AI as Respondent: Extending Factorial Survey Methods with LLMs

Chen Peng[1], Arnstein Aassve[1], and Nicolò Cavalli[1]

[1]DONDENA Center for Research on Social Dynamics and Public Policy, Bocconi University, Milan (IT)

November 1, 2025

Abstract

Large language models (LLMs) are increasingly used as "synthetic respondents" in social and population research, yet little is known about whether they reproduce the distributions and heterogeneity characteristic of human judgment. This study evaluates five state-of-the-art models (GPT-4, GPT-4.1, DeepSeek EN/CN, Claude 3.5) using validated factorial survey experiments on family ideals from China (n = 5,186) and the United States (n = 5,906). Models received the original vignette texts and rating instructions used in human surveys, while demographic personas were constructed from real respondent profiles to ensure realistic heterogeneity.

LLMs broadly replicate the direction and rank order of human effects—assigning higher ratings to families with better communication, greater community respect, and higher income—but deviate in three systematic ways. First, response variance is sharply compressed (σ = 1.1–2.1 vs. humans 2.4–2.6), producing over-deterministic predictions with little disagreement. Second, models display normative inflation: they overvalue relational harmony and respect, treating these social-relational cues as moral indicators of "ideal families." Third, subgroup differences are flattened. LLMs cannot uncover the gendered response pattern.

These deviations suggest that uncalibrated LLM simulations may overstate normative consensus and understate contested trade-offs, biasing cross-national and policy analyses. We propose a calibration workflow combining scale alignment, subgroup variance restoration, and human-anchored response priors. By grounding LLM evaluation in validated experimental data, this study reveals both the promise and current limits of using LLMs as credible respondents in demographic and attitudinal research.

1 Introduction

Large Language Models (LLM) increasingly serve as synthetic participants in social science research. Recent work suggests that LLM-generated synthetic data can be biased

that standard evaluation approaches may not easily capture. For instance, Argyle et al. (2023) demonstrate that LLM outputs reflect real-world biases, while Wang et al. (2025) reveal "group flattening": oversimplified representation of demographic groups that diminishes within-group heterogeneity. Hu et al. (2024) show that models inherently exhibit social identity biases, potentially reinforcing existing societal inequalities, and Shrestha et al. (2025) find that while LLMs capture broad aggregate trends, they systematically fail to replicate the nuance and precision characteristic of authentic human responses.

Current evaluation approaches focus on aggregate accuracy metrics, but for the social and political science, more sophisticated evaluations are needed. This paper proposes a metric that is made up of four components: a) magnitude accuracy, b) directional consistency of estimated coefficients, c) pattern preservation, and d) systematic scale bias. In order to demonstrate the usefulness of this evaluation metric, we focus in on a particular methodological approach commonly used in the social and political sciences: Factorial Experiments. This method is extremely important in political science, because it estimates nuanced attitude and behavior patterns. We use a particular study of family ideals vignettes to demonstrate the usefulness of the composite evaluation metric. Aassve et al. (2024) surveyed 20,141 respondents across eight countries. We use the resulting data from this survey to evaluate the extent LLMs can authentically replicate established human preference patterns when defined over a multi-dimensional concept. This provides a rigorous methodological foundation: we leverage validated cross-cultural factorial survey experiments (Auspurg and Hinz, 2015) that systematically vary family characteristics across ten dimensions (i.e. factors).

Our methodological innovation centers on experimental fidelity rather than prompt optimization. We preserve the exact vignette text and rating instructions presented to human respondents, adding only demographic persona information derived from authentic participant profiles. This approach eliminates prompt engineering confounds and provides unbiased assessment of model suitability for real research applications—testing whether LLMs can handle authentic research protocols rather than engineered tasks.

We employ five state-of-the-art LLMs (GPT-4, GPT-4.1, DeepSeek EN/CN, Claude

3.5) across two culturally distinct contexts—China and the United States—to examine whether LLMs can authentically reproduce human evaluative patterns in a validated factorial survey design. Our contributions are threefold.

First, we provide the first large-scale cross-national benchmark comparing human and model-generated responses under identical experimental conditions. By holding vignette text and rating instructions constant and introducing realistic demographic personas, we isolate the model's internal representational behavior from prompt engineering effects.

Second, we show that successful replication of social judgment tasks depends on capturing human variability—the range and structure of disagreement—rather than merely recovering average effects. LLMs reproduce directional patterns but systematically compress variance, overvalue relational harmony, and misrepresent subgroup heterogeneity.

Third, we document systematic and interpretable distortions: models consistently under-predict married individuals, neutralize gender asymmetries in work–family conflict, and inflate relational attributes such as family communication and community respect, effectively moralizing social cohesion. These deviations have direct implications for how population scholars interpret synthetic data, especially when studying attitudinal polarization or norm pluralism.

Taken together, our study offers an empirical template for evaluating LLMs as "synthetic respondents," demonstrates key biases invisible to standard accuracy metrics, and establishes practical guidelines for calibrating model outputs to preserve human-like heterogeneity essential for demographic inference.

2 Methodology

Experimental Design and Data. We analyze vignette experiments from China (n=5,186) and the United States (n=5,906) following Aassve et al. (2024). Respondents evaluated systematically varied family scenarios on a 0–10 scale of "successful family." Eight attributes were manipulated: (1) union status (married/cohabiting), (2) number of children (0–3+), (3) income (below/average/above), (4) family communication (poor/good), (5) grandparent contact (infrequent/frequent), (6) community respect (not respected/respected), (7) gender roles (traditional/commonplace/egalitarian), and (8) work–family conflict (male/female/neither/both conflicted). We estimated Average Component Marginal Effects (ACMEs) via fixed-effects models.

LLMs Prompt Protocol. LLMs received the exact vignette text and rating instructions as human respondents ("On a scale of 0–10, to what extent does this describe a successful family?"). To simulate realistic heterogeneity, we incorporated demographic personas directly from authentic survey profiles (e.g., age, education, income, family structure). This ensures evaluation reflects real-world respondent distributions rather than engineered prompts.

3 Results

3.1 Mean and Variances: Lack of Variability

Our analysis reveals a counter-intuitive finding that challenges standard evaluation approaches: successful replication requires capturing human disagreement patterns, not just central tendencies. Table 1 and Figure 1 demonstrate this fundamental limitation—human ratings exhibit natural variance (σ = 2.4-2.6) with full use of the 0-10 scale, while all LLMs show severely constrained distributions. DeepSeek models are more constrained (σ = 1.1-1.3, limited to 3-8 range), while GPT-4 comes closer to human variability (σ = 2.0).

