Who Speaks for Whom? LLM-Generated Survey Data as a Proxy for Public Opinion

Radhakrishnan Venkatakrishnan[1], **Travis Brodbeck**[1,2], **Michael D. Young**[1],
[1]University at Albany, [2]Siena University,
Correspondence: rvenkatakrishnan@albany.edu

Abstract

Technological advancements, such as Large Language Models (LLMs), offer a potential solution to the two-faceted problem facing social science researchers: rising costs and declining response rates. The use of artificial personas is a budding practice, where chatbots are given the demographic characteristics of the person they are supposed to role-play as and answer questions for researchers. Before scholars and practitioners augment or replace the data created by interviewing humans, it is essential to understand how well models perform in generating accurate, reliable, and robust data, with concerns that the training of LLMs results in a bias towards the norms of WEIRD cultures. We present a procedure for practitioners to use to evaluate the quality of their synthetic data by measuring Intra Class Correlation (ICC), Earth Mover Distance (EMD), Variance, Hedging, and demographic drivers of LLM output. We find that the models may generate plausible results in the aggregate, but these synthetic data do not exhibit the depth or nuance of human respondents. Secondarily, we find that despite having generated definitive answers on a ten-point scale, the reasoning provided by the LLM exhibited varying degrees of hedging that do not consistently align with the LLM's answer. The distortion of the results was not uniformly distributed; instead, the effects were more extreme for some demographic groups. Our findings suggest that the technology generating synthetic survey data may not be mature enough to address the increasing challenges of interviewing humans for public opinion research. Code and data are available in Github.[1]

1 Introduction

The evolution of LLMs underlying Artificial Intelligence (AI) tools suggests that the technology may be approaching the limits of the Turing Test (Reinbold, 2020; Bhatnagar, 2026), moving from simple imitation to sophisticated impersonation. Researchers face the question of how AI will change social science research, specifically in how it is conducted. Recent scholarship indicates that humans are already struggling to distinguish between human-authored and LLM-generated text (Kreps et al., 2022). This blurring of lines presents a fundamental challenge for social science researchers: if general audiences cannot discern the origin of content - human or AI, researchers may soon face an increasingly difficult task in distinguishing synthetic and authentic data. Before social science researchers embark on the utilization of synthetic or manufactured data, it must be reliably comparable to human responses, and we must understand how those results are generated.

Research budgets are stretched due to declining survey response rates (Eggleston, 2024) attributed to a variety of technological changes, such as the adoption of answering machines (Oldendick and Link, 1994), caller ID (Link and Oldendick, 1999), cell phones (Brick et al., 2007), and call screening conducted by AI agents on modern smartphones (Markus, 2025). Technological change is not the only cause for declining response rates as non-researchers, such as telemarketers, contributed to public's aversion to answering the phone (Link et al., 2006). These factors combined with the public's eroding trust in institutions like pollsters (Johnson et al., 2024) create an environment where conducting the "gold standard" of probabilistic telephone research is both more difficult and more expensive. As researchers moved to online surveys, they faced challenges in data quality between collection modes (Couper and Miller, 2008), similar to the differences in data between self-administered paper interviews and telephone interviews (Vannieuwenhuyze et al., 2010). Against the backdrop of rising costs and the ubiquity of internet access, non-probability panels and opt-in surveys are more commonly used for survey experiments and data

[1]rvenka31/llm-proxy-public-opinion-surveys

Proceedings of the 7th Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), pages 133–148
July 3, 2026 ©2026 Association for Computational Linguistics

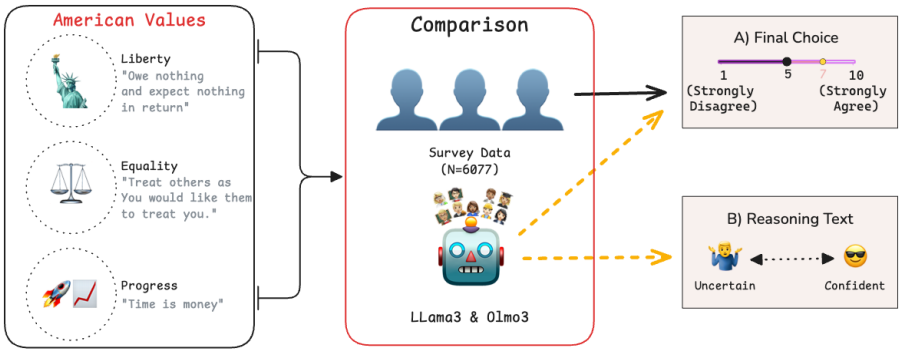

Figure 1: A visual overview of the research framework, illustrating the relationship between demographic inputs and representational fidelity compared on the American Values Survey between human responses and LLM responses. Human responses include only their Final choice, marked on a Likert scale from 1 to 10. LLM responses also include inference-time tokens called reasoning traces.

collection (Callegaro and DiSogra, 2008). Due to the costs of probability sampling designs, non-probability panels are increasingly used, especially when trying to interview hard-to-reach populations, often young people, men, and people of color (Pyrooz et al., 2025).

Just as widespread internet access once made online and non-probability sampling more common, the release of ChatGPT 3.5, along with other similar models, has given researchers the opportunity to leverage new means to answer questions in new ways (Hayashi, 2024). This opportunity puts researchers in a tough position: between the rock of methodological rigor and the hard place of rising costs fueled by declining response rates. Like online surveys, a method for collecting data more cheaply and quickly is tempting. Given this temptation, this paper investigates the viability of using synthetic responses generated by LLMs to measure public opinion.

Synthetic data are manufactured data designed to mimic real-world data by using techniques like deep learning and generative AI (Joshi et al., 2024). It is also referred to as "silicone samples", when tasked to mimic human participants in public opinion surveys (Argyle et al., 2023). Using synthetic data is appealing for several reasons. Unlike humans, LLMs are unlikely to suffer from interview fatigue where interpreting and answering a few dozen questions weighs on one's cognitive ability, potentially biasing future answers in the survey (Ghafourifard, 2024). Related to fatigue, humans satisfice by skipping optional questions, saying they don't know, or providing incomplete answers to finish the interview faster (Krosnick, 1991; Krosnick et al., 2002). LLMs are programmed to complete interviews as instructed, whereas humans may be interrupted, uninterested, skeptical of the prompts, and decide to terminate mid-interview. Humans can be difficult to reach at certain times of day or the week (Weeks et al., 1987), whereas LLMs can be summoned at any time. With the exception of computational costs, LLMs do not ask for financial incentives to participate in research, providing ample opportunity for experimentation. Humans can experience difficulty reading or hearing, whereas LLMs are not restricted by the senses. Humans either decide to participate in an interview out of their own personal motivations, whether that is to help the researcher, to advance research, or to receive an incentive (Hjortskov et al., 2023). LLMs, on the other hand, are programmed to follow instructions and act upon demand.

