BEHAVIORAL HISTORY OUTPERFORMS DESCRIPTIONS OF THE PERSON FOR LLM SYNTHETIC PERSONAS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) are increasingly used as synthetic personas representing survey respondents, to reduce the costs of survey research, market analysis, and policy experimentation. The validity of LLM-based synthetic personas as substitutes for particular respondents depends on whether they reproduce individuals' decisions. We examine what information helps synthetic respondents better predict each individual's later choices. We design experiments that evaluate synthetic respondents constructed with progressively richer information about each individual, beginning with no personal information, then adding demographics, personality traits, cognitive scores, and finally the respondents' earlier survey choices as behavioral history, yielding five conditions. We use a two-wave panel of 845 US adults who completed measures of 14 behavioral biases (spanning risk, time preferences, overconfidence, and reasoning), so each respondent's earlier answers provide a human test–retest benchmark. When earlier answers serve as behavioral history, all items that score the target bias are withheld. At the population level, the average number of biases per respondent in every condition is close to the human average (7.1–8.1 biases, against 7.1 for humans). This aggregate similarity, however, masks substantial differences in variance structure: persona descriptions recover only 53–67% of the amount of human between-person variation, whereas adding behavioral history restores the overall amount of variation to approximately the human level. At the individual level, predictive agreement is much weaker: description-based personas achieve only 7–12% of the informedness observed in human test–retest responses, while adding behavioral history raises this to 28%. Across demographic groups, the cumulative condition including behavioral history has the highest estimated informedness in all 17 groups, whereas description-based conditions provide little or no information for some groups. We also find that synthetic responses exhibit stronger education- and income-related differences than human responses. For LLM synthetic personas, a respondent's past answers add more to individual-level prediction than a description of who they are.

1 INTRODUCTION

LLM-based synthetic respondents could reduce the cost of survey research, market analysis, and policy experimentation (Korinek, 2023; Manning et al., 2024). Their validity, however, depends on what they are used for: a model may reproduce a population distribution without reproducing the decisions of the individuals who generated it. Substituting a model for a particular respondent therefore requires a stricter test: whether the model predicts that person's later choices. We ask how this person-level prediction changes as the model receives richer information about the target individual.

Behavioral-economics surveys are a demanding place to run this test. They measure behavioral biases: systematic departures from standard models of rational choice and accurate beliefs, such as favoring immediate rewards (present bias), weighting losses more than gains (loss aversion), or overrating one's own ability (overconfidence) (Stango & Zinman, 2023). We use "bias" only in this sense, not in its statistical or fairness sense. An LLM may give the normatively correct answer while

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.

failing to reproduce the target respondent's choice (Liu et al., 2025b). Because these biases shape people's decision-making, a model that gets the answers right, or matches the population on average, cannot stand in for a particular person unless it also reproduces that person's biases.

Prior research shows that LLMs can reproduce some subgroup distributions, experimental regularities, and aggregate patterns of economic behavior (Argyle et al., 2023; Aher et al., 2023; Filippas et al., 2024; Hu et al., 2026; Zhu et al., 2025; Bini et al., 2026; Ashokkumar et al., 2026). At the same time, persona prompting can compress human variation, distort demographic relationships, or introduce prompt artifacts and identity-specific reasoning errors (Bisbee et al., 2024; Domínguez-Olmedo et al., 2024; Gupta et al., 2024; Hu & Collier, 2024; Park et al., 2024b; Gao et al., 2025; Wang et al., 2025; Gui & Toubia, 2023). Personas also shift risk and cognitive-bias decisions, although these studies do not test the later choices of the specific people described (Iwamoto et al., 2026; Liu et al., 2025a; Yoshikawa et al., 2026). Person-matched studies further show that richer personal records can improve prediction, although substantial individual-level errors remain (Park et al., 2024a; Toubia et al., 2025; Choi et al., 2025; Kinzinger & Hartmann, 2026; Garzón et al., 2026). Appendix A reviews further related work.

Prior work evaluates group-level simulation (Hu et al., 2026), compares agents grounded in interviews, surveys, or longitudinal histories (Park et al., 2024a; Jia et al., 2026), and shows that apparent person-level accuracy can largely reflect item means rather than individual deviations, with simple statistical baselines often doing better (Peng et al., 2026; Ahn et al., 2026; Chen et al., 2026; Wang et al., 2026; Lukauskas & Šarkauskaitė, 2026). What remains unresolved is which kind of information about a person (demographics, personality, cognitive scores, or past answers) best helps a model reproduce that person's responses, and whether the model also captures how people differ from one another. We study this with a five-condition information ladder: the model, respondent panel, target questions, and scoring rules stay fixed while demographics, personality traits, cognitive scores, and behavioral history are added in turn. This lets us test whether richer information improves individual prediction, only restores the amount of between-person variation, or changes the demographic structure of responses.

Our evaluation uses a two-wave panel of 845 US adults who completed measures of 14 behavioral biases three years apart (Stango & Zinman, 2023). For each respondent, we build a synthetic respondent from that person's own survey record and ask the model to answer each item as that person would, then compare its answers with the person's actual choices. The five conditions differ only in how much of the record the model sees: nothing (A1), demographics (A2), personality traits (A3), cognitive scores (A4), and earlier answers to other survey items (A5). In A5, we withhold every earlier item used to score the target bias, preventing the model from directly observing the respondent's previous target outcome. Related non-target answers remain and may carry information about correlated behavior. Because the conditions add information in sequence, adjacent comparisons measure the incremental value of each block in that sequence, not its independent contribution.

We evaluate the conditions at the population and individual levels, comparing each synthetic response with the target person's later answer and using the person's own earlier answer as a human test–retest benchmark. Because the binary outcomes are imbalanced, informedness is our primary person-level metric: unlike raw accuracy, it does not reward predictions that merely follow the more common response. We also examine how predictive performance and response patterns vary across demographic groups.

Our central finding is that matching the population is not the same as matching the person. Descriptive persona information produces plausible aggregate counts but recovers little person-specific information and only part of the variation observed across human respondents. The largest improvement occurs in A5, the cumulative condition that adds behavioral history: it provides the strongest individual-level signal and restores approximately the human amount of between-person variation. Even so, A5 falls well short of human respondents: it does not reproduce the human prevalence profile, we find no evidence that it recovers the relationships among biases, and it recovers only about a quarter of the human test–retest signal.

2 EXPERIMENTAL SETUP

2.1 PARTICIPANTS AND OUTCOMES

Our data come from 845 adults in a US national panel who completed a behavioral-economics battery in 2014–15 (wave 1) and again in 2017 (wave 2) (Stango & Zinman, 2023). Wave 2 is always the prediction target. Wave 1 provides the human test–retest benchmark and the behavioral history used in A5.

We analyze the 14 biases that can be administered in text, each scored for every respondent as a binary verdict: 1 if the respondent's answers show the bias and 0 if they do not. They span present bias, inconsistent choice, risk, overconfidence, mathematical and statistical reasoning, and attention. Three of the source study's 17 biases require a graphical game or a next-day login and are excluded; Appendix B lists every included instrument and scoring rule.

2.2 PERSONA INFORMATION AND ADMINISTRATION

Each persona is built from the respondent's own wave-2 record: 16 demographic fields (such as age, gender, race, Hispanic origin, education, and income); Big Five personality scores from the BFI-10 (Rammstedt & John, 2007), which the panel administered; and four cognitive scores from short tests in the same survey, covering numeracy, financial literacy, fluid intelligence (a number-series test), and executive function (a timed color–word test). Table 1 defines the five nested conditions (A1–A5). A1 contains only the survey question, with no persona information, and A2–A4 add demographics, personality, and cognitive scores in turn. A5 also includes the respondent's wave-1 answers to other survey items. The whole item block used to score the target bias is removed, so the model never sees the respondent's earlier answers to that bias. Because the conditions are nested, results we attribute to behavioral history refer to A5, which adds earlier answers to the full persona description.

Table 1: **Five nested prompt conditions.** Each condition adds one information block to the previous condition. A5 includes respondent-specific behavioral evidence, but the target bias block is withheld.

	Condition	Prompt information
A1	No identity	the survey question alone
A2	Demographics	adds 16 wave-2 demographic fields
A3	Personality	adds the BFI-10 Big Five, with scale notes
A4	Cognitive	adds four cognitive scores, each with its scale
A5	Behavioral history	adds the person's wave-1 responses, excluding all items used to score the target bias

Each item is sent in a fresh context, with no conversation history, except where items must be answered together, such as when a construct is defined only under joint evaluation. This mirrors the separation of the original survey modules and avoids artifacts from accumulated context (Domínguez-Olmedo et al., 2024). When an item refers to an earlier answer, the prompt restates the answer the model gave in the same condition. All responses were generated with Claude Opus 4.8 in August 2026, with one generation per respondent, item, and condition. Appendix C documents the prompts, withholding rules, and administration.

2.3 SCORING AND EVALUATION

The same code scores human and model responses using the source study's published rules; no outcome is rescaled. The primary outcome is each respondent's raw wave-2 verdict, with no measurement-error correction on either side. Individual agreement is measured with informedness ($J = \mathrm{TPR} + \mathrm{TNR} - 1$), where the true-positive rate (TPR) and true-negative rate (TNR) are the shares of respondents with and without the bias whose verdict is predicted correctly. It equals 1 for perfect prediction and 0, on average, for a predictor that ignores the person, such as one that always gives the same answer or guesses at random. We average J over the 14 biases and report each condition as a share of the human test–retest value (Appendix D).

