WHEN SYNTHETIC USERS FAIL: A CROSS-DOMAIN BENCHMARK OF LLM-SIMULATED HUMAN SURVEY RESPONSES

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) are increasingly used as synthetic users: stand-ins for human respondents whose simulated answers inform product, policy, and market decisions. We ask when this substitution is valid and when it fails. Using a unified evaluation protocol, we study four models spanning two model families and an 8B-to-frontier capability range across two independent domains of real human responses: U.S. social attitudes from the General Social Survey (GSS) and cross-cultural values from the World Values Survey (WVS). We benchmark each model against non-LLM baselines fit on held-out human data. Across both domains and model families, we find two consistent failures under the demographic prompting and survey-simulation protocols we test. First, at the individual level, no LLM outperforms the strongest non-LLM baseline. Under proper scoring rules (log-loss and Brier score), every model assigns less probability to the observed human response than a demographic baseline; this gap persists under distance-aware scoring, and on WVS, every model also falls well below the baseline in exact-match accuracy. Second, the models over-determine demographics: they treat demographic identity as more predictive of attitudes than it is among real human responses, most clearly for the Claude models and Llama-70B, though our small model sample overstates this gap. In the two within-family size comparisons we test, the larger model remedies neither failure.

1 INTRODUCTION

Running a human study is slow and expensive, so a fast-growing body of work proposes replacing human participants with large language models (LLMs) prompted to behave as *synthetic users*: to pilot surveys, pre-test questionnaire items, estimate opinion distributions, and stand in for study participants. The appeal is obvious: a survey that takes months and thousands of dollars to field can, in principle, be approximated in minutes at the cost of a few API calls. These simulated respondents are increasingly wired into decision-support pipelines. Product teams query synthetic personas to prioritize features, marketers use them to pre-test messaging and estimate segment-level demand through "silicon samples" (Sarstedt et al., 2024; Li et al., 2024), and policy and organizational analysts use them to gauge how a population might react before committing resources.

In each case, the synthetic user manufactures evidence that a human sample would normally provide, and a downstream decision is made on that evidence. Like any evidence-producing component of a decision-support system, it should be verified and validated against its intended use before it is trusted (Shmueli & Koppius, 2011; O'Keefe & O'Leary, 1993): if synthetic users are wrong in a structured way, the resulting decisions inherit that error at scale. A practical evaluation framework, one that tells a team in advance whether synthetic-user evidence is trustworthy for the decision at hand, is therefore as important as the simulation method itself.

The central unresolved question is not whether LLMs can produce survey-shaped output (they trivially can), but *when the substitution is valid and when it fails*. The evidence so far conflicts: some studies report encouraging aggregate alignment (Argyle et al., 2023), while others warn that substitution is unsafe (Bisbee et al., 2024; Gao et al., 2025). Prior work has also largely answered the question one domain at a time, often reporting a single aggregate similarity number that can

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.

look reassuring even when the underlying simulation is poor. This leaves several basic questions unresolved. At the individual level, do LLM synthetic users predict human answers better than a simple demographic-conditional predictor fit on labeled human data (RQ1)? At the aggregate level, do they reproduce the population distribution of answers (RQ2)? Do they preserve the demographic structure of attitudes, or distort how predictive demographics are of responses (RQ3)? And are these conclusions stable across model family, capability, and output format (RQ4), and across domains (RQ5)?

We address these questions with one fixed protocol (the same models, prompts, decoding, sampling, and metrics) applied to two independent domains of real human responses: U.S. general social attitudes from the General Social Survey (GSS, 2016–2024 waves, 10 attitude questions) and cross-cultural values from the World Values Survey (WVS Wave 7, 63 countries, 16 ordinal value questions). We evaluate four models from two families spanning an 8B-to-frontier capability range under two prompt formats. A central feature of the evaluation is an explicit demographic-lookup baseline fit on held-out real respondents. Individual survey answers are not deterministic functions of demographics, so "the model predicts individuals with $X\%$ accuracy" is uninterpretable until it is compared with what a trivial predictor already captures.

Under demographic prompting, two failures replicate across *both* domains and *both* model families we tested. The first is a *lack of individual-level advantage*. Under proper scoring of distributional outputs (log-loss and Brier score), every model in both domains places less probability on the observed human response than the demographic lookup, and the deficit persists under distance-aware metrics. On exact match, every LLM on GSS at best ties the lookup and trails a learned logistic baseline; on WVS, every model is 11 to 22 percentage points less accurate than the lookup. The second is *demographic over-determination*: models treat demographic attributes as far more predictive of a person's answer than they are among real people. For example, politics explains only about 1.5% of the variation in real U.S. confidence in banks, yet Haiku 4.5 under the single-answer prompt behaves as though it explains roughly 67%. This exaggeration is significant against zero for most question–group pairs in both domains (a lenient test at this sample size; Limitations), and it also appears under a coding-invariant measure. In the two within-family size comparisons we test, neither failure is remedied by the larger model; the frontier models over-determine at least as strongly as the small 8B model, and often more.

Our contribution is a unified validation framework that brings together ingredients previously studied largely in isolation. It rests on a *cross-domain benchmark*: the same protocol applied to two independent domains of real human responses, U.S. social attitudes (GSS) and cross-cultural values (WVS), so that properties of LLM simulation can be separated from properties of a single dataset. A baseline-anchored individual evaluation then compares every model with non-LLM baselines fit on held-out human data, primarily under proper scoring, so that individual-level accuracy is read against the information the demographics already carry. A stereotyping index measures, on a bounded scale, how much more predictive a demographic attribute is of the answer in the model than among real people; with a coding-invariant companion based on Cramér's V, it turns the worry that LLMs *flatten* groups into its measurable opposite for identity-linked attitudes. We further add descriptive decision-level diagnostics for segment targeting (gap inflation, wrong-target, and spurious-split rates). Together, these tools provide a reproducible procedure that a practitioner can run before deploying a synthetic-user system (Section 5), and the negative findings show why it is needed.

2 Related Work

A fast-growing line of work treats LLMs as programmable populations, prompting them with personas or demographics to predict opinions and behaviors. The optimistic anchor is Argyle et al. (2023), who show that demographically conditioned prompts can reproduce aggregate patterns in U.S. political survey data, launching the "silicon sample" idea. A skeptical counter-current followed: LLM survey responses can misrepresent population and subgroup variation (Bisbee et al., 2024), align more with some demographic groups than others (Santurkar et al., 2023), and shift with option ordering and labeling (Dominguez-Olmedo et al., 2024).

A parallel effort makes survey and cross-cultural value simulation measurable at scale. WorldValuesBench derives a demographic-context-to-answer benchmark from the World Values Survey (Zhao et al., 2024), and one of our two domains reuses its question set and scales for comparability; So-

cioBench evaluates LLMs as survey-response predictors across countries and domains (Wang et al., 2025). Their only reference points are in-sample oracles or a random-guess floor, not a demographic predictor fit on held-out human data, so a reported accuracy is not evidence that the simulation adds information; each also draws on a single survey program (WVS; ISSP). Distribution-specialization work fine-tunes models to match country-level response distributions, improving group-level prediction while still struggling to generalize (Cao et al., 2025).

Persona assignment is not a neutral conditioning operation: it can amplify toxicity and stereotypes (Deshpande et al., 2023), the surface architecture of a prompt can induce methodological artifacts (Brucks & Toubia, 2025), and sociodemographic persona formulations materially change model behavior (Lutz et al., 2025). Learned representations carry human-like social biases (Caliskan et al., 2017), and the fairness literature establishes that aggregate performance can mask subgroup disparities (Buolamwini & Gebru, 2018). Our stereotyping index adapts this lens with a different target: rather than asking whether a model attaches a negative label to a group, we compare how predictive demographics are in the model versus in real humans.

Individual-level results are usually reported without non-LLM baselines, so it is unclear whether a model beats a demographic lookup, let alone a learned demographic model (Shmueli & Koppius, 2011). Subgroup analyses typically ask whether groups are "flattened," but do not quantify, on a human-comparable scale, how much demographics should explain versus how much the model makes them explain. In the spirit of the total-survey-error framework (Groves & Lyberg, 2010), we treat individual, aggregate, and subgroup validity as separate questions. Table 1 summarizes coverage (extended review in Appendices A.1–A.7).

Table 1: Closest prior work as characterized in Section 2. Baseline: non-LLM predictor fit on human data; Proper: individual proper scoring; Subgroup: attribute–answer association vs. humans; X-domain: several surveys, one protocol; Decision: decision impact. –: not in our characterization (not a claim that the work lacks it).

Work	Baseline	Proper	Subgroup	X-domain	Decision
Argyle et al. (2023)	–	–	partial	–	–
Bisbee et al. (2024)	–	–	partial	–	–
Santurkar et al. (2023)	–	–	partial	–	–
Zhao et al. (2024)	partial	–	–	–	–
Wang et al. (2025)	–	–	–	partial	–
Cao et al. (2025)	–	–	partial	partial	–
This work	✓	✓	✓	✓	✓

3 Data, Protocol, and Metrics

3.1 Domains and sampling

We use two public human-response datasets and restrict all claims to the scope we actually use. *GSS (U.S. social attitudes):* the GSS 1972–2024 cumulative file restricted to the 2016–2024 waves, whose recent waves give large per-subgroup cells and partially mitigate pretraining contamination. After removing respondents with missing values on any prompt demographic, the in-scope pool is 14,704 respondents and 85,898 (respondent, question) pairs over 10 questions; prompt demographics are age, sex, race, highest degree, region, political views, and party identification. *WVS (cross-cultural values):* WVS Wave 7 (inverted CSV, v6.0), retaining the 16 value questions from the WorldValuesBench probe set (Zhao et al., 2024) present in this release, with its wordings and ordinal scales. The in-scope pool is 91,774 respondents across 63 countries and 1,426,473 (respondent, question) pairs; demographics: age group, sex, education, settlement type, and country.

Our headline analyses need a reliable estimate of the human answer distribution within each (question, group) cell, so we sample the LLM evaluation set by cell rather than at random: about 100 respondents per question (GSS: 993 rows; WVS: 1,458 rows), allocated so that the main grouping is well covered (GSS: degree $\times$ race; WVS: country). The human answer distributions and the demographic baseline are always estimated from the full dataset, never from this small sample, so the human side rests on large, stable numbers, the model side on this small one (Limitations).