LLMs exhibit similar problematic behaviors to humans. As humans will lie, LLMs lie or hallucinate the facts (Farquhar et al., 2024). As humans may experience acquiescence bias, being more agreeable to minimize conflict (Davis et al., 2019), LLMs have shown affirmative or positivity bias, saying yes or agreeing with the human prompting the tool (Fanous et al., 2025). LLM outputs are subject to multiple sources of variability: prompt formulation, chain-of-thought instructions, and inference parameters such as temperature, top_p, and top_k — all of which control output randomness and can reduce reliability (Li et al., 2025; Wei et al., 2022). Similarly to noticing social cues and norms, LLMs often gravitate to the mean and can show less variation than would be observed with real data (Xie and Xie, 2025). Like humans having blind spots, the training data for underrepresented groups could bias output and lead to inaccurate syn-

thetic data that could harm research of marginalized groups (Foka et al., 2025; Santurkar et al., 2023). As humans learn throughout the day, obtaining new information, LLMs gain new information to update their foundational model and training data, or through fine-tuning and prompting, creating possible knowledge gaps or incomplete datasets that result in less accurate output.

Incorporating synthetic samples creates many opportunities for cost savings and experimentation, but also introduces significant risks regarding data quality and representativeness. This paper seeks to navigate this "new world" of survey methodology, specifically exploring the rise of using LLMs to create synthetic or artificial data to measure public opinion. Looking to elicit values and morals embedded in an LLM to compare against human respondents is an evolving research direction (Pistilli et al., 2024; Jiang et al., 2024). The World Values Survey (Haerpfer et al., 2022) collects responses from human participants worldwide and has been adapted for LLM evaluation in works such as (Zhao et al., 2024), which evaluated LLMs on the implicit and explicit values they express across different test settings. The OpinionQA dataset captures misalignment in steering LLMs with given persona on ATP questionnaires. (Santurkar et al., 2023) While there is growing evidence that LLMs exhibit WEIRD (Western, Educated, Industrialized, Rich, and Democratic) alignment (Zhou et al., 2025), very few works (Santurkar et al., 2023) have looked into this alignment. Yet the focus of these works was broad and not solely on American values, despite their centrality to the Western dimension of WEIRD. Since LLMs tools offer the potential for richer, conversational data collection, they require a rigorous framework for measurement and evaluation. For synthetic data to be viable, we must be able to: **(1) Quantify Model Behavior:** Develop metrics to measure the tendency of models to provide cautious, non-committal answers—and other stylistic artifacts. **(2) Define Appropriateness:** Establish benchmarks for when synthetic data is an acceptable proxy for human opinion and when it introduces unacceptable error. **(3) Assess Output:** Apply established metrics to ensure the validity of social science research without compromising quality.

Ultimately, while LLMs offer innovative paths for pre-testing and imputing missing data, their role in representing the collective "voice" of the public must be scrutinized. We must determine if we are accurately measuring opinion or simply reflecting a distorted version of the past. Therefore, before synthetic data can be considered a valid proxy, we must develop strict frameworks to measure its failures, focusing on its demographic stereotyping and artificial confidence. To that end, this research critically evaluates model behavior through the following questions

RQ1: To what extent do LLM responses align with human respondents across demographic groups?

RQ2: To what extent do demographic variables influence LLM responses relative to human responses?

RQ3: How does model selection, scale, prompting, and reasoning influence LLM performance?

We will analyze LLM responses across different demographic inputs to identify patterns of stereotyping and misrepresentation, establishing when synthetic data serves as an acceptable proxy for human opinion and when it does not. By systematically mapping failure points, we intend to help with the adoption decision of these tools in survey-based social science research.

2 Methodology

2.1 Data

To answer these questions, we simulate synthetic responses for individual human profiles based on the actual survey. Our source dataset is the American Values Survey ($N = 6,077$), comprising 34 ten-point Likert-scale value statements aggregated into three subscales: *Liberty*, *Equality*, and *Progress* (Gibson and Lipinski, 2021). The survey was conducted in 2021 by Sienna University [2] and it is detailed in Appendix §A.1. Among the demographic information captured, we select 9 demographic variables (Age, Race, Ethnicity, Gender, Employment, Education, Political Affiliation, State, and Voter Registration) for our experiment.

2.2 Models

We evaluate four open-access instruction-tuned models: Llama-3.1-8B-Instruct, Llama-3.3-70B-Instruct (Grattafiori et al., 2024), Olmo-3-7B-Instruct, and Olmo-3.1-32B-Instruct (Olmo et al., 2025). They were selected to explain two effects: parameter count (8B/7B vs 32B/70B) to test if size increases demographic sensitivity, and model

[2] https://sri.siena.edu/the-american-values-study/

family (Llama vs Olmo) to assess the generalization within open-access US-based models across architectures. For each human respondent h_i in the dataset, we construct a persona-based system prompt $P(h_i)$ that includes all 9 demographic attributes and generate $k = 5$ responses using a temperature of $\tau = 0.1$, which introduces minimal stochastic variation. Detailed prompt and hyperparameter settings are provided in the Appendix §A.2.

2.3 Prompts

We experiment with two prompt variants to assess the impact of "chain-of-thought" (CoT), which prompts the model to think step-by-step (Wei et al., 2022; Kojima et al., 2023).

Final Choice First (FCF): Prompting the model to provide a numerical score for each value question before articulating its reasoning trace.

Reasoning First (RF): Prompting the model to first generate a detailed reasoning trace before arriving at a numerical score for the value question.

This design allows us to quantify the effects of CoT and its order on the model's behavior. The language of the two prompts is identical, except for the order of the reasoning and final choice. The complete prompt is provided in the appendix Table 4.

2.4 Evaluation

We define three complementary metrics to quantify the model's performance in survey simulation.

1. **Behavioral Consistency:** Using Variance σ^2 and Intra Class Correlation(ICC), we capture the consistency of the model's score across iterations and for the same demographic profile. We use ICC(1,1) to assess the reliability of individual iterations and ICC(1,k) to assess consistency across the full set (Shrout and Fleiss, 1979).

2. **Alignment Quality:** We use a combination of *Wasserstein Distance/Earth Mover Distance (EMD)* and *Variance ratio* to assess how LLM scores align with human respondents and whether they vary in a similar vein to the human respondents for a given question. Using EMD, we measure the model's simulated score distribution against human ground truth for value constructs, thereby capturing representational distortion, as in (Zhao et al., 2024). Variance ratio, unlike iteration variance (captured by ICC), assesses whether the LLM differentiates between demographic groups to the extent that human respondents naturally differ on a value question. Together, these two measures characterize each question along two dimensions: how far the LLM deviates from human responses (EMD) and whether it responds to demographic variation proportionally (variance ratio).

3. **Hedging% ($H\%$):** Quantifies the model's epistemic uncertainty by analyzing the frequency of hedging language in the reasoning traces for the simulated scores. We quantify the model's uncertainty by analyzing the reasoning trace. Using a verified lexicon of hedging markers L_{hedge} (e.g., "assume", "even though", "appears to") from (Islam et al., 2020), we calculate the Hedging% $H\%$ as the proportion of words that are hedging words across the k iterations. There are 657 unique hedging markers grouped into three categories: hedge words, booster words, and hedging phrases. For this analysis, we group the three categories into a single category, i.e., hedging. Using Spacy's tokenization, we identify hedging in the reasoning trace by grouping words and phrases.

Together with the three dimensions, we can assess the model's behavioral rigidity, representational distortion, and epistemic amplification or suppression across different demographic profiles. The first two dimensions (consistency and alignment) capture the model's rigidity, ensuring that LLMs are sufficiently replicative, and explain representational distortion by focusing on the score distributions and their divergence from human responses. The third metric (Hedging%) captures the model's reasoning uncertainty during generation, which may reflect its confidence in the assigned scores and its awareness of potential biases or limitations in its knowledge.