We assess population resemblance with bias counts, a variance decomposition of the person-by-bias matrix, and a respondent-grouped classifier that distinguishes synthetic from human profiles.

We assess individual fidelity with per-bias agreement, person-by-bias residual correlation, and an identification AUC that asks whether a synthetic profile of continuous bias scores is closer to its own respondent than to another respondent. For both AUCs, 0.5 is chance: a classifier AUC near 0.5 means synthetic and human profiles are hard to tell apart, whereas a higher identification AUC means synthetic profiles are closer to their own respondent.

Samples differ by analysis because each needs different observed cells. The bias-count distribution uses a count sample of 696 respondents with all 14 biases available in every condition, and the variance decomposition uses the 518 respondents with a complete wave-2 matrix. Per-bias prevalence uses all available cases; per-bias agreement uses the common retest sample, on which the target, the human benchmark, and every condition are observed; and profile analyses use a primary sample of 829 respondents with at least 7 jointly observed biases. Appendix E gives the full inclusion rules.

3 Results

3.1 Aggregate counts conceal a different population

Across the five conditions, average bias counts are close to the human average, but A1–A4 produce less variation across respondents (Figure 1). Humans average 7.08 biases (SD = 1.90), while A5 has the same mean and a similar standard deviation (SD = 1.94). By contrast, A1 has an SD of only 0.73 and places 87% of respondents in two adjacent count bins, while A2–A4 also remain less dispersed than humans. Thus, only A5 closely matches both the human average and overall spread.

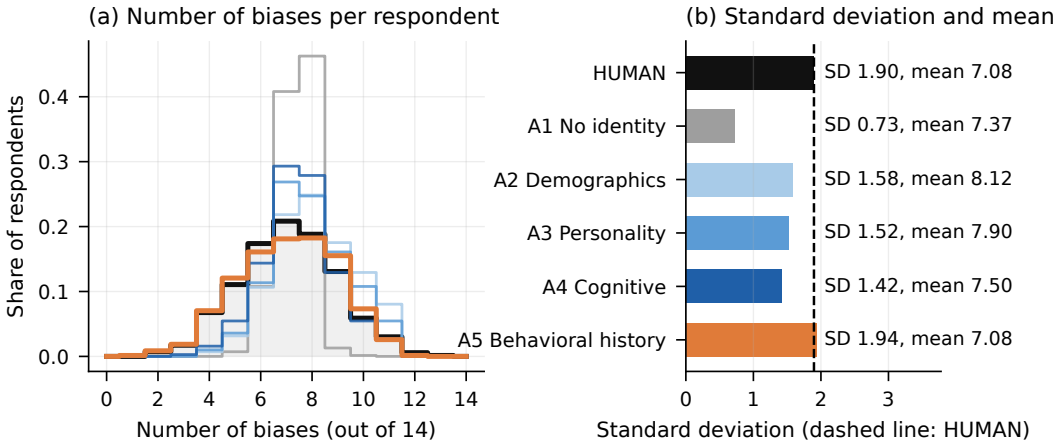


Figure 1: **How many of the 14 biases each respondent shows, and how the distribution differs by condition.** Panel (a) shows the bias-count distribution, and panel (b) reports each condition's standard deviation and mean; colors are the same in both panels. $n = 696$ respondents with all 14 biases available in every condition.

We decompose total variation in the person-by-bias matrix into differences between biases, differences between people, and residual variation specific to each person–bias combination. Each component is measured as its share of the total sum of squares (Table 2). Persona descriptions (A2–A4) recover only 53–67% of the human between-person variation, whereas behavioral history restores approximately the full amount. In A1, differences between biases explain 72.8 times as much variation as differences between people; this ratio remains between 12.2 and 15.7 in A2–A4, compared with 2.6 in humans. A5 comes closest among the synthetic conditions, with a ratio of 5.1 and a larger residual share than A1–A4. However, a Mantel test (Mantel, 1967), run without one mechanically confounded pair of biases (Appendix F), does not show that any condition recovers the human pattern of relationships among biases. Synthetic profiles also stay easy to tell apart from human ones. On the primary sample ($n = 829$), a classifier with folds grouped by respondent separates them with AUC 1.000 under A1 and 0.976 under A5, but cannot separate the two human waves (AUC 0.476; 0.510 on complete profiles, Table 12).

Table 2: **Variance decomposition of the person-by-bias matrix.** The complete $n \times B$ binary matrix is separated into between-bias, between-person, and residual shares of total variance (here $n = 518$ respondents and $B = 14$ biases).

Condition	Bias	Person	Residual	Person/HUMAN	Bias/Person
HUMAN (wave 2)	0.191	0.074	0.735	1.00	2.6
A1 No identity	0.834	0.011	0.155	0.15	72.8
A2 Demographics	0.605	0.049	0.345	0.67	12.2
A3 Personality	0.567	0.046	0.386	0.62	12.3
A4 Cognitive	0.614	0.039	0.347	0.53	15.7
A5 Behavioral history	0.395	0.078	0.527	1.05	5.1

Matching the average number of biases does not mean that synthetic respondents reproduce the human population, because overprediction of some biases can offset underprediction of others. We therefore compare the prevalence of each bias across conditions (Figure 2; values in Table 9). A1 produces highly polarized prevalence rates near zero or one, while A2–A5 show more variation but still differ substantially from humans. In every model condition, the gambler's fallacy is nearly absent and non-belief in the law of large numbers is nearly universal. A5 is closest to human prevalence on 9 of the 14 biases; other conditions come closer on loss aversion, certainty preference, overconfidence-level, gambler's fallacy, and exponential-growth bias (EGB) on assets.

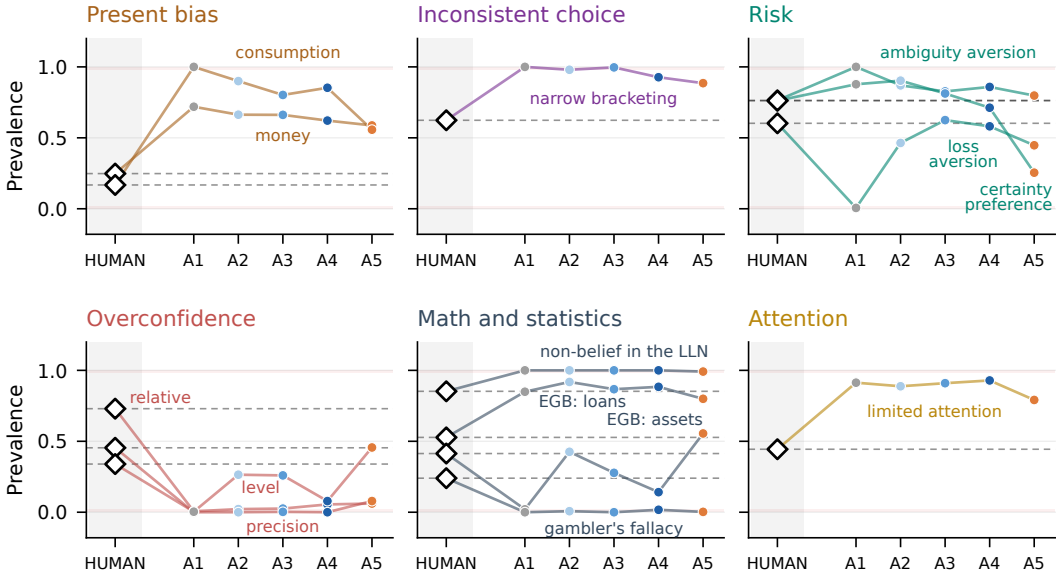


Figure 2: **Prevalence of each bias across the information conditions.** Each panel shows one bias family: HUMAN (grey column), then A1–A5. Dashed lines mark each bias's human wave-2 prevalence, so the vertical distance from a dashed line is the aggregate error. Shaded bands mark prevalence near zero or one. Each bias uses its available cases ($n = 629$–845). EGB is exponential-growth bias and LLN the law of large numbers.

3.2 Individual-level accuracy and informedness across conditions

Aggregate resemblance does not show whether a condition predicts the correct response for the corresponding person. As Table 3 shows, mean accuracy ranges only from 0.550 in A1 to 0.569 in A5, compared with 0.709 for the human retest. However, because bias prevalence ranges from 0.17 to 0.85, raw accuracy can be misleading: repeatedly predicting the more common verdict can produce high accuracy without capturing individual differences.

Informedness reveals much larger differences across conditions than raw accuracy. A1 has essentially no informedness ($J = 0.001$); A2–A4 reach only $J = 0.020$–0.033, or 7–12% of the hu-

man benchmark, while A5 reaches $J = 0.078$, or 28%. Behavioral history therefore provides the strongest individual-level signal, although it remains far below the human benchmark.

Table 3: **Behavioral history improves individual-level prediction.** Each synthetic verdict is compared with the same respondent's observed wave-2 verdict, and results are averaged across the 14 biases on each bias's common retest sample ($n = 526$–843). HUMAN compares the same person's wave-1 and wave-2 verdicts.

