	5.7
Other financial assets	0.667	0.746	17.4	10.6
Other debt	0.584	0.709	1.6	13.5
Mean	**0.661**	**0.799**	**9.9**	**7.4**

B.8 Model configurations for numerical answers

All three LLM columns in Table 2 answer the Kimi K2.5 type pool. The answering LLMs are DeepSeek V4 Flash, GPT-5.4, and Grok 4.1 Fast. Each has fifty type-conditioned answers and fifty repeated survey-style answers for each of 500 households. These are different runs from the seven-family holding comparison in Table 1. In particular, the GPT column is GPT-5.4, not GPT-5-mini. The former Flash self-generated-type and pooled-Flash columns are omitted; Grok now uses its Kimi-type run rather than its self-generated-type run.

The GPT-5.4 batch used the Codex proxy. A comparison repeated 200 type-conditioned responses through the direct API for five households. Substituting these responses into the original cell-level calculation changed each of the four log R-squared values by at most 0.0013. The small slice does not establish equivalence of the two access routes for the full sample. Table 2 retains the original GPT-5.4 batch, not the slice or a GPT-5-mini run.

B.9 The ten holding questions against the survey items

This subsection reports the supplementary run that the note to Table 1 points to. The ten holding questions were written from the asset categories of the data files rather than copied from the survey, and several of them differ in scope from the survey items, so the same types were asked a second time under the survey's own wording. Households, types and answering settings are held fixed across the two arms. The same 500 households are used, the same pool of fifty Kimi K2.5 descriptions for each household is reused without rewriting, and DeepSeek V4.1 Flash answers at temperature 0.7 with reasoning disabled, with the system message, the background block and the parsing rules taken from the Table 1 type-sampling line. The first arm reuses an existing run after a mechanical check that everything except the question text matches. The second arm contains 25,000 answers, with no question left unparsed. There is one answering family and one run for each arm.

The first arm asks the ten questions as the paper asks them.

```
Q1:  Do you directly own any stocks or shares in publicly
traded companies through a brokerage account?
Q2:  Do you have any money in checking or savings accounts?
Q3:  Does your employer offer a retirement plan (like a
401k) that you participate in?
Q4:  Do you have a private retirement account (like an IRA)
that you set up yourself?
Q5:  Do you own a business or have an ownership interest in
a business?
Q6:  Do you own your home (as opposed to renting)?
Q7:  Do you own a vehicle (car, truck, or motorcycle)?
Q8:  Do you own any real estate other than your primary
home?
Q9:  Do you have any other financial assets (bonds, life
insurance cash value, etc.)?
Q10:  Do you have any debts other than your mortgage,
vehicle loans, and student loans?
```

The second arm asks the survey items, copied from the 2019 codebook, with the interviewer's branch symbols expanded into one answerable sentence and no change to any definition of an asset or a debt.

```
W15  Do you (or anyone in your family living there) have any
shares of stock in publicly held
corporations, stock mutual funds, or investment trusts-not
including stocks in retirement
accounts?
W27a Do you (or anyone in your family living there) have
checking or savings accounts, including
money market accounts?
P1   Next, I need to get some information about pensions and
retirement plans.  Not including Social
Security Railroad Retirement, or I.R.A.s (Individual
Retirement Accounts) are you participating
in a pension or retirement plan through your present job, or
through your union?  If either you
or your spouse or partner participates, answer Yes.
W21  Do you (or your family living there) have any money in
private annuities or Individual
Retirement Accounts (I.R.A.s)?
W10  Do you (or anyone in your family living there)
currently own part or all of a farm or business?
A19  Do you (or anyone else in your family living there) own
the (apartment/mobile home/home), do you
pay rent, or what?      [answer options A. Own  B. Rent or
other]
F47  These next questions are about personal vehicles and
transportation.  Do you (or anyone else in
your family living there) own or lease a car or other
vehicle for personal use?
W1   The next questions I will be asking are designed to
give estimates of wealth of families.  In
these questions when I refer to the family, I mean your
family living there with you.  Do you (or
your family living there) have any real estate (other than
your main home), such as a second
home, land (not including farmland), rental real estate, or
money owed to you on a land
contract?
W27  Not including retirement accounts, do you (or anyone in
your family living there) have any money
in certificates of deposit, government bonds, or treasury
bills?
W33  Do you (or anyone in your family living there) have any
other savings or assets, such as cash
```

```
value in a life insurance policy, a valuable collection for
investment purposes, or rights in a
trust or estate that you haven't already told us about?
W38a Aside from the debts that we have already talked about,
(like any mortgage on your main home
(or/like) vehicle loans,) do you (or anyone in your family
living there) currently have any
credit card or store card debt?  Do not count new debt that
will be paid off this month.
W38b Do you (or anyone in your family living there)
currently have any other debts such as student
loans, medical or legal bills, or loans from relatives?
```

The second arm asks twelve items and scores ten. Table B9a gives the mapping and the four places where the item had to be adapted before its answer could be scored against the same outcome as the first arm.

Table B9a. The survey items behind the ten scored questions.

Question	Survey item	Codebook variable	Adaptation
Stocks	W15	ER73820	None
Checking or savings	W27a	ER73847	None
Employer retirement plan	P1	ER74015	One sentence added, telling the respondent to answer yes if either the respondent or the spouse or partner participates
IRA or private annuity	W21	ER73840	None
Farm or business	W10	ER73811	None
Home ownership	A19	ER72030	Item kept, answer options reduced from own, rent or neither to own against rent or other
Vehicle	F47	ER72856	None
Other real estate	W1	ER73797	None
Other assets	W33	ER73874	W27 asked before it and discarded, so that the phrase about what the respondent has already reported has the context the survey gives it
Other debts	W38a and W38b	ER73879 and ER73885 to ER73889	Both items asked, and the outcome is set to yes if either is answered yes

The adaptations exist because the outcome each arm is scored against is the same recorded variable, and that variable is built from more than one survey item in three cases. They are not neutral, since adding the spouse to the employer plan item, merging the two debt items and reducing the housing options each change the share who would answer yes, so the second arm is a survey-worded arm rather than the survey itself.

Ranking barely moves and the holding-rate bias falls. Over the ten questions the mean AUC rises by 0.005, with an interval from -0.001 to 0.012, while the mean absolute bias falls by 1.7 percentage points. The gap between our wording and the survey's definitions therefore shows up mainly in how many households answer yes and much less in which households answer yes. Four patterns appear question by question. The survey wording lowers the bias on the employer retirement plan, the retirement account, home ownership, checking or savings and other real estate. It overshoots on two questions, since other debts turns from underprediction into overprediction and the vehicle question turns from almost exact into overprediction. It changes almost nothing on stocks and on farm or business. On other financial assets it does not help, and the overprediction is slightly larger under the survey wording.

Three limits bound what this run shows. There is one answering family and one run for each arm, so the intervals reflect the sampling of households and not the variation between repeated answers, and nothing here says whether the other families of Table 1 would improve in the same way. The second arm keeps the four adaptations, so it is not the survey administered as written. On other financial

Table B9b. Ranking and holding-rate bias under the two wordings, all 500 households.

Question	Data (%)	Bias (pp)		Change (pp)	95% interval	AUC	
		Ours	Survey			Ours	Survey
Stocks	13.4	9.1	8.3	−0.8	−1.4 to −0.1	0.865	0.867
Checking or savings	84.6	7.5	1.7	−5.8	−6.4 to −5.2	0.799	0.803
Employer retirement plan	39.8	−6.4	0.9	7.3	6.6 to 8.0	0.855	0.864
IRA or private annuity	24.8	−6.6	3.2	9.8	8.8 to 11.0	0.809	0.813
Farm or business	11.0	−1.2	−0.8	0.4	0.1 to 0.8	0.744	0.773
Home ownership	61.0	−14.5	−6.4	8.2	7.5 to 8.9	0.877	0.884
Vehicle	80.4	−0.6	5.8	6.4	5.8 to 7.2	0.792	0.789
Other real estate	11.2	−4.8	−1.5	3.3	2.8 to 3.8	0.807	0.808
Other financial assets	11.0	16.3	18.8	2.5	1.5 to 3.4	0.738	0.734
Other debts	47.2	−7.5	10.4	17.9	16.5 to 19.4	0.709	0.716
Mean	**38.4**	**7.4**	**5.8**			**0.800**	**0.805**

Note. Bias is the predicted holding rate minus the observed rate, in percentage points. Differences are the survey wording minus our wording. Intervals are from 1,000 paired resamples of households, with the same resampling index applied to all ten questions. The Mean row averages the ten questions, and for bias it averages the absolute values.

Table B9c. The richest fifth, the six questions whose bias moves most and the item on other financial assets, 100 households.

Question	Data (%)	Bias (pp)		Change (pp)	95% interval
		Ours	Survey		
IRA or private annuity	63.0	−25.2	−2.4	22.8	20.6 to 24.8
Other debts	30.0	−15.2	1.8	17.0	14.5 to 19.7
Other real estate	34.0	−11.3	−0.1	11.2	9.5 to 13.0
Stocks	42.0	3.8	11.3	7.6	6.2 to 8.9
Employer retirement plan	46.0	−4.2	2.8	7.0	5.7 to 8.5
Home ownership	92.0	−4.2	0.7	4.9	3.6 to 6.4
Other financial assets	21.0	40.5	43.5	3.0	0.2 to 5.8

Note. The richest fifth is the top wealth group of the five the sample is drawn across, 100 households. Over its ten questions the mean AUC rises from 0.683 to 0.695, an increase of 0.012 with an interval from −0.011 to 0.034 over 993 usable resamples, and the mean absolute bias falls from 12.1 to 7.7 percentage points.

assets this version of the survey item did not remove the overprediction, and that does not show the overprediction to be unrelated to the wording, since only two wordings were compared.

B.10 WEIGHTING THE TYPE RESPONSES

This subsection reports the weighting test that Section 3's paragraph on what the types represent promises. Section 3 states there that each type's answer contributes one fiftieth of the prediction, and that the evaluation does not assume that the frequency with which a type is generated equals its share of the population. The question here is whether the plain average of the fifty answers can be improved by weighting them. One answering family is used, Claude Haiku 4.5, on the same 500 households and the same ten holding questions, and no new answers were collected, since every test reweights answers already on record.

The fifty types are not fifty distinct behaviors. Grouping a household's types by their complete pattern of ten yes-or-no answers, the fifty collapse to 17.7 distinct patterns on average, with a median of 18 and a range of 5 to 38, so 64.5 percent of the fifty slots repeat a pattern already present. The weighting question is therefore a real one rather than an empty one.

The first test gives each behavior group a total weight proportional to its multiplicity raised to a power. At a power of one every generated type counts once, which is the plain average the paper uses. Below one the weighting moves toward counting each distinct pattern once, and above one it moves toward the patterns the language model repeats, which is the direction that would help if generated frequency were the difficulty. Table B10 reports the sweep.

Table B10. Ranking and holding-rate deviation as the weight on repeated response patterns changes.

Weight power	Mean AUC	Mean absolute holding-rate deviation (pp)
0.0, each distinct pattern once	0.785	5.7
0.5	0.798	4.6
1.0, the plain average	0.801	4.2
1.5	0.799	4.7
2.0	0.797	5.0
3.0	0.794	5.5

Note. Behavior groups are the complete ten-question answer patterns. The deviation is the mean over the ten questions of the absolute difference between the predicted and the observed holding rate. Grouping the pooled types into sixteen clusters instead gives the same ordering.

The second test fits one set of group weights whose only purpose is to make the predicted population rates match the observed ones, and then applies those weights to every household. That drives the rate deviation to almost nothing and costs 2.5 AUC points in sample and 2.7 points under five-fold cross-validation over households. The third test is the most favourable case for weighting, since it fits for each question separately the weights that directly minimize the Brier score and is allowed to overfit. It gains 1.3 AUC points in sample and loses 1.0 point out of sample.

The plain average therefore sits at or near the best point of an explicit family of weightings that runs from counting distinct patterns to counting repeated ones. What this establishes is narrow, since none of the weightings tested improves both the population rates and the ranking out of sample, since the one that matches the rates almost exactly gives up ranking quality and the one fitted question by question does not carry out of sample. None of them identifies the distribution of behavioral types in the population either, which the survey does not observe.

B.11 HOW MANY TYPES THE RANKING NEEDS

This subsection reports the check on the number of types that Section 3's paragraph on what the types represent mentions. Fifty types for each household is a choice, so one answering family, DeepSeek V4 Pro with reasoning enabled, wrote and answered types for the households of the same PSID 2019 sample in several batches. The run reached 200 types for 359 households and 201 for one more, 132 households stopped at 120, and the remaining eight lie between 100 and 199.

Two comparisons follow from it. Over the 500 households the comparison file records and all ten questions, the ranking at 80 types and at 120 types differs by 0.003 on average and by at most 0.008, on other debt. On 359 households with at least 200 recorded types, six of the ten questions can be followed from 5 types to 200, and Table B11 gives that curve.

Table B11. AUC by number of types for each household, the 359 households with at least 200 recorded types when the curve was computed, six questions.

Types for each household	Stocks	Employer plan	Home	IRA	Vehicle	Business
5	0.705	0.751	0.821	0.677	0.693	0.543
10	0.773	0.785	0.837	0.717	0.733	0.532
20	0.765	0.814	0.856	0.744	0.743	0.662
30	0.771	0.831	0.855	0.754	0.751	0.653
50	0.781	0.844	0.863	0.776	0.760	0.651
80	0.776	0.853	0.864	0.770	0.759	0.661
100	0.788	0.853	0.866	0.770	0.760	0.670
120	0.795	0.851	0.867	0.777	0.761	0.677
150	0.798	0.850	0.864	0.769	0.763	0.675
180	0.794	0.854	0.864	0.768	0.762	0.667
200	0.797	0.855	0.864	0.769	0.761	0.663

Note. The 359 households are those with at least 200 recorded types when the curve was computed. They are the one household with 201 types, of which the first 200 are used, and 358 of the 359 households that end the run with exactly 200; the remaining one reached 200 only in a later top-up batch and is not in the curve. A household's value at each number of types is the share of yes answers among its first types in the order they were generated, and an answer that could not be parsed counts as no. The curve covers six of the ten holding questions, which are the six the run records at every number of types.