2.5 Regression

Using metrics derived from the evaluations and additional targets, we specify LLM score, Absolute Error (|LLM score - human score|), Signed Error (LLM score - human score), and Hedging% as dependent variables in the Ordinary Least Squares (OLS) regression analysis to identify the demographic drivers of the above dependent variables.

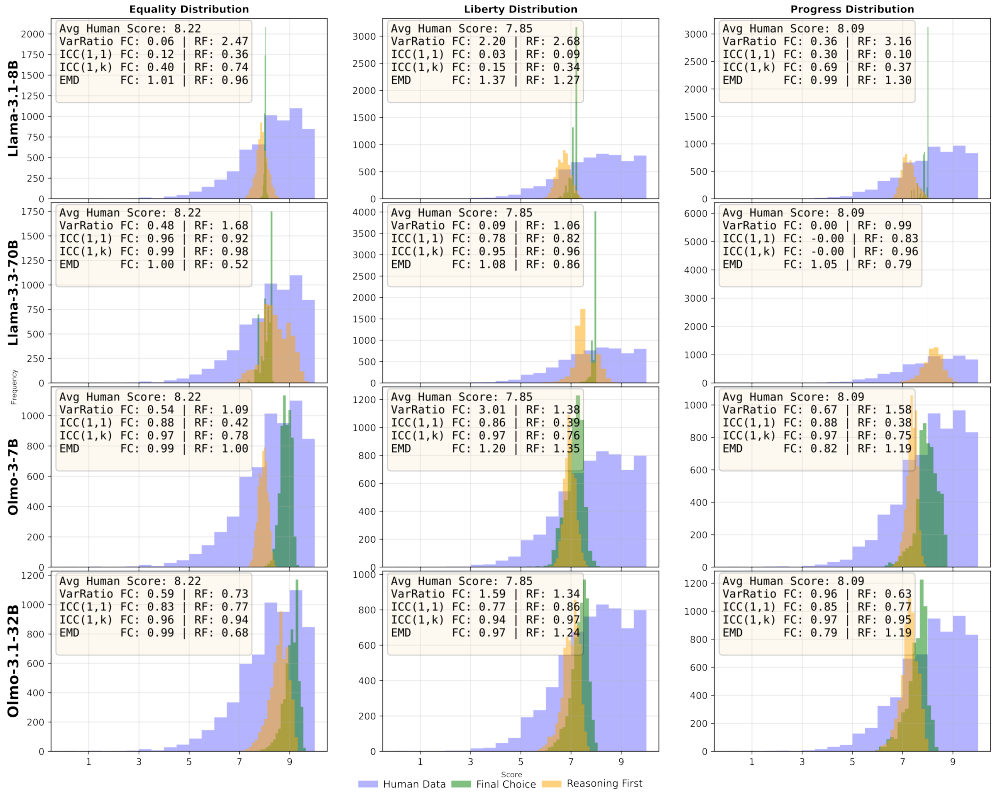

Figure 2: Score distribution between 4 LLMs in 2 Prompt Styles (FC and RF) and Human scores with metrics ICC(1,1), ICC(1,k), Wasserstian distance (EMD) and variance ratio

The regression model is specified as follows:

$$Y_{i,q} = \beta_0 + \sum_{d=1}^{D} b_d X_{i,d} + \alpha_q + \epsilon_{i,q} \qquad (1)$$

where $Y_{i,q}$ is the dependent variable (either LLM score, Abs. Error, Sign. Error or $H\%$) for respondent i and the question q pair. $X_{i,d}$ represents the d^{th} demographic feature for the respondent i, b_d is the corresponding unstandardized coefficient, α_q is the fixed effect of the value questions and $\epsilon_{i,q}$ is the error term. Categorical variables are encoded using one-hot encoding. Age and State are aggregated into groups for reducing dimensionality. Age is categorized into age_groups (18-29, 30-44, 45-64, 65+), and States are converted into regions (Northeast, Midwest, South, West). We report partial R^2 and coefficients (b) to quantify the explained variance in the outcome attributable to demographics after accounting for question fixed effects, and to quantify the magnitude and direction of the effects of demographic characteristics on the outcome. Additionally, we include the LLM's numeric output (final_choice) as a predictor of the dependent variable $H\%$ to assess whether the model's language correlates with its internal certainty. A significant effect would suggest the LLM's *persona* maintains stylistic consistency with its scores; specifically, a negative coefficient would indicate the model hedges less as its assigned scores increase.

3 Results

3.1 Behavioral Rigidity

ICC(1,1), ICC(1,k) that captures the single run and iterative reliability are shown in the Figure 3. The FCF strategy demonstrates high reliability overall. Llama-3.1-8B achieved the highest consistency, with ICC(1,1) = 0.94 and ICC(1,k) = 0.99, indicating near-perfect agreement across runs. Olmo-3-7B and Olmo-3.1-32B also showed strong reliability (ICC(1,1) = 0.96 and 0.93, respectively). Examining individual dimensions, reliability varied considerably. For the Liberty dimension, Llama-3.1-8B showed very low reliability (ICC(1,1) = 0.03). For the Equality dimension, most models performed well, with Llama-3.3-70B reaching ICC(1,1) = 0.96. The Progress dimension was more variable: Llama-3.3-70B yielded a near-zero ICC(1,1), indicating no reliable agreement, while Olmo-3-7B and Olmo-3.1-32B maintained good reliability (ICC(1,1) of 0.88 and 0.85, respectively). Under the RF strategy, overall reliability

was generally lower than in the FCF condition. Dimension-level patterns echoed those observed in the FCF condition. Overall, results indicate that the FCF prompting strategy yields higher inter-rater reliability than the RF strategy across models and dimensions. Larger models (Llama-3.3-70B, Olmo-3.1-32B) tend to produce more consistent scores regardless of strategy.

EMD, and variance ratio for each model, prompt style, and value dimension are shown in Figure 2. The prompt style affects the LLama and Olmo models differently. With FCF Llama-3.1-8B and Llama-3.3-70B has the highest EMD of 1.12 and 1.04. In contrast, Olmo-3.1-32B (EMD = 1.01) and Olmo-3-7B (EMD = 0.0.92) showed considerably lower overall deviation. With RF, the pattern is partially reversed. Llama-3.3-70B showed a marked 30% reduction in overall EMD (EMD = 0.72) compared to its FCF value. The Olmo models, however, increased with RF: Olmo-3-7B (EMD = 1.18) and Olmo-3.1-32B (EMD = 1.03) both rose substantially from their FCF values. For a breakdown by model family, construct, and prompt style, refer to the Appendix Tables 6 to 10.