3.2 MODELS AND PROMPTING

We evaluate four chat-style models through one common interface: Claude Haiku 4.5 and Claude Sonnet 4.6 (the *closed* family, accessed via an API) and Llama-3.1-8B and Llama-3.3-70B (the *open*, open-weight family), spanning a small 8B model to frontier-scale systems. Each sampled respondent is run under two prompt formats and two independent generation runs ("seeds," which differ only in sampling randomness). The *single-answer prompt* (Style A) shows the demographic profile and question and asks for one answer; the *distribution prompt* (Style C) asks for a probability for each answer option, as JSON. All calls use the same output parser, which flags any response it cannot read as a valid answer as *invalid/refused* (settings, prompts, and parser rules in Appendices B–E).

3.3 NON-LLM BASELINES

The individual-accuracy number is only interpretable against what a non-LLM predictor achieves on the same people. We build four baselines, all fit on a held-out 50% split of the in-scope pool (assigned by a hash of the respondent id) and evaluated on the evaluation-sample rows from the other split, so no baseline is fit and scored on the same respondents: the *question marginal* (the most common answer, ignoring demographics); the *demographic lookup*, our primary yardstick, which predicts from the conditional answer distribution given demographics and backs off from a fine to a coarse demographic cell to the marginal (for WVS the coarse cell is country alone, reported separately as a country-only baseline); *multinomial logistic regression* on the one-hot prompt demographics; and a *random forest* on the same features (300 trees). The lookup and logistic model also emit a full predicted distribution, the reference for proper scoring. The baselines use in-domain, labeled human responses that a zero-shot LLM is not given at inference time, so they are an evaluation reference, not a deployable substitute for a synthetic user: they measure how much individual-level information the demographics already carry. Following the predictive-validity tradition (Shmueli & Koppius, 2011), we ask whether a synthetic user adds information beyond that reference on held-out human data (details in Appendix F).

3.4 METRICS

We report individual, aggregate, and subgroup fidelity separately, mirroring the total-survey-error distinction between measurement and representation error (Groves & Lyberg, 2010) and multi-metric evaluation norms (Liang et al., 2023; Ribeiro et al., 2020). Every headline comparison against a baseline carries a paired-bootstrap 95% confidence interval: we resample the evaluation rows 500 times, scoring model and baseline on the *same* rows, and report the interval of the difference.

Individual fidelity (RQ1) is scored primarily with two proper scoring rules on the whole distribution-prompt output, *log-loss* and the *Brier score*, against the baseline's conditional distribution. We also report exact-match *accuracy* (for the distribution prompt, its arg-max option) and, because exact match is harsh on long ordinal scales, the *mean absolute error* (MAE) of the predicted answer and the *earth-mover's distance* (EMD) of the predicted distribution, each next to the baseline on the same people.

Aggregate fidelity (RQ2) compares the share of people choosing each option between the model and real humans using the *Jensen–Shannon (JS) divergence* (0 means identical; lower is better). Because we sample by cell, we also report a *population-reweighted* JS in which each cell's model distribution is weighted back to its share of the real population; agreement between the two rules out an artifact of the sampling design.

Subgroup determinism (RQ3) asks how strongly a demographic attribute *predicts* the answer, for real people and for the model. We measure it by η^2, the share of total answer variation explained by group membership (from 0, the attribute tells you nothing, to 1, it determines the answer), computed identically for humans and model. Our headline measure, the *stereotyping index*, is the difference

$$\Delta\eta^2 \;=\; \eta^2_{\text{model}} - \eta^2_{\text{human}},$$

positive when the model treats an attribute as *more* predictive of the answer than it is among real people, with a 95% bootstrap interval over 500 resamples of respondents. Because η^2 treats answer codes as numbers, we pair it with a *coding-invariant* companion, $\Delta V = V_{\text{model}} - V_{\text{human}}$, the difference in Cramér's V between group membership and the chosen answer, which applies to all questions including nominal ones.

Decision impact extends over-determination to the most common use of a synthetic user: *segment targeting*, where a team asks which demographic segment is most extreme on an attitude and how large the between-segment gap is. For each (question, demographic axis) we compute, identically for model and humans, each segment's share p_g on the high end of the scale and report (i) the *gap-inflation factor* $\text{gap}_{\text{model}}/\text{gap}_{\text{human}}$, with gap $= \max_g p_g - \min_g p_g$; (ii) the *wrong-target rate*, the fraction of pairs where the model's $\arg\max_g p_g$ differs from the human one; and (iii) the *spurious-split rate*, the fraction where humans show a negligible gap (≤ 0.10) but the model a large one (≥ 0.25). *Stability (RQ4)* is measured by invalid-output rates, agreement between decoding runs, and movement across family, size, and prompt wording. The stereotyping-index significance counts are essentially unchanged by a Benjamini–Hochberg FDR correction, and the RQ1 margins survive a respondent-clustered bootstrap (Appendix H); full metric definitions are in Appendix G.

4 RESULTS

Table 2: Cross-domain, cross-model results. Δbase = individual accuracy minus the demographic-lookup baseline (negative = worse than the baseline); paired-bootstrap 95% CIs for this margin are shown per model in Figure 1, and the stronger baselines, distance-aware, and proper-scoring comparisons in Table 3. JS = mean Jensen–Shannon divergence (lower is better; population-reweighted values agree, see Section 4.2). $\overline{\Delta\eta^2}$ = median stereotyping index (positive = over-determination); "sig" = number of question–group pairs whose 95% CI excludes zero, out of the total (a coding-invariant Cramér's V replication is in Section 4.3). Style A = single answer, Style C = distribution.

Domain	Model	Sty	Acc	Base	Δbase	JS	$\overline{\Delta\eta^2}$ (sig)
GSS	Haiku 4.5	A	0.587	0.589	−0.003	0.090	+0.048 (5/8)
GSS	Haiku 4.5	C	0.583	0.589	−0.006	0.011	+0.059 (9/12)
GSS	Sonnet 4.6	A	0.563	0.589	−0.026	0.082	+0.081 (10/12)
GSS	Sonnet 4.6	C	0.589	0.589	−0.001	0.017	+0.104 (9/12)
GSS	Llama 8B	A	0.496	0.589	−0.093	0.056	+0.026 (7/12)
GSS	Llama 8B	C	0.509	0.586	−0.077	0.023	+0.026 (3/7)†
GSS	Llama 70B	A	0.569	0.589	−0.021	0.072	+0.051 (6/8)
GSS	Llama 70B	C	0.582	0.589	−0.008	0.016	+0.031 (5/8)
WVS	Haiku 4.5	A	0.170	0.388	−0.218	0.382	+0.087 (30/32)
WVS	Haiku 4.5	C	0.269	0.404	−0.135	0.046	+0.056 (21/24)†
WVS	Sonnet 4.6	A	0.236	0.388	−0.152	0.274	+0.107 (31/32)
WVS	Sonnet 4.6	C	0.277	0.388	−0.111	0.037	+0.103 (26/28)
WVS	Llama 8B	A	0.179	0.388	−0.209	0.232	+0.030 (22/32)
WVS	Llama 8B	C	0.182	0.400	−0.218	0.130	+0.376 (1/1)†
WVS	Llama 70B	A	0.249	0.388	−0.139	0.293	+0.136 (31/32)
WVS	Llama 70B	C	0.267	0.388	−0.121	0.046	+0.054 (27/30)

† Style-C cells with elevated invalid rates (Section 4.5); interpret with caution. Llama-8B Style-C on WVS is effectively unusable (85% invalid, $n = 440$).

4.1 RQ1: UNDER PROPER SCORING NO LLM BEATS THE DEMOGRAPHIC BASELINE; ON VALUES ALL FALL WELL BELOW IT

Table 2 reports the full cross-model, cross-domain results and Table 3 the RQ1 scoring comparisons; Figure 1 gives paired-bootstrap confidence intervals on the accuracy margin over the baseline (raw accuracies in Figure 3). Our primary individual-level measure scores the whole Style C distribution with proper scoring rules, and shows that the model assigns *lower* probability to the true human answer than the baseline does: on GSS the model's log-loss is 0.84–1.03 versus 0.80–0.81 for the baseline on the same rows (Brier 0.52–0.62 vs. 0.50); on WVS 1.76–2.26 versus 1.70–1.85 (Brier 0.80–0.89 vs. 0.71–0.74). Distance-aware error on the ordinal questions agrees. On GSS the model's MAE is 0.57–0.68 versus 0.54 for the baseline; on WVS's long scales it is 1.83–2.05 versus 1.74. The distributional EMD agrees (0.76–0.83 vs. 0.73 on GSS; 1.94–2.30 vs. 1.67–2.02 on WVS).

On exact match, the individual accuracy of every model on GSS is at or below the demographic lookup baseline (0.589). The best margin any model achieves is −0.001 (Sonnet, Style C),

a paired-bootstrap tie (CI $[-0.028, 0.022]$); Llama-8B is 9 points worse (0.496 vs. 0.589, CI $[-0.120, -0.067]$). The lookup is not a weak straw man: a learned *logistic* baseline reaches 0.622 and a random forest 0.583, so the best LLM also trails the strongest non-LLM predictor. On GSS no margin is positive, with the deficits for Llama-8B (both prompts) and Sonnet's single-answer prompt clearing 95% significance and the remaining cells statistical ties at or below the baseline. On WVS the gap is far larger and uniformly negative: under the single-answer prompt every model is 13.9 to 21.8 accuracy points below the demographic baseline (e.g. Haiku Style A: 0.170 vs. 0.388), every one of these margins has a 95% CI that lies strictly below zero, and every model under either prompt is below the learned baselines as well. The arg-max of the distribution prompt also trails the lookup in every WVS cell (Table 2). So for predicting an individual's response, and especially a cross-cultural value response, a simple demographic predictor fit on in-domain human data beats all four LLMs; synthetic users provide *no* individual-level advantage over it under any of these scorings, and on values they are actively worse.