The steep part of the curve lies below 20 types and the curve is flat from 50 onward. Going from 50 types to 200 changes AUC by at most 0.016, on stocks, and moves individual retirement accounts slightly the other way. Both comparisons use one answering family and the second uses a subset of the households, so together they bound the choice of fifty rather than settle it.

B.12 Employer retirement plan balances by employment status

This subsection breaks down the employer retirement plan balances discussed in Section 3's paragraph on numerical answers. Section 3 states there that the simulated retired households report balances many times those in the data, and that among working households the average amounts are close to the data. Table B12 splits the 500 evaluation households by the employment status recorded in their background so that the two statements can be read separately. The observed columns follow the convention of Table 2, under which a household that declares a holding but leaves the balance blank is left out of the observed mean for that asset. The simulated columns come from the pairing in which Kimi K2.5 writes the descriptions and DeepSeek V4 Flash answers. A household's simulated holding rate is the share of its fifty answers reporting a balance above zero, and its simulated expected balance is the share answering yes multiplied by the mean balance when the answer is yes. Every one of the 500 households has at least five parsed answers, so the household counts apply to both sides of the table.

The employed households are the group Section 3 describes as close to the data, and the retired households are the group it describes as many times above it. Under the earlier convention that counted blank balances as zero the three ratios were 1.30, 18.1 and 0.80, so excluding the blanks brings the employed group closer to the data and leaves the retired group at nearly eighteen times. The third group is the one where the simulation underpredicts. This breakdown uses the 500 evaluation households rather than the cell means of the full survey sample on which Table 2 is scored, so it locates the error and does not decompose the score.

Table B12. Employer retirement plan holding and balance by employment status, 500 PSID 2019 households.

Employment	Households		Holding (%)		Mean balance ($)		Sim./obs.
	Total	Missing balance	Observed	Simulated	Observed	Simulated	
Employed	330	44	57.3	52.9	50,179	56,740	1.13
Retired	107	2	3.7	21.3	3,034	53,960	17.8
Neither employed nor retired	63	3	9.5	4.2	8,083	6,187	0.77

Note. Employment status is the field supplied in the household background. The blank column counts households that declare an employer retirement plan and leave the balance unrecorded, and those households are excluded from the observed mean for this asset. Simulated figures are fifty type-conditioned answers per household from Kimi K2.5 descriptions answered by DeepSeek V4 Flash. The third row combines the statuses not currently working due to disability (27 households), not currently working (22), a homemaker (9), a student (3), and on leave from work (2).

B.13 Three qualifications on Table 2

Table 2 comes with three qualifications. The benchmark is better at the first stage, with an AUC on whether a household holds each asset between 0.81 and 0.88 against 0.77 to 0.85 for type sampling, so the cell means combine participation with amounts and do not isolate how well amounts are predicted among holders. The comparison also depends on the scale, since on raw dollars, where the richest cells dominate, the benchmark is ahead of every pairing on stocks, employer retirement plans, and individual retirement accounts. Finally, neither method reaches the individual household, where the log R-squared conditional on holding is +0.15 on stocks, +0.27 on savings, and +0.07 on employer retirement plans for type sampling and negative on individual retirement accounts, and negative on all four for the benchmark. A supervised model trained directly on the cell means, which sets a ceiling for this comparison, is reported in B.14.

B.14 A supervised model trained directly on the cell means

This subsection reports the supervised model that the qualifications on Table 2 in B.13 promise. B.13 describes it as a model trained directly on the cell means, which sets a ceiling for that comparison. The LightGBM benchmark of Appendix C is trained on individual households and scored on cell means, so its training target and its evaluation target differ. The model of this subsection removes that difference by taking the cell mean itself as the training target, on the same cells, with the same evaluation metric, and with features restricted to those the household background supplies. It remains a supervised model and sees the survey outcome, whereas type sampling sees no survey outcome at all, so the comparison measures how close a route that uses no survey labels comes to a supervised ceiling at the same target.

The features are the within-cell means of seven fields, namely age, years of education, gender, marital status, number of children, household income, and net worth, with net worth taken as the same field the household background reports. A second feature set adds the within-cell share of households recorded as working. One row of training data is one cell. Evaluation is out of fold only, with five-fold cross-validation over cells repeated under five seeds and a leave-one-out pass, and the reported figure is the mean over seeds of the same log cell-mean R-squared that scores Table 2. Two training target spaces are computed, the log cell mean and the raw dollar cell mean, and the log target is the one reported below because it is aligned with the metric and therefore the more favourable treatment of the supervised side.

The published outputs of that model were computed when blank balances counted as zero. Table 2 now excludes them, so the model was retrained and rescored under the current convention. The recomputation first reproduced all forty eight published figures to four decimals before the convention was changed. Table B13 gives the result next to the Table 2 columns.

Changing the convention moved two of the four figures. Savings and employer retirement plans are the two assets with blank balances in the survey, and the cell-mean model loses a little on the first and

Table B13. Log R-squared of predicted against observed cell-mean dollar amounts, with a supervised model trained on the cell means.

	Type sampling			LightGBM training	
Asset	DS V4 Flash	GPT-5.4	Grok 4.1 Fast	Households	Cell means
Stocks	0.486	0.530	0.475	0.526	0.684
Savings	0.633	0.872	0.810	0.805	0.909
Employer retirement plans	0.087	−0.103	0.000	0.523	0.643
IRA	0.675	0.672	0.696	0.639	0.763

Note. Cells are the same as in Table 2. The three LLM columns and the cell-mean model cover 125 cells and the household-trained LightGBM covers 124, so the last two columns are not scored on an identical set. The cell-mean column is the five-seed mean of five-fold out-of-fold predictions with the log cell mean as the training target, and the standard deviations over seeds are 0.012 for stocks, 0.004 for savings, 0.036 for employer retirement plans, and 0.009 for individual retirement accounts. The leave-one-out pass gives 0.698, 0.922, 0.660, and 0.767. Adding the within-cell working share to the features gives 0.687, 0.904, 0.629, and 0.765. Training on the raw dollar cell mean instead gives −0.101, 0.682, 0.365, and 0.139.

gains on the second, from 0.910 to 0.909 and from 0.635 to 0.643. Stocks and individual retirement accounts carry no blank balance at all, and their figures are identical under the two conventions, which the recomputation verified by assertion rather than assumed. No cell was dropped for want of observed households.

The cell-mean model stands above every other column on all four assets, so the ceiling it sets is reached by neither route in Table 2. Measured against the nearest other column, the distance is smallest on savings and largest on stocks. Measured against the three LLM columns alone it is largest on employer retirement plans, where both supervised columns stand well above all three, because a model fitted on the survey recovers participation in a plan from the household's recorded fields and the simulation cannot, since whether an employer offers a plan is not among the eight state variables. Two qualifications bound the reading. The training target here is the quantity being scored and the training space is the metric's own, which is the treatment most favourable to the supervised side, and the same model trained on the raw dollar cell mean falls to a negative figure on stocks and to a small positive one on individual retirement accounts. The number of cells is also small, so five-fold splits are noisy, and the standard deviations and the leave-one-out figures are reported with each value. Computing the features from the households whose balance is observed rather than from all households in the cell leaves stocks and individual retirement accounts unchanged and moves the other two assets by less than 0.02. The figure is one untuned specification rather than a maximum over models or settings, so it marks what this supervised comparator reaches and not a proven upper bound.

B.15 How much variation remains within a cell

This subsection reports two facts about the PSID data that Sections 2 and 5 rely on. The first is how finely the eight background fields identify a household. The prompt lists age, gender, marital status, children, education, employment, household income and net worth, with the two money amounts rounded as they are printed in the prompt. Table B14a counts the distinct combinations of these fields among the 9,460 eligible households and among the 500 evaluation households. Rendered as in the prompt, the eight fields form 9,424 distinct states in the full sample, and 99.2% of households share their state with no other household. Dropping the two money amounts and keeping the six discrete fields leaves 3,628 states, and 19.9% of households are still alone. Coarsening age to five-year bands leaves 1,605 states and 7.6% alone. Among the 500 evaluation households the three shares are 99.2%, 79.4% and 40.6%. A background at this resolution identifies one household rather than a group. The quantity that an LLM answer would have to match, the distribution of outcomes among households with that background, is therefore not observable in the data at the prompt's own resolution. The paper coarsens the background to the cells of Table 2 for this reason.

The second fact is how much of the variation in the ten holding questions remains inside those cells. The cells are the net-worth deciles by age quintiles by income quintiles of Table 2. The cutoffs come

Table B14a. Distinct background states under three ways of combining the fields, full sample and evaluation sample.

Background fields	States	Singleton states	Households alone (%)	Largest state
Panel A. Full sample (9,460 households)				
All eight fields	9,424	9,388	99.2%	2
Six discrete fields	3,628	1,883	19.9%	23
Six fields, age in five-year bands	1,605	723	7.6%	84
Panel B. Evaluation sample (500 households)				
All eight fields	498	496	99.2%	2
Six discrete fields	445	397	79.4%	4
Six fields, age in five-year bands	316	203	40.6%	9

from the 500 evaluation households, and the decomposition uses the 7,002 full-sample households that fall in the 125 cells scored in Table 2. Table B14b splits each 0/1 outcome into a between-cell and a within-cell share by a one-way decomposition. The shares carry a method-of-moments random-effects correction that removes the sampling noise in the cell means, and the correction lowers the between-cell share by 0.004 to 0.015 relative to the uncorrected value. Across the ten questions the within-cell share averages 0.705, from 0.398 for home ownership to 0.905 for other financial assets. The last two columns give a reference from a separate exercise. A gradient-boosted model with all eight fields as inputs, scored on out-of-fold predictions from five-fold cross-validation over the 9,460 households, reaches a mean McFadden pseudo R2 of 0.234 and a mean AUC of 0.835. The two exercises use different samples and different metrics, so their numbers are not to be subtracted from one another. Both say that the background leaves most of the variation in holdings unexplained, and that is the variation which one answer per background cannot represent.

Table B14b. Between-cell and within-cell shares of the variance of the ten holding questions, 7,002 households in the 125 cells of Table 2, with a supervised reference on all 9,460 households.

		Variance share		Supervised	
Question	Holding rate	Between cells	Within cells	Pseudo R2	AUC
Stocks held directly in a brokerage account	10.5%	0.273	0.727	0.213	0.859
Checking or savings account	78.3%	0.276	0.724	0.279	0.839
Employer retirement plan (participates)	38.4%	0.357	0.643	0.295	0.852
Private retirement account (IRA)	19.5%	0.378	0.622	0.300	0.873
Business ownership	7.0%	0.134	0.866	0.061	0.784
Home ownership	50.7%	0.602	0.398	0.473	0.922
Vehicle	78.1%	0.397	0.603	0.304	0.834
Other real estate	9.1%	0.209	0.791	0.147	0.841
Other financial assets	9.0%	0.095	0.905	0.015	0.737
Other debts	48.8%	0.228	0.772	0.249	0.806
Mean of the ten questions	n.a.	0.295	0.705	0.234	0.835

Note. Variance shares are adjusted for sampling noise and use 7,002 households. Holding rates and supervised scores use 9,460 households. Pseudo R2 is McFadden's measure; scores use five-fold out-of-fold predictions.

The within-cell variance of the two tables above has a counterpart on the LLM side. Under type sampling each of the 500 evaluation households receives fifty type descriptions, and each description answers separately, so the fifty answers of one household have a variance of their own. Table B14c reports that variance for the seven answering families of Table 1, which all answer the Kimi K2.5 type pool, and for the three families of Table 2, next to the variance of the fifty-type average across households and the sum of the two. The sum is the total variance that the simulated households

would show if one type were drawn for each household instead of averaging. The real benchmarks are the total variance of the same 500 households and the within-cell variance of the 125-cell decomposition above, taken as the adjusted within-cell share times the total variance. Amounts are on the log scale of Table 2 with non-holders at zero, and households whose true amount is missing are excluded on both sides, 49 for the employer plan and 3 for savings. For holdings, averaged over the ten questions, the variance across types within a household is 0.93 to 1.04 times the real within-cell variance with a mean of 0.98, the single-draw total is 0.98 to 1.06 times the real total with a mean of 1.02, and the variance of the type average is 0.33 to 0.43 times the real total with a mean of 0.38. For amounts, averaged over the four items, the three ratios are 0.96 to 1.15 with a mean of 1.03, 1.02 to 1.05 with a mean of 1.03, and 0.42 to 0.50 with a mean of 0.46. Question by question the spread is wider, and over the ten questions the seven-family average of the first ratio runs from 0.67 to 1.65 and that of the second from 0.66 to 1.99. Two qualifications apply. The types condition on all eight background fields, while the cells condition on three coarsened fields, so the real within-cell variance is a residual after three fields only and the two sides do not condition on the same information. The comparison is also one of second moments only. It shows that the dispersion which averaging removes is of the same order as the dispersion among real households with similar backgrounds, and it does not evaluate how accurate a single drawn type would be for any household.

Table B14c. Dispersion of type responses compared with observed households.

	Simulated variance			Relative to data		
LLM	Within	Between	Total	Within	Total	Mean
Panel A. Holdings (ten-question mean)						
Haiku	0.0996	0.0685	0.1682	0.94	1.04	0.43
Kimi	0.1033	0.0637	0.1670	0.98	1.04	0.40
GPT-mini	0.1091	0.0612	0.1703	1.03	1.06	0.38
MiniMax	0.1098	0.0599	0.1697	1.04	1.05	0.37
Qwen	0.1009	0.0575	0.1585	0.96	0.98	0.36
Grok	0.0983	0.0631	0.1614	0.93	1.00	0.39
DS Flash	0.1062	0.0535	0.1598	1.01	0.99	0.33
Mean	0.1039	0.0611	0.1650	0.98	1.02	0.38
Panel B. Amounts (four-asset mean, log scale)						
DS Flash	2.3028	1.5195	3.8223	1.15	1.05	0.42
GPT-5.4	1.9608	1.7392	3.7000	0.98	1.02	0.48
Grok	1.9258	1.8063	3.7321	0.96	1.03	0.50
Mean	2.0631	1.6883	3.7515	1.03	1.03	0.46

Note. Within is variance across fifty types of one household; Between is variance of household means; Total is their sum. Relative to data: Within divides by real within-cell variance, while Total and Mean divide by real total variance. Real total and within-cell variances are 0.1612 and 0.1055 for holdings, and 3.6402 and 2.0097 for amounts. See the text for sample and conditioning differences.