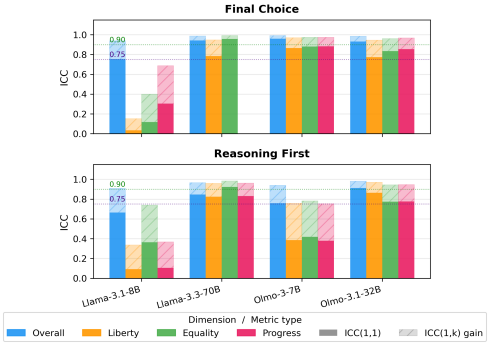

Figure 3: ICC(1,1) and ICC(1,k) where k=5 showing the overall and invidual dimension reliability

3.2 Regression

The regression analysis in Table 1 reveals that demographic variables explain LLM score outcomes at an average of 68% (range: 38%–91% across models and prompt combinations), compared to just 9% for human responses. For the outcome $H\%$, demographic variables explain on average 71% of the variation (range: 48%–89%). Absolute Error (15%) and Signed Error (21%) are more modestly explained by demographics alone. Across all models and prompt styles, predictors that appear in the top 20 in at least 6 out of 8 models and prompt combinations are included in 1. Political affiliation emerged as a consistent and substantively meaningful predictor across all four outcomes. Compared to Democrats, Republicans received lower LLM Scores, higher Hedge%, and greater Abs. Error, suggesting that LLMs assign less favorable scores, hedge more frequently, and deviate more from human scores when evaluating for a republican persona. Being a registered voter was associated with higher LLM Scores ($\beta = .23$, 100%), lower Hedge%, and lower Abs. Error ($\beta = -.13$, 100%). Employment status showed consistent effects. Unemployed respondents received substantially lower LLM Scores ($\beta = -.24$, 100%), higher Hedge%, and greater Abs. Error ($\beta = .14$, 100%). Part-time employed respondents similarly received lower LLM Scores. Education effects were notably uneven. Notably, respondents with only a grade-school education exhibited the largest coefficient Abs. Error and Sign. Error ($\beta = .45$, 100%;$\beta = .31$, 100%), despite not appearing to be consistent predictors of LLM Score or Hedge%. High-school educated demographic received lower LLM Scores and greater Absolute Error, but negative Signed Error ($\beta = -.22$, 100%), indicating systematic underestimation relative to human scores. Non-binary respondents showed substantially higher Abs. Error and Sign. Error ($\beta = .26$, 100%;$\beta = .33$, 100%), suggesting more distributional deviation in model responses, while transgender respondents were associated with notably over estimation in Sign. Error ($\beta = .53$, 100%) and lower Hedge% ($\beta = -.16$, 100%). Compared to the reference youngest age group (18–29), older respondents (30–44, 45–64, 65+) received lower Sign. Error scores indicating under estimation ($\beta = -.34$, 100%; $\beta = -.43$, 100%; $\beta = -.57$, 100%). Age was not a consistent predictor of Hedge%.

4 Discussion

4.1 Demographic Influence on Model Size and Prompt Style

Model size and prompt style interact — larger models (Llama-3.3-70B, OLMo variants) achieve good reliability (ICC(1,1) and ICC(1,k)) under FCF, but only larger models maintain acceptable reliability under RF. Llama-3.1-8B fails the reliability threshold under RF on most dimensions and should be treated as unsuitable for consistent scoring regardless of prompt style. Both OLMo-3-7B and OLMo-3.1-32B sit above 0.75 on most dimensions under

Table 1: Consistent significant demographic predictors of LLM response outcomes and their average coefficients

Predictor	**LLM Score**	**Abs. Error**	**Sign. Error**	**Hedge%**
R^2 *[min,max]*	*.68 [.38,.91]*	*.15 [.04,.22]*	*.21 [.06,.33]*	*.71 [.48,.89]*
Political affiliation (ref: Democrat)				
Republican	$-.12^{*}$ (88%)	$.13^{*}$ (100%)	$.01^{*}$ (100%)	$.01^{*}$ (100%)
Independent	$-.08^{*}$ (100%)	$.09^{*}$ (100%)	—	$.05^{*}$ (75%)
Other	$-.10^{*}$ (100%)	$.22^{*}$ (100%)	$.13^{*}$ (88%)	—
Voter registration				
Registered (Yes)	$.23^{*}$ (100%)	$-.13^{*}$ (100%)	—	$-.04^{*}$ (88%)
Employment (ref: Full-time)				
Unemployed	$-.24^{*}$ (100%)	$.14^{*}$ (100%)	—	$.05^{*}$ (75%)
Part-time	$-.11^{*}$ (100%)	—	—	$.03^{*}$ (100%)
Other	—	—	—	—
Education (ref: Bachelor's)				
Grade school	—	$.45^{*}$ (100%)	$.31^{*}$ (88%)	—
High school	$-.09^{*}$ (88%)	$.12^{*}$ (100%)	$-.22^{*}$ (100%)	$.01^{*}$ (75%)
Some coll/trade	—	$.06^{*}$ (88%)	$-.15^{*}$ (100%)	$.04^{*}$ (100%)
Graduate/Prof.	$.07^{*}$ (100%)	—	—	$.03^{*}$ (88%)
Race (ref: Other/non-listed)				
White/Cauc.	$.08^{*}$ (88%)	—	—	$.07^{*}$ (88%)
Asian	$.01^{*}$ (88%)	—	$-.26^{*}$ (100%)	$.06^{*}$ (75%)
Native American	$-.09^{*}$ (88%)	$.22^{*}$ (100%)	—	—
Black/Afr. Amer.	—	—	$-.20^{*}$ (100%)	—
Age group (ref: 18–29)				
30–44 years	—	$-.06^{**}$ (88%)	$-.34^{*}$ (100%)	—
45–64 years	$-.01^{*}$ (100%)	$-.12^{*}$ (100%)	$-.43^{*}$ (100%)	—
65+ years	$-.01^{*}$ (100%)	$-.16^{*}$ (100%)	$-.57^{*}$ (100%)	—
Gender (ref: Female)				
Male	$-.01^{*}$ (88%)	—	—	$.01^{*}$ (88%)
Non-binary	—	$.26^{*}$ (100%)	$.33^{*}$ (75%)	—
Transgender	—	—	$.53^{*}$ (100%)	$-.16^{**}$ (100%)
Hispanic ethnicity				
Hispanic (Yes)	—	—	$-.14^{*}$ (100%)	$-.04^{*}$ (75%)

Positive coefficients, Negative coefficients $^{*}p < .001$. $^{**}p < .01$.
(%) indicates occurance of the predictor in top 20 significant predictor for the 8 cases (4 Models and 2 Prompt types).
— not in top 20 in <75% or 6 out of the 8 cases.

FCF, and largely above 0.75 on Overall under RF 2. They're the most reliable scorers across both prompt styles.

Forcing the model to generate a "Chain of Thought" introduces additional context beyond the direct lookup of stereotypes and lowers the artificially high R^2. However, even with context, the LLM R^2 remains roughly 7x higher than the human baseline (See Appendix Table 2), indicating the bias is deep-seated.

LLaMA models exhibit a clear scaling effect, with the larger 70B model showing lower R^2 values than the smaller 8B model across both prompt styles, suggesting that larger models allow for greater response variance. The smaller OLMo 7B shows an extremely high R^2 (0.86–0.96), indicating responses are almost entirely driven by demographic inputs. While the larger OLMo 32B reduces this determinism under the FCF prompt style, it does not do so consistently under the RF prompt, suggesting that chain-of-thought style prompting may lead models to construct explicit demographic rationales, reinforcing rather than moderating stereotype-driven responses. Taken together, these patterns suggest that LLMs broadly simulate demographic archetypes rather than individual variation, with model size offering only partial mitigation. This tendency is most pronounced for specific constructs: for the Liberty scale, LLaMA-3.1-8B reached an R^2 of 0.96 (Refer Appendix Table 2), indicating that for questions involving freedom and government constraint, smaller models collapse almost entirely into stereotypical responses, leaving virtually no room for within-group variance.