This variability constraint represents a systematic failure across all tested models—even those with accurate mean predictions miss the natural human response patterns essential for valid social science applications. Models appear to be overly confident, clustering ratings in narrow ranges rather than reflecting authentic human disagreement about complex social judgments.

3.2 Gender differences in response variance

We compute, for human Ground Truth (GT) and the LLMs' simulations, the within-person standard deviatio (SD) of vignette ratings and classify it as either above or below the gender-specific median of the rating as successful family. We then regress the respondent's mean vignette rating on `gender` × response variability. The combined margins plots show a gendered pattern that LLMs largely fail to reproduce.

Human Ground Truth. Panel GT of Figure 2 shows that, female participants whose variability of responses are above-median give *substantially lower* mean family-success

ratings than women with below-median variability, whereas men exhibit a more modest change. This suggests higher variability is linked to lower rating of the vignette, in particular for women.

LLMs. Across Panels ds_en, ds_cn, gpt4, gpt4.1 and claude3.5 of Figure 2 , the slopes are near flat for both genders and the female–male gap is markedly attenuated. Panel claude3.5 of Figure 2 shows that Claude 3.5 is an outlier among the LLMs, displaying a reversed pattern for women (higher means at higher SD), but the overall LLM tendency is to *dampen* or *erase* the GT interaction.

In sum, Humans (GT) show a sizable gender difference in how decisional variability maps onto family ideals - women's ratings drop when within-person dispersion is high - whereas most LLMs produce little to no gender differentiation on this dimension. In short, LLMs *flatten* the GT gender interaction between variability and evaluation, with one model (Claude 3.5) deviating in the opposite direction.

3.3 Relative Importance of Factors for Family Ideal: AMCEs

Across models, LLMs reproduce the *direction* of core effects found in Human_GT for the U.S. sample (see Figure3 Table2)—i.e., positive valuation of good communication, community respect, and higher income, and negative valuation of lower income and childlessness. However, model fits for LLMs exhibit very high R^2 (0.89–0.97) compared to humans (≈ 0.51), indicating *over-deterministic* responses with attenuated disagreement/variance.

Moralized/prosocial attributes are overweighted Relative to Human_GT, all models inflate coefficients for *Communicates well* and *Respected in community*. In Table2, Human_GT estimates are ≈ 1.00 (communication) and 0.72 (community), whereas LLMs range from $\approx$ 1.39–3.66 (communication) and 1.11–1.53 (community). GPT-4 / GPT-4.1 display the largest inflation, turning normative preferences into near-deterministic signals.

Family structure and fertility cues Humans penalize cohabitation slightly (-0.11 in Table2); most LLMs mirror the sign but some, notably GPT-4.1 and Claude 3.5, *exaggerate* the penalty (down to -0.22 and -0.81, respectively). Childlessness is penalized by all agents, but its magnitude varies: some models under-penalize (e.g., DeepSeek EN/CN

−0.11 to −0.19), while GPT-4.1 over-penalizes (e.g., −0.50). LLMs also treat "more children" as uniformly positive, missing the human nuance (little separation between two vs. three children).

Socioeconomic gradient All models capture the negative effect of *Lower income* and the positive effect of *Higher income*. Yet sensitivity to deprivation is generally *attenuated* vs. Human_GT: LLMs typically produce smaller absolute penalties for low income and sometimes amplify the reward to higher income (e.g., DeepSeek EN).

Gender roles and work–family conflict LLMs preserve the human hierarchy that favors *Egalitarian* over *Traditional* roles, but tend to *amplify* the egalitarian premium (e.g., DeepSeek EN ≈ 0.86–0.97; Claude 3.5 ≈ 0.59–1.08). For work–family conflict, all models reproduce the ranking (*Neither conflicted* > *Male/Female conflicted*), yet smooth or flip smaller asymmetries: *Male conflicted* becomes uniformly positive; *Female conflicted* sometimes turns positive or weakly negative depending on model, effectively *neutralizing gender gaps* visible in human data.

Baseline shift and scale Intercepts for LLMs are much lower (roughly 2.4–3.6) than for humans (≈ 4.2–4.9), implying a lower baseline. Together with steeper coefficients on moralized traits, this yields a *steeper moral gradient*: "good" families get very high scores; "bad" families very low scores—stronger than in human ratings.

In sum, LLMs replicate the *logical structure* of human judgment (signs and rank ordering) but deviate in *moral intensity* (inflated prosocial traits), *variance structure* (over-determinism), and *gender asymmetry* (neutralized). They resemble a normatively idealized "average" rater rather than the heterogeneous distribution of human respondents.

3.4 Gender Differences in Family Ideals and LLM Representation

As shown in Tables 4–7, the ground-truth (GT) estimates reveal clear gendered distinctions in family ideals. Consistent with prior results (**??**), women place greater emphasis on interpersonal and process-oriented dimensions of family life, communication ($\beta = 1.04$, CI [1.00, 1.08]), partner harmony ($\beta = 0.63$, CI [0.56, 0.71]), and extended family contact

($\beta = 0.50$, CI [0.46, 0.54]), while attaching relatively less importance to formal markers such as marriage ($\beta = 0.16$, CI [0.12, 0.20]) and parenthood ($\beta = 0.29$, CI [0.23, 0.35]). Women also exhibit stronger preferences for egalitarian gender roles ($\beta = 0.33$, CI [0.27, 0.39]) compared to men ($\beta = 0.17$, CI [0.11, 0.23]), suggesting a conditional approach to family formation in which parenthood and marriage are valued when gender equality and emotional quality are ensured. Men, in contrast, maintain a relatively stronger attachment to parenthood and less emphasis on relational quality.

When comparing these patterns to the LLM outputs in Tables 4–7, several deviations emerge. All LLMs reproduce the broad structure of GT preferences but display consistent *gender asymmetry* in the interpretation of work–family conflict: they reward *Male conflicted* profiles across contexts but diverge sharply in treating *Female conflicted* ones. In particular, GPT-4.1 systematically penalizes *Female conflicted* (e.g., $\beta = -0.25$ to -0.30, all $p < 0.01$), while GPT-4 yields small positive effects ($\beta \approx 0.12$), and Claude and DeepSeek produce mixed or null results. These patterns suggest that LLMs disproportionately associate father role strain with responsibility, whereas similar conflict in mothers is interpreted as failure of balance or competence.

Furthermore, LLMs *amplify* the moral contrast between egalitarian and traditional gender roles: the positive effect of egalitarianism is roughly doubled relative to GT, while traditional roles are either weakly negative or inconsistent. This indicates a normative flattening of gender ideology, models exaggerate the polarization between "modern egalitarian" and "traditional" families that human respondents perceive more gradationally.