	HUMAN retest	A1 No identity	A2 Demographics	A3 Personality	A4 Cognitive	A5 Behavioral history
Mean accuracy	0.709	0.550	0.558	0.564	0.568	**0.569**
Mean informedness J	0.283	0.001	0.020	0.024	0.033	**0.078**
Share of HUMAN	1.00	0.00	0.07	0.08	0.12	**0.28**
Mean $\lvert\mathrm{TPR} - \mathrm{TNR}\rvert$	0.26	0.90	0.70	0.71	0.72	**0.53**
Ratio to HUMAN	1.00	3.42	2.65	2.69	2.74	**2.00**

We also examine $|\mathrm{TPR} - \mathrm{TNR}|$, which measures how unevenly a condition performs across the two outcome classes (0 means both are predicted equally well). This gap is 0.90 for A1 and 0.53 for A5, compared with 0.26 for the human benchmark. Although behavioral history reduces this imbalance, A5's predictions remain much less balanced than the human benchmark.

Table 4: **Mean informedness J by demographic group.** We compute $J = \mathrm{TPR} + \mathrm{TNR} - 1$ within each group-by-bias cell and then average across estimable biases. HUMAN is the group's wave-1–wave-2 benchmark, and n is the median valid sample across included biases. Other is the source panel's own category, not a merger of small groups. The last column is A5 minus A4; bold values exceed the mean gain across groups (0.039).

Attribute	Group	n	HUMAN retest	A1 No identity	A2 Demographics	A3 Personality	A4 Cognitive	A5 Behavioral history	A5−A4
income	low	308	0.269	0.010	0.018	0.020	0.025	0.071	**0.046**
	mid	308	0.285	−0.012	−0.002	0.003	0.037	0.089	**0.052**
	high	192	0.263	0.010	0.019	0.027	0.027	0.030	0.003
education	high school or less	129	0.255	0.001	0.001	0.020	0.047	0.057	0.010
	some college	289	0.283	−0.001	0.002	0.005	0.022	0.059	0.037
	bachelor's	223	0.281	−0.005	0.006	0.023	0.039	0.087	**0.048**
	postgraduate	165	0.275	0.011	0.007	−0.005	0.014	0.059	**0.045**
ethnicity	Black	86	0.206	0.017	−0.002	0.040	0.016	0.045	0.029
	Other	55	0.267	−0.031	0.046	0.048	0.035	0.116	**0.081**
	White	637	0.290	0.000	0.023	0.026	0.035	0.077	**0.042**
age	young	279	0.254	0.007	0.042	0.033	0.024	0.055	0.031
	middle	285	0.282	−0.003	0.007	0.016	0.031	0.100	**0.069**
	older	246	0.306	−0.001	0.004	0.016	0.049	0.071	0.022
gender	Female	514	0.275	0.001	0.016	0.016	0.032	0.080	**0.048**
	Male	297	0.284	0.003	0.023	0.034	0.030	0.064	0.034
origin	Hispanic	105	0.212	−0.027	0.054	0.028	0.053	0.065	0.012
	not Hispanic	705	0.290	0.005	0.013	0.022	0.030	0.079	**0.049**

We next calculate informedness separately for each demographic group (Table 4). A5 has positive informedness and the highest point estimate among the model conditions in all 17 reported groups, whereas the description-based conditions yield values near zero or below zero for some groups. When informedness is averaged equally across the 17 demographic groups, it rises from approximately zero in A1 to 0.016 in A2, 0.022 in A3, 0.032 in A4, and 0.071 in A5, compared with a human group-average benchmark of 0.269 (Figure 3). A5 therefore reaches 26% of this benchmark.

Behavioral history helps some groups more than others. Relative to A4, A5 raises informedness by 0.039 on average across the 17 groups, and by more than this in 9 of them (bold in Table 4),

including postgraduate respondents, for whom A3 and A4 give the lowest informedness of any group (-0.005 and 0.014). The gain is smallest for high-income (0.003), high-school-or-less (0.010), and Hispanic (0.012) respondents. The human benchmark itself also differs across groups, from 0.206 for Black and 0.212 for Hispanic respondents to 0.306 for older respondents. Measured against its own benchmark, A5 reaches 21–35% in 15 of 17 groups, and more for Hispanic (31%) than non-Hispanic respondents (27%); the exceptions are high-income (12%) and Other respondents (43%, $n = 55$). A5 also begins to rank groups the way humans do (rank correlation 0.49 with HUMAN, against 0.20 for A4).

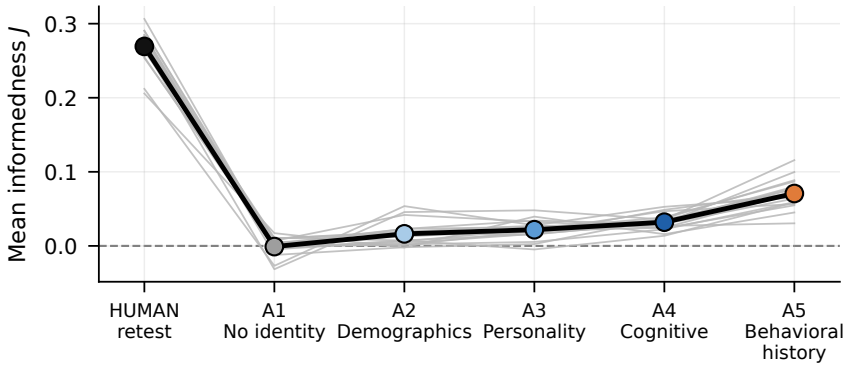


Figure 3: **Behavioral history yields the highest informedness across demographic groups.** Each thin line shows mean informedness J for one of the 17 demographic groups, and the bold line shows the average across groups. HUMAN retest is each group's wave-1–wave-2 benchmark; the dashed line marks $J = 0$.

To examine bias-specific differences across demographic groups, we compare A1–A5 across 14 biases and 15 subgroup levels (the groups of Table 4 without the two Hispanic-origin levels), producing 210 possible bias-by-subgroup cells (Table 5). We exclude three cells, two with fewer than 40 valid observations and one with fewer than five observations in one outcome class, leaving 207 estimable comparisons. A5 has the highest informedness in 104 cells (50.2%), but even the best synthetic condition remains below the human benchmark in 173 cells (84%).

Table 5: **Best- and worst-performing conditions by bias and demographic subgroup.** For each of the 207 estimable bias-by-subgroup cells, A1–A5 are ranked by informedness J. "Best in" and "worst in" report how often each condition has the highest or lowest point estimate.

Condition	Best in	%	Worst in	%
A1 No identity	22	10.6%	**90**	**43.5%**
A2 Demographics	19	9.2%	48	23.2%
A3 Personality	19	9.2%	25	12.1%
A4 Cognitive	43	20.8%	24	11.6%
A5 Behavioral history	**104**	**50.2%**	20	9.7%

Profile-level checks lead to the same conclusion (Appendix G). On the primary sample ($n = 829$), after person and bias means are removed, A5's residual correlation with human responses is 0.070, or 24% of the human test–retest value of 0.287; A4 reaches 0.023, or 8% (Table 12). A5 also reaches an identification AUC of only 0.552, compared with 0.718 for the human retest ($n = 828$; Table 13). Reassigning A5 outputs among demographically matched respondents reduces informedness from 0.078 to 0.026 ($p < 0.001$), indicating that the gain depends on matching outputs to the intended respondent, although this test does not isolate why behavioral history helps. The gains also hold on continuous scores (Appendix H) and vary across respondents (Appendix I).

3.3 Persona conditions reshape demographic disparities

We compare two forms of demographic disparity: differences in prediction accuracy across groups and differences in the prevalence of biases attributed to those groups (Figure 4; values in Table 10).

Persona conditioning reshapes these differences rather than uniformly amplifying or suppressing them. Income and education disparities exceed their human references in accuracy under every condition and in attributed prevalence under every condition with a persona profile (A2–A5). Age and gender do not follow this pattern. The gender disparity is suppressed almost everywhere: every condition with a persona profile shows a smaller gender gap in accuracy than humans, and in attributed prevalence only A4 comes close to the human gap (0.043 versus 0.046) and only A5 exceeds it (0.053). For age, richer conditions reduce the accuracy disparity, but A2–A5 increase the disparity in attributed prevalence. Conditions can also reverse a human pattern: loss aversion rises modestly with income among human respondents but falls with income under every condition with a persona profile (Appendix J).

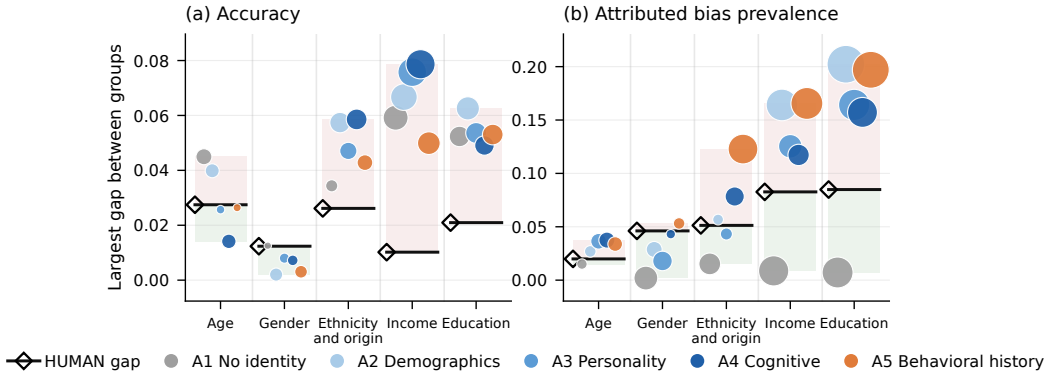


Figure 4: **Persona conditions reshape demographic disparities relative to humans.** Each point shows the largest gap across an attribute's groups in prediction accuracy (a) or attributed bias prevalence (b); the black line marks the human gap. Red shading marks gaps larger than the human gap and green shading smaller ones, and larger markers lie farther from the human gap.