4.2 Dimension effect

Liberty emerged as the most divergent dimension between human respondents and LLMs, with EMD values peaking at 1.17 and the highest variance ratio of 1.67 (Figure 2). Equality generally yielded the lowest EMD values (0.89) and variance ratio (0.96), particularly for larger models within the same family. Progress showed intermediate EMD (1.02) and variance ratio (1.05), though notable spikes were observed for Llama-3.1-8B (EMD = 1.30 under RF), suggesting that reasoning strategies do not uniformly reduce distributional bias and may amplify it for certain model-dimension combinations. While Llama-3.3-70B demonstrated the highest overall alignment under RF (EMD = 0.72), models such as Olmo-3.1-32B showed that reasoning can exacerbate divergence in specific dimensions, particularly Liberty and Progress. The lack of inferential variation across dimensions raises concerns about adopting a single model family or prompt style with confidence.

4.3 Hedging

Hedging is a known LLM tendency to avoid commitment, which would naturally distort survey simulation. The variation in hedging across demographics was an unexpected result, yet has semblance to the findings from (Santurkar et al., 2023). The bias was most pronounced across a few demographics and select questions. Whenever the model is prompted with "Non-binary" or "Transgender," it immediately enters a "Safety Mode" characterized by high hedging. This is evident from the strong negative coefficients for these predictors in the hedging regression. This could be due to the post-training safety fine-tuning that penalizes the model for making strong statements about marginalized identities, leading to a default response of hedging when these identities are present in the input. This could highlight the cautionary process, which differs from typical model behavior arising from under-representation in the training data or confusion about the topic. Instead, it reflects a deviant behavior often associated with sensitive topics in

fairness and safety research. The model's response is not necessarily driven by confusion about the topic itself, but rather by a learned behavior to avoid making definitive statements about certain identities, which results in increased hedging. The quality of synthetic data is distorted when LLM output is influenced by functions such as "Safety Mode".

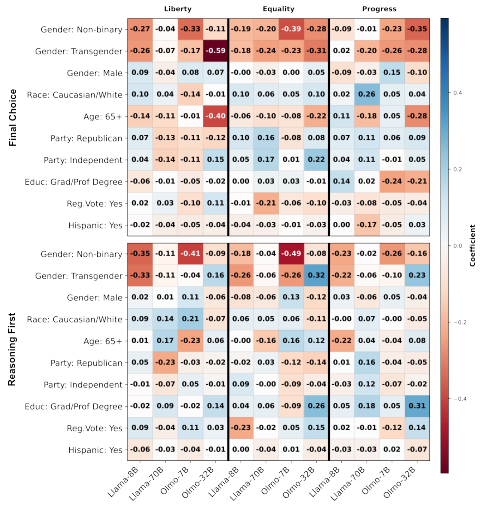

Figure 4: Hedging predictors by demographics. Including significant (p<0.05) predictors excluding states.

4.4 Demographic Caricature

LLMs act as mean-finders, generalizing toward the average member of the population. Our results from Table 1 suggest the WEIRD claim is only partially justified for educated individuals . Similar to (Santurkar et al., 2023), the response caricatures are aligned well with liberal and educated. Additionally, the LLM does not follow the edges of opinion but consistently seeks the center. Despite the inclusion of demographic context in the artificial persona, LLMs remain fixated on certain perceived monoliths (Unemployed, Grade School, Non-binary, and Transgender). The models mostly ignore the other demographic traits that make up the intersectionality of the individual, thereby restricting potential variation in synthetic output.

The results suggest that the LLM treats "Registered Voter" as a proxy for "Good/High Value Citizen," mechanically boosting the scores across the board. LLMs underweight the environmental and longitudinal factors like geography and age, opting instead to amplify "identity signals" like voter registration or employment status that align with the most frequent and often most biased patterns in their training sets. Individual constructs show pronounced stereotypes. Regarding the Progress construct, the demographic indicator educ_Grade school emerges as a substantial negative driver (coefficients ranging from -0.43 to -0.50). The model appears to use "low education" as a heuristic, effectively decreasing scores on complex topics such as economics and infrastructure. Similarly, partyid_Republican serves as a dominant negative predictor for Equality ($\beta = -0.40$)). While empirical human data often show lower scores for this group, the model fails to replicate the substantial internal variance found in human populations. Instead, the LLM flattens the diverse spectrum of Republican thought—from Libertarianism to Populism—into a monolithic "Anti-Equality" coefficient. This highlights a critical failure in representational fidelity: the model lacks the granular capacity to distinguish between specific ideological motivations, such as fiscal opposition to taxes versus social opposition to equity programs, reinforcing the findings of (González Barman et al., 2025) in an experimental setting focused on eliciting diverse opinions.

5 Conclusion

This study provides both a procedure and substantive results of evaluating the performance of select open-source LLMs as synthetic respondents in survey research. While LLMs offer ample experimental opportunities and methods for generating hypotheses, the results of this study suggest that they are currently inadequate substitutes for data collected from human respondents. On the one hand, the results generated by the different models and prompting strategies were highly consistent across iterations, whereas the substantive output from the artificial data was less robust. The synthetic data from the models and prompting styles used resulted in extremely narrow variances in the data that fail to capture the gradation of actual human responses. Beyond the numerical results, the reasoning provided by the LLMs constituted a noncommittal, unclear rationale that did not consistently support their answers to the question. We approach the adoption of LLM respondents in public opinion survey data with deep skepticism, positing that what appears to be human-like opinion is often just a highly probable text completion, stripped of genuine human nuance. If researchers indiscriminately adopt these tools and incorporate synthetic data in published results, we risk measuring a distorted version of training data and other noise rather than

dynamic public sentiment.

Limitations

The aggregate results and questions associated with this study were made publicly available in 2021 and may have been included in the models' training data, potentially contaminating the synthetic data and confounding its quality. This risk could not be directly measured or controlled for in our analysis. Furthermore, our prompt does not explicitly account for the effect of this 2021 data, which was not included in the replication survey. The use of a ten-point integer scale (0–10) is another limitation, as it differs from more conventional response formats such as binary, four-point, or three-point scales, which may limit the generalizability of our evaluation metrics. Another limitation identified during the analysis was the conversion of respondents' age values from integers to age groups for regression analysis, which limits the specificity for making claims. Additionally, we don't isolate the post-training effect by comparing our Instruct models with their base-model counterparts, which induces a noticeable performance shift, as noted by (Santurkar et al., 2023). We encourage researchers to consider these elements in future replications of this study.

6 Acknowledgments

We acknowledge the support of Siena Research Institute in allowing us to use the American Values Survey. We acknowledge the support of Social Science Automation in facilitating the presentation of this research. We thank Eddie Smith and Pierce Johnson for their valuable observations on the LLM output responses. We thank the reviewers for pointing out related works that strengthened our findings.

References

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 31(3):337–351.

Yash Bhatnagar. 2026. Breaking the turing test: Testing the relevance of the turing test against modern llms. *International Journal for Research in Engineering Application Management*, pages 1–.

J Michael Brick, Pat D Brick, Sarah Dipko, Stanley Presser, Clyde Tucker, and Yangyang Yuan. 2007. Cell phone survey feasibility in the us: Sampling and calling cell numbers versus landline numbers. *Public Opinion Quarterly*, 71(1):23–39.

Mario Callegaro and Charles DiSogra. 2008. Computing response metrics for online panels. *Public opinion quarterly*, 72(5):1008–1032.

Mick P Couper and Peter V Miller. 2008. Web survey methods: Introduction. *Public opinion quarterly*, 72(5):831–835.