Section 3 compares type sampling with a supervised benchmark on the same 500 households. The benchmark is the gradient-boosted classifier that supplies the last two columns of Table B14b. Its inputs are the eight background fields, that is age, gender, marital status, children, education, employment, and the exact household income and net worth, and it is trained in five stratified folds over the 9,460 eligible households with seed 42. Each of the 500 evaluation households is scored by the fold that did not train on it, so no household contributes to its own prediction, and the AUC is computed on the 500 records of the evaluation file, in which two households are each drawn twice, exactly as Table 1 does. Table B14d sets this out-of-fold AUC beside Table 1's seven-family means. Over the ten questions the benchmark reaches 0.807 against 0.800 for type sampling and 0.663 for the survey-style prompt. It is higher on eight questions, on other real estate only beyond the third decimal, marginally lower on savings, and clearly lower on other financial assets, where it also falls below the survey-style prompt.

Three caveats apply. The benchmark sees exact dollar amounts where the LLM sees the rounded strings of the background template. Net worth is a sum that mechanically contains the balances of several of the assets asked about, so part of the benchmark's ranking is arithmetic rather than learned. The 500 households are a stratified sample with equal numbers in each wealth, education

and age group, so their AUC is a same-household comparison and not the full-sample 0.835 of Table B14b, and on the 498 distinct households the mean is 0.808. The rows for the 500 households in the source file behind Table B14b are a different exercise, five folds inside the 500 alone with a mean of 0.761, and they are not used here.

The last column applies the holding-rate error of Table 1 to the same out-of-fold probabilities, the absolute difference in percentage points between the mean predicted probability over the 500 households and their observed holding rate. It averages 1.3 points over the ten questions, from 0.3 on other debt to 2.5 on business, against Table 1's errors of 6.4 for type sampling and 11.8 for the survey-style prompt. The signed differences are negative on eight questions and positive only on vehicle (1.8 points) and other debt (0.3 points), with a mean of minus 0.9 points, so the benchmark sits slightly below the holding rates of the evaluation sample. That sign is what a population-trained benchmark scored on a stratified sample can produce, since the sample holds most assets at higher rates than the full sample, and on the 9,460 households the same signed differences are mixed in sign and stay within 1.3 points.

Table B14d. Out-of-fold AUC and holding-rate error of the supervised benchmark on the 500 evaluation households, beside the seven-family AUC means of Table 1.

	AUC			
Question	Supervised	Type sampling	Survey style	Supervised error (pp)
Stocks	0.878	0.861	0.777	0.9
Savings	0.801	0.804	0.572	0.6
Employer plan such as a 401(k)	0.874	0.858	0.780	1.2
IRA	0.827	0.806	0.711	1.5
Business	0.785	0.748	0.533	2.5
Home	0.888	0.881	0.788	1.9
Vehicle	0.819	0.797	0.643	1.8
Other real estate	0.800	0.800	0.557	0.8
Other financial assets	0.644	0.741	0.673	1.9
Other debt	0.758	0.701	0.598	0.3
Mean of the ten questions	**0.807**	**0.800**	**0.663**	**1.3**

B.16 Bootstrap intervals for Tables 1 and 2

This subsection reports bootstrap 95% intervals for the entries of Tables 1 and 2 and repeats their point estimates. For Table 1 the unit of resampling is the household. Each of 2,000 draws takes 500 households with replacement, recomputes the AUC and the holding-rate error of every LLM family on every question, and averages the seven families, and the interval runs from the 2.5th to the 97.5th percentile. Both methods use the same draws, so their difference is paired, and for the ten-question mean AUC the difference between type sampling and the survey-style method has interval [0.123, 0.148]. Because the error is an absolute value, its interval is asymmetric where the error is near zero. For Table 2 the unit is the cell. Each draw takes the 125 scored cells with replacement and recomputes the log R-squared of every column, and the LightGBM column uses the 124 drawn cells it scores. The seed is 42 in both tables.

Because both methods are scored on the same 2,000 resamples of the 500 households, the gap between them on each question is a paired quantity, and Table B15c reports the seven-family mean of that gap with its interval. The AUC gap is positive on every question and every one of the ten intervals excludes zero. The holding-rate error gap is negative with an interval that excludes zero on 8 questions, and the intervals for the employer plan and home ownership include zero.

B.17 Asking the survey question fifty times, and who writes the types

This subsection reports the two comparisons behind the four means quoted in Section 3. The first asks whether repetition alone closes the gap between the survey-style question and type sampling. For DS Flash, Grok, and Kimi (the K2.6 version, whereas Table 1 uses K2.5), the same survey-style question is put to each household fifty times at the temperature used in Table 1, and the share of yes answers is the prediction. Asking once is draw 0 of the same batch, so the two rows share the LLM version, prompt, temperature, and households, and the only difference is one draw against

Table B15a. Table 1 with bootstrap 95% intervals in brackets.

	MAE (pp)		AUC	
Question	Survey style	Type sampling	Survey style	Type sampling
Stocks	12.8 [9.9, 16.0]	2.4 [2.1, 4.6]	0.777 [0.732, 0.820]	0.861 [0.811, 0.904]
Savings	12.3 [9.5, 15.4]	3.5 [3.2, 4.2]	0.572 [0.542, 0.608]	0.804 [0.750, 0.854]
Employer plan such as a 401(k)	10.6 [7.8, 14.4]	9.1 [5.7, 12.6]	0.780 [0.749, 0.810]	0.858 [0.824, 0.888]
IRA	13.2 [10.3, 17.0]	4.1 [2.4, 7.0]	0.711 [0.673, 0.749]	0.806 [0.761, 0.849]
Business	9.6 [7.1, 12.4]	1.5 [1.2, 3.5]	0.533 [0.510, 0.561]	0.748 [0.674, 0.813]
Home	10.9 [9.3, 12.9]	12.1 [9.6, 15.4]	0.788 [0.759, 0.816]	0.881 [0.851, 0.910]
Vehicle	10.5 [7.4, 13.5]	5.7 [5.0, 7.0]	0.643 [0.605, 0.681]	0.797 [0.743, 0.844]
Other real estate	5.3 [3.5, 7.8]	4.1 [1.6, 6.9]	0.557 [0.521, 0.594]	0.800 [0.737, 0.851]
Other financial assets	24.4 [20.7, 28.1]	16.1 [13.0, 18.9]	0.673 [0.614, 0.729]	0.741 [0.667, 0.811]
Other debt	8.7 [7.7, 10.5]	5.3 [4.8, 6.4]	0.598 [0.565, 0.632]	0.701 [0.658, 0.745]
Mean	**11.8 [10.9, 12.9]**	**6.4 [5.9, 7.2]**	**0.663 [0.649, 0.678]**	**0.800 [0.781, 0.818]**

Table B15b. Table 2 with bootstrap 95% intervals in brackets.

Method / LLM	Stocks	Savings	Employer plan	IRA
Asked once (three-LLM mean)	−0.403 [−0.894, −0.040]	0.420 [0.140, 0.613]	−0.211 [−0.520, 0.078]	−0.132 [−0.661, 0.213]
TS: DS V4 Flash	0.486 [0.264, 0.656]	0.633 [0.450, 0.760]	0.087 [−0.133, 0.300]	0.675 [0.494, 0.804]
TS: GPT-5.4	0.530 [0.316, 0.691]	0.872 [0.814, 0.912]	−0.103 [−0.345, 0.148]	0.672 [0.525, 0.784]
TS: Grok 4.1 Fast	0.475 [0.245, 0.641]	0.810 [0.728, 0.867]	0.000 [−0.250, 0.261]	0.696 [0.500, 0.835]
LightGBM	0.526 [0.316, 0.678]	0.805 [0.747, 0.852]	0.523 [0.382, 0.650]	0.639 [0.489, 0.766]

Table B15c. Paired gaps between type sampling and the survey-style method by question, with bootstrap 95% intervals in brackets.

Question	AUC gap	Holding-rate error gap (pp)
Stocks	0.084 [0.059, 0.108]	−10.5 [−12.5, −7.6]
Savings	0.232 [0.181, 0.279]	−8.8 [−11.3, −5.7]
Employer plan such as a 401(k)	0.077 [0.063, 0.093]	−1.5 [−8.4, 4.4]
IRA	0.095 [0.074, 0.115]	−9.1 [−14.5, −3.7]
Business	0.215 [0.148, 0.275]	−8.1 [−9.2, −5.1]
Home	0.093 [0.079, 0.107]	1.2 [−0.5, 3.0]
Vehicle	0.153 [0.115, 0.192]	−4.7 [−8.1, −0.6]
Other real estate	0.243 [0.190, 0.291]	−1.1 [−2.2, −0.2]
Other financial assets	0.068 [0.033, 0.101]	−8.4 [−10.1, −6.7]
Other debt	0.103 [0.082, 0.123]	−3.4 [−5.1, −2.3]
Mean	**0.136** **[0.123, 0.148]**	**−5.4** **[−6.6, −4.2]**

Note. Gaps are type sampling minus the survey-style method; brackets give paired bootstrap 95% intervals.

the average of fifty. The second asks whether the results of Table 1 depend on Kimi writing every type. For DS Flash 4.1 and Grok, the answering LLM answers a pool of types written by its own family and the pool written by Kimi K2.5. The self-written pool for DS Flash 4.1 was written by DS Flash and is answered at temperature 1.0 against 0.7 for the Kimi pool, and Grok answers its own pool at 0.6 against 1.0 for the Kimi pool. The self-written pools are nominally fifty types, and the number of items that vote per household ranges from 50 to 51 for DS Flash 4.1 and from 2 to 50 for Grok, whereas the Kimi pools have exactly fifty. Each question is scored on the households for which both rows have a valid prediction, which is all 500 in every case. The error and the AUC are computed for each LLM family and then averaged over the families in the row, and the Mean row averages the ten unrounded values. Tables B17a and B17b give the ten questions. Table B17c gives the counterpart for Table 2, the log R-squared of cell means when the survey-style amount question is asked fifty times and the fifty answers are averaged per household, for the three LLMs of Table 2, whose asking-once column is the first valid draw of the same fifty.

C Supervised learning benchmark for asset balances

The LightGBM benchmark (Ke et al., 2017) is trained on roughly 9,100 PSID 2019 households outside the evaluation set. Its inputs are seven of the eight state variables supplied to the language models: age, gender, marital status, children, education, household income, and net worth. Employment is excluded. Net worth uses the same definition as the simulated household background, the `NetWorthWithHome_2019` field.

For each asset, the benchmark fits two models. A classifier is trained on the holding indicator and predicts the probability of holding the asset. A separate regressor is trained on the balances of households recorded as holders and predicts the balance conditional on holding. Holders whose balance is missing in the survey are left out of this regressor's training rather than counted as zero. The product of the two predictions gives the expected balance, including nonholders' zeros. Both components use LightGBM; the second component is not a separate linear regression.

Table B17a. Holding-rate error in percentage points under repetition and under each type writer.

	Repetition		Type writer	
Question	Once	Fifty times	Own family	Kimi
Stocks	17.2	16.9	6.3	4.6
Savings	12.1	11.7	5.4	5.3
Employer plan such as a 401(k)	6.9	7.5	12.0	6.1
IRA	8.7	9.1	4.7	4.0
Business	10.0	10.0	4.7	1.2
Home	8.9	8.7	18.9	14.8
Vehicle	6.7	6.9	6.4	3.7
Other real estate	7.5	7.4	5.2	4.2
Other financial assets	19.3	19.3	19.9	19.8
Other debt	16.4	15.5	7.5	5.9
Mean	**11.4**	**11.3**	**9.1**	**7.0**

Note. Repetition averages DS Flash, Grok, and Kimi; the type-writer comparison averages DS Flash 4.1 and Grok. Versions, temperatures, and type-pool differences are described above.

Table B17b. AUC under repetition and under each type writer.

	Repetition		Type writer	
Question	Once	Fifty times	Own family	Kimi
Stocks	0.769	0.816	0.845	0.863
Savings	0.572	0.626	0.779	0.807
Employer plan such as a 401(k)	0.775	0.823	0.855	0.856
IRA	0.724	0.763	0.811	0.812
Business	0.522	0.555	0.688	0.756
Home	0.801	0.854	0.868	0.877
Vehicle	0.680	0.733	0.772	0.791
Other real estate	0.540	0.564	0.753	0.808
Other financial assets	0.656	0.697	0.724	0.737
Other debt	0.606	0.627	0.707	0.710
Mean	**0.664**	**0.706**	**0.780**	**0.802**

Note. Repetition averages DS Flash, Grok, and Kimi; the type-writer comparison averages DS Flash 4.1 and Grok. Versions, temperatures, and type-pool differences are described above.

Table B17c. Log R-squared of cell-mean dollar amounts when the survey-style question is asked fifty times.

Asset	DS Flash	GPT-5.4	Grok	Mean of three LLMs
Stocks	0.379	0.219	−0.308	0.097
Savings	0.477	0.905	0.785	0.722
Employer retirement plans	0.046	−0.299	−0.269	−0.174
IRA	0.288	0.541	0.195	0.341

The reported benchmark sets negative conditional-balance predictions to zero before multiplying by the predicted holding probability. Thus, its predicted balance is `p(holding | state) × max(predicted balance conditional on holding, 0)`. This is a post-prediction constraint in the implementation, not a requirement of LightGBM. The clipping comparison keeps the fitted models and evaluation cells fixed; the unclipped predictions and the comparison results are retained in the source files.

D FURTHER EXPLORATIONS: WEALTH AND AGE GROUPS, OTHER SURVEYS, AND A LATER WAVE

This section answers the two places where Section 5 points to Appendix D. D.1 reports the comparison by wealth and by age group. D.2 and D.3 report the three additional applications, to Japan, to China and to the 2023 PSID wave, each of which uses one answering LLM family. Table D2a gives the means for all four survey waves and Table D2b gives the questions one by one for Japan and China. In each of the three surveys the household background carries eight fields of the same kind as the main experiment, with financial assets in place of net worth for Japan and China. The question sets differ in composition and the surveys differ in year, so these figures record how far the method has been carried and do not compare the three countries.