According to Shrestha et al. (2025) , such distortions are expected: LLMs often generate *out-group imitations* rather than in-group representations and *flatten* identity heterogeneity by collapsing diverse perspectives into a single normative script. In this case, the LLMs appear to "speak for" gender groups rather than represent them, overemphasizing gender-role coherence and underrepresenting within-gender variation that human data clearly preserve.

In sum, while GT data reveal that women's family ideals prioritize communication and equality, LLM-generated judgments tend to oversimplify these gendered patterns—rewarding male work–family conflict, penalizing female conflict, and exaggerating ideological polarization. These findings underscore the caution raised by Shrestha et al. (2025) : when LLMs are used to simulate human preferences, their representations can

reproduce societal asymmetries and flatten nuanced gender differences rather than authentically mirror them.

4 Discussion and Implications

From mean accuracy to distributional fidelity. Our results suggest that "replication" in social judgement tasks hinges less on mean accuracy than on whether models reproduce *distributions*: the shape of responses, their dispersion within persons, and their heterogeneity across groups. LLMs largely recover the *direction* and rank ordering of Average Component Marginal Effects (ACMEs), but compress variance and overweight moralized/prosocial cues. This yields high R^2 with narrow residuals, a lower baseline intercept, and muted subgroup interactions: a profile consistent with safety tuning and generic helpfulness priors rather than human-like heterogeneity.

Substantive inference at risk of moral inflation. Overweighting of communication and community respect raises a concrete risk: LLM outputs describe a normatively idealized society with excess consensus on "virtue" attributes. In applied work, this can produce (i) overstated returns to soft norms (e.g., communication quality) and (ii) understated trade-offs (e.g., income deprivation, role conflict). Uncalibrated simulations may thus overestimate policy support or the ease of norm change.

Heterogeneity matters. Variance compression smooths away genuine disagreement, the very signal on which population research depends. In demography, small subgroup differences often accumulate into macro-level patterns, such as fertility intentions by socioeconomic status. When these gradients are flattened, counterfactual projections become systematically biased. In political science, where conjoint experiments are widely used to study voter decision-making, heterogeneity by partisanship is often central. For example, how partisans weigh candidate gender, ethnicity, or immigration stance differently (Hainmueller et al., 2015). If such subgroup interactions are muted, the resulting model risks erasing the cleavages that structure political preferences. Capturing heterogeneity is therefore not a technical refinement but a prerequisite for valid inference in understanding value complexity and social change.

A measurement-model perspective. Viewed through the lens of measurement, the 0-10 ratings are bounded, ordinal-like responses shaped by (i) a latent stance and (ii) a respondent-specific response style (central tendency, extremity avoidance, dispersion). LLMs emulate (i) but not (ii). Two implications follow: (1) *Scale calibration* can be formalized as a monotone map (e.g., isotonic or beta regression on the bounded scale) plus a variance rescaling; (2) *Response-style priors* can be learned from a few human anchors and then used to condition subsequent LLM predictions, partially restoring within-person dispersion.

Cross-national and policy analytics. Because variance compression can mimic cross-country convergence, cross-national contrasts should be reported *pre* and *post* calibration, with sensitivity to dispersion corrections. For policy simulations (e.g., messaging on gender equality), require evidence that subgroup interactions replicate human structure; otherwise, communicate that the exercise estimates reactions of an "optimistic average rater" rather than a population.

Ethics, transparency, and governance. LLM-based social measurement inherits model-side priors and safety filters. We recommend: (i) publishing a brief "calibration plan" alongside code; (ii) logging model version/seed/temperature; (iii) running change-detection tests whenever the provider updates models; and (iv) documenting where calibration meaningfully alters conclusions. These steps are minimal to keep synthetic-respondent studies auditable and reproducible.

5 Limitations and Further Work

Limitations

Scope. We study one multi-attribute domain (family ideals) in two contexts (USA, China). While this design is rich and policy-relevant, external validity to other domains (e.g., immigration, public health) and settings remains to be established.

Protocol realism. We use closed-ended ratings with fixed instructions. Real surveys include open-ended justifications, time pressure, satisficing, interviewer effects, and device frictions that our protocol only partially mimics. Personas capture sociodemographics

but not latent traits (e.g., Big Five, local norms), which are known to influence response styles.

Model-side priors. Observed moral inflation and variance compression likely reflect pretraining mixtures, safety tuning, and decoding defaults. These properties can change with silent provider updates. Conclusions therefore depend on versioning and should be re-verified over time.

Future Work

1. **Response-style priors (few-shot human anchors).** Elicit each respondent's first K vignette ratings to estimate individual tendencies such as mean level, extremity avoidance, and dispersion. Condition subsequent LLM predictions on these priors—via prompt conditioning or post-hoc shrinkage—to restore realistic within-person variance and interaction patterns.

2. **Noise injection under sign and rank constraints.** Introduce calibrated residuals ϵ so that $\text{sign}(\widehat{\beta})$ and the rank ordering of factor levels remain preserved, while human-like variance and kurtosis are recovered. Compare Gaussian jitter on the latent scale with tempered sampling (temperature/top-p) matched to empirical dispersion.

3. **Adversarial and counterfactual vignettes.** Design stress tests by flipping one attribute at a time near ambiguous boundaries (e.g., moderate income, dual work–family conflict). Examine how sensitive each model is to these controlled perturbations, identifying dimensions where LLMs systematically over- or underweight attributes relative to humans.

4. **Cross-model ensembling.** Combine models with complementary biases (e.g., one underweights income penalties, another overemphasizes relational harmony) to expand response diversity and reduce idiosyncratic distortions. Evaluate whether ensembles yield more realistic distributions and subgroup heterogeneity.

5. **Prompt and persona refinement.** Use explicit role prompts to elicit gendered or subgroup-specific perspectives (e.g., generate separate male and female responses). Enrich persona construction with additional survey-based attributes—such as response

time, satisfaction levels, or household composition—to better approximate real respondent variability and survey behavior.