Rachel E Davis, Timothy P Johnson, Sunghee Lee, and Christopher Werner. 2019. Why do latino survey respondents acquiesce? respondent and interviewer characteristics as determinants of cultural patterns of acquiescence among latino survey respondents. *Cross-Cultural Research*, 53(1):87–115.

Jonathan Eggleston. 2024. Frequent survey requests and declining response rates: evidence from the 2020 census and household surveys. *Journal of Survey Statistics and Methodology*, 12(5):1138–1156.

Aaron Fanous, Jacob Goldberg, Ank Agarwal, Joanna Lin, Anson Zhou, Sonnet Xu, Vasiliki Bikia, Roxana Daneshjou, and Sanmi Koyejo. 2025. Syceval: Evaluating llm sycophancy. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pages 893–900.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Anna Foka, Gabriele Griffin, Dalia Ortiz Pablo, Paulina Rajkowska, and Sushruth Badri. 2025. Tracing the bias loop: Ai, cultural heritage and bias-mitigating in practice. *AI & SOCIETY*, 40(8):5835–5847.

Mansour Ghafourifard. 2024. Survey fatigue in questionnaire based research: The issues and solutions. *Journal of caring sciences*, 13(4):214–215.

Chris Gibson and Daniel Lipinski. 2021. Americans, deeply divided, yet share core values of equality, liberty progress.

Kristian González Barman, Simon Lohse, and Henk W de Regt. 2025. Reinforcement learning from human feedback in llms: Whose culture, whose values, whose perspectives? k. gonzález barman et al. *Philosophy & Technology*, 38(2):35.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard

Ponarin, Bjorn Puranen, and 1 others. 2022. World values survey: Round seven-country-pooled datafile version 5.0.

Yoichi Hayashi. 2024. Prospects for revolutionary and popular ai technology following the launch of chatgpt in 2023.

Morten Hjortskov, Christian Bøtcher Jacobsen, and Anne Mette Kjeldsen. 2023. Choir of believers? experimental and longitudinal evidence on survey participation, response bias, and public service motivation. *International Public Management Journal*, 26(2):281–304.

Jumayel Islam, Lu Xiao, and Robert E Mercer. 2020. A lexicon-based approach for detecting hedges in informal text. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3109–3113.

Liwei Jiang, Taylor Sorensen, Sydney Levine, and Yejin Choi. 2024. Can Language Models Reason about Individualistic Human Values and Preferences? *Preprint*, arXiv:2410.03868.

Timothy P Johnson, Henning Silber, and Jill E Darling. 2024. Public perceptions of pollsters in the united states: experimental evidence. *Social Science Quarterly*, 105(1):114–127.

Indu Joshi, Marcel Grimmer, Christian Rathgeb, Christoph Busch, Francois Bremond, and Antitza Dantcheva. 2024. Synthetic data in human analysis: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7):4957–4976.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large Language Models are Zero-Shot Reasoners. *arXiv preprint*. ArXiv:2205.11916 [cs].

Sarah Kreps, R Miles McCain, and Miles Brundage. 2022. All the news that's fit to fabricate: Ai-generated text as a tool of media misinformation. *Journal of experimental political science*, 9(1):104–117.

Jon A Krosnick. 1991. Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied cognitive psychology*, 5(3):213–236.

Jon A Krosnick, Allyson L Holbrook, Matthew K Berent, Richard T Carson, W Michael Hanemann, Raymond J Kopp, Robert Cameron Mitchell, Stanley Presser, Paul A Ruud, V Kerry Smith, and 1 others. 2002. The impact of" no opinion" response options on data quality: non-attitude reduction or an invitation to satisfice? *Public Opinion Quarterly*, 66(3):371–403.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Lujun Li, Lama Sleem, Geoffrey Nichil, Radu State, and 1 others. 2025. Exploring the impact of temperature on large language models: Hot or cold? *Procedia Computer Science*, 264:242–251.

Michael W Link, Ali H Mokdad, Dale Kulp, and Ashley Hyon. 2006. Has the national do not call registry helped or hurt state-level response rates? a time series analysis. *International Journal of Public Opinion Quarterly*, 70(5):794–809.

Michael W Link and Robert W Oldendick. 1999. Call screening: Is it really a problem for survey research? *The Public Opinion Quarterly*, 63(4):577–589.

Andy Markus. 2025. AT&T tests new AI digital receptionist. https://about.att.com/blogs/2025/ai-digital-receptionist.html.

Robert W Oldendick and Michael W Link. 1994. The answering machine generation: who are they and what problem do they pose for survey research? *Public Opinion Quarterly*, 58(2):264–273.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, and 1 others. 2025. Olmo 3. *arXiv preprint arXiv:2512.13961*.

Giada Pistilli, Alina Leidinger, Yacine Jernite, Atoosa Kasirzadeh, Alexandra Sasha Luccioni, and Margaret Mitchell. 2024. Civics: Building a dataset for examining culturally-informed values in large language models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 1132–1144.

David C Pyrooz, James A Densley, and Jose Antonio Sanchez. 2025. Are online opt-in panels viable data sources on hard-to-reach populations? population and relational inferences on gang membership in the united states. *International Criminology*, pages 1–17.

Patric M Reinbold. 2020. Taking artificial intelligence beyond the turing test. *Wis. L. Rev.*, page 873.

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

Patrick E Shrout and Joseph L Fleiss. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420.

Jorre Vannieuwenhuyze, Geert Loosveldt, and Geert Molenberghs. 2010. A method for evaluating mode effects in mixed-mode surveys. *Public opinion quarterly*, 74(5):1027–1045.

Michael F Weeks, Richard A Kulka, and Stephanie A Pierson. 1987. Optimal call scheduling for a telephone survey. *Public Opinion Quarterly*, pages 540–549.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, and 1 others. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Yu Xie and Yueqi Xie. 2025. Variance reduction in output from generative ai. *arXiv preprint arXiv:2503.01033*.

Wenlong Zhao, Debanjan Mondal, Niket Tandon, Danica Dillion, Kurt Gray, and Yuling Gu. 2024. 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)*, pages 17696–17706, Torino, Italia. ELRA and ICCL.

Ke Zhou, Marios Constantinides, and Daniele Quercia. 2025. Should llms be weird? exploring weirdness and human rights in large language models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pages 2808–2820.

A Appendix

A.1 American Values Survey Data

The American Values Survey was conducted by Siena College Research Institute (SCRI) in 2021. The respondents were from 50 states and the District of Columbia. Respondents were asked how much they agree with 34 value statements covering core American values on Liberty, Equality, and Progress. For our experiment we only kept questions related to the three value constructs and removed questions not relevant to the values. The survey was administered online, and initial attention checks were conducted to ensure high-quality responses. Irrespective of demographics and inclinations, human respondents express a core sense of embodying American values. This dataset offers a unique opportunity to evaluate LLMs, as it combines rich demographic information with clearly demarcated group identities that, despite strong internal affiliations, are united by a shared commitment to a nation's core values. The actual value statements are shown in the Table 3.