D.1 HOUSEHOLD FINANCE BY WEALTH AND AGE

This subsection reports the comparison by wealth and by age that Section 5's paragraph on household finance points to. Tables D1a and D1b set the two ways of asking beside two supervised references trained on the survey, one given the same eight state variables the language models receive and one with net worth removed. Tables D1c and D1d turn to the amounts, where the three answering LLMs of Table 2 are set beside a single supervised reference whose inputs are not the same as either of the two used for ranking. All figures are computed within a group, so households are only ever compared with others in the same wealth quintile or the same age band, and figures from different groups are not comparable with the whole-sample figures of Table 1.

Table D1a. Ranking and holding-rate error within wealth quintiles, 100 households in each.

	AUC				MAE (pp)			
Wealth quintile	TS	Once	ML + NW	ML - NW	TS	Once	ML + NW	ML - NW
First, lowest net worth	0.796	0.653	0.857	0.786	4.4	8.3	2.4	10.3
Second	0.705	0.578	0.731	0.685	6.9	10.9	2.9	4.6
Third	0.683	0.566	0.677	0.677	9.4	13.2	2.1	4.9
Fourth	0.776	0.578	0.778	0.755	7.9	18.7	3.0	7.6
Fifth, highest net worth	0.720	0.567	0.679	0.744	12.0	20.4	2.3	14.9

Note. TS is type sampling; Once is asking once; ML + NW and ML - NW are supervised models with and without net worth.
Note. AUC is averaged over the six questions that have at least five holders and at least five non-holders in every one of the five quintiles, and over the seven answering LLM families for the first two columns. The error is the mean absolute difference between the predicted and the observed holding rate, in percentage points, over all ten questions and over the same seven families. The supervised references are single LightGBM models trained on the survey outside the evaluation sample.

Type sampling ranks households best in the lowest quintile and worst in the third, and the top quintile falls between them. The holding-rate error is largest in the top quintile under both ways of asking, and under type sampling it does not rise quintile by quintile, since the fourth is lower than the third. B.9 reports, for the same top quintile and for one answering family, how much of that error moves when the questions are taken from the survey.

Ranking under type sampling is lowest in the youngest band and close to its highest in the two oldest, and the holding-rate error varies little across the four. Neither measure shows the oldest households as the hard case.

Table D1b. Ranking and holding-rate error within age bands.

		AUC				MAE (pp)			
Age	N	TS	Once	ML + NW	ML - NW	TS	Once	ML + NW	ML - NW
18 to 34	125	0.763	0.632	0.785	0.713	7.7	11.1	3.3	7.0
35 to 49	125	0.798	0.642	0.796	0.718	7.3	13.1	2.9	4.4
50 to 64	125	0.812	0.671	0.809	0.723	6.5	13.6	2.6	2.7
65 and over	125	0.807	0.675	0.813	0.730	7.4	12.9	1.6	3.6

Note. The bands are the four the sample is drawn across. AUC here uses all ten questions, since every one of them has at least five holders and at least five non-holders in every band. Columns are defined as in Table D1a.

Table D1c. Coefficient of variation of the share of net worth held in each asset, holders only.

Wealth quintile	Asset	N	DeepSeek V4 Flash	GPT-5.4	Grok 4.1 Fast	Supervised reference	Observed
Fifth	Stocks	39	0.58	0.46	0.44	0.58	1.24
Fifth	Savings	88	0.55	0.33	0.39	0.73	1.40
Fifth	Employer retirement plan	34	0.42	0.35	0.32	1.04	0.88
Fifth	IRA	59	0.57	0.40	0.41	0.50	0.91
Fourth	Stocks	10	0.58	0.32	0.42	0.93	0.71
Fourth	Savings	85	0.48	0.34	0.39	0.54	1.24
Fourth	Employer retirement plan	41	0.36	0.22	0.21	0.79	1.11
Fourth	IRA	28	0.47	0.30	0.35	0.29	0.87
Third	Stocks	11	0.54	0.54	0.59	0.80	1.17
Third	Savings	86	0.40	0.29	0.35	0.56	1.19
Third	Employer retirement plan	32	0.47	0.32	0.24	0.81	1.08
Third	IRA	16	0.40	0.32	0.30	0.72	0.55

Note. The share is the asset's balance divided by net worth, computed only on households recorded as holding that asset, and the same households are used for all five columns in a row. A smaller coefficient of variation means the spread across households is narrower.

The three answering LLMs sit below the observed coefficient of variation in all twelve cells, so the simulated shares of net worth are more alike across households than the recorded ones on every asset and in all three quintiles shown. The supervised reference is below the observed value in nine of the twelve cells and above it in three, namely the employer plan in the top quintile, stocks in the fourth and the retirement account in the third, so the compression is not confined to the language models. The widest gap between the observed value and the narrowest of the three is on savings in the top quintile.

Table D1d. Median simulated savings balance as a multiple of the median observed balance, holders only.

Age band	Holders	Observed median ($)	DeepSeek V4 Flash	GPT-5.4	Grok 4.1 Fast	Supervised reference
18 to 34	97	10,000	1.32	1.69	1.64	1.72
35 to 49	104	8,000	1.60	1.93	1.66	1.54
50 to 64	100	9,000	2.69	2.49	2.66	1.49
65 and over	102	6,000	4.39	4.41	4.80	2.41

Note. The multiple is the median predicted balance divided by the median observed balance, among households recorded as holding a savings account, and the four predicted columns use the same households. The amount convention is the mean of the answers that report a balance above zero.

The multiple rises with age for all three answering LLMs and does not rise for the supervised reference, and in the oldest band the three sit at roughly twice the reference. The reference overstates the oldest band as well, so what separates the three is the rise with age and the size of the multiple rather than its direction. The scan below screened this multiple and did not carry it past its later requirements.

Two things bound what these tables can be read to mean. The supervised reference that is given net worth calculates it by adding up the very balances the ten questions ask about, so its holding-rate error, which stays between 2.1 and 3.0 percentage points across the five quintiles and between 1.6 and 3.3 across the four age bands, cannot be separated from that arithmetic relation, and it is not a clean measure of how well a group can be done. The two supervised references also disagree about the top quintile, since the one with net worth puts it in the bottom two of the five while the one without net worth puts it third of five and above the second and third quintiles, so nothing here supports the claim that wealthy households are inherently harder to rank.

The four tables above look at two fields chosen in advance. Section 5 also reports a scan that takes every recorded field in turn, and Table D1e gives what it found. Eight fields were scanned, namely age band, gender, marital status, number of children, education, employment, household income quintile and net worth quintile, and within each field three families of measures were computed, the within-group ranking, the within-group holding-rate error, and the median of the simulated amount against the median of the observed amount on each of the four assets. Groups smaller than 30 households were merged in advance, and each measure was compared with the average of all eligible groups of the same field under 5,000 random reassignments of the group labels. A combination of field, group and measure counts as found only when five requirements hold together, that the gap runs in the unfavourable direction, that it survives the false-discovery control, that it also survives a conservative version of that control which assumes no structure among the tests, that it exceeds the threshold set within its own field, and that a bootstrap interval over households excludes zero. Fifteen combinations pass the initial screening, nine of those pass the conservative control, seven have bootstrap intervals that exclude zero, and four satisfy all five.

Everything found lies on the holding rates and runs along the income or the wealth distribution. No group passes on ranking and none passes on the median amounts, so within this sample the scan gives no group whose ranking or whose amount levels fall short of the supervised references by more than the field allows. Two of the four are measured against the reference that is given net worth and two against the reference without it, and no combination is found against both, so which reference is used decides the answer in every case. The rounded-amount rerun leaves all four in place, which says that these four do not depend on the extra precision the references receive, and it says nothing about how much that precision contributes.

Table D1e. The four combinations of field, group and measure that satisfy every requirement of the scan.

Field	Quintile	N	Reference	Gap (pp)	Cutoff	Deviation: 95% interval	Rounded gap
Income	Fifth	103	With net worth	7.45	1.65	0.21 to 4.13	7.04
Net worth	Fifth	100	With net worth	9.71	1.67	1.30 to 5.55	9.04
Net worth	Second	100	Without net worth	2.26	1.75	0.32 to 4.46	2.36
Net worth	Third	100	Without net worth	4.50	1.75	3.12 to 5.94	4.50

Note. The gap is in percentage points of holding-rate error, and it is the error of type sampling minus the error of the named supervised reference on the same households. The threshold and the bootstrap interval refer to the deviation of this gap from the mean gap across all eligible groups of the same field, including the group itself. They do not refer to the gap level in the preceding column. Intervals resample households and are reported after selection by the permutation test, without adjustment for that selection or for multiple comparisons. The last column repeats the test against a reference whose income and net worth are rounded the way the prompt rounds them, and that reference is matched one for one to the reference named in the same row.

Several combinations were screened out at the later requirements, and they are worth naming because they are the places a reader would look first. In education the lowest group was screened on both its ranking and its holding-rate error against the reference with net worth and passed neither the conservative control nor the bootstrap check. The group that is neither employed nor retired was screened on ranking and its bootstrap interval excluded zero, but it did not pass the conservative control. Households with no children were screened on the holding-rate error against the reference without net worth with the same outcome. The youngest band was screened on the employer retirement plan amount and failed both checks. The second and third income quintiles were screened on ranking, and there the gap runs in the favourable direction, since type sampling is ahead of the reference without net worth rather than behind it. Failing a requirement here means only that the gap cannot be told apart from the group labels being shuffled at this sample size, not that the group is fine.

The savings of the oldest households belong here as well. The multiple in Table D1d was screened in the family that looks at the level the LLMs themselves reach, where the oldest band was flagged and survived the false-discovery control without surviving the conservative one. The excess of that multiple over the supervised reference has a p value of 0.033 and does not survive the false-discovery control at all, so Table D1d records how far the three sit from the reference and the scan does not establish that the distance is larger than the field allows.

Three limits bound the whole scan. The eight fields are related to one another, since retirement overlaps heavily with the oldest band and education is correlated with income, so a finding on one field may belong to another. The distribution of the fields that the sample was not stratified on is a by-product of drawing 500 households across net worth, education and age, so it does not describe the population. A combination that does not pass means only that the gap cannot be separated from a random reassignment of the labels at 500 households under these controls, and the power of the tests was not estimated.

D.2 Japan and China

This subsection covers the two additional surveys that the same sentence names. The Japanese data are the 2019 wave of a public opinion survey on household financial behavior, which the Social Science Japan Data Archive distributes as dataset 1639. Kimi K2.5 writes and answers fifty descriptions for each of 500 households. The wave carries nine holding items and the four reported in Table D2a are the financial asset items that also appear in the Chinese wave. The Chinese data are the 2011 wave of the China Household Finance Survey, with Grok 4.1 Fast writing and answering, and the same four items are reported out of seven.

Two entries in Table D2a need a word. China has a survey-style comparison within the family the table reports, since a no-type run exists for Grok 4.1 Fast as well as for Claude Haiku 4.5. Japan has no such comparison, because its only no-type run was answered by Claude Haiku 4.5 while the reported type-sampling arm was answered by Kimi K2.5, so the cells say that no arm of the same family exists. For the 2019 PSID wave an earlier appendix recorded a survey-style mean AUC of

Table D2a. Means across the reported holding questions, by survey wave.

Survey and wave	N	Items	LLM	AUC		MAE (pp)	
				survey style	type sampling	survey style	type sampling
Japan, household financial behavior, 2019	500	4 of 9	Kimi K2.5	no same-family arm	0.756	no same-family arm	8.0
China, CHFS, 2011	500	4 of 7	Grok 4.1 Fast	0.757	0.915	7.6	7.1
United States, PSID, 2019	500	10	DeepSeek V4 Pro	0.661	0.799	9.9	7.4
United States, PSID, 2023	500	10	DeepSeek V4 Pro	0.656	0.769	12.7	6.3

Note. Households are the evaluation sample of each wave. For Japan and China the reported questions are the four financial holding items, and Table D2b lists all nine and all seven. MAE is the mean absolute difference between predicted and observed holding rates in percentage points. Type sampling uses fifty descriptions per household written and answered by the family named in the row.

0.666. The file the manuscript now treats as authoritative gives 0.661 for that family and method, and Table D2a uses that value.

Table D2b. All holding questions of the Japanese and Chinese waves.

Survey	Question	Observed rate	Predicted rate	AUC	In the reported mean
Japan 2019	Stocks held directly	0.222	0.191	0.729	Yes
Japan 2019	Investment trusts	0.152	0.246	0.740	Yes
Japan 2019	Deposits	0.964	0.837	0.787	Yes
Japan 2019	Bonds	0.044	0.111	0.767	Yes
Japan 2019	Insurance products	0.696	0.176	0.763	No
Japan 2019	Home ownership	0.742	0.397	0.647	No
Japan 2019	Any borrowing	0.476	0.109	0.658	No
Japan 2019	NISA account	0.110	0.187	0.636	No
Japan 2019	iDeCo account	0.032	0.129	0.592	No
China 2011	Stocks	0.106	0.099	0.890	Yes
China 2011	Funds	0.052	0.132	0.843	Yes
China 2011	Bank deposits	0.588	0.730	0.968	Yes
China 2011	Bonds	0.004	0.061	0.959	Yes
China 2011	Commercial insurance	0.176	0.029	0.535	No
China 2011	Home ownership	0.916	0.613	0.694	No
China 2011	Any loan	0.092	0.149	0.613	No

Note. Observed and predicted rates are shares of the 500 households of each wave. The four items entering each reported mean are the financial asset items common to the two waves. Insurance products, home ownership, borrowing, and the two Japanese tax-favoured accounts are outside those means.

Table D2b shows three things the single mean hides. For Japan the four items in the mean are not the four best ranked, since the insurance item ranks third of the nine and is outside the mean, while the two items with a counterpart in the PSID holding list, borrowing and home ownership, rank sixth and seventh. For China the four in the mean are the four best ranked, so the three items outside it are the three on which the ranking is weakest. Extreme observed rates carry part of these figures, since the Chinese bonds item has two holders among the 500 households. Among the four items in the Chinese mean the deposit item shows the largest departure of the predicted rate from the observed one, and across all seven the largest departure is on home ownership.

Other answering families were run on the same households of the two waves, and the family reported here is not the strongest of them. On the four items of the Japanese wave Grok 4.1 Fast reaches a mean AUC of 0.770 against the 0.756 of the family reported, and on the four items of the Chinese wave Claude Haiku 4.5 reaches 0.921 against the 0.915 of the family reported.