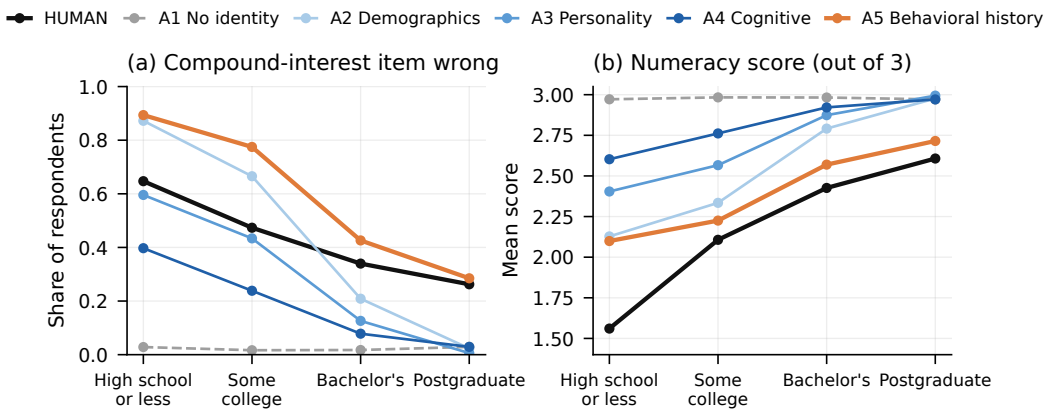


Figure 5: **Synthetic respondents misrepresent how skill varies with education.** (a) Share answering the compound-interest item wrongly. (b) Average number of the three numeracy questions answered correctly (0–3).

The compound-interest item separates the model's ability to solve a question from its ability to reproduce human responses (Figure 5; values in Table 11). A1 answers correctly for nearly every respondent regardless of education, whereas A2, A3, and A5 introduce a steeper education gradient than the human one, and A2–A4 make postgraduate personas almost uniformly correct. A5 restores a postgraduate error rate close to that of human respondents (28% versus 26%) but overpredicts errors in the other education groups, so behavioral history does not recover the human pattern. Panel (b) shows a different distortion for numeracy: every condition answers more questions correctly than the respondents it represents, most of all for those with a high school education or less (1.56 correct for humans, against 2.10–2.97 for the conditions). For postgraduate personas, A1–A4 answer almost

every question correctly (2.97–2.99, against 2.61 for humans), whereas A5 stays closer (2.72) and comes closest in every education group.

4 DISCUSSION

Synthetic bias counts are more strongly structured by observed respondent attributes than human bias counts, and the clearest difference is socioeconomic: education and income account for 74% of the variation explained under A2, 42% under A4, and 49% under A5, compared with 18% in the human data, where cognitive scores are the largest contributor (Table 6). Behavioral history improves individual prediction but leaves this link almost intact: adding it to A4 lowers the variation explained by respondent attributes from 47.5% to 42.9%, mainly through smaller cognitive, personality, and age contributions, while the education and income contributions barely change (19.8 to 21.1 points, against 2.9 for humans). The history available here is limited to a single earlier survey; richer or more recent behavioral records may weaken this link, which we see as the main direction for future work.

Table 6: **Decomposition of explained variation in human and synthetic bias counts.** Cells are Shapley contributions to R^2 (percentage points) from regressing each column's bias count on respondent attributes (757 respondents; Shapley, 1953). Blocks absent from a condition's prompt (the Big Five and cognitive scores for A2; the cognitive scores for A3) measure correlation with the synthetic answers, not use of that information. Stars mark jointly significant blocks (Wald test with Benjamini–Hochberg correction, Benjamini & Hochberg, 1995; $^{*}q < .05$, $^{**}q < .01$, $^{***}q < .001$); every condition's education-and-income share exceeds the human share ($q < .001$).

Construct	HUMAN (wave 2)	A2 Demographics	A3 Personality	A4 Cognitive	A5 Behavioral history
Age	0.09	0.21	0.34	1.01	0.12
Gender	**1.56***	**0.87***	0.45	**1.65***	**1.71***
Ethnicity & Hispanic origin	0.84	0.59	0.47	1.10	1.33
Education	1.64	**22.69*****	**16.91*****	**12.53*****	**13.80*****
Income	1.29	**11.31*****	**8.89*****	**7.25****	**7.26****
Big Five (5 traits)	**1.93***	0.80	**3.81*****	**3.34*****	1.44
Cognitive scores (4)	**8.65*****	**8.15***	4.82	**20.12*****	**15.91*****
R^2 from the seven blocks	16.0	44.6	35.7	47.0	41.6
Total R^2 (18 attributes; App. E)	16.5	46.2	37.2	47.5	42.9
Education and income, % of total R^2	18%	74%	69%	42%	49%

5 LIMITATIONS

This study examines one model, one US panel, and 14 text-administered tasks, so the findings may not generalize to other settings. Each response was generated once per respondent, item, and condition, so measured between-person variation may include run-to-run randomness. The public data may have appeared in training, and the design cannot distinguish learned population rates from memorized individual records (Ludwig et al., 2026; Lopez-Lira et al., 2025).

6 CONCLUSION

Aggregate resemblance is not sufficient evidence that an LLM can substitute for an individual respondent. Demographic, personality, and cognitive descriptions produce plausible population averages but carry little person-specific information and can distort differences between groups. Adding behavioral history recovers more of the between-person variation and improves individual prediction, yet most of the human test–retest signal is still missing. These results suggest that synthetic-respondent studies may benefit from behavioral evidence and should validate person-level agreement before treating model responses as substitutes for individual respondents.

Reproducibility statement

The appendices report the bias definitions, prompt protocol, scoring rules, sample construction, estimators, continuous-score checks, and coverage limits. Human and model responses are scored with the same rules, and each analysis states its sample in the text, a caption, or the appendix. The supplementary material contains the fixed model responses, an environment file, and the analysis scripts, with a README that maps each table and figure to its script; rebuilding the paper does not require new model calls. The de-identified source survey data are available from the original study's public replication package, and the supplement explains how to obtain them.

Ethics statement

This study uses de-identified secondary data from a published replication package and collects no new human-subject data. The personas contain structured survey fields without names or free text, and the target outcome is withheld from every prediction prompt. The demographic analyses describe a research simulation and should not be read as evidence about a deployed decision system.

AI use statement

Large language models were used in several ways in this work. Most centrally, Claude Opus 4.8 generated the synthetic survey responses that the paper studies (Appendix C). We also used LLM-based tools to help search the related literature, to write and check analysis code, and to edit the writing and format the manuscript. The authors checked every cited work against its source, verified the code and all reported numbers, and revised and approved all text, and they take full responsibility for the content of the 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*, volume 202 of *Proceedings of Machine Learning Research*, pp. 337–371, 2023.

Daehwan Ahn, Chengfeng Mao, and Dokyun Lee. Item-mean surrogates: Why richer persona data fail to improve LLMs as human surrogates. *arXiv preprint arXiv:2608.29455*, 2026.

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.

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.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

Pietro Bini, Lin William Cong, Xing Huang, and Lawrence J. Jin. Behavioral economics of AI: LLM biases and corrections. Working Paper 34745, National Bureau of Economic Research, 2026.

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

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

Julien Boelaert, Samuel Coavoux, Étienne Ollion, Ivaylo Petev, and Patrick Präg. Machine bias. How do generative language models answer opinion polls? *Sociological Methods & Research*, 54(3):1156–1196, 2025.

Paige Bollen, Joe Higton, and Melissa Sands. Nationally representative, locally misaligned: The biases of generative artificial intelligence in neighborhood perception. *Political Analysis*, 34(3): 479–487, 2026.

Melanie Brucks and Olivier Toubia. Prompt architecture induces methodological artifacts in large language models. *PLOS ONE*, 20(4):e0319159, 2025.

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.

Zihan Chen, Di Zhu, and Lei Nico Zheng. When synthetic users fail: A cross-domain benchmark of LLM-simulated human survey responses. *arXiv preprint arXiv:2607.26348*, 2026.

Junhyuk Choi, Hyeonchu Park, Haemin Lee, Hyebeen Shin, Hyun Joung Jin, and Bugeun Kim. Pay what LLM wants: Can LLM simulate economics experiment with 522 real-human persona? *arXiv preprint arXiv:2508.03262*, 2025.

André F. Cruz, Moritz Hardt, and Celestine Mendler-Dünner. Evaluating language models as risk scores. In *Advances in Neural Information Processing Systems*, volume 37, pp. 97378–97407, 2024.

Ziyan Cui, Ning Li, and Huaikang Zhou. A large-scale replication of scenario-based experiments in psychology and management using large language models. *Nature Computational Science*, 5(8): 627–634, 2025.

Ricardo Domínguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. Questioning the survey responses of large language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 45850–45878, 2024.