Table 1: Summary Statistics Reveal Gap In Response Variances

Model	China (n=5,186)			USA (n=5,906)		
	Mean	Std	Range	Mean	Std	Range
Ground Truth	5.27	**2.40**	0-10	5.88	**2.64**	0-10
DeepSeek (CN)	5.63	**1.10**	3-8	5.63	**1.15**	3-9
DeepSeek (EN)	5.45	**1.17**	3-9	5.37	**1.29**	3-9
GPT-4	5.52	**2.02**	2-9	5.52	**2.07**	2-10
Claude 3.5	4.48	**1.34**	2-9	5.00	**1.56**	2-10

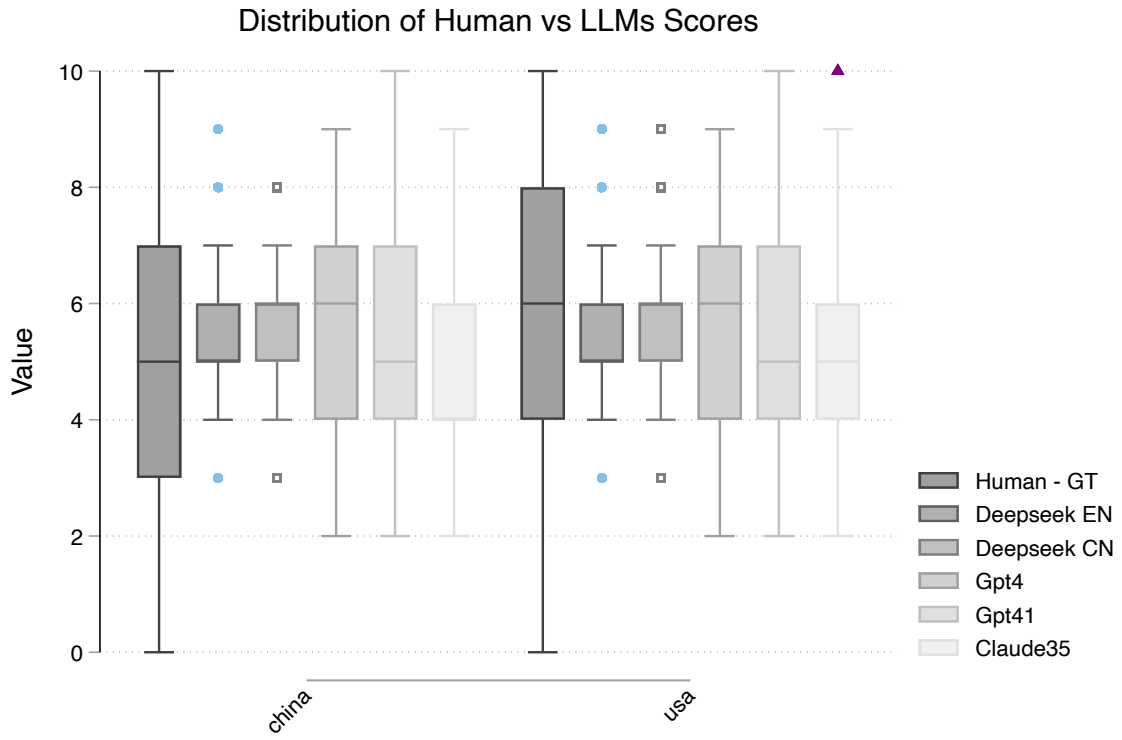

Figure 1: Distribution of Ground Truth and LLM Ratings

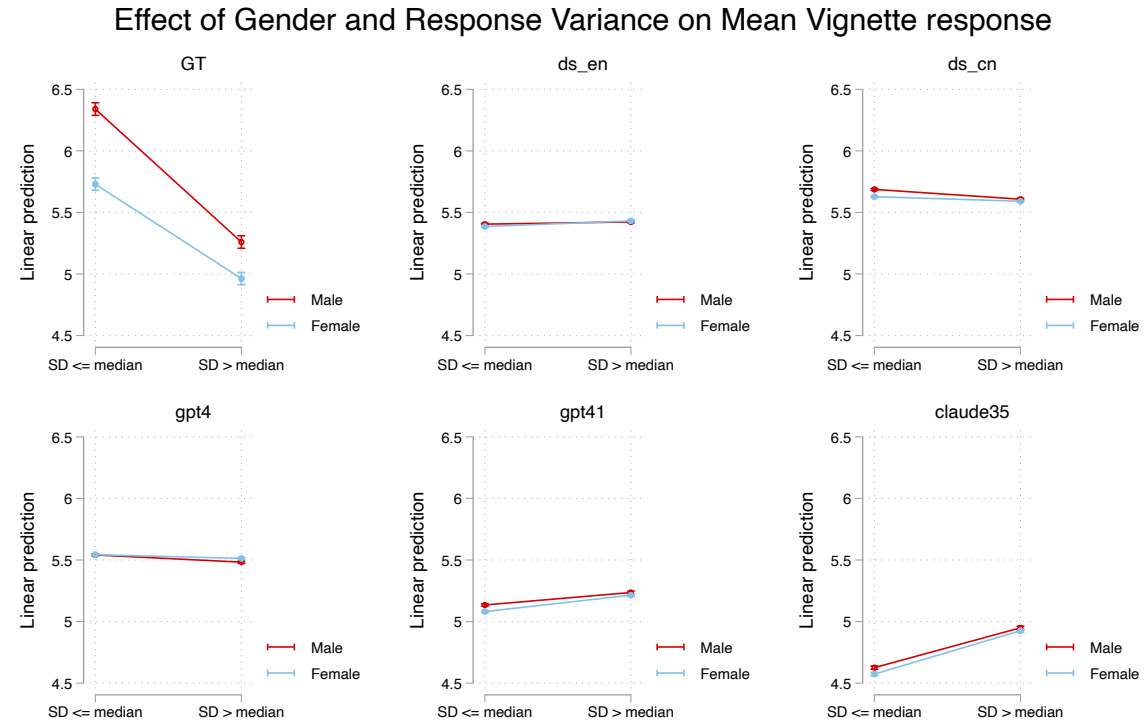

Figure 2: Gender Difference in Response Style

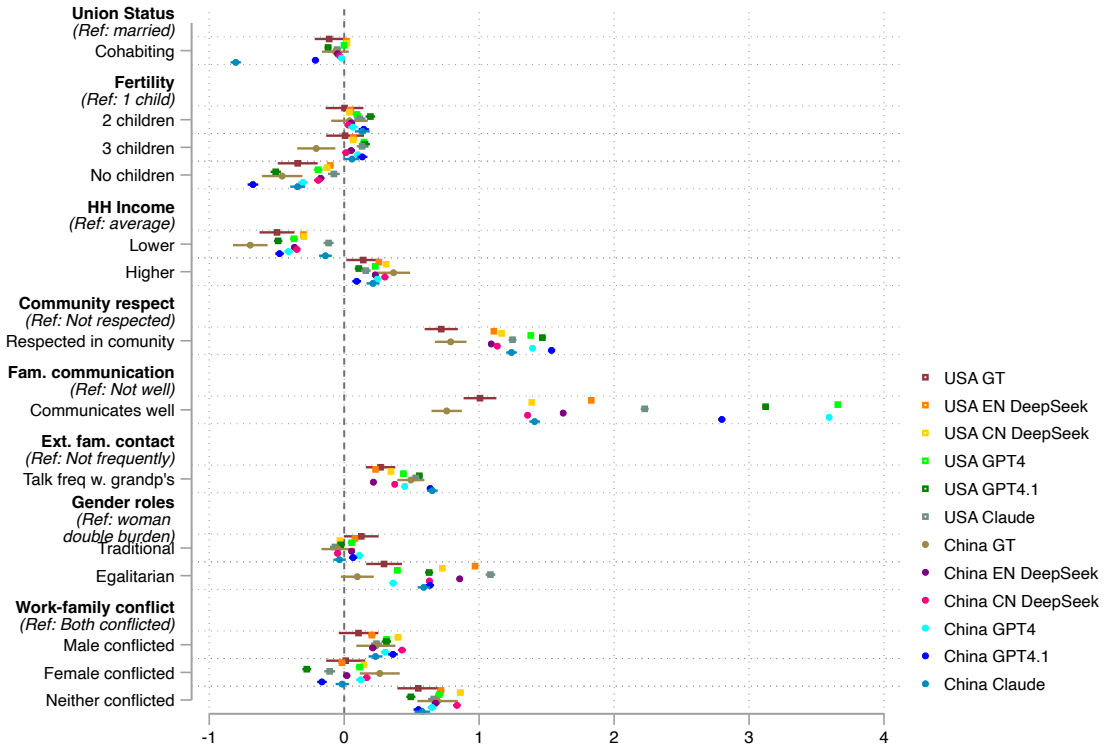

Figure 3: AMCEs - Fixed Effect Models between Human Ground Truth (GT) and LLMs

Table 2: Average Marginal Component Effects Fixed Effect Model - USA

	(1) Human$_G T$ b/se	(2) deepseek_en_usa b/se	(3) deepseek_cn_usa b/se	(4) gpt4_usa b/se	(5) gpt41_usa b/se	(6) claude35_usa b/se
Cohabiting	-0.11*	0.02	0.02*	-0.00	-0.12**	-0.05**
	(0.05)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)
2 children	-0.01	0.05**	0.04**	0.09**	0.20**	0.12**
	(0.07)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)
3 children	-0.00	0.07**	0.07**	0.14**	0.15**	0.14**
	(0.08)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)
No children	-0.36**	-0.11**	-0.13**	-0.19**	-0.50**	-0.07**
	(0.08)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)
Lower	-0.50**	-0.30**	-0.30**	-0.37**	-0.49**	-0.12**
	(0.06)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)
Higher	0.14*	0.26**	0.31**	0.23**	0.11**	0.16**
	(0.06)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)
Respected in comunity	0.72**	1.11**	1.17**	1.38**	1.47**	1.25**
	(0.05)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)
Communicates well	1.00**	1.83**	1.39**	3.66**	3.12**	2.22**
	(0.05)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Talk freq w. grandp's	0.27**	0.23**	0.34**	0.44**	0.56**	0.53**
	(0.05)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)
Traditional	0.13*	0.08**	-0.03*	0.06**	-0.02	-0.07**
	(0.06)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)
Egalitarian	0.29**	0.97**	0.73**	0.40**	0.63**	1.08**
	(0.07)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)
Male conflicted	0.10	0.20**	0.40**	0.31**	0.33**	0.24**
	(0.07)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)
Female conflicted	-0.00	-0.02	0.15**	0.12**	-0.27**	-0.11**