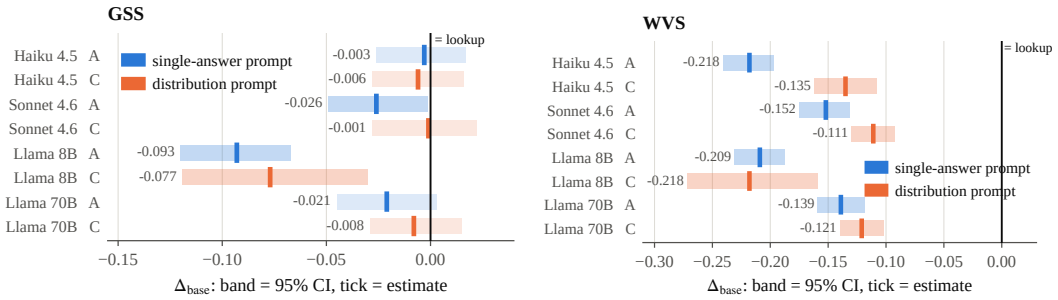


Figure 1: Individual-accuracy margin over the demographic baseline (Δ_{base}) with paired-bootstrap 95% confidence intervals, GSS (left) and WVS (right). Negative means worse than the baseline. Every interval lies below zero except the two GSS ties, and the WVS deficits are large and tight.

Table 3: RQ1 scoring comparisons: the individual-level deficit holds under proper scoring, distance-aware scoring, and against stronger baselines. *Top:* pooled individual accuracy of the four non-LLM baselines (the best LLM is shown for reference). *Bottom:* per-model distance-aware error on ordinal questions and proper scores, each next to the baseline on the same rows (all lower is better): MAE = mean absolute error of the single-answer prediction; EMD = earth-mover's distance of the full predicted distribution (Style C) from the true answer, in scale steps; log-loss and Brier are evaluated on the distribution prompt at the true answer. Each cell shows model / baseline (m/b) on the same rows. In every cell the model is at best tied and usually worse than the baseline.

	Marginal	Dem. lookup	Logistic	Random forest	Best LLM
GSS accuracy	0.568	0.589	**0.622**	0.583	0.589
WVS accuracy	0.348	0.388	**0.393**	0.366	0.277

Domain	Model	MAE (m/b)	EMD (m/b)	log-loss (m/b)	Brier (m/b)
GSS	Haiku 4.5	0.57 / 0.54	0.78 / 0.73	0.84 / 0.81	0.52 / 0.50
GSS	Sonnet 4.6	0.60 / 0.54	0.78 / 0.73	0.84 / 0.81	0.52 / 0.50
GSS	Llama 8B	0.68 / 0.54	0.83 / 0.74	1.03 / 0.80	0.62 / 0.50
GSS	Llama 70B	0.59 / 0.54	0.76 / 0.73	0.86 / 0.81	0.53 / 0.50
WVS	Haiku 4.5	1.86 / 1.74	1.94 / 1.67	1.76 / 1.70	0.80 / 0.71
WVS	Sonnet 4.6	1.84 / 1.74	2.22 / 2.02	1.88 / 1.85	0.81 / 0.74
WVS	Llama 8B	2.05 / 1.74	2.30 / 1.87	2.26 / 1.85	0.89 / 0.73
WVS	Llama 70B	1.83 / 1.74	2.27 / 2.02	1.94 / 1.85	0.82 / 0.74

4.2 RQ2: Aggregate distributions are reproduced reasonably, especially under Style C

Reproducing the group's overall answer *spread* is much easier than guessing individuals, and asking directly for a distribution (Style C) does this markedly better than asking for a single answer (Style A). On GSS, Style C reaches 0.011–0.023 JS (essentially the human distribution), versus 0.056–0.090 for Style A; WVS shows the same ordering (Style C $\approx$ 0.037–0.046 for the models that produce valid output, versus 0.23–0.38 for Style A). Reweighting each cell back to its human population share leaves the JS divergence essentially unchanged (e.g. Haiku GSS Style C 0.0108 $\rightarrow$ 0.0109; Sonnet WVS Style C 0.0371 $\rightarrow$ 0.0373), so the model is compared against the true population distribution, not against our sampling design.

4.3 RQ3: LLMs over-determine demographics (stereotyping), not flatten them

A common worry, given that LLM survey responses vary less than real ones (Bisbee et al., 2024), is that LLMs *flatten* groups, washing out the differences between them. For identity-linked attitudes we find the opposite. The median stereotyping index is *positive* in every reliable model–prompt cell in both domains, and its confidence interval excludes zero (a test against zero; Limitations) for most question–group combinations (Table 2; grouping axes in Appendix C): on these measures, models treat demographics as more decisive of the answer than they are among real people. The effect is dramatic for politically- and identity-linked questions. On GSS, a person's political views explain about 10% of the real variation in gender-role attitudes, but in the two Claude models under the single-answer prompt they appear to explain 60–69% (39–75% across the two Llama models); for confidence in banks, political views explain only about 1.5% of the real variation yet about 67% in Haiku under the single-answer prompt (and up to 81% across all model–prompt cells), roughly a forty-fold exaggeration. The simulated population is therefore not a blurred copy of the real one; it is a *caricature*, in which who you are dictates what you think more tightly than it does in reality (question-level results in Figure 4).

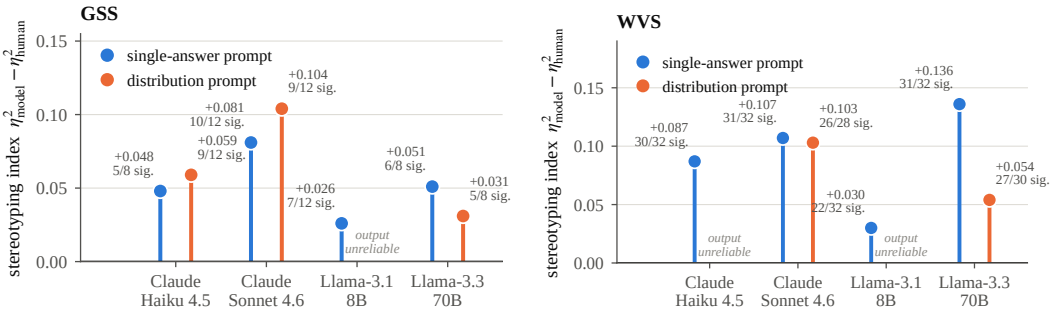


Figure 2: Stereotyping index ($\eta^2_{\text{model}} - \eta^2_{\text{human}}$) by model and prompt style, GSS (left) and WVS (right); higher means the model treats demographics as more decisive of the answer than they are among real people. All reliable bars are positive in both domains, smallest for Llama-8B (RQ3), and the larger model in each family does not reduce the effect (RQ4). Distribution-prompt bars with high invalid rates are omitted ("output unreliable"; Section 4.5).

This does not depend on how we code the answers. The coding-invariant Cramér's V tells the same story on all questions, including the nominal GSS items the η^2 index omits. The median ΔV is positive for both Claude models in both domains (GSS Sonnet Style A $+0.152$, Haiku $+0.108$; WVS Sonnet Style A $+0.160$, Llama-70B $+0.166$), and its bootstrap CI excludes zero for a majority of question–group pairs (e.g. 22 of 40 for Sonnet on GSS, 23 of 32 on WVS). Both indices (and the model-side percentages above) set a small per-question model sample against full-table human values; as η^2 and V are biased upward at small n, they overstate the model–human gap in magnitude and in significance counts (Limitations).

4.4 DECISION IMPACT: WHAT OVER-DETERMINATION WOULD IMPLY FOR SEGMENT TARGETING

If a team read segment structure off the synthetic users rather than off real humans, the targeting read would differ in three ways (Figure 5, appendix). First, *the between-segment gap* a decision keys on *is larger when read off the model*: for the frontier model (Sonnet, single-answer) the median gap is $2.3\times$ the true human gap on GSS and $2.5\times$ on WVS, and every model's median exceeds the human gap (medians 1.1–$2.3\times$ on GSS, 2.0–$4.7\times$ on WVS). Second, *the model's top segment often differs from the human one*: for 50% of GSS (question, axis) pairs and 72% of WVS pairs (Sonnet single-answer). Third, the model *shows large segment splits where real people show negligible ones*: on WVS, 28% of pairs (Sonnet single-answer; up to 41% for Haiku and Llama-70B) fall in the spurious-split regime, where humans differ across segments by ≤ 10 points while the model differs by ≥ 25. At this sample size arg-max and threshold rates are sensitive to sampling noise, and a same-size human subsample would itself show gap inflation and arg-max disagreement against the full table, so these rates cannot yet be attributed to the model. We therefore report them as descriptive diagnostics, not estimates of model error, with a same-size human reference part of the protocol (Limitations).

Two cases illustrate the pattern. Among U.S. respondents, confidence in banks barely varies with political leaning (a "great deal" of confidence ranges only 13–24% across the seven political categories), yet Haiku (single-answer prompt) treats politics as roughly 67% determinative of the answer; a team reading the simulation would conclude, wrongly, that trust in banks is sharply polarized. On WVS, attitudes to government responsibility (Q108) differ across education levels by under 5 points among real people but by 74 points in the model ($16\times$; model-side CI $[0.60, 0.86]$). If confirmed against a same-size human reference, a distortion toward exaggerated group difference would not average out across decisions; it would push a team toward over-segmenting.

4.5 RQ4: THE LARGER MODELS TESTED DO NOT FIX IT; OUTPUTS ARE UNSTABLE TO FORMAT AND PROMPT SURFACE

First, *in the two within-family size comparisons we test, the larger model does not remedy either failure*. Within the closed family, the frontier model (Sonnet 4.6) over-determines *more* than the smaller Haiku 4.5 (median $\Delta\eta^2$ $+0.104$ vs. $+0.059$ on GSS Style C; $+0.103$ vs. $+0.056$ on WVS Style C) and does not beat the baseline on individuals; within the open family, Llama-70B has the highest stereotyping index on WVS ($+0.136$).

Second, *format stability is a real failure mode for smaller models*: the distribution prompt over long numeric scales produced high invalid-output rates, with Llama-8B at 85% invalid on WVS (rendering its Style-C numbers unusable, $n = 440$) and 57% on GSS, and Haiku at 40% on WVS, while Sonnet and Llama-70B remained at 0%. Counting invalid outputs as failures only widens the gap to the baseline (Appendix H).