Apostolos Filippas, John J. Horton, and Benjamin S. Manning. Large language models as simulated economic agents: What can we learn from Homo Silicus? In *Proceedings of the 25th ACM Conference on Economics and Computation*, pp. 614–615, 2024.

Leon Fröhling, Jens Rupprecht, Markus Strohmaier, and Claudia Wagner. When persona attributes improve population alignment in large language models. *arXiv preprint arXiv:2609.02526*, 2026.

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

Rubén Garzón, Pauline Baron, Vincent Grari, Jonne Kamphorst, Michael Bernstein, and Marcin Detyniecki. From demographics to survey anchors: Evaluating LLM agents for modeling retirement attitudes. *arXiv preprint arXiv:2605.16303*, 2026.

George Gui and Olivier Toubia. The challenge of using LLMs to simulate human behavior: A causal inference perspective. *arXiv preprint arXiv:2312.15524*, 2023.

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. Bias runs deep: Implicit reasoning biases in persona-assigned LLMs. In *International Conference on Learning Representations*, 2024.

John Hartley, Conor Hamill, Dale Seddon, Devesh Batra, Ramin Okhrati, and Raad Khraishi. How personality traits shape LLM risk-taking behaviour. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 21068–21092, 2025.

Tiancheng Hu and Nigel Collier. Quantifying the persona effect in LLM simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10289–10307, 2024.

Tiancheng Hu, Joachim Baumann, Lorenzo Lupo, Nigel Collier, Dirk Hovy, and Paul Röttger. SimBench: Benchmarking the ability of large language models to simulate human behaviors. In *International Conference on Learning Representations*, 2026.

Ryota Iwamoto, Takunori Ishihara, and Takanori Ida. Comparing risk preferences and reference dependence in humans and AI: A persona-based approach with fine-tuning. *Journal of Behavioral and Experimental Economics*, 124:102634, 2026.

Mumin Jia, Yilin Chen, Divya Sharma, and Jairo Diaz-Rodriguez. When can digital personas reliably approximate human survey findings? *arXiv preprint arXiv:2605.10659*, 2026.

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

Anton Korinek. Generative AI for economic research: Use cases and implications for economists. *Journal of Economic Literature*, 61(4):1281–1317, 2023.

Ang Li, Haozhe Chen, Hongseok Namkoong, and Tianyi Peng. LLM generated persona is a promise with a catch. In *Advances in Neural Information Processing Systems*, volume 38, pp. 73875–73896, 2025.

Jiaxin Liu, Yixuan Tang, Yi Yang, and Kar Yan Tam. Evaluating and aligning human economic risk preferences in LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 18185–18199, 2025a.

Ryan Liu, Jiayi Geng, Joshua C. Peterson, Ilia Sucholutsky, and Thomas L. Griffiths. Large language models assume people are more rational than we really are. In *International Conference on Learning Representations*, 2025b.

Alejandro Lopez-Lira, Yuehua Tang, and Mingyin Zhu. The memorization problem: Can we trust LLMs' economic forecasts? *arXiv preprint arXiv:2504.14765*, 2025.

Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. Large language models: An applied econometric framework. *Annual Review of Economics*, 18(1):283–316, 2026.

Mantas Lukauskas and Viktorija Šarkauskaitė. Plausible but not valid: A psychometric audit of LLMs as synthetic survey respondents. *arXiv preprint arXiv:2608.14606*, 2026.

Benjamin S. Manning and John J. Horton. General social agents. Working Paper 34937, National Bureau of Economic Research, 2026.

Benjamin S. Manning, Kehang Zhu, and John J. Horton. Automated social science: Language models as scientist and subjects. Working Paper 32381, National Bureau of Economic Research, 2024.

Nathan Mantel. The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27(2):209–220, 1967.

Qiaozhu Mei, Yutong Xie, Walter Yuan, and Matthew O. Jackson. A Turing test of whether AI chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences*, 121(9):e2313925121, 2024.

Jiří Milička, Anna Marklová, Klára VanSlambrouck, Eva Pospíšilová, Jana Šimsová, Samuel Harvan, and Ondřej Drobil. Large language models are able to downplay their cognitive abilities to fit the persona they simulate. *PLOS ONE*, 19(3):e0298522, 2024.

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

Peter S. Park, Philipp Schoenegger, and Chongyang Zhu. Diminished diversity-of-thought in a standard large language model. *Behavior Research Methods*, 56(6):5754–5770, 2024b.

Tianyi Peng, Melanie Brucks, George Gui, Daniel J. Merlau, Grace Jiarui Fan, Malek Ben Sliman, Eric J. Johnson, Abdullah Althenayyan, Silvia Bellezza, Dante Donati, Hortense Fong, Elizabeth Friedman, Ariana Guevara, Mohamed Hussein, Kinshuk Jerath, Bruce Kogut, Akshit Kumar, Kristen Lane, Hannah Li, Vicki Morwitz, Oded Netzer, Patryk Perkowski, and Olivier Toubia. Digital twins are funhouse mirrors: Five systematic distortions. *Science Advances*, 12 (36):eaeh8260, 2026.

Beatrice Rammstedt and Oliver P. John. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*, 41(1):203–212, 2007.

Jillian Ross, Yoon Kim, and Andrew W. Lo. LLM economicus? Mapping the behavioral biases of LLMs via utility theory. In *Conference on Language Modeling*, 2024.

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*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29971–30004, 2023.

Indira Sen, Marlene Lutz, Elisa Rogers, David Garcia, and Markus Strohmaier. Missing the margins: A systematic literature review on the demographic representativeness of LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 24263–24289, 2025.

Lloyd S. Shapley. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker (eds.), *Contributions to the Theory of Games II*, volume 28 of *Annals of Mathematics Studies*, pp. 307–317. Princeton University Press, 1953.

Yuanming Shi and Andreas Haupt. The collapse of heterogeneity in Silicon Philosophers. In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency*, pp. 6280–6304, 2026.

Victor Stango and Jonathan Zinman. We are all behavioural, more, or less: A taxonomy of consumer decision-making. *The Review of Economic Studies*, 90(3):1470–1498, 2023.

Erika Elizabeth Taday Morocho, Lorenzo Cima, Tiziano Fagni, Marco Avvenuti, and Stefano Cresci. Assessing the reliability of persona-conditioned LLMs as synthetic survey respondents. In *Companion Proceedings of the ACM Web Conference 2026*, pp. 320–329, 2026.

Olivier Toubia, George Gui, Tianyi Peng, Daniel J. Merlau, Ang Li, and Haozhe Chen. Database report: Twin-2K-500: A data set for building digital twins of over 2,000 people based on their answers to over 500 questions. *Marketing Science*, 44(6):1446–1455, 2025.

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.

Steven Wang, Kyle Hunt, Shaojie Tang, and Kenneth Joseph. When can LLM digital twins reduce human measurement? From behavioral fidelity to statistical substitutability. *arXiv preprint arXiv:2609.07987*, 2026.

Yunze Xiao, Vivienne J. Zhang, Chenghao Yang, Ningshan Ma, Weihao Xuan, and Jen-tse Huang. The Chameleon's limit: Investigating persona collapse and homogenization in large language models. *arXiv preprint arXiv:2604.24698*, 2026.

Katsumasa Yoshikawa, Junya Takayama, and Takato Yamazaki. Persona-aware evaluation of cognitive bias in LLMs: From benchmark to applied decision-making. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pp. 4213–4225, 2026.

William J. Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950.

Jian-Qiao Zhu, Haijiang Yan, and Thomas L. Griffiths. Language models trained to do arithmetic predict human risky and intertemporal choice. In *International Conference on Learning Representations*, 2025.

A Additional related work

Simulating people with LLMs. Large-scale replications show that LLMs reproduce many scenario-based experimental effects (Cui et al., 2025), yet demographic prompts give the model no reliable basis for within-sample heterogeneity (Manning & Horton, 2026). Several studies document how persona-conditioned responses narrow: answers show strong question-specific bias with low variance (Boelaert et al., 2025), judgments become over-correlated across domains (Shi & Haupt, 2026), and the variation that remains tracks coarse demographic stereotypes rather than individual differences (Xiao et al., 2026). Persona prompting does not reliably improve alignment with survey distributions (Taday Morocho et al., 2026), and whether persona attributes help depends on how much humans vary on a question (Fröhling et al., 2026).

Behavioral biases in LLMs. Earlier studies test LLMs on economic rationality, strategic behavior, and behavioral biases without matching answers to individual respondents (Chen et al., 2023; Mei et al., 2024; Ross et al., 2024), or show that assigned personality traits shift LLM risk-taking (Hartley et al., 2025); Binz et al. (2025) fine-tune a model to predict human choices in psychology experiments. Here, each verdict is scored against the same respondent's answers.

Representation and calibration. Whether LLM outputs represent demographic groups evenly is a known concern (Santurkar et al., 2023; Sen et al., 2025): a model can match national aggregates while misrepresenting subgroups (Bollen et al., 2026), and LLM risk scores for individuals can carry predictive signal yet be widely miscalibrated (Cruz et al., 2024).

B The 14 biases

Table 7 lists the instrument and scoring rule for each outcome. Every binary verdict uses the same sign convention: 1 denotes the biased response. Human and model answers are processed by the same scoring code using the source study's rules (Stango & Zinman, 2023).