D.3 THE 2023 PSID WAVE

This subsection covers the later wave of the main survey, which the same sentence names. It uses the same ten holding questions and the same background wording as the main experiment, and DeepSeek V4 Pro writes and answers its own descriptions in both waves. Between the two waves the holding-rate error under type sampling falls while its ranking loses ground, and both measures worsen under the survey-style method, so over the four years the two ways of asking move apart on holding-rate error and closer together on ranking. The employer retirement item carries part of the ranking loss, and the 2023 wave changed the sample definition for that item.

D.4 POLITICAL ATTITUDES AND TWO SURVEY EXPERIMENTS

Section 5 extends the method from household finance to political attitudes and to two survey experiments. This subsection gives the samples, the descriptions, and the numbers behind the sentences there.

Political attitudes. The respondents are 500 people from the 2022 General Social Survey, and the questions are 18 items with two answer options each. Four are political attitudes, the 2020 presidential vote, abortion for any reason, a police permit before buying a gun, and the death penalty. Two are social attitudes, whether government should reduce income differences and whether same-sex relations are wrong. The remaining twelve ask about the respondent's own circumstances, self-employment, health, satisfaction with finances, whether finances have improved, social class, happiness, spanking, union membership, the roles of men and women, job satisfaction, whether life is exciting, and whether there is a gun in the home. Each respondent is described in two ways, by the five fields that the survey shares with the household description of Section 3, age, marital status, education, work status, and income, and by twelve fields that add gender, race, religion, region, place of upbringing, party identification, and political views. Three methods answer from each description. Type sampling writes fifty types per respondent and takes one answer from each, asking once takes a single answer from the description alone, and the supervised benchmark is LightGBM trained on the survey's other respondents with the same fields and scored on the 500. DS Flash writes the types and gives the answers. Table D4 reports the mean AUC by question group.

Table D4. AUC on 18 General Social Survey questions for 500 respondents, by description and method.

	Five fields			Twelve fields		
	Type sampling	Asking once	LightGBM	Type sampling	Asking once	LightGBM
All 18 questions	0.62	0.57	0.61	0.71	0.64	0.71
Political attitudes (4)	0.58	0.54	0.57	0.80	0.73	0.80
Social attitudes (2)	0.59	0.55	0.58	0.81	0.74	0.82
Own circumstances (12)	0.63	0.59	0.63	0.66	0.59	0.67

Note. Each entry is the unweighted mean of per-question AUC over the questions in the row, scored on the respondents who answered them. With twelve fields type sampling exceeds asking once on every one of the 18 questions, by 0.03 to 0.17, and the presidential vote reaches 0.94 with type sampling and 0.73 when party identification and political views are removed from the twelve fields.

Howard's experiment. The study, from the archive of (Ashokkumar et al., 2026), asked 1,288 respondents of a probability panel to forecast their spending in the coming week. Respondents were assigned at random to a control group that forecast directly (452), a group that first wrote three reasons why the coming week would differ from a usual week (413), and a group that first wrote three reasons why it would be similar (423). We simulated 500 of these respondents, each described by the background fields the panel records, rendered in the household template of Section 3 without the net-worth line and with income at the midpoint of its bracket, with fifty types written by Kimi K2.6 and answers given by DS Flash. The simulated respondents wrote the three reasons before forecasting, as the real ones did. Table D5 reports the forecast bias, the log difference between the forecast for the coming week and the reported spending of the past week, for the real respondents and for three ways of simulating them, our type sampling, asking once from the same description, and the procedure of the archive, which describes the manipulation to the LLM instead of having it write the reasons.

Table D5. Forecast bias in Howard's experiment, real respondents and three simulations.

Group	Real respondents	Type sampling	Asking once	Archive procedure
Different week	0.18	0.31	0.27	−0.07
Similar week	−0.04	−0.02	−0.02	−0.17

Note. Entries are mean log forecast bias relative to the control group. In dollars the effect of the different-week manipulation is 148 for the real respondents, 166 for type sampling, 122 for asking once, and 22 for the archive procedure. At the level of individual respondents, the simulated forecasts correlate 0.30 with the real ones and the simulated past-week spending 0.30, and the AUC for willingness to pay for a dinner is 0.69. The archive scores only two attitude items for this study, so the forecast bias is not part of its own evaluation.

Kennedy's experiment. The study, from the same archive, showed 2,595 respondents one of four descriptions of a neighboring family that varied the family's environmental behavior and its wealth, and asked eight questions about the family. We reused the 500 simulated respondents of Howard's experiment as a panel, since their types were written from the background fields alone, and had each of them answer the eight questions under all four descriptions. The comparison is at the level of the 36 treatment contrasts the study reports, because the simulated panel and the real sample are different people. The simulated contrasts correlate 0.88 with the real ones and have the same sign in 32 of 36, which is also the number the archive's own ensemble of predictions gets right, although the two miss different contrasts. The simulated effects are larger than the real ones, about twice as large for the effects of environmental behavior on liking and friendship, and the mean absolute error is 0.050 against 0.034 for the ensemble, whose predictions are shrunk toward zero.

Table D6 reports the comparison of the 36 contrasts, and Table D7 the subgroup comparison.

Table D6. The 36 treatment contrasts of Kennedy's experiment, simulated panel against the real sample.

Source	Correlation with real contrasts	Same sign	Mean absolute error
Simulated panel, unweighted	0.88	32 of 36	0.050
Simulated panel, reweighted to the study's sample	0.89	32 of 36	0.049
Archive ensemble of predictions	0.87	32 of 36	0.034
Archive, one LLM alone	0.82	32 of 36	0.037

Note. Every row misses four contrasts, but not the same four. Reweighting matches the panel's composition to the study's sample and changes little, so composition is not what separates the panel from the real respondents.

Table D7. Subgroup patterns of the contrasts, simulated panel against the real sample.

Grouping field	Cells	Correlation with real subgroup deviations	Same sign
Fields in the description (income, age, education, gender, employment, marital status, children)	180	−0.27	37 percent
Fields not in the description (race, party identification, ideology)	90	−0.02	51 percent

Note. A cell is one field group, one question, and one kind of effect, and the entry compares the deviation of each subgroup's contrast from the mean contrast across the subgroups of that field. For the fields the description carries, the simulated deviations are as large as the real ones but smooth and monotone in the field, while the real ones are not; for the fields it does not carry, the simulated deviations are near zero.

Within groups of respondents defined by the fields the description carries, the pattern of the real effects is not reproduced, with a correlation of −0.27 over 180 cells, and for groups defined by

fields the description does not carry, race, party identification, and ideology, the simulated effects hardly vary at all.

D.5 A FICTIONAL CRISIS AND A PLACEBO: REASONING OR RECALL

Section 5 asks whether the agreement with the surveys could be recall of training data. This subsection reports a direct test run on the household description of Section 3. Each of 200 PSID households was described three more times, with a paragraph of life history added to the description while the same fifty types answered under each version. The first paragraph recounts the household's experience of the 2008 financial crisis, written for its age, with asset losses, a spell of unemployment, and tight credit. The second recounts a fictional crisis, the Torlund financial collapse, a name that appears nowhere, with the same economic consequences and no year or real marker. The third is a placebo, the serious illness and full recovery of someone close to the household, with no economic content. The effect of each history is the change in the type-averaged holding probability against the description without it, and the test asks whether the fictional crisis reproduces the effect of the real one. An LLM that recites what it has read about 2008 would respond to the real crisis and not to the fictional one, whereas an LLM that reasons from the economic content would respond to both alike. The families are DS Flash, Kimi, and Grok, and Table D8 reports the effects on three holdings.

Table D8. Effects of a real, a fictional, and a placebo life history on holding probabilities, 200 households.

Family	Holding	2008 crisis	Fictional crisis	Placebo	Fictional over real
DS Flash	Home ownership	0.026	0.026	−0.006	1.02 [0.89, 1.17]
DS Flash	Retirement accounts	0.027	0.021	0.003	0.77 [0.61, 0.91]
DS Flash	Stocks	−0.005	−0.003	−0.001	0.69 [0.35, 1.09]
Kimi	Home ownership	0.039	0.032	0.009	0.79 [0.63, 0.94]
Kimi	Retirement accounts	0.026	0.016	0.012	0.27 [−0.28, 0.61]
Kimi	Stocks	0.006	0.003	0.005	not identified
Grok	Home ownership	0.007	0.038	0.020	not identified
Grok	Retirement accounts	0.011	−0.108	−0.116	not identified
Grok	Stocks	−0.004	−0.012	−0.010	not identified

* Entries are changes in the type-averaged holding probability relative to the same household without the added history, in probability points, over 200 households and fifty types each. The last column is the fictional effect over the real effect, with the placebo effect subtracted from both where it is distinguishable from zero, and a 95 percent household bootstrap interval from 1,000 draws. A ratio is reported only where the real effect, after any placebo correction, is distinguishable from zero, keeps its sign, and is larger in size than the placebo effect, a rule added at analysis time and disclosed as such; Grok's placebo effects are as large as its real effects, so its ratios are not identified, and Kimi's stock effect is too small to measure. The pre-registered reading takes a ratio above 0.7 as reasoning-dominant and below 0.3 as recall-dominant.

E THE LOTTERY EXPERIMENT

E.1 THE ESTIMATOR

Employment status varies in two ways in the estimator, through the PSID shares $\pi_i(g)$ that weight the possible initial statuses and through the shares $\Pi_i^c(g, s)$ of types that themselves report each status under a given description.

Let i index households and $g \in \{W, R, O\}$ the employment statuses, working, retired, and otherwise not working, where the last includes looking for work. Without a lottery the household's status in

2019 is uncertain. The weight $\pi_i(g)$ is the share of PSID households in household i's 2009 cell that hold status g in 2019. These weights enter only the estimator and are never shown to the LLM. For each initial status g the household is described without the lottery ($c = 0$) and with it ($c = 1$), and its types answer under each description. The types are written separately for the two conditions, and each reports its own status. Their answers define a transition matrix and a value for each cell. The transition matrix $\Pi_i^c(g, s)$ is the share of household i's types with a usable answer to the question that, described with initial status g under condition c, report status s, so each row sums to one. The value $v_i^c(g, s)$ is the average outcome among those types, taken over positive amounts for balances. The expected outcome of household i under condition c given initial status g is therefore

$$E_i^c(g) = \sum_s \Pi_i^c(g, s)\, v_i^c(g, s),$$

The difference $E_i^1(g) - E_i^0(g)$ satisfies the following identity. Its first term is the change within employment statuses at fixed shares, and its second is movement between them at fixed values.

$$E_i^1(g) - E_i^0(g) = \sum_s \Pi_i^0(g, s)\left[v_i^1(g, s) - v_i^0(g, s)\right] + \sum_s \left[\Pi_i^1(g, s) - \Pi_i^0(g, s)\right] v_i^1(g, s).$$

The estimator we report keeps the change within employment statuses only for $s = g$, the types that report the status in the description. It values movement between them at the destination status's reference level $\bar{v}(s)$, common to all households, rather than at the movers' own answers. Since the two conditions have separately written types, movement here means a change in the shares reporting each status, not a transition observed for an individual type.

$$\hat{\tau}_i^g = \Pi_i^0(g, g)\left[v_i^1(g, g) - v_i^0(g, g)\right] + \sum_s \left[\Pi_i^1(g, s) - \Pi_i^0(g, s)\right] \bar{v}(s).$$

For hours, $\bar{v}(W)$ is the average hours of the types that report working when described as working without the lottery, pooled across households, and $\bar{v}(R) = \bar{v}(O) = 0$. For the employment indicators $\bar{v}(s)$ is one for the indicated status and zero otherwise. For holdings and balances $\bar{v}(s)$ is the average of the outcome among the types that report status s when described with initial status s without the lottery, pooled across households and over positive amounts for balances. The household's effect is the expectation of $\hat{\tau}_i^g$ over its initial status,

$$\hat{\tau}_i = \sum_g \pi_i(g)\, \hat{\tau}_i^g,$$

and Section 4 reports the mean of $\hat{\tau}_i$ across households, with the within-status part and the between-status part given by the means of the first and the second term of $\hat{\tau}_i^g$ weighted in the same way. In forming $\pi_i(g)$ the PSID categories disabled, homemaker, looking for work, and student are combined into O and the three probabilities are normalized to sum to one.

The transition shares and the values are formed over the answers whose reported status and whose outcome can both be read, so their denominator is at most fifty and each row of Π_i^c still sums to one across the three statuses. For a balance an answer counts when it reports holding the item together with an amount that is not negative, and it counts as zero when it reports not holding the item, while $v_i^c(g, s)$ is the mean of the counted answers that give a positive amount.

The reference level $\bar{v}(s)$ is measured on the untreated answers of the whole sample rather than on the household itself. For the indicator that a type reports working, and likewise for the indicator that a type reports being retired, $v_i^1(g, g)$ and $v_i^0(g, g)$ are equal by construction, so those two outcomes have no within-status part.

A household enters the calculation for an outcome only when each of its six arms, three employment statuses crossed with two conditions, returns at least five answers whose reported employment status can be read. The floor is applied to those answers rather than to the answers that give the outcome itself. All two hundred households clear it for DeepSeek V4-Flash, and 197 clear it for Kimi K2.5,

which is the sample behind the hours estimate reported in Section 4. Observed data on lottery winners cannot be split in this way, because a winner is recorded in one employment status only and the counterfactual state is not observed.

Each household has one description for each of the three employment statuses under each condition, and each of these six descriptions is posed to the household's fifty types. The transition share $\Pi_i^c(g, s)$ is estimated as the number of those types that report status s in answer to the description with status g under condition c, divided by the number of their answers whose reported status can be read. Households with fewer than five such answers in any of the six arms are dropped, which leaves 200 households for DS Flash and 197 for Kimi, and an arm holds on average between 49.8 and 50 readable answers in both families. Every answer with a readable status also gives a readable number of hours, so these are the shares that enter the hours effect. Table E1a averages the shares at the 150,000 dollar annuity over households and weights each household's row for status g by its projected probability $\pi_i(g)$, as the between-status part of the estimator does. With these weights, the change in the share of types reporting working in each row, weighted by the average projected probabilities in the note to the table and multiplied by $\bar{v}(W)$ of 40.1 hours for DS Flash and 41.0 for Kimi, gives -0.394 and -0.736 hours, which are 0.110 and 0.196 of the hours effects of -3.58 and -3.75 hours in Table 3, so the table reproduces the between-status part of the hours effect.