Third, *predictions are unstable to prompt-surface choices that should be irrelevant*, consistent with prior reports of option-order and persona-formulation sensitivity (Dominguez-Olmedo et al., 2024; Brucks & Toubia, 2025; Lutz et al., 2025). On GSS, merely reversing option order flips 9.8% (Sonnet) to 10.8% (Haiku) of single-answer predictions, and reframing the identical demographics as a persona sentence flips a further $\sim 9\%$, both roughly 4–$10\times$ the seed-noise floor. Synthetic-user validity is thus not only a model property but a prompt/interface property (Appendix H).

4.6 RQ5: BOTH FAILURES TRANSFER ACROSS DOMAINS

Both failures hold in both domains (Table 2; over-determination weakest for Llama-8B), and the individual-level failure is *stronger* in the cross-cultural domain (an 11–22 point deficit vs. a rough tie). We read them as properties of demographic prompting with these models, not of one survey or prompt.

5 DISCUSSION

The two failures may share a cause: the models we test, prompted with demographics, appear to map a profile to a stereotyped "typical" response, consistent with evidence that word embeddings (Caliskan et al., 2017) and large language models (Bender et al., 2021) encode social associations and that persona conditioning steers models toward stereotyped abstractions (Deshpande et al., 2023). That mapping is too coarse to beat a demographic lookup at the individual level (and worse than it on the 10-point WVS scales), yet too tight, treating demographics as more predictive than they are. A synthetic-user study can therefore look successful on an aggregate metric, even one improved by distribution-specialized training (Cao et al., 2025), while being invalid for individual- or subgroup-level use.

Implications and a validation protocol A team that only checks aggregate fidelity, the common practice, can be misled twice: decisions keyed to individuals or narrow segments rest on individual-level fidelity the models do not have, and decisions that read the structure of opinion off the simulation would inherit its over-determination (Section 4.4). If confirmed against a same-size human reference, such error would be structured and would not average out. We therefore recommend reporting individual and aggregate fidelity separately, against non-LLM baselines on held-out human data with proper or distance-aware scoring; reporting subgroup determinism ($\Delta\eta^2$ or ΔV) and checking gap inflation, wrong-target, and spurious-split rates against a same-size held-out human subsample before acting on a segment-level read; reporting invalid rates; not assuming a larger model is a safer synthetic user; and restricting validity claims to the scope tested (full checklist in the appendix).

Limitations Our claims are scoped to GSS respondents 2016–2024 and WVS Wave 7 across 63 countries, on the questions listed, with two prompt formats and two seeds, and to demographic prompting under survey-simulation protocols. Richer persona construction (Li et al., 2025), few-shot conditioning, or fine-tuning could shift the individual-fidelity and stereotyping numbers; testing that is the point of a fixed, reusable protocol. The subgroup and decision-impact statistics set a $\approx$100-respondent model sample per question against full-table human values; since η^2, Cramér's V, and segment gaps are biased upward at small n, this is anti-conservative, inflating magnitudes and significance counts (tested against zero, not a same-n null), so they overstate the model–human difference; the right reference is a same-size human subsample. Survey weights, pretraining contamination, and the gap between stated opinions and behavior are covered in the appendix.

6 CONCLUSION

Across two independent domains and two model families, and under the demographic-prompting protocols we test, LLM synthetic users show two failures: at the individual level every model places less probability on the true human answer than non-LLM baselines under proper scoring (and on cross-cultural values falls well below them in accuracy), and the models over-determine demographics, most clearly for the Claude models and Llama-70B, although the small model sample overstates its size. The individual-level deficit also holds under distance-aware scoring, the over-determination also appears under a coding-invariant measure, and in neither within-family size comparison does the larger model fix either failure. Beyond the negative result, we contribute a reusable evaluation framework (the cross-domain benchmark, the non-LLM baseline suite, the stereotyping index with its coding-invariant companion, and the validation protocol) with which a team can decide, before deployment, whether an LLM synthetic user is fit for purpose or whether, given labeled human data, a trivial demographic predictor would serve at least as well.

REPRODUCIBILITY STATEMENT

Model identities, decoding settings, and seeds are given in Appendix B; every question, answer scale, and coding in Appendix C; the prompt templates in Appendix D; parser rules and invalid-output handling in Appendix E; baseline construction and back-off in Appendix F; and metric definitions and estimation choices in Appendix G, with the robustness checks in Appendix H. Both human datasets are public (GSS cumulative file; WVS Wave 7, v6.0), and the construction of the LLM evaluation sample is described in Section 3.1. Data-build, inference, and analysis scripts will be fully released after the review process.

AI USE STATEMENT

Generative AI tools assisted with manuscript preparation, including LaTeX formatting, adaptation to the ICLR submission template, and consistency checks between manuscript versions. The use of LLMs to generate synthetic survey responses is described in the experimental protocol and Appendix B.

REFERENCES

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

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.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623, 2021.

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.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

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

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pp. 77–91. PMLR, 2018.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

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

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. Toxicity in ChatGPT: Analyzing persona-assigned language models. In *Findings of the association for computational linguistics: EMNLP 2023*, pp. 1236–1270, 2023.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. Can AI language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600, 2023.

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

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.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64 (12):86–92, 2021.

Robert M Groves and Lars Lyberg. Total survey error: Past, present, and future. *Public opinion quarterly*, 74(5):849–879, 2010.

Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. Evaluating large language models in generating synthetic HCI research data: a case study. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pp. 1–19, 2023.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

John J Horton, Apostolos Filippas, and Benjamin S Manning. Large language models as simulated economic agents: What can we learn from Homo Silicus? Technical report, National Bureau of Economic Research, 2023.

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

Peiyao Li, Noah Castelo, Zsolt Katona, and Miklos Sarvary. Frontiers: Determining the validity of large language models for automated perceptual analysis. *Marketing Science*, 43(2):254–266, 2024.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023.

Zhicheng Lin. Six fallacies in substituting large language models for human participants. *Advances in Methods and Practices in Psychological Science*, 8(3):25152459251357566, 2025.

Marlene Lutz, Indira Sen, Georg Ahnert, Elisa Rogers, and Markus Strohmaier. The prompt makes the person(a): A systematic evaluation of sociodemographic persona prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 23212–23237. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.1261.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 220–229, 2019.

Xinyi Mou, Xuanwen Ding, Qi He, Liang Wang, Jingcong Liang, Xinnong Zhang, Libo Sun, Jiayu Lin, Jie Zhou, Xuanjing Huang, and Zhongyu Wei. From individual to society: A survey on social simulation driven by large language model-based agents. *ACM Computing Surveys*, 58(11):1–41, 2026.

Robert M O'Keefe and Daniel E O'Leary. Expert system verification and validation: a survey and tutorial. *Artificial intelligence review*, 7(1):3–42, 1993.

Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th annual ACM symposium on user interface software and technology*, pp. 1–18, 2022.

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

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R Bowman. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2086–2105, 2022.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4902–4912, 2020.

Aurora Saibene, Michela Assale, and Marta Giltri. Expert systems: Definitions, advantages and issues in medical field applications. *Expert Systems with Applications*, 177:114900, 2021.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *International conference on machine learning*, pp. 29971–30004. PMLR, 2023.

Marko Sarstedt, Susanne J Adler, Lea Rau, and Bernd Schmitt. Using large language models to generate silicon samples in consumer and marketing research: Challenges, opportunities, and guidelines. *Psychology & Marketing*, 41(6):1254–1270, 2024.

Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 59–68, 2019.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 3407–3412, 2019.

Galit Shmueli and Otto R Koppius. Predictive analytics in information systems research. *MIS quarterly*, 35(3):553–572, 2011.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on machine learning research*, 2023.

Jia Wang, Ziyu Zhao, Tingjuntao Ni, and Zhongyu Wei. SocioBench: Modeling human behavior in sociological surveys with large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 26257–26289, 2025.

Wenlong Zhao, Debanjan Mondal, Niket Tandon, Danica Dillion, Kurt Gray, and Yuling Gu. WorldValuesBench: A large-scale benchmark dataset for multi-cultural value awareness of language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 17696–17706, 2024.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. Can large language models transform computational social science? *Computational Linguistics*, 50(1): 237–291, 2024.

A Extended related work

Our study sits at the intersection of six literatures. We review each in turn, then state the gap that motivates a baseline-anchored, cross-domain validation framework.

A.1 LLMs as simulated survey respondents and participants

A fast-growing line of work treats LLMs as programmable populations, prompting them with personas or demographics to predict opinions and behaviors. The optimistic anchor is Argyle et al. (2023), who show that demographically conditioned prompts can reproduce aggregate patterns in U.S. political survey data, launching the "silicon sample" idea; related work extends the substitution to behavioral experiments (Aher et al., 2023), economic agents (Horton et al., 2023), and the generation of synthetic research data in HCI and product settings (Hämäläinen et al., 2023). A skeptical counter-current quickly followed. Bisbee et al. (2024) show that LLM-generated survey responses can misrepresent population and subgroup variation; Santurkar et al. (2023) find that model "opinions" align more with some demographic groups than others rather than neutrally representing a population; and Dominguez-Olmedo et al. (2024) show that model survey answers are sensitive to ordering and labeling, and so should not be read as direct measurements of population beliefs. Position pieces in cognitive science and psychology sharpen the caution: replacing human participants risks confusing linguistic plausibility with validity (Dillion et al., 2023; Gao et al., 2025; Lin, 2025). The practical problem for this paper is precisely this gap between survey-shaped output and valid evidence.