	(0.08)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)
Neither conflicted	0.55**	0.71**	0.86**	0.72**	0.49**	0.66**
	(0.08)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)
_cons	4.85**	3.22**	3.59**	2.38**	2.51**	2.43**
	(0.10)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)
Observations	5906	5906	5906	5906	5906	5906
R^2	0.514	0.932	0.916	0.966	0.945	0.888

Note: Participant FE, vignette order FE. * p<0.05, ** p<0.01

Table 3: Average Marginal Component Effects Fixed Effect Model - China

	(1) Human$_G T$ b/se	(2) deepseek_en_usa b/se	(3) deepseek_cn_usa b/se	(4) gpt4_usa b/se	(5) gpt41_usa b/se	(6) claude35_usa b/se
Cohabiting	-0.06	-0.05**	-0.03**	-0.02	-0.22**	-0.81**
	(0.05)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)
2 children	0.06	0.05**	0.03	0.06**	0.14**	0.13**
	(0.07)	(0.01)	(0.02)	(0.02)	(0.02)	(0.03)
3 children	-0.19*	0.05**	0.01	0.09**	0.13**	0.05
	(0.07)	(0.02)	(0.01)	(0.02)	(0.02)	(0.03)
No children	-0.44**	-0.17**	-0.19**	-0.30**	-0.68**	-0.36**
	(0.07)	(0.01)	(0.02)	(0.02)	(0.02)	(0.03)
Lower	-0.70**	-0.37**	-0.35**	-0.41**	-0.48**	-0.14**
	(0.06)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)
Higher	0.37**	0.23**	0.30**	0.25**	0.09**	0.21**
	(0.06)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)
Respected in comunity	0.79**	1.09**	1.13**	1.39**	1.53**	1.24**
	(0.05)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)
Communicates well	0.76**	1.62**	1.36**	3.59**	2.80**	1.41**
	(0.05)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)
Talk freq w. grandp's	0.49**	0.21**	0.37**	0.45**	0.64**	0.65**
	(0.05)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)
Traditional	-0.05	0.06**	-0.05**	0.11**	0.07**	-0.04
	(0.06)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)
Egalitarian	0.10	0.86**	0.64**	0.36**	0.64**	0.59**
	(0.06)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)
Male conflicted	0.25**	0.21**	0.42**	0.30**	0.38**	0.23**
	(0.07)	(0.01)	(0.02)	(0.02)	(0.02)	(0.03)
Female conflicted	0.28**	0.01	0.18**	0.13**	-0.16**	-0.02
	(0.08)	(0.01)	(0.02)	(0.02)	(0.02)	(0.03)
Neither conflicted	0.70**	0.67**	0.83**	0.66**	0.56**	0.58**
	(0.07)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)
_cons	4.22**	3.54**	3.71**	2.47**	2.56**	2.87**
	(0.10)	(0.02)	(0.02)	(0.02)	(0.02)	(0.04)
Observations	4935	4935	4935	4935	4935	4935
R^2	0.539	0.920	0.909	0.964	0.941	0.783