Table E1a. Employment-status transition shares at the 150,000 dollar annuity.

		Control			Treated		
LLM	Described status	Working	Retired	Other	Working	Retired	Other
DS Flash	Working	0.991	0.000	0.008	0.970	0.007	0.022
DS Flash	Retired	0.093	0.876	0.030	0.098	0.879	0.023
DS Flash	Other	0.047	0.015	0.938	0.060	0.025	0.915
Kimi	Working	0.998	0.000	0.002	0.980	0.005	0.015
Kimi	Retired	0.087	0.835	0.078	0.067	0.848	0.085
Kimi	Other	0.046	0.007	0.947	0.032	0.006	0.961

Note. Other means otherwise not working. Rows give the employment status g set in the description and columns the status s that the types report, in the description without the prize (control, $c = 0$) and with it (treated, $c = 1$). Each entry is $\sum_i \pi_i(g)\,\Pi_i^c(g, s) / \sum_i \pi_i(g)$ over 200 households for DS Flash and 197 for Kimi. The average projected probabilities of working, retired, and otherwise not working are 0.595, 0.288, and 0.117 in DS Flash and 0.598, 0.284, and 0.118 in Kimi. Each row sums to one before rounding, and four rows sum to 0.999 after rounding. Computed from the rounded entries and probabilities, the reconciliation in the text gives 0.107 and 0.198 in place of 0.110 and 0.196.

When a household is described as working, almost all of its types report working without the prize, and the prize lowers that share by about two percentage points in both families. An average that weights every household equally differs mainly in the retired row's shift toward reporting work, because that shift is concentrated in households with a small projected probability of retirement, so equal-weight averages combined with the average projected probabilities do not reproduce the between-status part of the hours effect.

E.2 The household history in the prompt and its source years

The 200 households are drawn from the PSID households interviewed in every wave from 2007 to 2019, and the main text describes them by the two waves it uses, 2009 and 2019. Income enters the design in two places. The 2009 income that defines a household's cell for projecting its 2019 state without a lottery is the 2009 interview value. The shared part of each household's description, by contrast, states its income and net worth ten years before 2019 as its situation in 2009. The net worth is the value recorded at the household's 2009 interview. The income is the value recorded at its 2007 interview, expressed in 2019 dollars, which the persona file built for this experiment carried as the pre-lottery baseline, although the 2009 interview income is also in that file and is the value used to project the household's 2019 state without a lottery. Across the 200 households, the median income is 72,219 dollars at the 2007 interview and 66,914 dollars at the 2009 interview, the means are 83,957 and 85,432 dollars, and the median absolute change between the two interviews is 18 percent. Nearly half of the households changed by more than 20 percent and 22 percent of them by more than 50 percent. The sentence is identical in the control and the treated description, so both conditions receive the same history. Whether replacing it with the 2009 interview income would

change the estimated effects has not been examined. The age in the prompt is the 2019 age minus ten, which differs from the age recorded at the 2009 interview for 33 of the 200 households because of interview timing. The projected 2019 income and net worth come from the 2019 outcomes of PSID households in the household's 2009 age, income, and wealth cell, and when that cell holds fewer than ten households the projection uses the cell defined by age and income alone, which occurs for 19 of the 200 households.

E.3 Net-worth answers entered in dollars rather than in thousands

This subsection gives the unit rule used for Table E9, under which flagged net-worth answers are divided by 1,000.

Every amount question asks for thousands of dollars, and the background paragraph of the prompt prints the projected figures in dollars. Part of the answers copy a printed dollar figure into a thousands field, which reads a household with 55,510 dollars of projected net worth as holding 55,510,000 dollars. Two rules flag such an answer to the net-worth question. The first is internal, and it fires when the four asset amounts in the same answer sum to a positive figure, the net-worth answer is at least 3,000 in thousands, which is the whole nominal prize, and the net-worth answer is at least thirty times that sum. The second compares the answer with the dollar figures that the prompt itself printed, and it fires when the answer is within five parts in a thousand of one of them and that figure is at least 1,000 dollars in absolute value. The main specification divides a flagged answer by 1,000, and the sensitivity drops the flagged cell.

Table E1. Net-worth answers flagged by the unit rule, by LLM family.

			Flagged			Treated		
LLM	N	Answers read	Internal	Printed figure	Total	Answers	Flagged	Flagged (%)
DeepSeek V4-Flash	188	40,000	6	45	46	20,000	46	0.23%
Kimi K2.5	188	59,671	4	9	9	29,851	9	0.03%
Grok 4.1 Fast	94	29,989	467	950	960	14,993	957	6.38%
GPT-5-mini	188	59,720	1	41	41	29,740	40	0.13%
Claude Haiku 4.5	188	59,984	3	72	72	29,992	71	0.24%

Note. The households column is the sample that Table E9 uses for the family, and the answer counts cover every answer read in the batches that sample draws on. The two rules overlap, so the total is below their sum. Answers are pooled over the three employment assignments and the two conditions, and each amount field is read at its last occurrence in the answer. DeepSeek V4-Flash and Grok 4.1 Fast read fewer answers than the other families, because one of their batches does not contain the net-worth question.

Only Grok 4.1 Fast is affected at a scale that moves Table E9. Of its 960 flagged net-worth answers, 950 equal the projected net worth that the prompt prints for the untreated trajectory, so dividing them by 1,000 leaves an answer in which the prize has added nothing to reported wealth, and the remaining 10 are flagged by the internal rule alone. The rule reaches the other amount fields as well. For Grok 4.1 Fast it flags one retirement balance and twelve savings balances, and it flags fifteen saving amounts through the net-worth answer of the same response. It flags no amount field of the other four families.

Twelve of Grok's flagged answers write the conversion out, for example as a dollar figure divided by one thousand followed by the result. The reading used throughout takes the last occurrence of the field label and then the first number after it, which for those answers is the figure before the conversion. Taking the figure after the equals sign instead leaves both of Grok's Table E9 statistics unchanged and moves the flagged count from 960 to 948, and it changes nothing for the other four families.

E.4 Control consistency and treatment recognition

This subsection gives the questions, samples, and definitions of control consistency and treatment recognition as stated in Section 4.

Table E2. Table E9's treatment columns under the three treatments of a flagged answer.

		Median gain			Within band		
LLM	N	Original	Rescaled	Dropped	Original	Rescaled	Dropped
DeepSeek V4-Flash	188	28.47%	27.39%	27.44%	99.47%	100.00%	100.00%
Kimi K2.5	188	14.11%	14.07%	14.07%	87.23%	87.23%	87.23%
Grok 4.1 Fast	94	93.48%	17.22%	19.24%	40.43%	86.17%	87.23%
GPT-5-mini	188	5.11%	5.02%	5.02%	50.53%	50.00%	50.00%
Claude Haiku 4.5	188	6.17%	6.17%	6.17%	54.26%	54.26%	54.26%

Note. Original retains flagged answers; Rescaled divides them by 1,000; Dropped excludes flagged cells.
Note. Table E9 prints the column divided by 1,000. Median gain is the median across households of the reported net-worth increase as a share of the 1,500,000 dollars received, and Within band is the share of households whose increase lies between 5% and 100% of that cash.

Control consistency is checked on the untreated answers under the working assignment. The comparison pool is the PSID households in the 2017 to 2019 balanced extract whose age is between 25 and 75 and whose age, net worth, income, and education are all recorded, which leaves 8,400 households. Each household and each pool member is placed in a cell defined by the net-worth quintile of the pool and one of six age brackets, and within the household's own cell the twenty nearest pool members are taken by Euclidean distance on four standardised features, which are age in 2019, the inverse hyperbolic sine of net worth, the logarithm of income, and years of education. The household's simulated value is the mean over its fifty types, and for an amount it is the mean over the answers that report holding the item and give a positive figure. The benchmark for an amount is the same positive-only mean over the twenty neighbours, and the benchmark for a holding is their holding rate.

The supervised reference beside which Table E9's control columns are read is a different object from the benchmark of Appendix C. It is a LightGBM pair trained on the 9,369 households of the same balanced extract that are not among the two hundred, with contemporaneous 2019 covariates, so it is an upper bound on what supervised learning can extract from this survey rather than a like-for-like competitor of the simulated answers. On the four holdings it is scored against the same neighbourhood benchmark as the families, and on retirement-account holding it reaches 0.888 while the five families reach 0.853 to 0.882.

In the implementation cited below, the lottery prompt puts the following question block to the types after the household's background and the type description. It is reproduced verbatim, and the net-worth question and the factual questions that treatment recognition uses are repeated on their own below, where their coverage across batches is also recorded.

```
Q1.  In a typical week, how many hours do you currently
work?

Q2.  What were you doing most of last week?
A) working                    B) with a job but not at work
C) keeping house              D) retired
E) looking for work           F) unable to work / disabled
G) other

Q3.  What was your current spouse doing most of last week?
A) working         B) keeping house   C) retired
D) other           E) no spouse

Q4.  Do you currently own or rent your home?
A) own    B) rent

For each asset, first answer Y/N whether you have it.  If
Yes, then answer the approximate $ amount in thousands.  If
No, write 0 for the amount.

Q5a:  Do you own your home?  (Y = own, N = rent or don't
have a home)
```

```
Q5b:  If Yes, approximate current value of home in thousands
of dollars.

Q6a:  Do you have a mortgage on your primary home?
Q6b:  If Yes, approximate current mortgage balance in
thousands of dollars.

Q7a:  Do you have any retirement accounts (IRAs, 401Ks,
pensions)?
Q7b:  If Yes, approximate total in retirement accounts in
thousands of dollars.

Q8a:  Do you have any stocks, bonds, or mutual funds (NOT
counting retirement accounts)?
Q8b:  If Yes, approximate total in thousands of dollars.

Q9a:  Do you have any savings accounts or CDs?
Q9b:  If Yes, approximate total in savings accounts or CDs
in thousands of dollars.

Q10.  What percentage of your after-tax income will you save
or invest THIS year?
A) 0%    B) 1-5%    C) 6-10%    D) 11-20%    E) 21-30%    F)
more than 30%

Q11.  What percentage do you plan to save NEXT year?
A) 0%    B) 1-5%    C) 6-10%    D) 11-20%    E) 21-30%    F)
more than 30%

Q12.  Are you generally happy?
A) Yes   B) No

Q13.  What is your approximate current total net worth, in
thousands of dollars?
(Net worth = total assets [home value, retirement accounts,
stocks, savings,
vehicles, business value, other real estate] minus total
debts [mortgage
balance, credit card debt, loans].  May be negative.)

The next 8 questions are adapted from the U.S. Panel Study
of Income Dynamics (PSID).
Answer each as Yes or No based on your current situation.

P1.  Do you (or anyone in your family living there) have any
shares of stock in publicly held corporations, stock mutual
funds, or investment trusts -- NOT including stocks in
retirement accounts?
P2.  Do you (or anyone in your family living there) have
checking or savings accounts, including money market
accounts?
P3.  Do you (or anyone in your family living there) have
CDs, government savings bonds, or Treasury bills?
P4.  Do you (or anyone in your family living there) have an
Individual Retirement Account (IRA) that you set up
yourself?
P5.  Do you (or anyone in your family living there) have any
real estate OTHER than your main home -- such as a second
home, land, rental real estate, or money owed to you on a
land contract?
P6.  Do you (or anyone in your family living there)
currently own part or all of a farm or business?
P7.  Do you (or anyone in your family living there)
currently have credit card debt?
```

```
P8.  Since 2014, have you (or anyone in your family living
there) received an inheritance, in money or property?

Q14.  Approximately how much money have you saved (set aside
as new savings or
investments) over the PAST 2 YEARS, in thousands of dollars?
(0 if none; can be negative if you spent down savings.)

Q15a.  Have you received any lottery winnings or annuity
payments since 2009?  (Y/N)
Q15b.  If Yes, total cash already received over the years so
far, in thousands of dollars?
(0 if N or none.)
Q15c.  If Yes, approximate yearly payment amount in
thousands of dollars?
(0 if N or none.)

Format your answer EXACTLY as:
Q1_hrs:  <number>
Q2:  X  Q3:  X  Q4:  X
Q5a:  Y/N  Q5b:  <number>
Q6a:  Y/N  Q6b:  <number>
Q7a:  Y/N  Q7b:  <number>
Q8a:  Y/N  Q8b:  <number>
Q9a:  Y/N  Q9b:  <number>
Q10:  X  Q11:  X  Q12:  X  Q13:  <number>
P1:  Y/N  P2:  Y/N  P3:  Y/N  P4:  Y/N
P5:  Y/N  P6:  Y/N  P7:  Y/N  P8:  Y/N
Q14:  <number>
Q15a:  Y/N  Q15b:  <number>  Q15c:  <number>
```

Treatment recognition is checked on the net-worth question. Its wording and the wording of the factual questions are reproduced below as the implementation states them.

```
Q13.  What is your approximate current total net worth, in
thousands of dollars?
(Net worth = total assets [home value, retirement accounts,
stocks, savings,
vehicles, business value, other real estate] minus total
debts [mortgage
balance, credit card debt, loans].  May be negative.)

Q15a.  Have you received any lottery winnings or annuity
payments since 2009?  (Y/N)
Q15b.  If Yes, total cash already received over the years so
far, in thousands of dollars?
(0 if N or none.)
Q15c.  If Yes, approximate yearly payment amount in
thousands of dollars?
(0 if N or none.)
```

In the untreated condition the household's mean answer is compared with the projected net worth printed in the prompt, and a ratio near one is the prerequisite for reading any difference between the conditions. The median ratio is 1.031 for DeepSeek V4-Flash, 1.000 for Kimi K2.5, Grok 4.1 Fast, and GPT-5-mini, and 1.115 for Claude Haiku 4.5. In the treated condition the mean answer is converted to dollars, the same projection is subtracted, and the difference is divided by the 1,500,000 dollars received over ten annual payments. A share near zero means that the annuity is being treated as income that is spent as it arrives, a share near one means that the whole of the cash received is counted as wealth, and a share above one exceeds the cash on hand. The 5% to 100% band was chosen on this sample with those two readings in mind, and it is not an accounting identity, since consumption, asset returns, and new debt all move net worth. A single answer outside the band therefore does not by itself show that the treatment was not registered.