A.2 Survey-simulation benchmarks and cross-cultural value modeling

A parallel effort builds benchmarks that make survey and cross-cultural value simulation measurable at scale. WorldValuesBench derives a large demographic-context-to-answer benchmark from the World Values Survey (Zhao et al., 2024); SocioBench evaluates LLMs as survey-response predictors across countries and sociological domains (Wang et al., 2025); and distribution-specialization work fine-tunes models to match country-level response distributions, improving group-level prediction while still struggling to generalize (Cao et al., 2025). One of our two domains reuses the WorldValuesBench question set and scales for comparability. Our work differs from these benchmarks on three axes that matter for a validity claim. First, *reference point*: WorldValuesBench reports only in-sample oracle reference points (a uniform and a per-group majority answer) on group-level distributions, and SocioBench only a random-guess floor; neither compares against a demographic predictor fit on held-out human data at the individual level, so a reported accuracy cannot be read as evidence that the simulation adds information; our demographic-lookup and learned baselines supply that reference, and against it the individual-level advantage disappears. Second, *level of analysis*: these benchmarks (and the distribution-specialization work of Cao et al. (2025), which improves group-level fit through fine-tuning) target aggregate or country-level distributions, which our RQ2 confirms models reproduce reasonably; we show that aggregate success coexists with individual-level failure and with a subgroup distortion the aggregate metric cannot see. Third, *cross-domain transfer*: each benchmark draws on a single survey program (WVS; ISSP), so a positive or negative result could be an artifact of that program; by holding one protocol fixed across a U.S. attitude survey and a 63-country value survey we can separate properties of LLM simulation from properties of a dataset. In short, we complement calibrated value and persona benchmarks by asking whether simulated users add individual-level information beyond a demographic baseline, whether the demographic structure of answers is faithful, and whether the answer transfers across domains.

A.3 Persona prompting, silicon samples, and prompt artifacts

Persona and demographic prompting is the dominant mechanism for constructing synthetic users, and it is used well beyond political surveys: in marketing and consumer research, where "silicon samples" are proposed as low-cost respondent pools (Sarstedt et al., 2024), and in automated perceptual analysis for product decisions (Li et al., 2024). Crucially, persona assignment is not a neutral conditioning operation. It can amplify toxicity and stereotypes (Deshpande et al., 2023), the surface architecture of a prompt can induce methodological artifacts (Brucks & Toubia, 2025), sociode-

mographic persona formulations materially change model behavior (Lutz et al., 2025), and persona generation itself carries quality and validity pitfalls (Li et al., 2025). This literature motivates treating prompt-surface stability and demographic over-determination (our RQ3–RQ4) as first-order validity conditions rather than minor robustness checks.

A.4 Social simulation and LLM agents

Synthetic survey respondents are one instance of a broader movement to simulate humans with LLMs. Generative agents endow LLMs with memory, reflection, and planning to produce believable individual and social behavior (Park et al., 2023), building on earlier populated prototypes for social computing systems (Park et al., 2022); recent surveys map the field from individual agents to whole-society simulation (Mou et al., 2026) and assess the promise and epistemic risks of LLMs for computational social science (Ziems et al., 2024). This work mostly evaluates believability or downstream system behavior; we contribute a quantitative, human-anchored validity benchmark for one high-stakes class of these simulations.

A.5 Algorithmic stereotyping, fairness, and subgroup validity

Our over-determination finding connects to a long line of work on social bias in learned representations: human-like biases in embeddings (Caliskan et al., 2017; Bolukbasi et al., 2016), stereotyped associations in generation (Sheng et al., 2019), and representational harms of large models (Bender et al., 2021). The fairness literature further establishes that aggregate performance can mask subgroup harm: intersectional accuracy disparities (Buolamwini & Gebru, 2018), group-conditioned error criteria (Hardt et al., 2016), hand-built bias benchmarks for constrained-choice QA (Parrish et al., 2022), and the reminder that fairness must be assessed within the sociotechnical context of use (Selbst et al., 2019). Our stereotyping index adapts this subgroup lens but differs in target: rather than asking whether a model attaches a negative label to a group, we compare *how predictive* demographics are of an answer in the model versus in real humans, turning stereotyping into a human-comparable validity metric.

A.6 Predictive validity and evaluation practice

Synthetic users are evidence-producing components inside decision pipelines, and such components require verification and validation against intended use. Information-systems research treats predictive validity (out-of-sample performance against baselines) as a first-class evaluation goal distinct from explanatory fit (Shmueli & Koppius, 2011); the decision-support and expert-systems tradition has long emphasized verification, validation, and evaluation of intelligent artifacts (O'Keefe & O'Leary, 1993; Saibene et al., 2021). Multi-metric reporting norms from ML evaluation reinforce the same discipline: holistic evaluation across accuracy, calibration, robustness, and fairness (Liang et al., 2023), behavioral testing beyond aggregate accuracy (Ribeiro et al., 2020), broad capability benchmarks (Srivastava et al., 2023), and standardized documentation of models and datasets (Mitchell et al., 2019; Gebru et al., 2021). Finally, survey methodology's total-survey-error framework distinguishes representation from measurement error (Groves & Lyberg, 2010), which motivates treating individual, aggregate, and subgroup validity as separate questions: an LLM synthetic user introduces a new error source that can look benign in aggregate while failing at the individual or subgroup level.

A.7 What is missing

This prior work is strong but fragmented: each study typically fixes one domain, one or two models, and one evaluation lens, which makes it hard to tell whether a reported failure (or success) is a property of LLM simulation or of the specific setup. Two gaps matter for practice. First, individual-level results are usually reported without *non-LLM baselines*, so it is unclear whether the model beats a demographic lookup, let alone a learned demographic model (Shmueli & Koppius, 2011), and whether any apparent gap is merely an artifact of harsh exact-match scoring on ordinal scales. Second, subgroup analyses typically ask whether groups are "flattened," but do not quantify, on a human-comparable scale, how much demographics *should* explain versus how much the model makes them explain. We close both gaps by anchoring individual fidelity to a baseline suite scored

with distance-aware and proper-scoring metrics, and by quantifying subgroup determinism with both a variance-based and a coding-invariant association measure. Holding one protocol fixed across two domains and four models, we then test whether the conclusions transfer.

B MODEL CALL SETTINGS

All four models (Claude Haiku 4.5, Claude Sonnet 4.6, Llama-3.1-8B-Instruct, and Llama-3.3-70B-Instruct) are called through a single common chat-completion interface. Decoding uses temperature 1.0 (the model's default sampling; where a model rejects an explicit temperature field it is omitted and the provider default applies) and a token cap of 100. Each (model, prompt format, respondent, question) cell is generated under two independent runs ("seeds" 0 and 1); these differ only in sampling randomness and are used to estimate the run-to-run noise floor in RQ4.

C EXACT QUESTIONS AND ANSWER SCALES

Tables 4 and 5 give every question verbatim, with its answer options and integer coding, whether it is treated as ordinal or nominal, and which metrics it enters. GSS wordings and value labels are the standard GSS codebook text; WVS wordings and ordinal scales follow WorldValuesBench for comparability. An item is ordinal when a mean answer is meaningful; only ordinal items enter the η^2 stereotyping index and the distance-aware metrics (MAE/EMD), while all items enter accuracy, JS divergence, and the coding-invariant Cramér's V (V). GSS prompt demographics are age, sex, race, highest degree, region, political views, and party identification (subgroup axes entering the subgroup metrics: race, degree, region, political views); WVS prompt demographics are age group, sex, education level, settlement type, and country (subgroup axes entering the reported subgroup metrics: education and age group; country cells do not meet the minimum-cell filter in the per-question evaluation sample). In both domains sex has two levels, fewer than the three an axis needs to enter the subgroup metrics, so it is a prompt demographic only.

Table 4: GSS questions: wording, answer options (code=label), type, and metrics entered. "all" = acc, JS, MAE/EMD, log-loss/Brier, η^2, V.

Var	Question wording	Options (code=label)	Type	Metrics
`happy`	Taken all together, how would you say things are these days – would you say that you are very happy, pretty happy, or not too happy?	1=very happy; 2=pretty happy; 3=not too happy	ordinal	all
`trust`	Generally speaking, would you say that most people can be trusted or that you can't be too careful in dealing with people?	1=most people can be trusted; 2=can't be too careful; 3=depends	nominal	acc, JS, V
`fair`	Do you think most people would try to take advantage of you if they got a chance, or would they try to be fair?	1=would take advantage of you; 2=would try to be fair; 3=depends	nominal	acc, JS, V
`helpful`	Would you say that most of the time people try to be helpful, or that they are mostly just looking out for themselves?	1=try to be helpful; 2=looking out for themselves; 3=depends	nominal	acc, JS, V
`cappun`	Do you favor or oppose the death penalty for persons convicted of murder?	1=favor; 2=oppose	nominal	acc, JS, V
`grass`	Do you think the use of marijuana should be made legal or not?	1=should be legal; 2=should not be legal	nominal	acc, JS, V

Table 4 continued

Var	Question wording	Options (code=label)	Type	Metrics
abany	Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion if the woman wants it for any reason.	1=yes; 2=no	nominal	acc, JS, V
gunlaw	Would you favor or oppose a law which would require a person to obtain a police permit before he or she could buy a gun?	1=favor; 2=oppose	nominal	acc, JS, V
fefam	Do you agree or disagree with this statement: It is much better for everyone involved if the man is the achiever outside the home and the woman takes care of the home and family?	1=strongly agree; 2=agree; 3=disagree; 4=strongly disagree	ordinal	all
confinan	How much confidence do you have in the people running banks and financial institutions – a great deal, only some, or hardly any?	1=a great deal; 2=only some; 3=hardly any	ordinal	all

Table 5: WVS questions (WorldValuesBench probe set): wording, scale, type, and metrics entered. All are ordinal integer scales.