The table reports two sample sizes. n is the number of valid wave-2 responses used for prevalence. n_{pair} is the subset that also has a valid wave-1 response and a valid prediction from every condition; it is the common retest sample used for the test–retest benchmark. Certainty preference has the largest gap between n and n_{pair} because its scoring rule excludes respondents who switch more than once in the elicitation lists.

Three outcomes from the source study are omitted. Two require a graphical budget-allocation game, and one requires a next-day login. All counts in this paper are therefore recomputed over the same 14 text-administered outcomes rather than copied from the source study's 17-outcome totals.

Table 7: **The 14 biases, scoring rules, and coverage.** Rate is the human wave-2 prevalence. EGB is exponential-growth bias and LLN the law of large numbers.

Bias	Instrument	Biased if	Rate	n	n_{pair}
Present bias					
Present bias — money	24 convex time-budget allocations	less patient when the earlier date is today	0.248	819	782
Present bias — consumption	two snack choices, now vs later	unhealthy now *and* healthy later	0.168	829	821
Inconsistent choice					
Narrow bracketing	Rabin–Weizsäcker pairs, shown jointly	dominated pair, or not the both-safe pair	0.624	837	820
Risk					
Loss aversion	Fehr–Goette single vs six \$80/\$50 gambles	rejects one or both	0.602	845	843
Certainty preference	two probability-equivalent lists	Callen certainty premium > 0	0.762	629	526
Ambiguity aversion	Ellsberg two-bag choice	chooses the known bag	0.764	821	818
Overconfidence					
Overconfidence — level	predicted vs actual numeracy score	predicted $>$ actual	0.339	813	799
Overconfidence — precision	two subjective probability distributions	places 100% on a single bin	0.454	775	748
Overconfidence — relative	self-ranked percentile before the test	self-rank $>$ realized rank	0.730	818	817
Math and statistics					
Non-belief in the LLN	1,000 coin flips, central interval	$\Pr(481\text{–}519) < 78$	0.852	819	809
Gambler's fallacy	next flip after a run	answer < 50	0.240	817	815
EGB — loans	monthly payment on a stated loan	implied APR $>$ stated APR	0.527	783	736
EGB — assets	\$200 at 10% compounded for two years	answer < 242	0.414	735	688
Attention					
Limited attention	four regret/inattention items	any flagged response	0.444	829	819

C PROMPT CONSTRUCTION AND ADMINISTRATION

The prompt text below is verbatim; persona values in the examples are invented.

Prompt. Each prompt had a system section, holding an instruction preamble and the respondent profile, and a user section, holding one survey item and its response format. The task statement and the empty profile of the no-identity condition follow Peng et al. (2026); the section structure and guidelines are taken from, and three of the instructions adapted from, Li et al. (2025). Because prompt order, labels, and framing can themselves bias LLM answers (Brucks & Toubia, 2025), items are worded and formatted identically in every condition. The preamble was identical in A1–A4:

```
### TASK ###
You are simulating one real survey respondent. The 'Respondent Profile' below describes a
participant in a study of household financial decision-making; it may contain their
background, their personality and ability scores, and their answers to an earlier set of
questions. Answer the 'New Survey Question' exactly as that person would answer it.

### GUIDELINES ###
1. Be Faithful to the Persona: Ensure your answer is consistent with the persona's data.
2. Focus on Relevant Aspects: Center your reasoning on the relevant factors that would
   influence the persona's opinion on that topic.
3. Be Objective: Avoid injecting personal bias or overly politically correct views that
   may not align with the persona's standpoint.

### INSTRUCTIONS ###
```

```
- Give exactly one answer per question, in the format that question asks for.
- If more than one answer is plausible for this person, select among them in proportion to
  how likely each is. Do not default to the most common or the most sensible answer.
- Always answer, even in unclear cases - treat it as a forced-choice survey.
- Reason internally. Do not include your reasoning in the response.

# Respondent Profile
```

A5 adds two sentences after the task paragraph: *"The 'Respondent Profile' includes this person's responses to an earlier set of questions. Use them as evidence of how the persona answers questions of this kind."*

Respondent profile. In A1 the profile reads `[No respondent profile provided.]` A2 adds the 16 wave-2 demographic fields, shown as 13 labeled lines because three of the fields are numeric codes behind the education and income labels. A missing field is omitted rather than imputed.

```
## Demographics
- Age: 54
- Gender: Male
- Race: White
- Hispanic or Latino origin: No
- Highest level of education completed: Some college, no degree
- Total household income last year, before taxes: $40,000 to $49,999
- Employment status: Working now
- Work arrangement: Works for someone else
- Occupational group: Office and Administrative Support
- Marital status: Married or living with a partner
- Housing: Owns
- Number of other people living in the household: 2
- Born in the United States: Yes
```

A3 adds the BFI-10 personality block (Rammstedt & John, 2007); missing traits are omitted rather than imputed.

```
## Personality

This person completed the BFI-10, a standard ten-item measure of the five major personality
traits. Each trait is measured by two statements - one worded positively, one worded
negatively - which the person rated on a five-point scale from "Disagree strongly" to
"Agree strongly". The two ratings are combined, reversing the negatively worded one, into a
single score from 1.0 to 5.0. A score of 3.0 is neutral; higher means the trait is more
strongly present, lower means less.

- Openness - curiosity, imagination, receptiveness to new experiences and ideas.
  Rated on: "has an active imagination" and "has few artistic interests".
  Score: 3.5 out of 5

  [the remaining four traits follow in the same form]
```

A4 adds the four cognitive scores, each reported against its scale.

```
## Measured cognitive skills

This person took four short tests as part of the same survey. Their results:

- Numeracy - two questions about division and percentages.
  Score: 2 out of 2 correct.

- Financial literacy - three standard questions on compound interest,
  inflation, and investment diversification.
  Score: 3 out of 3 correct.

- Fluid intelligence - a number-series test in which the respondent
  supplies the next number in a sequence.
  Score: 10 out of 15 correct.

- Executive function - a timed test requiring the respondent to name the
  ink colour of a colour word while ignoring the word itself, as many as
  possible in two minutes.
  Score: 61 correct in two minutes. The average for this survey's
  respondents was 70.
```

LLMs can answer below their own ability to fit a simulated persona (Milička et al., 2024); Figure 5 tests whether stated scores have this effect here. The profile's numeracy score counts the two items on division and percentages; Figure 5(b) also counts the compound-interest item shown in panel (a).

A5 adds the respondent's wave-1 answers, excluding the withheld items described below, in this form:

```
## How this person has answered questions before - use as evidence

These are this person's own responses, recorded on an earlier occasion. They show how this
person reads questions, what they tend to prefer, and how carefully they work things out.

Q7. Suppose you had $100 in a savings account and the interest rate was 2% per year. After
5 years, how much do you think you would have in the account if you left the money to grow?
  1. More than $102
  2. Exactly $102
  3. Less than $102
  4. Don't know
Answer: 1 (More than $102)
```

Survey item and response format. Each user section carried one item and a response format:

```
# New Survey Question

Q85. Imagine that we had a computer 'flip' a fair coin: a fair coin is one with a 50-50
chance of coming up heads or tails. Say that the computer flips the coin 10 times. The first
9 are all heads. What are the chances, in percentage terms, that the 10th flip will be a
head?

Format Instructions:
Respond with only a JSON object in exactly this form, and nothing else:
{"Q85": <answer>}
```

Item wording follows the source instrument (Stango & Zinman, 2023), adapted for text administration; the income-change distribution (Q28b) asks about realized rather than expected change. Each Rabin–Weizsäcker task is shown on a single screen with an instruction to consider both decisions together. Items that refer to earlier answers (Q27, Q28, Q30 and Q88) restate those answers from the same condition, and conditional follow-up items are asked only when triggered.

Withholding rules. The prediction target never enters a prompt: wave-2 answers and all derived bias variables are excluded from every condition, and profiles contain structured fields only, with no names or free text. In A4 and A5, a cognitive score is hidden on the screens that would reveal its answer: numeracy on Q4–Q6, Q27, Q28 and Q28b; financial literacy on Q7–Q9; and fluid intelligence on the number series, Q29 and Q30. The A5 history omits the entire item block used to score the target bias.

Administration. Each item was answered in a separate, single-turn call to a Claude Code subagent running Claude Opus 4.8 (`claude-opus-4-8`) in August 2026, with a fresh context and no conversation history, except where items must be answered together (§2). The system and user sections were sent together, followed by the closing instruction:

```
(Answer as this persona. You are a plain-text respondent: use no tools, reason internally,
and output ONLY the JSON object requested, nothing else.)
```

Claude Opus 4.8 does not accept sampling parameters, so none were set. One reply was generated for each respondent, item, and condition, and the answer was read from the JSON object in the reply.

D ESTIMATORS

An individual-level score is easy to interpret only on a scale where a predictor that ignores the person scores zero, so we define each estimator explicitly.

For a bias b with human wave-2 verdict $Y \in \{0, 1\}$ and a predictor $\hat{Y}$:

$$\text{accuracy} = \Pr(\hat{Y} = Y), \tag{1}$$

$$\text{TPR} = \Pr(\hat{Y}{=}1 \mid Y{=}1), \qquad \text{TNR} = \Pr(\hat{Y}{=}0 \mid Y{=}0), \tag{2}$$

$$J = \text{TPR} + \text{TNR} - 1. \tag{3}$$

Informedness, also known as Youden's index (Youden, 1950), has a null value of 0, whereas raw accuracy has no single null value: a majority-class predictor scores $\max(p_b, 1-p_b)$, while a random draw at the population rate scores $p_b^2 + (1-p_b)^2$. Both scores depend only on the prevalence p_b.