A.2 Prompting LLMs

The LLM responses were forced to be in a structured format $JSON$ to facilitate downstream analysis of the reasoning traces and final scores. The detailed prompts are provided in Table 4. The prompts are also similar to the actual survey questions, with only the input and output instructions appended at the beginning and end of the survey text for validity. The actual survey did not collect reasoning traces, so we do not have human ground truth for the reasoning component. Since our experiment requires large-scale inference, we use the VLLM library (Kwon et al., 2023). To generate multiple responses to the same prompt, we use a temperature of $\tau = 0.1$, since the ideal deterministic temperature $\tau = 0$ doesn't conveniently allow multiple iterations for the same input prompt through the VLLM library. We dont limit the number of tokens generated by model. Apart from the temperature $k = 5$ i.e. number of samples per input is the only hyperparameter we modified. Rest of the values were kept default. All four LLMs are loaded from Hugging Face's official model weight collection. We don't use any quantization. All the models are run on NVIDIA A100 GPUs. We use 2 GPUs for the smaller 7B and 8B models and use 4 GPUs for the larger 32B and 70B models.

A.3 Evaluation Metrics

We tabulate the metrics used to compare the LLM responses with human responses. The ideal and poor values for each metric and their resultant interpretations are tabulated for easier understanding in Table 5. Behavioral consistency is a prerequisite for alignment evaluation — only models with sufficient ICC reliability are meaningfully assessed on alignment quality metrics. They could indicate consistent distortion or consistent alignment. A high ICC, alongside a low σ^2, is a signature of consistency and confidence, treating the demographic profile as a fixed caricature rather than an opinion distribution. We include a summary of the 5 metrics (ICC(1,1), ICC(1,k), variance ratio, EMD, and Hedging%) used in the experiment, broken down by Model, Construct, and prompt styles.

Table 2: Mean R^2 [min, max] by scope.

Scope	LLM	Abs.Error	Sign.Error	Hedge%	Human
Overall	.68 [.38, .91]	.15 [.04, .22]	.21 [.06, .33]	**.71** [.48, .89]	.09
By value construct:					
Liberty	**.76** [.22, .96]	.21 [.03, .38]	.26 [.04, .46]	.61 [.28, .87]	.06
Equality	.68 [.45, .91]	.08 [.03, .16]	.12 [.06, .30]	**.72** [.39, .93]	.12
Progress	.67 [.00, .90]	.10 [.02, .21]	.17 [.05, .38]	.62 [.45, .83]	.07

Note. Mean R^2 across dimensions; [min, max] in brackets. Abs.Error is Absolute Error and Sign.Error is Signed Error. Hedge% is where the average hedging token present in the reasoning text is the target. Human R^2 has zero variance within each construct. **Bold** indicates $R^2 \geq 0.70$.

Label	Value Statement
Equality_1	All people are equal, regardless of race, ethnicity, gender, physical appearance, or any other personal characteristic.
Equality_2	Treat others as you would like them to treat you.
Equality_3	No one is above the law.
Equality_5	The religious beliefs and practices of all people should be both protected and respected.
Equality_6	Any injustice to a single person is an injustice to all.
Equality_7	We are all, all of us, in this life together and we should look out for the well-being of everyone else.
Equality_8	People may disagree but that is no excuse for being disagreeable.
Equality_9	No person is complete if they do not give of themselves in service to others.
Equality_10	Before making a judgement about someone else, try to walk a mile in their shoes.
Equality_11	In order for us all to live together, each of us has to make concessions.
Equality_12	Not everyone starts off with the same set of tools or skills, sometimes we need to level the playing field by giving some people a head start.
Equality_13	Because we only have one planet, protecting our environment is a priority.
Equality_14	Steps must be taken to protect people from those who lie and cheat.
Equality_15	Each of us should have an equal chance to be successful.
Equality_16	Every American has the right and responsibility to vote.
Liberty_1	No one, not even the government, should be able to restrict another's pursuit of happiness.
Liberty_4	The benefits of investing capital and hard work rightfully belong to the entrepreneur that accepted the risk.
Liberty_5	You only live once: seek to experience all that life has to offer.
Liberty_6	No one should tell me how to live, how to love or what to think.
Liberty_11	Each of us has the power to pull ourselves up by our bootstraps, that is, to take control of our own destiny.
Liberty_12	What you know is more important than who you know.
Liberty_13	Each of us is free to follow our own unique path in life.
Liberty_14	Stand on your own two feet without reliance on other people, organizations or the government.
Liberty_15	Owe nothing and expect nothing in return.
Liberty_16	Everyone can speak their mind in public regardless of the viewpoint without fear of punishment.
Progress_1	Success comes to those that dedicate themselves to making the most of their abilities.
Progress_2	Advances in areas like health, technology, business, or personal development, rely on the careful application of science.
Progress_3	Give a person a fish, and you'll feed them for a day. Teach a person to fish, and you've fed them for a lifetime.
Progress_4	A penny saved is a penny earned.
Progress_5	Having specific goals, whether those goals involve personal, economic, artistic or societal achievements, is the best way to proceed in life.
Progress_6	It is important to achieve something specific and measurable each and every day.
Progress_7	Every problem has a solution.
Progress_8	Tomorrow always holds the possibility of being a better day.
Progress_9	Time is money.

Table 3: Survey constructs and full value statements used in the evaluation.

Final Choice First (FCF) Prompt Template

Your demographic information is given in JSON format
{'state': 'Rhode Island', 'gender': 'Female', 'age': 60.0, 'hispanic': 'No', 'race': 'Caucasian/White', 'education': 'Graduate or Professional degree', 'registered to vote': 'Yes', 'party id': 'Democrat', 'employment': 'Employed full-time'}

First, you will see a statement that may be familiar to you. Take a moment and consider it in terms of how you live your life. Are these words that you live by? All the time? Or some of the time? Would others say that you embody these statements, that your actions, or that the things you say are a reflection of these words often, sometimes, not very often, or perhaps not at all? We are all different people. Some statements may be completely us, others partially, and some not us at all. Try as best as you can to evaluate yourself, your thoughts, your actions, as well as your beliefs on each statement. As you evaluate yourself on each statement, you can score yourself anywhere between 0 and 10.

0: You disagree with the statement, are not guided by it, and no one would ever say that the statement reflects how you live your life.
5: The statement may be one that you endorse, but you don't always live your life with it in mind.
10: You believe the statement, and not only aspire to live that way, but you do.

STATEMENT:
Each of us is free to follow our own unique path in life.

INSTRUCTION:
Return ONLY valid JSON with exactly these keys:
- "final_choice": integer (0–10)
- "reasoning": a first-person justification for how you arrived at your score. Do NOT write placeholders like "..." or "[insert reasoning here]".

OUTPUT:

Reasoning First (RF) Prompt Template

Your demographic information is given in JSON format
{'state': 'Rhode Island', 'gender': 'Female', 'age': 60.0, 'hispanic': 'No', 'race': 'Caucasian/White', 'education': 'Graduate or Professional degree', 'registered to vote': 'Yes', 'party id': 'Democrat', 'employment': 'Employed full-time'}

First, you will see a statement that may be familiar to you. Take a moment and consider it in terms of how you live your life. Are these words that you live by? All the time? Or some of the time? Would others say that you embody these statements, that your actions, or that the things you say are a reflection of these words often, sometimes, not very often, or perhaps not at all? We are all different people. Some statements may be completely us, others partially, and some not us at all. Try as best as you can to evaluate yourself, your thoughts, your actions, as well as your beliefs on each statement. As you evaluate yourself on each statement, you can score yourself anywhere between 0 and 10.

0: You disagree with the statement, are not guided by it, and no one would ever say that the statement reflects how you live your life.
5: The statement may be one that you endorse, but you don't always live your life with it in mind.
10: You believe the statement, and not only aspire to live that way, but you do.