Var	Question wording	Scale	Type	Metrics
Q48	On a scale of 1 to 10, 1 meaning 'None at all' and 10 meaning 'A great deal', how much freedom of choice and control over your life do you feel you have?	1–10	ord.	all
Q106	On a scale of 1 to 10, 1 meaning 'Incomes should be made more equal' and 10 meaning 'There should be greater incentives for individual effort', where would you place your view?	1–10	ord.	all
Q107	On a scale of 1 to 10, 1 meaning 'Private ownership of business should be increased' and 10 meaning 'Government ownership of business should be increased', where would you place your view?	1–10	ord.	all
Q108	On a scale of 1 to 10, 1 meaning 'The government should take more responsibility to ensure that everyone is provided for' and 10 meaning 'People should take more responsibility to provide for themselves', where would you place your view?	1–10	ord.	all
Q112	On a scale of 1 to 10, 1 meaning 'No corruption at all' and 10 meaning 'Abundant corruption', how much corruption do you think there is in your country?	1–10	ord.	all
Q113	On a scale of 1 to 4, 1 meaning 'None of them' and 4 meaning 'All of them', how many state authorities do you think are involved in corruption?	1–4	ord.	all
Q114	On a scale of 1 to 4, 1 meaning 'None of them' and 4 meaning 'All of them', how many business executives do you think are involved in corruption?	1–4	ord.	all
Q121	On a scale of 1 to 5, 1 meaning 'Very bad' and 5 meaning 'Very good', what impact do you think immigrants have on the development of your country?	1–5	ord.	all
Q122	On a scale of 0 to 2, 0 meaning 'Disagree' and 2 meaning 'Agree', do you agree that immigration fills useful jobs in the labour market?	0–2	ord.	all
Q123	On a scale of 0 to 2, 0 meaning 'Disagree' and 2 meaning 'Agree', do you agree that immigration increases the crime rate?	0–2	ord.	all

Table 5 continued

Var	Question wording	Scale	Type	Metrics
Q158	On a scale of 1 to 10, 1 meaning 'Completely disagree' and 10 meaning 'Completely agree', how much do you agree that science and technology make our lives healthier, easier, and more comfortable?	1–10	ord.	all
Q159	On a scale of 1 to 10, 1 meaning 'Completely disagree' and 10 meaning 'Completely agree', how much do you agree that because of science and technology there will be more opportunities for the next generation?	1–10	ord.	all
Q160	On a scale of 1 to 10, 1 meaning 'Completely disagree' and 10 meaning 'Completely agree', how much do you agree that we depend too much on science and not enough on faith?	1–10	ord.	all
Q164	On a scale of 1 to 10, 1 meaning 'Not at all important' and 10 meaning 'Very important', how important is God in your life?	1–10	ord.	all
Q177	On a scale of 1 to 10, 1 meaning 'Never justifiable' and 10 meaning 'Always justifiable', how justifiable do you think it is to claim government benefits to which you are not entitled?	1–10	ord.	all
Q178	On a scale of 1 to 10, 1 meaning 'Never justifiable' and 10 meaning 'Always justifiable', how justifiable do you think it is to avoid paying a fare on public transport?	1–10	ord.	all

Here "all" denotes accuracy, JS, MAE/EMD, log-loss/Brier, η^2, and V, since every WVS item is ordinal.

D PROMPT TEMPLATES AND DEMOGRAPHIC FORMATTING

Demographics are rendered as a bulleted profile, one line per attribute, using human-readable value labels (e.g. "`- Political views (liberal-conservative): conservative`"); for the persona perturbation (RQ4) the same attributes are rewritten as a single natural-language sentence ("*a person with ...*"). The two prompt formats share this demographic block and differ only in the response instruction:

Single-answer prompt (Style A).

```
You are simulating a single {respondent noun} with the
following demographic profile:
{demographics}
Based only on this profile, predict how this specific person
would answer the following survey question.
Question:
{question text}
Answer options:
{options}
Respond with only the single {option letter | number} (e.g.
A). No explanation.
```

Distribution prompt (Style C).

```
Consider a {respondent noun} with the following demographic
profile:
{demographics}
Estimate the probability that this respondent would choose
each answer option for the question below.  Probabilities
must sum to 1.
Question:
{question text}
Answer options:
{options}
```

```
Return only valid JSON mapping each answer {key} to a
probability, using exactly these keys: {keys}. No other
text.
```

Answer options use short letters (A, B, . . .) when the options have text labels (GSS), and the scale number itself when options are numeric points on an ordinal scale (WVS), which avoids the eight-letter limit on the longest scales. In every case the prompt carries an explicit key→answer-code mapping that the parser uses to recover the underlying code.

E Parser rules and invalid-output handling

Outputs are parsed deterministically. For the single-answer prompt with letter keys, the parser takes the first standalone capital letter that is a valid option key; with numeric keys it takes the first integer token that is a valid key, matching multi-digit tokens so that "10" is not read as "1". For the distribution prompt, the parser extracts the first balanced JSON object, reads a probability for each declared key (absent keys default to 0), and normalizes to sum to one; any response that is empty, contains no valid key, fails JSON parsing, or yields a non-positive total is flagged *invalid/refused*. Invalid outputs are excluded from the accuracy and distribution metrics (a conditional-on-valid accounting) and counted in the invalid-rate stability metric (RQ4). Because invalids may not be missing at random, Section 4.5 also reports the opposite bound (each invalid counted as a failed prediction), and the two accountings coincide wherever the invalid rate is near zero. Prose-wrapped answers are tolerated as long as a valid key is recoverable.

F Baseline construction and back-off

All baselines are fit on the 50% "fit" fold (respondents with an even hash of their id) and evaluated on the evaluation-sample rows, which are drawn from the "eval" fold, so no respondent is used for both. The *question-marginal* baseline predicts the single most frequent fit-fold answer for the question. The *demographic-lookup* baseline estimates a conditional answer distribution for each demographic cell and predicts its mode, backing off when a cell has fewer than 20 fit-fold respondents: for GSS the back-off is {degree, race, sex, political views} → {degree, race} → question marginal; for WVS it is {country, education, age group} → {country} → question marginal, and we report the {country} level separately as a country-only baseline. The *logistic* baseline is a multinomial logistic regression on the one-hot-encoded prompt demographics, fit per question (≤ 2000 iterations); the *random-forest* baseline uses 300 trees on the same features. The lookup and logistic baselines also emit a full predicted distribution, used as the reference for the log-loss and Brier comparisons.

G Metric details and estimation choices

This appendix gives the full metric definitions summarized in Section 3.4 and the four estimation choices whose checks are recorded in Appendix H.

We use four families of metric, one per research question. Reporting individual, aggregate, and subgroup fidelity separately mirrors the total-survey-error distinction between measurement and representation error (Groves & Lyberg, 2010) and the multi-metric evaluation norms now standard for language models (Liang et al., 2023; Ribeiro et al., 2020): a single aggregate number hides exactly the failures we care about. Each metric is stated in plain terms first, then defined precisely. Every headline comparison against a baseline carries a *paired-bootstrap 95% confidence interval*: we resample the evaluation-sample rows 500 times, scoring model and baseline on the *same* rows in each resample, and report the interval of the difference, so the reader can tell a real margin from sampling noise.

Individual fidelity (RQ1): can the model guess one person's answer? We report *accuracy* (the fraction of individuals whose exact answer the model predicts correctly) next to each baseline's accuracy on the same people. For the distribution prompt (Style C), the model's single "guess" is the option it assigned the highest probability (the arg-max of its distribution). Exact-match accuracy is deliberately harsh on long ordinal scales, so for the ordinal questions we also report two *distance-aware* metrics that give partial credit for being close. The *mean absolute error* (MAE) is the average

gap, in scale points, between the single predicted answer and the true one. The *earth-mover's distance* (EMD) uses the whole predicted distribution from the distribution prompt: it is the expected number of scale steps needed to move the predicted probability mass onto the true answer, and so rewards a distribution that concentrates near the truth even when its mode is wrong. Both are reported for the model and for the baseline on the same rows.

Aggregate fidelity (RQ2): does the model reproduce the group's answer spread? Rather than any one person, we compare the overall *share* of people choosing each option (the answer distribution) between the model and real humans. We summarize the gap between two distributions with the *Jensen–Shannon (JS) divergence*, a standard 0-to-1 measure of how different two probability distributions are; 0 means identical, and larger means further apart, so lower is better. Because we sample by cell rather than at random (Section 3.1), the raw pooled model distribution reflects our sampling design, not the human population. We therefore also report a *population-reweighted* JS, in which each cell's model distribution is weighted back to that cell's share of the real population before comparing to the true (full-table) human distribution. If the two versions agree, the aggregate-fidelity result is not an artifact of the sampling design.

For the distribution prompt (Style C), arg-max accuracy discards most of the predicted distribution, so we additionally score the *whole* distribution with two proper scoring rules evaluated at the true human answer: the *log-loss* (negative log-likelihood the model places on the true answer) and the *Brier score* (squared error of the predicted distribution against the one-hot true answer). Lower is better, and both are reported against the baseline's conditional distribution on the same rows; a proper score is the right way to ask whether the distribution puts mass where the human actually answered.

Subgroup determinism (RQ3): does the model treat identity as destiny? Fairness research has long argued that aggregate performance can hide subgroup harm and must be assessed group by group (Buolamwini & Gebru, 2018; Hardt et al., 2016). We adopt that lens but ask a sharper, human-anchored question: how strongly a demographic attribute (say, political leaning) *predicts* the answer, for real people and for the model. The natural measure is the share of the total variation in answers that is explained by which group a person belongs to. Statisticians call this η^2 ("eta-squared"), the ratio of the variation *between* groups to the total variation (between plus within groups); it runs from 0 (the attribute tells you nothing about the answer) to 1 (the attribute determines the answer completely). We compute η^2 the same way for humans and for the model. Our headline measure, the *stereotyping index*, is simply the difference,

$$\Delta\eta^2 \;=\; \eta^2_{\text{model}} - \eta^2_{\text{human}},$$

which is positive when the model treats a demographic attribute as *more* predictive of the answer than it actually is among real people, that is, when the model stereotypes. Because it is a difference of two bounded quantities computed identically, it is not distorted by which prompt style produced the model's answers, and it captures the full spread within each group rather than only the group averages. We attach a 95% confidence interval to each value by resampling respondents 500 times (a bootstrap), so we can tell whether the index differs from zero (a lenient test at this sample size; Limitations).

η^2 treats the answer codes as numbers, which is natural for the ordinal questions but would be an arbitrary choice for nominal ones. To guard against the objection that the finding depends on how answers are coded, we add a *coding-invariant* companion measure that uses no numeric codes at all: Cramér's V, the standard association strength between two categorical variables (here, group membership and the chosen answer), computed identically for humans and for the model from the group×answer contingency table. Its stereotyping analogue is $\Delta V = V_{\text{model}} - V_{\text{human}}$, again with a bootstrap 95% CI. Because it is purely nominal, we can report it for *all* questions, including the nominal GSS items that the η^2 index omits. If both the numeric $\Delta\eta^2$ and the coding-free ΔV are positive, over-determination is not an artifact of the answer coding.