For a condition's value v, its share of the human benchmark is $(v - \text{null})/(\text{benchmark} - \text{null})$. The subtraction is essential: without it, a predictor at the null would still receive a large share of the human benchmark. Table 3 therefore reports raw accuracy descriptively but does not normalize it as a share of the benchmark; the comparisons rest on the chance-corrected rows.

E SAMPLES

Different analyses require different observed cells, so no single sample is appropriate for every result. Each analysis states its sample in the text, a caption, or this appendix.

Available case ($n = 629$–845 per bias). This sample includes everyone with a valid wave-2 answer for the bias. It is used for prevalence and other marginal distributions, which do not require a wave-1 response.

Common retest ($n = 526$–843 per bias). This sample additionally requires a valid wave-1 answer and a valid prediction from every condition. It supports the primary accuracy and informedness comparisons because the human benchmark and all model conditions are defined on the same people.

Counts ($n = 696$). The bias-count distribution (Figure 1) uses respondents with all 14 outcomes available in every condition, so each count is a whole number out of 14 and every condition is compared on the same people; 3.0% of human wave-2 cells are filled from the same person's wave-1 answer. The human mean is 7.08 (SD 1.90). Following the source study instead, with each person's share of valid outcomes multiplied by 14 on all 845 respondents, gives the same pattern: human 7.15 (SD 1.91), A1 SD 0.73, A2–A4 SD 1.53–1.62, and A5 7.32 (SD 1.93). Group prevalence (panel (b) of Figure 4) uses this all-respondent version.

Primary sample ($n = 829$). The primary classifier and residual analyses require at least 7 of 14 biases to be valid in both human waves and every model condition. Each retained respondent is compared on their common observed cells. Raising the requirement to 8 biases ($n = 818$) or 10 biases ($n = 794$) changes no headline estimate by more than 0.011. Identification uses its own validity rule and retains 828 profiles; the demographic-matched version retains 788.

Complete-profile sensitivity checks. Requiring all 14 binary outcomes leaves 389 profiles for residual and classifier analyses and 386 for identification. Excluding certainty preference leaves 532 and 531 profiles, respectively.

Balanced ($n = 518$). The variance decomposition requires all 14 outcomes to be valid for the human target and every model condition. This complete rectangle avoids assigning missingness to a person or bias component. The balanced human mean is 6.92; means from different samples answer different questions and are not combined.

Auxiliary samples. Four analyses need a validity rule of their own. The block-Shapley decomposition in Table 6 requires complete *predictors* rather than complete outcomes; requiring all 18 attributes (nine demographic variables: age, gender, White, Hispanic origin, education, income, employment, marital status, and household size; the five traits; and the four cognitive scores) retains 757 respondents; the other 88 lack at least one attribute, almost always a Big Five trait or a cognitive score. The respondent-level comparison (Table 15) retains 837 respondents for binary accuracy and 821 for graded error, which excludes present bias for consumption (no graded score) and certainty preference. Group accuracy in panel (a) of Figure 4 requires at least 8 valid biases and retains 818 respondents. Group informedness (Table 4 and Figure 3) instead pools each bias's own common-retest sample and so rests on the 845 respondents scorable for at least one bias.

Table 4 and Table 5 omit Asian (22) and American Indian (12) respondents, whose group-by-bias cells all have fewer than 40 valid observations. Figure 4 instead combines race and Hispanic origin into four groups: White non-Hispanic (592), Black non-Hispanic (92), Hispanic of any race (111), and other non-Hispanic (50, including Asian and American Indian respondents). Income groups are below \$40,000 (328), \$40,000–\$99,999 (318), and \$100,000 or more (198), age groups are terciles,

and education groups are high school or less, some college (including associate degrees), bachelor's, and postgraduate.

F COVERAGE LIMITS AND SENSITIVITY CHECKS

Why we report single attributes and not intersectional groups. §3.3 compares groups one attribute at a time. We did not report intersectional cells for two reasons. First, coverage is uneven: requiring at least five personas in each income × ethnicity × age × education × gender cell keeps 95% of White personas, 45% of Black personas, and none of the Asian, American Indian, or Other personas, and no Black-male cell remains. Second, the extreme cells are unstable: among the 60 cells with at least five people, the split-half reliability of the cell ordering is only 0.26–0.32, so the largest and smallest estimates largely reflect noise. Single-attribute gradients avoid both problems.

One mechanically confounded pair is excluded from the Mantel test. Overconfidence-level equals predicted minus actual numeracy, and the actual score includes the same item used to measure exponential-growth bias on assets. The pair is therefore correlated by construction and is the largest entry in both the human and the synthetic correlation matrices. With the pair included, four of five conditions appear to match the human correlation structure at $p < 0.05$; with it excluded, none do (Table 8). §3.1 reports the latter result.

Table 8: **Mantel test of the correlation structure among biases.** r is the correlation between the human and synthetic between-bias Spearman correlation matrices, computed on the continuous (percentile) bias scores, with a two-sided p from 5,000 permutations of the bias labels. Biases whose synthetic scores do not vary have undefined correlations, so A1 and A2 use fewer pairs.

	All pairs			Without the confounded pair		
Condition	Pairs	r	p	Pairs	r	p
A1	55	0.623	0.004	54	0.209	0.138
A2	78	0.176	0.288	77	−0.129	0.462
A3	91	0.363	0.016	90	−0.033	0.820
A4	91	0.278	0.046	90	0.010	0.950
A5	91	0.429	0.006	90	0.244	0.118

G ADDITIONAL TABLES

This appendix reports the values behind Figures 2, 4, and 5 (Tables 9–11) and the profile-level checks cited in §3.1 and §3.2. Table 12 gives the classifier and residual-correlation results on the primary sample and on complete 14-bias profiles (389 respondents for classifier and residual analyses, 386 for identification), and Table 13 gives identification on the primary sample. Under the complete-profile rule the results are similar: A5 recovers 20% of the human residual correlation, compared with 24% in the primary sample; its classifier AUC is 0.976, and its all-pool identification AUC changes from 0.552 to 0.531. Excluding certainty preference, whose multiple-switcher rule leaves the fewest valid answers, gives a 13-bias complete sample of 532 respondents, on which A5 recovers 23% of the human residual correlation, with a classifier AUC of 0.970 and an identification AUC of 0.544.

Reassignment test. Respondents are grouped into strata by gender, age tercile, two income groups (below and above $40,000), and three education bands (below high school, high school to associate degree, bachelor's or higher); strata with fewer than six respondents are dropped, leaving 516–823 respondents per bias. Within each stratum, A5 verdicts are permuted among respondents and informedness is recomputed; over 2,000 permutations the mean is 0.026 (95% interval 0.010–0.042), against 0.078 observed ($p = 0.0005$).

Table 9: **Prevalence of each bias by condition (values behind Figure 2).** Each bias uses its available cases ($n = 629$–845). Bold marks the condition closest to HUMAN.

Bias	HUMAN	A1	A2	A3	A4	A5
Present bias — money	0.248	0.719	0.663	0.662	0.621	**0.587**
Present bias — consumption	0.168	1.000	0.900	0.802	0.853	**0.557**
Narrow bracketing	0.624	1.000	0.980	0.996	0.927	**0.885**
Loss aversion	0.602	0.006	0.464	0.625	**0.581**	0.447
Certainty preference	0.762	0.878	0.903	0.812	**0.712**	0.254
Ambiguity aversion	0.764	1.000	0.870	0.827	0.859	**0.798**
Overconfidence — level	0.339	0.004	**0.264**	0.260	0.079	0.456
Overconfidence — precision	0.454	0.000	0.000	0.003	0.000	**0.079**
Overconfidence — relative	0.730	0.006	0.022	0.026	0.055	**0.061**
Non-belief in the LLN	0.852	1.000	1.000	1.000	1.000	**0.991**
Gambler's fallacy	0.240	0.000	0.007	0.000	**0.017**	0.002
EGB — loans	0.527	0.849	0.918	0.867	0.884	**0.799**
EGB — assets	0.414	0.020	**0.426**	0.278	0.141	0.555
Limited attention	0.444	0.913	0.888	0.910	0.929	**0.791**

Table 10: **Largest gap between groups (values behind Figure 4).** Each cell is the highest minus the lowest group value within an attribute: (a) accuracy, on respondents with at least 8 valid biases ($n = 818$); (b) attributed bias prevalence, averaged over the 14 biases, on all 845 respondents. Ethnicity groups are defined in Appendix E.

Attribute	HUMAN	A1	A2	A3	A4	A5
(a) Accuracy						
Age	0.027	0.045	0.040	0.026	0.014	0.026
Gender	0.012	0.013	0.002	0.008	0.007	0.003
Ethnicity & Hispanic origin	0.026	0.034	0.057	0.047	0.059	0.043
Income	0.010	0.059	0.067	0.076	0.079	0.050
Education	0.021	0.052	0.063	0.054	0.049	0.053
(b) Attributed bias prevalence						
Age	0.020	0.015	0.027	0.036	0.038	0.034
Gender	0.046	0.002	0.029	0.018	0.043	0.053
Ethnicity & Hispanic origin	0.051	0.015	0.057	0.043	0.078	0.123
Income	0.083	0.009	0.164	0.125	0.117	0.166
Education	0.085	0.007	0.202	0.164	0.157	0.197

Table 11: **Compound-interest errors and numeracy by education (values behind Figure 5).** (a) Share answering the compound-interest item wrongly. (b) Mean number of the three numeracy questions answered correctly (0–3). n is the number of respondents in each group.