Note: Participant FE, vignette order FE. * p<0.05, ** p<0.01

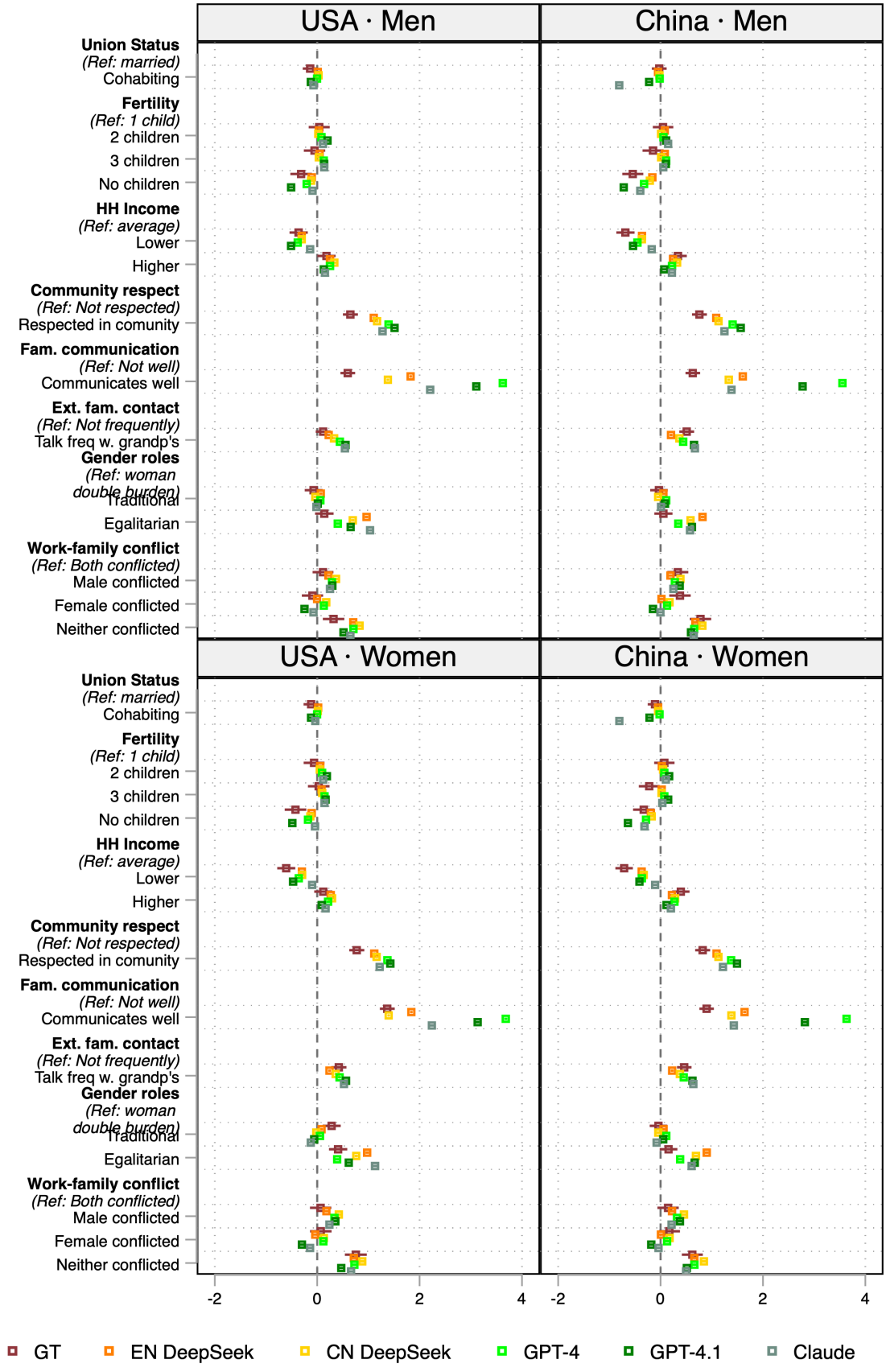

Figure 4: AMCEs - Fixed Effect Models between Human Ground Truth (GT) and LLMs by Gender

Table 4: Average Marginal Component Effects Fixed Effect Model — USA · Men

	(1)	(2)	(3)	(4)	(5)	(6)
	Human GT	deepseek_en_usa	deepseek_cn_usa	gpt4_usa	gpt41_usa	claude35_usa
	b/se	b/se	b/se	b/se	b/se	b/se
Cohabiting	-0.14	0.01	0.03*	-0.00	-0.12**	-0.07**
	(0.07)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)
2 children	0.04	0.03	0.02	0.08**	0.20**	0.11**
	(0.10)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)
3 children	-0.05	0.05*	0.03	0.13**	0.14**	0.14**
	(0.11)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)
No children	-0.31**	-0.11**	-0.12**	-0.21**	-0.51**	-0.09**
	(0.11)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)
Lower	-0.36**	-0.30**	-0.30**	-0.38**	-0.51**	-0.14**
	(0.09)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)
Higher	0.18*	0.25**	0.34**	0.25**	0.13**	0.15**
	(0.09)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)
Respected in comunity	0.65**	1.10**	1.17**	1.39**	1.51**	1.28**
	(0.07)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)
Communicates well	0.60**	1.82**	1.38**	3.63**	3.11**	2.21**
	(0.07)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)
Talk freq w. grandp's	0.12	0.22**	0.33**	0.44**	0.55**	0.54**
	(0.07)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)
Traditional	-0.07	0.07**	-0.04*	0.06**	0.01	-0.01
	(0.09)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)
Egalitarian	0.14	0.96**	0.69**	0.40**	0.65**	1.03**
	(0.09)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)
Male conflicted	0.11	0.22**	0.37**	0.29**	0.30**	0.25**
	(0.10)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)
Female conflicted	-0.09	-0.00	0.18**	0.13**	-0.25**	-0.07*
	(0.11)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)
Neither conflicted	0.32**	0.70**	0.83**	0.71**	0.51**	0.65**
	(0.11)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)
_cons	5.69**	3.24**	3.64**	2.39**	2.50**	2.45**
	(0.15)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)
Observations	2810	2810	2810	2810	2810	2810
R^2	0.518	0.932	0.916	0.964	0.948	0.896