Decision impact: what would over-determination imply for a decision that acts on it? The stereotyping index measures the distortion in the model's world; a practitioner needs to know what that distortion does to a decision. We therefore extend it to the terms of the single most common use of a synthetic user in practice: *segment targeting*. A product, marketing, or policy team asks which demographic segment is most extreme on an attitude, and how large the between-segment gap is, in order to target the top segment, size a niche, or claim "this attitude splits along ." For

each (question, demographic axis) we compute, identically for the model and for real humans, the *segment share* p_g (the fraction of segment g in the upper half of the answer scale: answer codes at or above the midpoint of the question's code range, which for a two-option item is the larger code), and read three quantities a targeting decision keys on: the *between-segment gap* $\max_g p_g - \min_g p_g$; the *target segment* $\arg\max_g p_g$; and, when the model's target differs from the human target, the *targeting cost* (the attitude a team forgoes, in real human share, by acting on the model's segment instead of the true one). From these we report (i) the *gap-inflation factor* $\text{gap}_{\text{model}}/\text{gap}_{\text{human}}$, how much larger the gap is when read off the model; (ii) the *wrong-target rate*, the fraction of (question, axis) pairs where the model would point a team to a different segment than the human data (when segments tie for the largest model share, the $\arg\max$ takes the first tied segment in data order, so ties are resolved deterministically rather than by a substantive rule); and (iii) the *spurious-split rate*, the fraction where humans show a negligible gap (≤ 0.10) but the model shows a large one (≥ 0.25), a segment split a team would "discover" and act on that is negligible in the full human table. Human shares use the full in-scope table; the model gap carries a 95% bootstrap CI over respondents.

Stability (RQ4): do the conclusions survive irrelevant changes? We report the rate of invalid or refused outputs (answers the parser cannot read as a valid choice), agreement between two decoding runs, and how results move across model family, model size, and prompt wording.

Four choices in the estimation deserve explicit justification, since each is a place a reader might worry the headline effects are an artifact of the analysis rather than of the models. We resolve each on the data itself and report the checks in Section 4.5 and Appendix H.

Confidence level. We use 95% intervals throughout. The individual-level deficit (RQ1) is large relative to its uncertainty, so moving from a 90% to a stricter 95% interval leaves every RQ1 headline unchanged: on WVS every individual-accuracy margin remains strictly below zero, and on GSS every model remains a statistical tie-or-worse against the baseline. The stereotyping index also remains significant for most question–group pairs (Table 2), but these RQ3 tests are against zero, not against the same-n human expectation, and so inherit the small-sample asymmetry noted in the Limitations.

Multiple comparisons. The stereotyping index is tested over many (question, axis) pairs, so a per-pair interval does not control the false-discovery rate. We therefore also apply a Benjamini–Hochberg FDR correction at $q = 0.05$ across all pairs within a model and prompt style. Correction barely changes the count of significant pairs (e.g. Sonnet single-answer: $32/32 \rightarrow 32/32$ on WVS, $10/12 \rightarrow 10/12$ on GSS; Appendix H), so "significant for the majority of pairs" is not an artifact of uncorrected multiplicity; the null tested, however, is zero rather than the small-sample expectation, so the count still inherits the asymmetry noted in the Limitations.

Respondent clustering. A minority of respondents answer more than one question (7.6% on GSS, 1.8% on WVS), so rows are not fully independent. We recompute the RQ1 margin with a *cluster* bootstrap that resamples respondents rather than rows; the 95% intervals are essentially unchanged and every WVS deficit still excludes zero (Appendix H).

Survey weights. GSS and WVS ship official design/population weights (`wtssps`, `W_WEIGHT`). We estimate human targets unweighted, and justify that choice by recomputing, with and without weights on the full in-scope table, the two human-side quantities the claims rest on: the population answer distribution (the RQ2 target) and the human η^2 (the RQ3 baseline that the stereotyping index subtracts). Weighting moves the human distribution by a mean Jensen–Shannon divergence of ≤ 0.0003 and the human η^2 by at most 0.007, both one to three orders of magnitude smaller than the effects we report (JS gaps of 0.04–0.38; stereotyping indices of $+0.05$ to $+0.7$ as reported, uncorrected for small-sample bias). Within the comparisons checked here, weighting does not account for either failure, so we report the simpler unweighted analysis and document the check in Appendix H.

H ESTIMATION ROBUSTNESS

This appendix records the four estimation checks summarized in Section G; all are computed on the same cached outputs as the main results, with no additional model calls.

Multiple comparisons (FDR). For each (question, axis) pair entering the η^2 stereotyping index we compute a one-sided bootstrap p-value for $H_0 : \Delta\eta^2 \leq 0$ and apply a Benjamini–Hochberg cor-

rection at $q = 0.05$ across all pairs within a model and prompt style. The number of significant, positively-signed pairs is nearly identical before and after correction: for the single-answer prompt, Sonnet is $32/32$ (WVS) and $10/12$ (GSS) both before and after correction; Haiku $31/32 \rightarrow 31/32$ (WVS); Llama-70B $31/32 \rightarrow 31/32$ (WVS) and $7/8 \rightarrow 7/8$ (GSS). Because this test is one-sided, its uncorrected counts can slightly exceed the two-sided interval counts in Table 2. The significance count is therefore not an artifact of uncorrected multiple testing, but the null is zero, not the same-n human expectation, so it still overstates the model–human difference (Limitations).

Respondent-clustered bootstrap. The individual-fidelity sample contains a small share of respondents who answer more than one question (7.6% on GSS, 1.8% on WVS). Recomputing the RQ1 accuracy margin over the baseline with a cluster bootstrap (resampling respondents, then taking all of their rows) leaves the 95% intervals essentially unchanged from the row-level intervals: on WVS every margin still excludes zero (e.g. Sonnet single-answer -0.152, $[-0.180, -0.121]$; Haiku -0.218, $[-0.248, -0.187]$), and on GSS the Llama-8B deficits remain significant while the Claude cells remain ties, exactly as in the row-level analysis. Within-respondent correlation does not drive the result.

Survey-weight robustness. GSS and WVS provide official weights (`wtssps`, coverage 100%; `W_WEIGHT`, coverage 100% on the in-scope pool). Applying them to the two human-side quantities the claims depend on changes them negligibly: the human population answer distribution moves by a mean Jensen–Shannon divergence of 0.0001 (GSS) and < 0.0001 (WVS), with a maximum over questions of 0.0003; the human η^2 that the stereotyping index subtracts moves by a mean absolute 0.0016 (GSS) and 0.0012 (WVS), with a maximum of 0.0065. These are one to three orders of magnitude below the effects we report, so the unweighted analysis is adequate for the population-distribution and subgroup-baseline comparisons these checks cover.

Sample adequacy. Individual-fidelity margins pool all evaluation rows per domain ($n = 1{,}986$ scored (respondent, question) pairs on GSS, $n = 2{,}916$ on WVS for the clean cells) rather than the ≈ 100 respondents of any single question, and every headline comparison carries a paired-bootstrap 95% interval computed on those rows; for RQ1 the intervals, not a nominal per-question count, are what license the claims. This pooling does not extend to the subgroup and decision-impact statistics, which are computed per question on that small model sample against full-table human values; there the sample-size asymmetry overstates the model–human difference (Limitations).

Invalid-output accounting (RQ4). The accuracy numbers elsewhere in the paper are *conditional on valid output*, which flatters the high-invalid cells because invalid responses may not be missing at random. We therefore report both bounds. Where invalid rates are near zero (all Claude and Llama-70B cells), the two coincide, so those results are unaffected. Where they are high, treating each invalid output as a failed prediction, the appropriate accounting when a synthetic user must return a usable answer, widens the deficit sharply: Llama-8B's Style-C accuracy falls from 0.51 to 0.22 on GSS and from 0.18 to 0.03 on WVS, and Haiku's WVS Style-C from 0.27 to 0.16. Because the non-LLM baselines never emit invalid output, penalizing invalids can only widen the LLM's gap to the baseline, never close it; the individual-level conclusion is robust to either accounting.

Prompt-surface perturbations (RQ4). Third, *predictions are unstable to prompt-surface choices that should be irrelevant*. Prior work shows that survey-style LLM answers are sensitive to option ordering and labeling (Dominguez-Olmedo et al., 2024) and that prompt architecture and persona formulation can themselves induce artifacts (Brucks & Toubia, 2025; Lutz et al., 2025); we quantify this for our setting. On the two clean closed models we measured the *prediction-flip rate* (the fraction of single-answer predictions that change) under two perturbations that leave the actual question and answer set unchanged: reversing the display order of the answer options, and reframing the identical demographics as a natural-language persona sentence rather than a bulleted profile. We benchmark these against a decoding-noise floor: the flip rate between two decoding seeds of the *same* prompt. On GSS, merely reversing option order flips 9.8% (Sonnet) to 10.8% (Haiku) of predictions, and the persona reframing flips a further $\sim 9\%$, both roughly 4–$10\times$ the seed-noise floor (1.1% and 2.6% respectively). An option-order spot-check on WVS shows the same effect more strongly (14.3% flips for Sonnet, 22.6% for Haiku). Option order also moves aggregate fidelity (e.g. Haiku GSS accuracy $0.587 \rightarrow 0.606$), so a study's conclusions can depend on an arbitrary presentation choice. Together with the capability and format results, this means synthetic-user validity is not only a model property but a prompt/interface property.

I ADDITIONAL RESULTS

Figure 3 shows raw individual accuracy against the naive demographic baseline (the margins with confidence intervals are in Figure 1).