STATEMENT:
Each of us is free to follow our own unique path in life.

INSTRUCTION:
Return ONLY valid JSON with exactly these keys:
- "reasoning": a first-person justification for how you arrived at your score. Do NOT write placeholders like "..." or "[insert reasoning here]".
- "final_choice": integer (0–10)

OUTPUT:

Table 4: FCF and RF prompt templates. *Italicised* text denotes the demographic precursor prepended to the prompt. The only difference between the two styles is the order of the JSON keys in the instruction.

Metric	Range / Value	Interpretation
ICC	Ideal: Closer to 1	The model's responses demonstrate high stability and are consistent and reliable for that specific persona.
	Poor: Values below 0.5	The model is generating inconsistent responses and appears to be sensitive to random noise in the prompt.
EMD	Ideal: Closer to 0	The model's score distribution shows high representational accuracy and closely matches human score distributions.
	Poor: Higher values	The model score distribution is significantly distorted from that of the human group's actual recorded values.
Variance Ratio	Ideal: Near 1.0	The model differentiates between personas to the same natural extent that humans do in real-world data.
	Poor: Below 1 or Above 1	The model either gives similar scores regardless of demographics or exaggerates differences across those demographics.
Hedging%($H\%$)	Ideal: Lower (Contextual)	The model provides a clear and decisive stance with very little ambiguity regarding its reasoning.
	Poor: High percentages	The model displays high uncertainty and is playing it safe to avoid taking a firm stance on the topic.

Table 5: LLM Evaluation Metrics for demographic consistency and alignment with human respondents on public opinion survey

Table 6: **Intraclass Correlation Coefficient (ICC(1,1)) Reliability Summary**
Higher ICC(1,1) values indicate greater single-run reliability across simulated responses. Values near 1.0 suggest highly uniform (potentially homogenized) outputs across individual ratings. Highest and Lowest values for every category are highlighted.

Main Effects Overview

	Avg. ICC(1,1)
By Model	
Olmo-3.1-32B	**0.84**
Llama-3.3-70B	0.76
Olmo-3-7B	0.69
Llama-3.1-8B	0.30
By Construct	
Equality	**0.66**
Liberty	0.58
Progress	0.52
By Reasoning Style	
Final Choice	**0.68**
Reasoning First	0.62

Interactions Overview

Model × Reasoning Style

Model	FC Style	RF Style
Llama-3.3-70B	0.67	0.86
Olmo-3.1-32B	0.85	0.83
Olmo-3-7B	**0.90**	0.48
Llama-3.1-8B	0.30	0.31

Model × Construct

Model	Equality	Liberty	Progress
Llama-3.3-70B	**0.94**	0.80	0.41
Olmo-3.1-32B	0.80	0.82	0.81
Olmo-3-7B	0.65	0.62	0.63
Llama-3.1-8B	0.24	0.06	0.20

Table 7: **Intraclass Correlation Coefficient (ICC(1,k)) Reliability Summary**
Higher ICC(1,k) values denote greater internal consistency and reliability among profiles for average simulated responses. Highest and Lowest values for every category are highlighted.

Main Effects Overview

	Avg. ICC(1,k)
By Model	
Olmo-3.1-32B	**0.96**
Olmo-3-7B	0.89
Llama-3.3-70B	0.85
Llama-3.1-8B	0.57
By Construct	
Equality	**0.85**
Liberty	0.75
Progress	0.71
By Reasoning Style	
Reasoning First	**0.83**
Final Choice	0.80

Interactions Overview

Model × Reasoning Style

Model	FC Style	RF Style
Llama-3.3-70B	0.73	**0.97**
Olmo-3.1-32B	0.96	0.96
Olmo-3-7B	0.98	0.81
Llama-3.1-8B	0.54	0.59

Model × Construct

Model	Equality	Liberty	Progress
Llama-3.3-70B	**0.99**	0.95	0.48
Olmo-3.1-32B	0.95	0.96	0.96
Olmo-3-7B	0.88	0.86	0.86
Llama-3.1-8B	0.57	0.24	0.53

Table 8: **Variance Ratio Summary**
Variance Ratio of the LLM responses to human baseline. Highest and Lowest values for every category are highlighted.

Main Effects Overview

	Avg. Var Ratio
By Model	
Llama-3.1-8B	**1.83**
Olmo-3-7B	1.42
Olmo-3.1-32B	1.05
Llama-3.3-70B	0.73
By Construct	
Liberty	**1.67**
Progress	1.05
Equality	0.96
By Reasoning Style	
Reasoning First	**1.60**
Final Choice	0.92

Interactions Overview

Model × Reasoning Style

Model	FC Style	RF Style
Llama-3.1-8B	0.89	**2.78**
Olmo-3-7B	1.47	1.36
Olmo-3.1-32B	1.13	0.97
Llama-3.3-70B	0.20	1.27

Model × Construct

Model	Equality	Liberty	Progress
Llama-3.1-8B	1.26	**2.44**	1.76
Olmo-3-7B	0.82	2.19	1.13
Olmo-3.1-32B	0.66	1.47	0.80
Llama-3.3-70B	1.08	0.57	0.50

Table 9: **EMD Summary**
EMD between LLM responses and reference human distribution. Highest and Lowest values for every category are highlighted.

Main Effects Overview

	Avg. EMD
By Model	
Llama-3.1-8B	**1.15**
Olmo-3-7B	1.09
Olmo-3.1-32B	0.98
Llama-3.3-70B	0.88
By Construct	
Liberty	**1.17**
Progress	1.02
Equality	0.89
By Reasoning Style	
Reasoning First	**1.03**
Final Choice	1.02

Interactions Overview

Model × Reasoning Style

Model	FC Style	RF Style
Llama-3.1-8B	1.12	**1.18**
Llama-3.3-70B	1.04	0.72
Olmo-3-7B	1.01	1.18
Olmo-3.1-32B	0.92	1.03

Model × Construct

Model	Equality	Liberty	Progress
Llama-3.1-8B	0.99	**1.32**	1.14
Olmo-3-7B	1.00	1.27	1.01
Olmo-3.1-32B	0.83	1.11	0.99
Llama-3.3-70B	0.76	0.97	0.92

Table 10: **Hedging Percentage Summary**
Hedging % measures the proportion of hedging tokens in the generated reasoning traces indicating uncertain or non-committal language. Highest and Lowest values for every category are highlighted.

Main Effects Overview

	Avg. Hedging %
By Model	
Llama-3.1-8B	**6.73**
Llama-3.3-70B	6.51
Olmo-3-7B	6.50
Olmo-3.1-32B	5.22
By Construct	
Equality	**6.59**
Liberty	6.14
Progress	5.94
By Reasoning Style	
Final Choice	**6.73**
Reasoning First	5.75

Interactions Overview

Model × Reasoning Style

Model	FC Style	RF Style
Llama-3.1-8B	**6.94**	6.52
Llama-3.3-70B	6.94	6.08
Olmo-3-7B	6.23	6.77
Olmo-3.1-32B	6.82	3.62

Model × Construct

Model	Equality	Liberty	Progress
Llama-3.1-8B	7.10	**7.24**	5.77
Llama-3.3-70B	7.09	6.07	6.30
Olmo-3-7B	6.81	6.34	6.32
Olmo-3.1-32B	4.89	4.89	5.36