The factual questions are not present in every batch. Across the treated arms of the two phases and the three assignments, 29,851 of Kimi K2.5's recorded responses contain them, while none of the 30,000 recorded responses of DeepSeek V4-Flash, none of Grok 4.1 Fast's 30,000, none of GPT-5-mini's 29,995, and none of Claude Haiku 4.5's 30,000 do. Among Kimi K2.5's 29,851 treated answers, 99.79% state that winnings or annuity payments have been received, 98.06% give the cash received as exactly 1,500 thousand dollars, and 99.68% give the yearly payment as exactly 150 thousand dollars. A separate DeepSeek V4-Flash batch does put the factual questions to two households with fifty answers each, which is not a panel and is not comparable with these shares. A pilot of fifty Kimi households showed that accurate answers to the factual questions can coexist with many types reporting net worth close to the untreated projection, which is the reason for separating the two parts of treatment recognition.

A separate comparison chooses the neighbours from the household's own reported net worth instead of the projection, with projected income in the untreated condition and projected income plus the annual payment in the treated condition, and it asks whether the asset answers resemble those of real households at the resulting income and wealth. Under the treated condition the fit in logs is negative on all five amounts for DeepSeek V4-Flash, from minus 0.13 on stocks outside retirement accounts to minus 3.50 on house value, and it is negative on four of the five for Kimi K2.5, whose retirement balance reaches 0.05. The reference households here are ordinary PSID households at similar income and wealth rather than observed winners, so these gaps describe how the simulated winners differ from ordinary households in the same state, and they neither establish a wrong response to winning nor a distinctive correct one.

E.5 Internal consistency

This subsection gives the definitions and the full set of figures behind internal consistency, which Section 4 reports for stock balances and for saving.

Internal consistency is examined on the answers under the working assignment. The first test pairs the stock balance and the net-worth answer inside one response, keeps the responses in which both are positive, and asks how often the stock balance exceeds the reported net worth. The second annualises the amount saved over the past two years, and it asks how often that annual figure exceeds the household's projected annual income plus the annual prize payment. Because a household that received 1,500,000 dollars ten years earlier could have earned a return on it, the budget is widened in two further versions by five and by ten percent of the cash received.

Table E3. Internal consistency, by LLM family and condition.

		Stocks		Saving	Saving above budget		
LLM	Condition	N	Above net worth	N	No return	5% return	10% return
DeepSeek V4-Flash	Untreated	2,100	0.62%	10,000	0.01%	0.01%	0.01%
DeepSeek V4-Flash	Treated	4,082	1.47%	10,000	0.07%	0.06%	0.01%
Kimi K2.5	Untreated	2,473	0.53%	9,918	0.21%	0.21%	0.21%
Kimi K2.5	Treated	4,404	4.02%	9,953	1.24%	0.87%	0.75%
Grok 4.1 Fast	Untreated	1,239	0.32%	5,000	2.96%	2.96%	2.96%
Grok 4.1 Fast	Treated	1,992	9.54%	4,997	11.65%	5.42%	2.98%
GPT-5-mini	Untreated	2,998	0.10%	9,997	0.07%	0.07%	0.07%
GPT-5-mini	Treated	4,285	1.45%	9,907	0.01%	0.00%	0.00%
Claude Haiku 4.5	Untreated	3,009	1.00%	9,997	0.25%	0.25%	0.25%
Claude Haiku 4.5	Treated	5,616	11.54%	9,995	1.29%	0.88%	0.68%

Note. Answers are those given under the working assignment in the two phases of the experiment, with the unit rule's main specification applied. The saving columns use the projected annual income of the household, and the prize payment of 150,000 dollars enters only in the treated condition.

Two readings of these columns need care. The saving test uses projected income rather than recorded income, and a household could have other resources such as an inheritance, so an answer above the budget is a comparative diagnostic and not a certain contradiction. The stock test is affected by the unit rule, and among Grok 4.1 Fast's treated answers the share falls from 9.54% to 5.40% once

the flagged responses are set aside, whereas for Claude Haiku 4.5 the two figures are 11.54% and 11.41%. The stock discrepancies in Claude Haiku 4.5 therefore remain after excluding answers flagged by the unit rule.

Among treated answers whose stock balance exceeds their reported net worth, most repeat the description's no-lottery net worth to within 2%, including 619 of 648 in Claude Haiku 4.5. Table E3a also compares the reported total with the sum of home value, retirement accounts, stocks and savings, less the mortgage. Missing components enter that sum as zero. These flags identify stock balances above net worth, not the unit-rule flags discussed above.

Table E3a. Reported totals in treated answers with stocks above net worth.

LLM	Flagged N	At prompt value	Parts at least stocks	Parts at least twice the total
DeepSeek V4-Flash	60	46	60	57
Kimi K2.5	177	146	177	171
Grok 4.1 Fast	190	174	190	186
GPT-5-mini	62	55	57	52
Claude Haiku 4.5	648	619	647	637

Note. Entries are answer counts. "At prompt value" means within 2% of the positive no-lottery net worth in the description. Parts are the listed assets less the mortgage; total is reported net worth. The sample and unit correction are the same as in Table E3.

E.6 Hours answers to the same type description across families

This subsection reports the agreement behind Section 4's statement that the families' hours answers to the same type description are highly correlated.

The same fifty type descriptions per household are answered by all five families, so the hours answers can be compared description by description. Removing the household means first leaves the variation between descriptions within a household, which is the object of interest here, and the comparison is made on the answers under the working assignment.

Table E4. Agreement of hours answers across families for the same type description.

Quantity	Untreated	Treated
Correlation between two families, household means removed, mean over the ten pairs	0.849	0.876
Range of that correlation over the ten pairs	0.819 to 0.868	0.843 to 0.904
Same family asked a second time, household means removed	0.900	0.894
Intraclass correlation of the description, household means removed	0.846	0.872
Resampling interval for that intraclass correlation	0.835 to 0.857	0.863 to 0.883
Share of descriptions on which all five families give the same hours	24.8%	29.4%
Share on which the highest and lowest answers differ by at most five hours	61.8%	61.9%
Share of descriptions that state a weekly hours figure	8.1%	6.2%
Intraclass correlation among the descriptions that do not state one	0.844	0.876

Note. The intraclass correlation is the description component divided by the sum of the description and residual components of a five-component decomposition into family, household, family by household, description within household, and residual. The interval comes from 200 household-level resamples. The second-run benchmark is DeepSeek V4-Flash's first phase answered twice.

The household-level hours effects are also close across families. The five mean effects run from minus 4.892 to minus 3.828 hours over the two hundred households, their largest spread is 1.065 hours with a resampling interval of 0.917 to 1.286, the pairwise correlations of the effects run from 0.814 to 0.939, and a decomposition of the effects assigns 84.6% to the household, 3.2% to the

family, and 12.2% to the residual. Reported net worth behaves differently. On the hundred households whose batch contains the net-worth question, its effect decomposes into 22.9% household and 50.2% family, and its pairwise correlations run from 0.122 to 0.774.

E.7 The 2x2 supplementary experiment

The treated condition of the main experiment changes two things at once. The answering prompt gains two sentences that state the prize, and the fifty types are the ones written for the treated history. A supplementary experiment crosses the presence or absence of the prize sentences with the untreated or treated type pool. It reuses the main experiment's type pools without rewriting them, and DeepSeek V4.1 Flash answers all four combinations for the two hundred households, fifty types, and three employment assignments, which is 120,000 answers. The two sentences are reproduced below.

```
That same year (2009), you also won a major lottery prize --
a 20-year annuity
paying $150,000 per year (total $3,000,000 over the 20-year
payout).  You have
been receiving these annual payments since 2009 and have 10
more years remaining.
```

Table E5. Contributions of the prize sentences and of the treated types, DeepSeek V4.1 Flash.

Contrast	Weekly hours	Working share	Net worth (000 USD)	Two-year saving (000 USD)
Prize sentences, untreated types	−0.71	+0.002	+435	+82
Prize sentences, treated types	−0.68	+0.002	+319	+76
Treated types, no prize sentences	−2.62	−0.011	+175	+39
Treated types, with prize sentences	−2.57	−0.011	+58	+32
Both changes together	−3.30	−0.010	+494	+114
Interaction	+0.03	0.000	−116	−6

Note. Two hundred households, the estimator of E.1 with the three assignments weighted, the unit rule's main specification, and 1,000 household-level resamples. The interval for the hours total is −3.65 to −2.94, and the interaction on hours lies between −0.08 and +0.13.

The change of types accounts for roughly four fifths of the reduction in hours, the two factors are close to additive there, and the fall in the share reporting working sits almost entirely on the types. Reported net worth and saving run the other way, since most of their increase follows the prize sentences and the interaction between the two factors is negative. The comparison cannot say at which stage a response is formed, because about 92% of the treated type descriptions already mention the lottery or the annuity against about 3% of the untreated ones, and answering under the treated types alone raises the share of answers reporting that winnings have been received by 0.53. The experiment uses DeepSeek V4.1 Flash rather than the main experiment's V4 Flash, so these shares do not decompose the main results.

E.8 Details behind Table 3

The original asset estimates in Table 3 are from Table 6, specification VIII, of the published article, which covers winners of annual prizes below 100,000 dollars and reports effects per dollar of accumulated payments of 0.158 (standard error 0.056) for total financial assets, 0.183 (0.061) for financial assets outside retirement accounts, and 0.037 (0.027) for house value. For these three outcomes, the simulated values in Table 3 lie within one standard error of the original estimates except for Kimi's financial assets outside retirement accounts. Retirement accounts are not part of this comparison. The mortgage comparison uses the working-paper version of the study, in which the increase in mortgage debt was 72 to 77 percent of the increase in house value, against 13 and 15 percent in DS Flash and Kimi. The hours estimate of the working paper, a reduction of 9.3 hours a week with a standard error of 2.6, compares winners of about 80,000 dollars a year with winners of about 15,000, both in 1986 dollars, whereas the simulation compares winners of 150,000 dollars a year with non-winners. Valuing the simulated prize in 1986 dollars gives about 76,600 dollars at

2009 prices and about 64,300 at 2019 prices, and per dollar of annual prize the simulated reductions of 3.58 hours in DS Flash and 3.75 in Kimi are 0.33 to 0.41 of the working paper's estimate and 0.74 to 0.92 of its estimate that is linear in the prize. Kimi's estimates use the 197 households with valid answers under all three employment assignments. Changes in the shares of households in each employment state account for 11 percent of the hours effect in DS Flash and 20 percent in Kimi.

E.9 Prize tiers and payout timing

This subsection reports the prize tiers and the payout timing behind Section 4's paragraph on several prizes, including the households that answer at all four prizes.

The same households were also given annuities of 25,000 and 75,000 dollars a year and a lump sum of 3,000,000 dollars paid in 2009. The lump sum has the same nominal total as the 150,000 dollar annuity, and by 2019 it has delivered twice as much cash, so it is a comparison of payout timing rather than a fourth step in prize size. The tier comparison uses the households for which each family produces an effect at all four prizes, which is two hundred for DeepSeek V4-Flash and 195 for Kimi K2.5.

Table E6. Responses across the four prizes.

LLM	Prize	Cash received by 2019	Weekly hours	Share of the hours effect from changes in reported states	Financial assets per dollar of cash received
DeepSeek V4-Flash	25,000 a year	250,000	−2.04	6.9%	0.094
DeepSeek V4-Flash	75,000 a year	750,000	−3.01	10.4%	0.162
DeepSeek V4-Flash	150,000 a year	1,500,000	−3.58	11.0%	0.204
DeepSeek V4-Flash	3,000,000 at once	3,000,000	−2.95	21.7%	0.321
Kimi K2.5	25,000 a year	250,000	−2.16	16.4%	0.076
Kimi K2.5	75,000 a year	750,000	−3.23	18.2%	0.142
Kimi K2.5	150,000 a year	1,500,000	−3.73	19.6%	0.142
Kimi K2.5	3,000,000 at once	3,000,000	−3.41	26.2%	0.394

Note. Two hundred households for DeepSeek V4-Flash and 195 for Kimi K2.5, the estimator of E.1 with the three assignments weighted, and the unit rule's main specification. Financial assets are the sum of the retirement, stock, and savings balances. Kimi K2.5's figure at 0.142 is 0.142200 at the 75,000 dollar annuity and 0.141561 at 150,000, so its per-dollar response does not rise between the two larger annuities. At 150,000 dollars a year, the hours effect on Kimi K2.5's own full sample of 197 households is −3.75, which is the figure Section 4 reports for the main experiment.

Among Swedish lottery winners, the point estimates of the effect on labor earnings for the lump-sum and installment versions of one lottery have a ratio of 0.88, and a joint test across four lottery types does not reject equal effects (Cesarini et al., 2017), which offers a descriptive reference rather than a validation of the payout-timing effect.

Two limits apply to the smaller prizes. Applying the band of treatment recognition unchanged at the smallest annuity leaves Kimi K2.5's median gain at 4.09% and the share of households inside the band at 43.6% against 87.2% at the largest annuity, with a further 9.0% of households reporting less wealth than the untreated projection. DeepSeek V4-Flash stays above 22% at all four prizes.

When the band for the wealth gain, which is calibrated at the 150,000-dollar annuity, is applied to other prizes, whether a family meets treatment recognition also depends on the size of the treatment, since at the smallest annuity the median gain for Kimi falls below the band's lower bound. Whether that band is valid across prizes has not been examined, so the responses to the smaller prizes remain descriptive until the criteria are examined again at each prize.

Table E6a. Lottery effects across prizes on the same households.

LLM and prize	Financial assets	Outside retirement	Retirement	House value	Mortgage share	Hours	Employment share
DS Flash, 25,000 a year	0.094	0.083	0.010	0.040	0.022	−2.04	0.069
DS Flash, 75,000 a year	0.162	0.125	0.037	0.055	0.150	−3.01	0.104
DS Flash, 150,000 a year	0.204	0.160	0.044	0.047	0.128	−3.58	0.110
DS Flash, 3,000,000 at once	0.321	0.263	0.058	0.038	0.015	−2.95	0.217
Kimi, 25,000 a year	0.076	0.051	0.025	0.014	0.114	−2.16	0.164
Kimi, 75,000 a year	0.142	0.118	0.024	0.039	0.197	−3.23	0.182
Kimi, 150,000 a year	0.142	0.114	0.027	0.033	0.155	−3.73	0.196
Kimi, 3,000,000 at once	0.394	0.301	0.093	0.032	−0.090	−3.41	0.262

Note. The same households are used at all four prizes, as in Table E6: 200 for DS Flash and 195 for Kimi. The first four columns are per dollar received, hours are per week, and the two share columns are ratios defined in Table 3. Kimi's 150,000-a-year row differs from Table 3 in some third-decimal cells because Table 3 uses the 197 households available at that prize. Kimi's lump-sum mortgage share is negative because mortgage debt falls while house value rises.