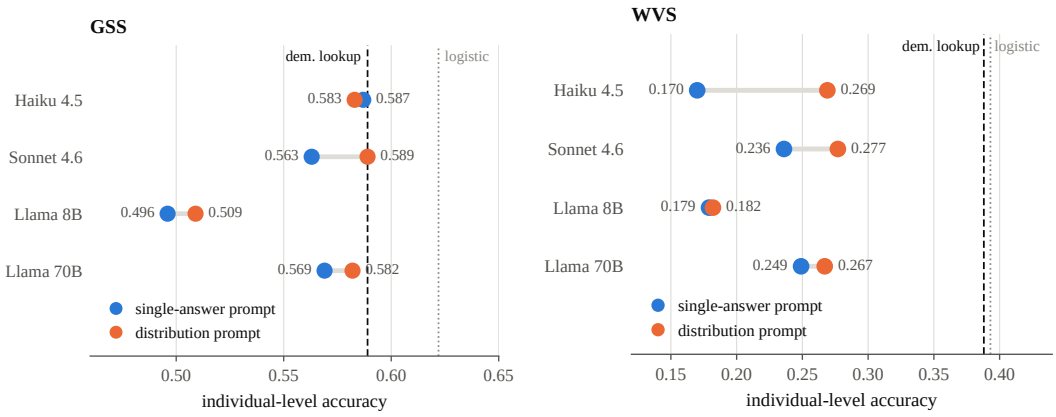


Figure 3: Individual accuracy vs. the naive demographic baseline (dashed line), GSS (left) and WVS (right). No model exceeds the baseline on GSS; every model falls well below it on WVS. The individual-level failure (RQ1) holds in both domains and is more severe for cross-cultural values.

On WVS, grouping by education and age group (16 questions $\times$ 2 axes $=$ 32 pairs; country cells are too sparse in the per-question evaluation sample to pass the minimum-cell filter and do not enter the subgroup metrics), nearly all 32 question–axis combinations have a stereotyping-index interval above zero (a test against zero; Limitations) for the two Claude models and the larger Llama. Figure 4 resolves the stereotyping index to the individual-question level.

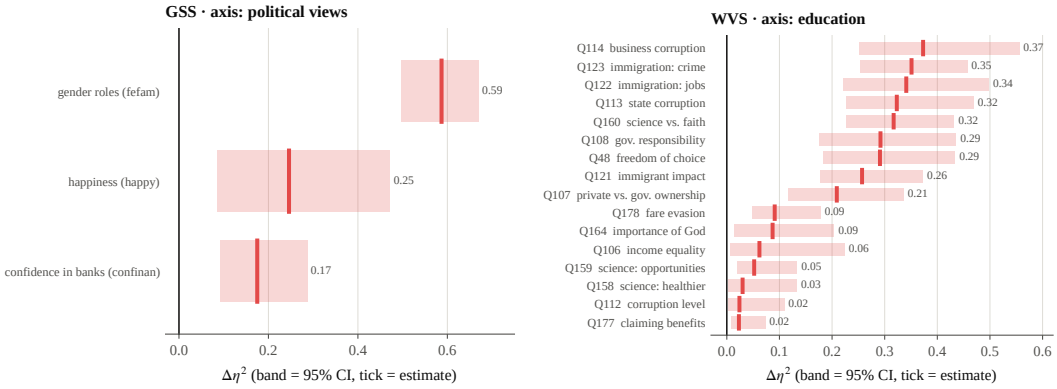


Figure 4: Question-level over-determination for the frontier model (Sonnet 4.6, single-answer prompt) with bootstrap 95% CIs, along the primary grouping axis (GSS: political views; WVS: education). Over-determination is not a model-average artifact: it holds question by question, with confidence intervals above zero (a test against zero) for the strongest items.

Figure 5 plots the decision-impact analysis of Section 4.4.

J EXTENDED DISCUSSION

The two failures discussed in this paper appear to share a cause: the models we test, prompted with demographics, appear to answer by mapping a profile to a stereotyped "typical" response for that profile. This is consistent with evidence that word embeddings (Caliskan et al., 2017) and large language models (Bender et al., 2021) encode social associations and that persona conditioning

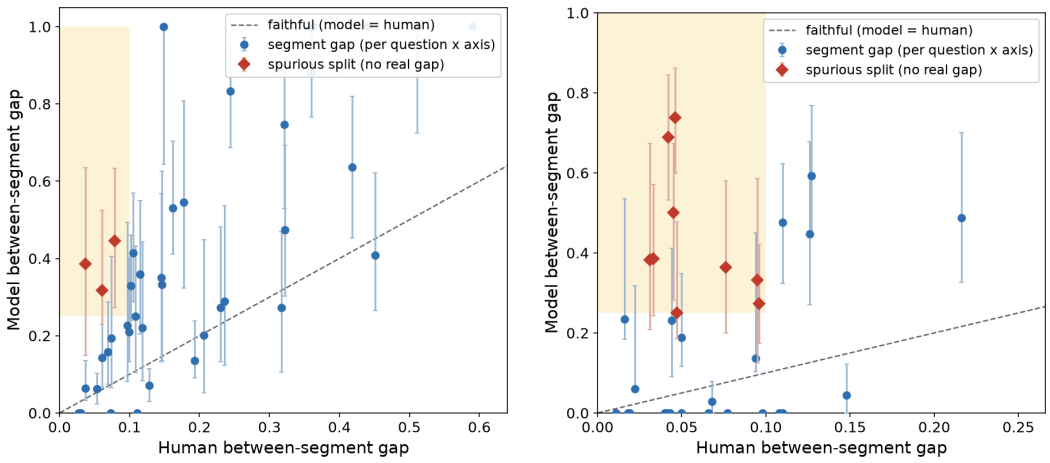


Figure 5: Decision impact of over-determination for the frontier model (Sonnet 4.6, single-answer prompt): the between-segment gap a targeting decision reads off the model (vertical axis, with 95% bootstrap CI) against the true human gap (horizontal axis), one point per (question, demographic axis), GSS (left) and WVS (right). The dashed line is faithful reading (model = human); points above it are gaps larger as read off the model than in the human table. Shaded: the *spurious-split* zone (human gap ≤ 0.10 but model gap ≥ 0.25), the segment differences a team would act on that are negligible in the full human table. Most points lie above the faithful line.

steers models toward stereotyped abstractions rather than nuanced individuals (Deshpande et al., 2023). That mapping is too coarse to beat a demographic lookup at the individual level, and it is worse than the lookup when the answer space is large, as in the 10-point WVS scales. Yet it is also too tight: it treats demographics as more predictive of attitudes than they truly are. The practical corollary is that a synthetic-user study can look successful on an aggregate-distribution metric, even one improved by distribution-specialized training (Cao et al., 2025), while being invalid for any individual- or subgroup-level use, and while exaggerating the demographic structure of opinion. Because the aggregate metric is the one most often reported, this is a live risk for the field.

Implications for decision support These two failures map onto how synthetic users are used to support decisions. A team that only checks aggregate fidelity, the common practice, can be misled in two ways at once. Any decision keyed to *individuals or narrow segments* (personalizing an offer, sizing a niche, or predicting how a specific demographic will react) rests on individual-level fidelity the models do not have; where labeled human data for the target population exist, a demographic lookup would serve at least as well and often better. And any decision that reads the *structure* of opinion off the simulation ("this attitude splits sharply along political lines," "this segment is monolithic") would inherit the models' over-determination, which, as read off the model, makes group differences look larger than they are in real people. Our decision-impact analysis (Section 4.4) describes what this would imply for segment targeting: read off the models, between-segment gaps are larger, the top segment often differs from the human one, and some segment splits are negligible in the full human table. These rates compare a small model sample with full-table human values, so they overstate the model–human difference and are best read as descriptive. If confirmed against a same-size human reference, the danger would be not random noise, which averages out, but *structured* error, which propagates into the decision. The framework we propose is meant to catch exactly this before the evidence is trusted: it turns "did the aggregate look right?" into a set of checks a team can run against held-out human data to decide, for their specific question, whether synthetic-user evidence is safe.

A validation protocol From these results we distill a checklist for researchers considering LLM synthetic users: (1) report individual-level and aggregate-level fidelity *separately*; (2) always benchmark individual fidelity against non-LLM baselines computed on real held-out data (at least a demographic lookup, ideally a learned demographic model) using distance-aware or proper-scoring metrics on ordinal scales; an LLM that does not beat them adds no individual-level information; (3)

report subgroup determinism (e.g. $\Delta\eta^2$, or a coding-invariant association measure), not just group means, to detect stereotyping; (4) before acting on any segment-level read, check the decision-impact quantities (gap inflation, wrong-target rate, and spurious-split rate) against a same-size held-out human subsample, since a distortion invisible in an aggregate metric can still flip a targeting decision; (5) report invalid/refusal rates, especially for distribution prompts and smaller models; (6) do not assume a larger or more capable model is a safer synthetic user; and (7) restrict validity claims to the population, domain, and level of analysis actually tested.

Limitations Our claims are scoped to the data we used: GSS respondents 2016–2024 and WVS Wave 7 across 63 countries, on the specific questions listed, with two prompt formats and two seeds, and to demographic prompting under survey-simulation protocols, not to every elicitation strategy. The LLM-evaluated sample is about 100 respondents per question, but the individual-fidelity claims pool all rows in a domain ($n \approx 2{,}000$ on GSS, $n \approx 2{,}900$ on WVS) and every headline comparison carries a paired-bootstrap 95% interval, so those conclusions rest on those intervals rather than on any single question's cell. The subgroup and decision-impact statistics, by contrast, are computed per question and compare that small model sample with full-table human values; because η^2, Cramér's V, and between-segment gaps are biased upward at small n, they overstate the model–human difference, and the appropriate reference is a same-size human subsample. We do not claim these numbers characterize every model, prompt, or elicitation strategy; richer persona construction (Li et al., 2025), few-shot conditioning, or fine-tuning could shift the individual-fidelity and stereotyping numbers, and testing that is exactly the value of a fixed, reusable protocol. We estimate human targets without survey weights; because applying the official GSS and WVS weights moves those targets by amounts one to three orders of magnitude smaller than the effects we report (Appendix H), this choice does not affect the population-distribution or subgroup-baseline comparisons. Older public surveys such as these may appear in model pretraining data; we mitigate this by scoping GSS to recent waves and by noting that memorization would, if anything, *help* aggregate fidelity while leaving the individual-level and over-determination failures intact. Finally, survey answers are stated opinions, not behavior; extending the protocol to behavioral tasks and to interactive multi-agent simulations (Park et al., 2023; Mou et al., 2026) is future work.