Education	n	HUMAN	A1	A2	A3	A4	A5
(a) Compound-interest item wrong							
High school or less	141	0.647	0.028	0.872	0.596	0.397	0.894
Some college	302	0.473	0.017	0.666	0.434	0.238	0.775
Bachelor's	230	0.340	0.017	0.209	0.126	0.078	0.426
Postgraduate	172	0.262	0.029	0.023	0.006	0.029	0.285
(b) Numeracy score							
High school or less	141	1.56	2.97	2.13	2.40	2.60	2.10
Some college	302	2.11	2.98	2.33	2.57	2.76	2.23
Bachelor's	230	2.43	2.98	2.79	2.87	2.92	2.57
Postgraduate	172	2.61	2.97	2.98	2.99	2.97	2.72

Table 12: **Profile-level checks on the primary and complete-profile samples.** Classifier AUC measures how easily a respondent-grouped classifier separates synthetic from human bias profiles; 0.5 is indistinguishable and closer to 0.5 is better. Residual correlation measures agreement with human wave 2 after removing person and bias means; higher is better. HUMAN compares wave 1 with wave 2. Brackets are 95% respondent-bootstrap intervals. The primary sample has $n = 829$ and the complete-profile sample $n = 389$. Dashes mark values not reported.

	Classifier AUC		Residual correlation	
Condition	Primary	Complete	Primary	Complete
HUMAN	0.476 [0.451, 0.500]	0.510 [0.476, 0.544]	0.287 [0.266, 0.308]	0.297 [0.267, 0.325]
A1	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	−0.004 [−0.026, 0.017]	—
A2	0.996 [0.993, 0.998]	—	0.008 [−0.013, 0.029]	—
A3	0.994 [0.991, 0.996]	—	0.022 [0.001, 0.044]	—
A4	0.994 [0.991, 0.996]	—	0.023 [0.002, 0.043]	0.040 [0.012, 0.067]
A5	0.976 [0.970, 0.982]	0.976 [0.967, 0.984]	0.070 [0.049, 0.090]	0.059 [0.029, 0.091]

Table 13: **Identification on the primary sample.** Identification AUC is the probability that the correct person's profile is closer than an alternative, with half credit for tied distances. All-pool comparisons use $n = 828$ profiles; demographic-matched comparisons use $n = 788$. Brackets are 95% respondent-bootstrap intervals. These are the estimates discussed in §3.2; the complete-profile sensitivity estimate is reported in the text of this appendix.

Condition	All-pool AUC [95% interval]	Matched AUC [95% interval]
HUMAN	0.718 [0.699, 0.737]	0.706 [0.685, 0.726]
A1	0.514 [0.495, 0.534]	0.513 [0.492, 0.535]
A2	0.518 [0.498, 0.538]	0.513 [0.491, 0.534]
A3	0.524 [0.505, 0.543]	0.520 [0.499, 0.542]
A4	0.533 [0.514, 0.553]	0.530 [0.509, 0.551]
A5	0.552 [0.534, 0.572]	0.538 [0.514, 0.560]

H CONTINUOUS SCORES

Thirteen of the 14 biases have a graded score underneath the binary verdict (present bias on consumption is a pure two-item construct and has none). Table 14 reports, per bias, how well each side's continuous score discriminates the respondent's wave-2 verdict (AUC, null 0.5). AUC does not depend on prevalence, so it cannot be inflated the way raw accuracy is.

Orientation. For non-belief in the law of large numbers, gambler's fallacy, and exponential-growth bias on assets, lower scores indicate more bias. We orient these scores before computing AUC.

Table 14: **Agreement on the continuous scores underlying 13 verdicts.** AUC measures how well each condition's continuous score discriminates the respondent's wave-2 binary verdict; HUMAN uses the person's wave-1 score. A5 reaches 0.33 of the human benchmark. Per-bias common samples range from $n = 526$ to $n = 843$.

Bias	HUMAN	A1	A2	A3	A4	A5
Present bias — money	0.485	0.507	0.462	0.523	0.479	0.512
Loss aversion	0.694	0.502	0.459	0.505	0.464	0.548
Narrow bracketing	0.659	0.515	0.511	0.532	0.558	0.545
Certainty preference	0.564	0.505	0.463	0.434	0.541	0.521
Ambiguity aversion	0.628	0.534	0.475	0.520	0.462	0.537
Overconfidence — level	0.663	0.503	0.554	0.545	0.506	0.602
Overconfidence — precision	0.732	0.500	0.500	0.499	0.500	0.523
Overconfidence — relative	0.801	0.482	0.414	0.468	0.458	0.490
Non-belief in the LLN	0.722	0.498	0.538	0.592	0.601	0.671
Gambler's fallacy	0.681	0.500	0.515	0.500	0.506	0.502
EGB — loans	0.601	0.483	0.476	0.494	0.495	0.501
EGB — assets	0.751	0.502	0.614	0.575	0.586	0.694
Limited attention	0.700	0.502	0.582	0.563	0.566	0.574
Mean	0.668	0.503	0.505	0.519	0.517	0.555
Share of HUMAN	1.00	0.02	0.03	0.11	0.10	0.33

Variance on the continuous scale. Repeating the decomposition in §3.1 with oriented, mid-rank tiles on the balanced sample ($n = 518$) gives a human between-person share of 0.097, compared with 0.074 for the binary verdicts. Thus, binarization hides some person-level variation in the human data. Relative to the human share, A5 falls from $1.05\times$ on binary verdicts to $0.85\times$ on continuous scores; the other conditions range from $0.23\times$ to $0.69\times$ on the continuous scale.

I Distribution of prediction gains across respondents

The mean gains in §3.2 do not show who benefits. We therefore compare A5 with A1 and A4 for each respondent, requiring at least seven tasks with valid targets and predictions. Binary accuracy is the share of matched wave-2 verdicts. Graded error is the mean absolute model–human difference, normalized by each task's scale, across the 12 graded biases other than certainty preference, which has the fewest valid answers. Relative to A4, A5 lowers graded error for 59% of respondents and raises it for 41%; the mean reduction is 0.015 (95% interval 0.011–0.020). Relative to A1, the graded-error split is even and the mean interval includes zero. For binary accuracy, A5 outperforms A1 for 46% of respondents and underperforms for 33%, with a mean gain of 2.5 percentage points (95% interval 1.5–3.5). The 95% interval for its mean binary gain over A4 includes zero. These person-weighted comparisons differ from the bias-weighted informedness in Table 3. Because each response was generated once, we cannot assess whether a given respondent's gain would persist across runs.

Table 15: **Respondent-level changes under the behavioral-history condition (A5).** Better, Same, and Worse are the percentages of respondents for whom A5 has lower, equal, or higher prediction error than the comparison condition. Binary gains are percentage-point changes in accuracy across 14 biases; graded gains are reductions in normalized absolute error across 12 biases. Intervals are 95% respondent-bootstrap intervals from 2,000 draws.

Measure	A5 compared with	n	Better	Same	Worse	Mean gain [95% interval]
Binary accuracy	A1	837	46.2	21.0	32.7	2.48 [1.45, 3.49]
Binary accuracy	A4	837	41.3	20.4	38.2	0.69 [−0.32, 1.62]
Graded error	A1	821	50.2	0.0	49.8	0.0036 [−0.0010, 0.0082]
Graded error	A4	821	59.2	0.0	40.8	0.0153 [0.0108, 0.0199]

J LOSS AVERSION BY INCOME AND EDUCATION

In the human data, loss aversion increases modestly with income, and with education up to a bachelor's degree, although adjacent intervals overlap (Table 16). Every condition with a persona profile (A2–A5) reverses this gradient. The reversal is steepest under A2 and weakest under A5, but it remains present under A5. Under A1, the model almost never rejects either gamble.

Table 16: **Loss-aversion prevalence by income and education.** HUMAN is the wave-2 rate with a 95% bootstrap interval. Income and education groups are those used in Figure 4 (Appendix E). Overlapping human intervals mean that the human gradient should be treated as descriptive.

Attribute	Group	n	HUMAN	A1	A2	A3	A4	A5
Income	low income	328	0.555 [0.503, 0.607]	0.009	0.826	0.869	0.820	0.509
	middle income	318	0.604 [0.547, 0.657]	0.006	0.330	0.582	0.516	0.450
	high income	198	0.677 [0.611, 0.742]	0.000	0.081	0.293	0.293	0.343
Education	high school or less	141	0.518 [0.440, 0.596]	0.000	0.766	0.872	0.865	0.560
	some college	302	0.596 [0.540, 0.649]	0.010	0.609	0.732	0.639	0.460
	bachelor's	230	0.643 [0.578, 0.704]	0.004	0.309	0.522	0.483	0.448
	postgraduate	172	0.628 [0.558, 0.703]	0.006	0.169	0.372	0.378	0.331