Note: Participant FE, vignette order FE. * p<0.05, ** p<0.01

Table 5: Average Marginal Component Effects Fixed Effect Model — USA · Women

	(1) Human GT b/se	(2) deepseek_en_usa b/se	(3) deepseek_cn_usa b/se	(4) gpt4_usa b/se	(5) gpt41_usa b/se	(6) claude35_usa b/se
Cohabiting	-0.12	0.02	0.01	-0.00	-0.12**	-0.04
	(0.08)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)
2 children	-0.06	0.06**	0.05**	0.10**	0.19**	0.12**
	(0.10)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)
3 children	0.03	0.09**	0.10**	0.14**	0.16**	0.14**
	(0.11)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)
No children	-0.42**	-0.11**	-0.13**	-0.18**	-0.49**	-0.04
	(0.11)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)
Lower	-0.60**	-0.30**	-0.29**	-0.36**	-0.47**	-0.10**
	(0.09)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)
Higher	0.12	0.26**	0.29**	0.21**	0.09**	0.16**
	(0.09)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)
Respected in comunity	0.77**	1.12**	1.16**	1.37**	1.43**	1.22**
	(0.08)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)
Communicates well	1.37**	1.84**	1.40**	3.68**	3.13**	2.24**
	(0.07)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)
Talk freq w. grandp's	0.42**	0.24**	0.36**	0.44**	0.56**	0.52**
	(0.08)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)
Traditional	0.28**	0.09**	-0.02	0.05**	-0.06**	-0.12**
	(0.09)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)
Egalitarian	0.41**	0.98**	0.76**	0.39**	0.62**	1.13**
	(0.09)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)
Male conflicted	0.07	0.18**	0.42**	0.34**	0.36**	0.24**
	(0.11)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)
Female conflicted	0.07	-0.04*	0.13**	0.12**	-0.30**	-0.14**
	(0.11)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)
Neither conflicted	0.75**	0.72**	0.88**	0.73**	0.47**	0.66**
	(0.11)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)
_cons	4.13**	3.20**	3.54**	2.37**	2.51**	2.41**
	(0.14)	(0.02)	(0.03)	(0.03)	(0.03)	(0.04)
Observations	3096	3096	3096	3096	3096	3096
R^2	0.510	0.933	0.918	0.967	0.944	0.884

Note: Participant FE, vignette order FE. * p<0.05, ** p<0.01

Table 6: Average Marginal Component Effects Fixed Effect Model — China · Men

	(1) Human GT b/se	(2) deepseek_en_usa b/se	(3) deepseek_cn_usa b/se	(4) gpt4_usa b/se	(5) gpt41_usa b/se	(6) claude35_usa b/se
Cohabiting	-0.03	-0.05**	-0.02	-0.02	-0.22**	-0.81**
	(0.07)	(0.01)	(0.01)	(0.02)	(0.02)	(0.03)
2 children	0.05	0.08**	0.01	0.06*	0.11**	0.15**
	(0.10)	(0.02)	(0.02)	(0.02)	(0.03)	(0.04)
3 children	-0.15	0.08**	0.01	0.11**	0.10**	0.05
	(0.10)	(0.02)	(0.02)	(0.02)	(0.03)	(0.04)
No children	-0.54**	-0.16**	-0.21**	-0.32**	-0.72**	-0.40**
	(0.10)	(0.02)	(0.02)	(0.03)	(0.03)	(0.04)
Lower	-0.69**	-0.36**	-0.36**	-0.45**	-0.54**	-0.17**
	(0.09)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)
Higher	0.34**	0.24**	0.32**	0.22**	0.07**	0.22**
	(0.09)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)
Respected in comunity	0.76**	1.09**	1.13**	1.41**	1.57**	1.25**
	(0.07)	(0.01)	(0.01)	(0.02)	(0.02)	(0.03)
Communicates well	0.63**	1.61**	1.33**	3.55**	2.77**	1.39**
	(0.07)	(0.01)	(0.01)	(0.02)	(0.02)	(0.03)
Talk freq w. grandp's	0.51**	0.20**	0.37**	0.44**	0.65**	0.67**
	(0.07)	(0.01)	(0.01)	(0.02)	(0.02)	(0.03)
Traditional	-0.03	0.06**	-0.05**	0.11**	0.09**	0.01
	(0.09)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)
Egalitarian	0.06	0.82**	0.59**	0.35**	0.61**	0.58**
	(0.09)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)
Male conflicted	0.34**	0.19**	0.39**	0.28**	0.38**	0.25**
	(0.10)	(0.02)	(0.02)	(0.02)	(0.03)	(0.04)
Female conflicted	0.38**	0.02	0.18**	0.13**	-0.15**	-0.00
	(0.11)	(0.02)	(0.02)	(0.02)	(0.03)	(0.04)
Neither conflicted	0.78**	0.68**	0.82**	0.66**	0.60**	0.65**
	(0.11)	(0.02)	(0.02)	(0.02)	(0.03)	(0.04)
_cons	4.31**	3.55**	3.78**	2.50**	2.60**	2.88**
	(0.15)	(0.03)	(0.03)	(0.03)	(0.03)	(0.05)
Observations	2541	2541	2541	2541	2541	2541
R^2	0.521	0.921	0.910	0.965	0.944	0.790