Table E7. The lump sum against the annuity of the same nominal total.

LLM	Outcome	Annuity of 150,000 a year	Lump sum of 3,000,000	Ratio
DeepSeek V4-Flash	Weekly hours	−3.58	−2.95	0.82
DeepSeek V4-Flash	Financial assets per dollar of nominal total	0.102	0.321	3.14
DeepSeek V4-Flash	Reported net worth per dollar of nominal total	0.130	0.405	3.12
DeepSeek V4-Flash	Saving over the past two years per dollar of nominal total	0.012	0.005	0.40
DeepSeek V4-Flash	Probability of holding a mortgage	+0.056	−0.044	
Kimi K2.5	Weekly hours	−3.73	−3.41	0.91
Kimi K2.5	Financial assets per dollar of nominal total	0.071	0.394	5.57
Kimi K2.5	Reported net worth per dollar of nominal total	0.083	0.383	4.62
Kimi K2.5	Saving over the past two years per dollar of nominal total	0.026	0.007	0.28
Kimi K2.5	Probability of holding a mortgage	+0.035	−0.170	

Note. The nominal total is 3,000,000 dollars under both prizes. The ratio is left blank for mortgage holding, because the two effects have opposite signs and their ratio is not informative. The difference in mortgage holding between the two prizes lies between −0.112 and −0.089 for DeepSeek V4-Flash and between −0.226 and −0.183 for Kimi K2.5. What is measured is the average probability that an answer reports a mortgage, not an observed repayment.

E.10 Definitions, samples, and aggregation in Table E9

This subsection gives the definitions, samples, and aggregation behind every column of Table E9.

The two control columns are computed on the untreated answers under the working assignment. The R squared column is the fit, in logarithms, between the household's conditional amount and the neighbourhood benchmark of E.4, and the column reports the range over home value, retirement accounts, stocks outside retirement accounts, and savings or CDs. The range in Table E9 excludes mortgage balances. All five balances are reported in Table E8a below. The AUC column measures how well the share of a household's types that answer yes ranks households against the neighbourhood holding rate, which is turned into a yes or no at one half, and the column reports the range over the four holdings. The AUC uses all two hundred households, and the R squared uses the households for which the family's own answer and the benchmark are both positive.

Table E8. Households in the five balance comparisons.

LLM	House value	Mortgage	Retirement	Stocks	Savings / CDs
DeepSeek V4-Flash	184	171	154	103	195
Kimi K2.5	184	166	154	103	195
Grok 4.1 Fast	183	142	154	103	195
GPT-5-mini	184	178	154	103	195
Claude Haiku 4.5	183	173	154	104	195

Note. A household enters a column when its own answer and the neighbourhood benchmark are both positive. The four holding columns use all two hundred households for every family.

Table E8a. Control balance R squared, including mortgages.

LLM	Home	Mortgage	Retirement	Stocks	Savings / CDs
DeepSeek V4-Flash	0.705	0.439	0.764	0.407	0.831
Kimi K2.5	0.742	0.057	0.784	0.420	0.794
Grok 4.1 Fast	0.594	−0.276	0.765	0.415	0.685
GPT-5-mini	0.754	−0.353	0.659	0.437	0.728
Claude Haiku 4.5	0.629	0.427	0.753	0.340	0.767
Supervised reference	−0.507	0.103	0.033	−0.376	−0.172

Note. R squared is computed in logs on positive balances. Stocks exclude retirement accounts. The supervised reference is defined in E.4.

Table E9. Control consistency and treatment recognition, by LLM family.

	Control		Treatment	
LLM	Balance R-squared	Holding AUC	Median gain	Within band
DS Flash	0.407–0.831	0.882–0.978	27%	100%
Kimi	0.420–0.794	0.875–0.980	14%	87%
Grok	0.415–0.765	0.881–0.976	17%	86%
GPT-mini	0.437–0.754	0.853–0.978	5%	50%
Haiku	0.340–0.767	0.880–0.974	6%	54%

Note. Ranges cover four balances (log R-squared), with all five balances including mortgages in Table E8a, and four holdings (AUC). Median gain is the increase in reported wealth relative to the cash received, and Within band is the share of households whose gain lies between 5 and 100 percent of that cash. This subsection gives the samples and definitions, and E.3 gives the treatment of net-worth answers entered in dollars.

Stocks outside retirement accounts give the lowest R squared in the range in Table E9 for every family. Mortgage balances, excluded from that range, have R squared values from −0.353 to 0.439 across the five families, against 0.103 for the supervised reference. Three corrections were applied to the published figures. Amounts written with a thousands separator are now read in full, and the yes or no answer is now paired with the amount inside the same response. These two corrections leave most control estimates unchanged. The truth for owning the home one lives in is now the survey's own ownership answer rather than ownership of real estate other than the main home, which raises the home AUC of all five families and lifts the lower end of the AUC ranges.

Table E9 retains the positive-balance calculations. A supplementary comparison also reports zero-inclusive balances using the same neighbour sets.

The two treatment columns average every net-worth answer that a household's batches supply, without weighting the three employment assignments and without requiring all three to be present. Coverage differs across families. Kimi K2.5, GPT-5-mini, and Claude Haiku 4.5 answer the net-worth question under all three assignments in both phases, Grok 4.1 Fast does so in its second phase only, and DeepSeek V4-Flash does so under all three assignments in its second phase but under the working assignment alone in the supplementary batch that supplies its first phase. The columns then use the households whose projected net worth is recorded and not zero. Twelve of the two hundred households have a projected net worth coded as zero, which leaves 188. Grok 4.1 Fast has 94, because its first batch does not contain the net-worth question.

E.11 Design notes: why the household is described this way

The description gives the household's state at two dates, and each part is there for a reason. The background at the time of the prize is drawn from the household's records and is the starting point of the experiment. Appendix E.2 states the source years and the age convention. The 2019 state without a lottery is supplied as well, because a description that gave only the 2009 state and left the ten years in between to the LLM produced less accurate asset-holding answers in pilot comparisons. In pilots on 100 households, replacing a description anchored in 2009 by one anchored at 2019 raised the AUC on stock holding in the control condition by 0.15 with Claude Sonnet 4.5 and by 0.13 with GPT-5.4. With Sonnet, four of the five asset outcomes improved.

The 2019 income and net worth are the medians of the household's 2009 cell rather than its own 2019 record because that record is one of many paths the household could have taken and the description is meant to give the state it would likely have reached. A separate contemporaneous comparison on 500 households provides related evidence. Haiku's answers ranked households better against labels constructed from the asset-holding rates of similar households than against their individual realized holdings, by 0.08 to 0.11 in AUC across the five assets. Those labels were constructed from twenty neighbours rather than from the 2009 cells used here.

The treated description states the same no-lottery 2019 state, since the state with the prize is what the experiment asks for and cannot be given. The household is told its circumstances without the prize and is left to say what the prize changed. This is not a request for counterfactual reasoning. A winner interviewed in 2019 knows his own circumstances apart from the prize, what he could earn and what he had put aside, and makes his decisions against them. The character the LLM plays knows nothing of itself beyond the description, so the no-lottery state supplies that self-knowledge from the panel. It supplies the part of the household's 2019 circumstances that the prize did not create. Whether the answers then given with the prize are reasonable is what the three criteria of Section 4 examine.

F Type responses across household states

F.1 Shuffle experiment

The shuffle experiment holds a type description fixed and pairs it with different household states to examine how its answers change. In the Random condition, fifty descriptions are sampled from the types previously generated for households across all five wealth quintiles. Each description is paired with each of fifty PSID households, ten from each wealth quintile, and answers the same ten holding questions. Kimi K2.5 Instant generates the descriptions and answers the questions. The pairing replaces the household's full recorded background, not wealth alone, so it does not isolate the effect of changing wealth while holding other characteristics fixed.

For each question, we classify a description by its answers across the fifty households. A description is classified as almost always answering no if its yes rate is at most 5%, and almost always answering yes if it is at least 95%. A swing type has a yes rate between those thresholds and Youden's J greater than 0.10, where J is the true-positive rate minus the false-positive rate against observed household holdings. This last requirement distinguishes changes that help discriminate holders from

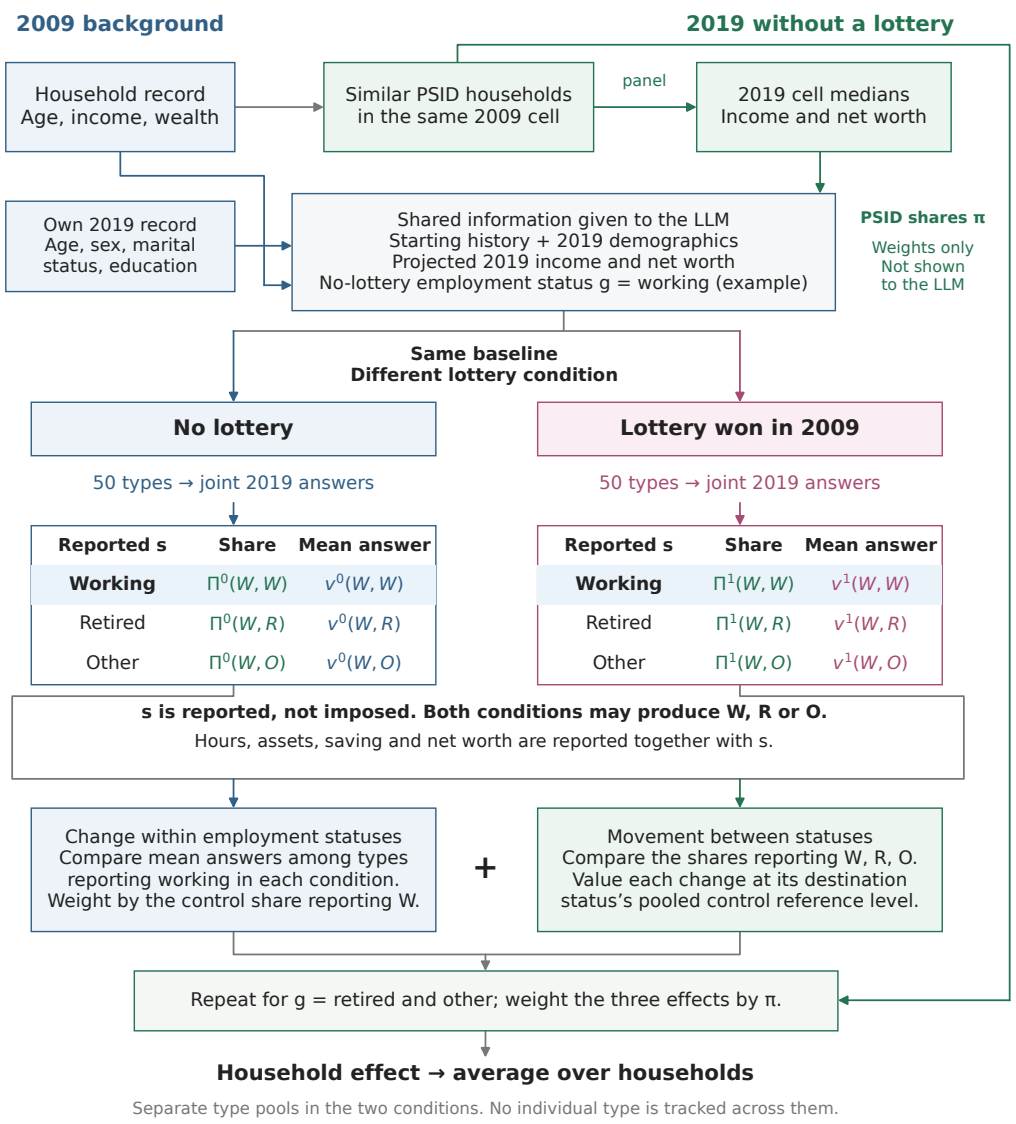

Figure E1. Given information, reported decisions and the lottery effect. The figure expands the case in which the description specifies working as the no-lottery employment status. W, R and O denote working, retired and other. Superscripts 0 and 1 denote the control and treated conditions. Each condition generates fifty types separately. Their reported status shares Π and mean answers v enter the two components defined in E.1. The PSID shares π weight the three assigned-status effects and are never shown to the LLM. Source years and projection rules are given in E.2.

nonholders from changes with little predictive value. Descriptions between the yes-rate thresholds that do not meet it remain unclassified.

Table F1. Responses of fifty descriptions to the stocks question across fifty household states.

Response pattern	Descriptions	Share
Almost always answers no	42	84%
Changes answer and meets the discrimination criterion	3	6%
Almost always answers yes	2	4%
Changes answer but does not meet the discrimination criterion	3	6%

The mean predicted stockholding rate in this experiment is 8.1%, compared with an observed rate of 6%. The many descriptions that mostly answer no lower the predicted rate, while descriptions whose answers differ across households contribute to ranking. This is a description of how the aggregate prediction is formed; it does not establish why the generated types have the proportions needed to reproduce observed behavior.

F.2 Classification across questions

We repeat the classification for other questions using the same fifty descriptions. The swing share is 6% for stocks, 18% for employer retirement accounts, and 68% for homeownership. These are shares of descriptions, not household holding rates. The classifications describe responses to a particular question and need not be the same across questions. The descriptions are generated without being given the subsequent question list. These observations do not explain how the language model generates the observed mixture of response patterns.

F.3 Replacing the descriptions

A separate experiment uses a library of 26 descriptions selected for informative changes in their answers across household backgrounds on several questions. This library predicts a stockholding rate of about 21%, compared with 6% observed, while stockholding AUC varies from 0.358 to 0.599 across three runs. Selection is across questions: under the classification used in Table F1, 15 of the 26 descriptions still almost always answer no to the stocks question. A further replacement with descriptions of public figures gives stockholding AUC of 0.42. These comparisons show sensitivity to the selected descriptions, but do not separately identify the contribution of each response category.