Note: Participant FE, vignette order FE. * p<0.05, ** p<0.01

Table 7: Average Marginal Component Effects Fixed Effect Model — China · Women

	(1) Human GT b/se	(2) deepseek_en_usa b/se	(3) deepseek_cn_usa b/se	(4) gpt4_usa b/se	(5) gpt41_usa b/se	(6) claude35_usa b/se
Cohabiting	-0.11	-0.05**	-0.05**	-0.02	-0.22**	-0.80**
	(0.07)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)
2 children	0.07	0.03	0.05*	0.07**	0.17**	0.10*
	(0.10)	(0.02)	(0.02)	(0.02)	(0.03)	(0.04)
3 children	-0.22*	0.02	0.02	0.07**	0.15**	0.04
	(0.10)	(0.02)	(0.02)	(0.02)	(0.03)	(0.04)
No children	-0.33**	-0.19**	-0.17**	-0.28**	-0.64**	-0.31**
	(0.11)	(0.02)	(0.02)	(0.03)	(0.03)	(0.04)
Lower	-0.71**	-0.37**	-0.33**	-0.36**	-0.41**	-0.11**
	(0.09)	(0.02)	(0.02)	(0.02)	(0.02)	(0.04)
Higher	0.40**	0.22**	0.28**	0.27**	0.12**	0.20**
	(0.09)	(0.02)	(0.02)	(0.02)	(0.02)	(0.04)
Respected in comunity	0.82**	1.09**	1.13**	1.38**	1.49**	1.22**
	(0.07)	(0.01)	(0.02)	(0.02)	(0.02)	(0.03)
Communicates well	0.90**	1.64**	1.38**	3.63**	2.82**	1.43**
	(0.07)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)
Talk freq w. grandp's	0.46**	0.23**	0.38**	0.45**	0.62**	0.64**
	(0.07)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)
Traditional	-0.04	0.06**	-0.04*	0.11**	0.05*	-0.08*
	(0.09)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)
Egalitarian	0.16	0.90**	0.69**	0.38**	0.67**	0.61**
	(0.09)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)
Male conflicted	0.15	0.22**	0.46**	0.32**	0.38**	0.22**
	(0.11)	(0.02)	(0.02)	(0.02)	(0.03)	(0.04)
Female conflicted	0.17	0.01	0.18**	0.13**	-0.18**	-0.04
	(0.10)	(0.02)	(0.02)	(0.03)	(0.03)	(0.04)
Neither conflicted	0.62**	0.66**	0.85**	0.66**	0.52**	0.50**
	(0.10)	(0.02)	(0.02)	(0.02)	(0.03)	(0.04)
_cons	4.11**	3.54**	3.64**	2.44**	2.52**	2.87**
	(0.14)	(0.03)	(0.03)	(0.03)	(0.04)	(0.05)
Observations	2394	2394	2394	2394	2394	2394
R^2	0.562	0.920	0.909	0.963	0.940	0.779

Note: Participant FE, vignette order FE. * p<0.05, ** p<0.01

References

A. Aassve, A. Adserà, P. Y. Chang, L. Mencarini, H. Park, C. Peng, S. Plach, J. M. Raymo, S. Wang, and W.-J. J. Yeung. Family ideals in an era of low fertility. *Proceedings of the National Academy of Sciences*, 121(6):e2311847121, 2024. doi: 10.1073/pnas.2311847121.

A. Agresti. *An Introduction to Categorical Data Analysis*. John Wiley & Sons, 3rd edition, 2018.

L. P. Argyle, E. C. Busby, N. Fulda, J. Gubler, C. Rytting, and D. Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3): 317–334, 2023. Preprint available at arXiv:2209.06899, 2022.

K. Auspurg and T. Hinz. *Factorial Survey Experiments*. Quantitative Applications in the Social Sciences. SAGE Publications, Thousand Oaks, CA, 2015.

J. Hainmueller, D. Hangartner, and T. Yamamoto. Candidates' ethnic identity and voter choice in britain. *American Journal of Political Science*, 59(3):529–548, 2015. doi: 10.1111/ajps.12168.

T. Hu, N. Levy, M. Lazar, S. Harrell, K. Vafa, C. A. Hidalgo, and H. Shirado. Generative language models exhibit social identity biases. *Nature Computational Science*, 4(11): 741–751, 2024. doi: 10.1038/s43588-024-00741-1.

P. Shrestha, D. Krpan, F. Koaik, R. Schnider, D. Sayess, and M. S. Binbaz. Beyond weird: Can synthetic survey participants substitute for humans in global policy research? *Behavioral Science & Policy*, 2025.

A. Wang, J. Morgenstern, and J. P. Dickerson. Large language models as replacements for human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, 2025. doi: 10.1038/s42256-025-00986-z.

Appendix

LLMs Replication Details

Model Specifications. We evaluated five state-of-the-art models: GPT-4 (March 2023, 8K-32K context), GPT-4.1 (April 2025, 1M context), DeepSeek V3 EN/CN (December 2024, 128K context), and Claude 3.5 Sonnet (June 2024, 200K context). Models were selected to cover capability spectrum, cultural alignment, and cost considerations.

Persona Construction. Demographic personas were derived directly from authentic participant profiles in the original PNAS dataset, maintaining realistic combinations of demographic characteristics. Each persona includes age, education level, income, family structure, employment status, and cultural context drawn from actual survey respondents.

Statistical Testing Framework. Demographic bias detection employed Fisher's exact tests across 567 comparisons with Bonferroni correction (p-values multiplied by number of tests). We used 2×2 contingency tables (Agresti, 2018) comparing each demographic group's representation in bias categories versus background population, with odds ratios calculated as OR = ad/bc.

Sample prompt

Listing 1: User Prompt Configuration

```
user_prompt = (
 "**Task**: You will be presented with six descriptions of
    different families. "
"Please answer the questions following the description of each
   family. "
"Please take your time picturing each family while reading the
   description.\n"
f"{vignette}\n"
"**Question**: Based on this description, on a scale of 0 (0 =
   Strongly Disagree) to 10 (10 = Strongly Agree) "
"to what extent do you agree or disagree with the following
   statement? "
'"This describes a successful family." '
"Combining the description and your own characteristics, please
   respond with a single number between 0-10, no explanation just
    number.\n**Answer**: "
)
```

Examples of Persona and Vignette

PERSONA:

"You are a respondent in the LifePoints platform who answers surveys for money. You are currently answering a comparative survey about family ideals that is conducted in 8 countries including the USA, Italy, Spain, Norway, urban China, Singapore, Japan, and Korea. You are a participant located in China. You have the following characteristics.. You are a 41.0 year-old male. currently living in big city and its suburbs. Your race is 999. and your ethnicity is 999. You do not belong to any religious denomination. You are currently married. You have been in a consensual union before. and have been married for 2003 years. You cohabited with your spouse before marriage. Your highest level of education is ISCED level 6. Your current work situation is: Employed. working as Profes-

sional and technical. with a Fixed-term contract. You have never left your parents' home. Your household consists of 3 people. You have 1 brother. and are the firstborn. You have 2.0 children. and definitely no plan to have children in the next 3 years. You have no gender preference if you could have only one child. Your monthly household income after taxes is equivalent to 2,800.00 euros. You spent 48.30 seconds in reading, judging and rating this description.. On a scale from 0 to 10, you rated satisfaction with your life a 8."

VIGNETTE:

"This is the first family description. In the following you will find a description of Lisa and Robert. Lisa and Robert are both around 45 years old. Lisa and Robert are married. Lisa and Robert have three children. Lisa and Robert's combined income is higher than the country average. The family is not well respected in their community. Lisa, Robert, and their children discuss their daily life frequently and feel comfortable expressing their feelings and raising disagreements with each other. Lisa, Robert, and their children talk with both Lisa's and Robert's parents infrequently. Both Lisa and Robert work full-time. Lisa takes care of most of the family and household responsibilities. Robert feels conflicted between his career and the possibility to help out with family responsibilities, and Lisa also feels conflicted between her family responsibilities and her career."