Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Large Language Models as Synthetic Respondents for Economic Preferences: A Global Audit

Fadong Chen

fadongchen@zju.edu.cn

Zhejiang University https://orcid.org/0000-0003-3934-580X

Qijun Wang

Zhejiang University

Lei Wang

Zhejiang University

Sudeep Bhatia

University of Pennsylvania

Article

Keywords: large language models, economic preferences, synthetic participants, global surveys, cross-cultural heterogeneity

Posted Date: February 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8775326/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations:
Ethics approval: The Institutional Review Board of the Neuro-Management Lab at Zhejiang University determined that our study was exempt.Participant consent: We used Global Preferences Survey (GPS) data as the human preference benchmark and did not collect any new human participant data for this study; therefore, participant consent was not required.

There is **NO** Competing Interest.

Version of Record: A version of this preprint was published at Nature Human Behaviour on October 12th, 2026. See the published version at https://doi.org/10.1038/s41562-026-02611-4.

Large Language Models as Synthetic Respondents for Economic Preferences: A Global Audit

Fadong Chen[a,b,1], Qijun Wang[a,1], Lei Wang[a,b], Sudeep Bhatia[c,*]

[a] School of Management, Neuromanagement Laboratory, Zhejiang University, Hangzhou, 310058, China

[b] The State Key Laboratory of Brain-Machine Intelligence, Zhejiang University, Hangzhou, 310058, China

[c] Department of Psychology, University of Pennsylvania, Philadelphia, PA 19104

[1] Fadong Chen and Qijun Wang contributed equally to this work.

[*] Correspondence to: fadongchen@zju.edu.cn

Large Language Models as Synthetic Respondents for Economic Preferences: A Global Audit

Abstract

Large-scale surveys are central to the study of economic preferences but are costly and slow to conduct. Recent advances in large language models (LLMs) raise the question of whether synthetic respondents can substitute for human participants. We address this question by running a virtual version of the Global Preference Survey (GPS), a large cross-national instrument that elicits six experimentally validated economic preferences in nationally representative samples from 74 countries and 42 languages. Using a one-to-one, persona-based prompting protocol that embeds respondents' demographics, location, local language, and country-specific incentives, eight leading LLMs predict human economic preferences for approximately 70,000 individuals in the original GPS. We find systematic but modest alignment with human data at the individual level, with LLMs capturing many of the demographic gradients in gender, age, and cognitive ability. However, alignment deteriorates with aggregation. LLMs only partially recover cultural heterogeneity and show limited correspondence with country-level preference structure. Probing sources of misalignment, we show that country-level performance covaries with economic development and digital infrastructure, but not with population size or the number of internet users. These results indicate that while LLMs can mirror certain forms of individual- and group-level heterogeneity in economic preferences, they are not reliable substitutes for representative surveys, macro-level inference, or policy evaluation. LLM-based synthetic respondents may nonetheless aid early-stage hypothesis generation, study-design diagnostics, provided their use is accompanied by careful validation and attention to developmental and cultural biases.

Keywords: large language models; economic preferences; synthetic participants; global surveys; cross-cultural heterogeneity

Introduction

Large-scale surveys are an important tool in economics. By directly eliciting latent preferences, they complement revealed-preference approaches and support both causal and descriptive inference at the population scale (Morgan 1975; Stantcheva 2023). However, implementing such surveys is expensive and slow. Researchers need to construct sampling frames, train enumerators, provide respondent incentives, and coordinate complex field operations. Recent advances in large language models (LLMs) have prompted an intriguing alternative: LLMs can be used as synthetic respondents to complete survey modules or experimental tasks (Argyle et al. 2023; Cui, Li, and Zhou 2025; Dillion et al. 2023; Grossmann et al. 2023; Horton 2023; Meng 2024; Qin, Huang, and Ding 2024; Wu, Xi, and Xie 2025). This approach can lower costs and enables faster iterations without human recruitment.

A growing literature has begun to explore the use of LLMs as synthetic participants across the social sciences. At the same time, concerns remain about using LLMs as surrogates for human respondents, given inherent algorithm biases such as the flattening of identity groups (Wang, Morgenstern, and Dickerson 2025), systematic skew toward Western, Educated, Industrialized, Rich, and Democratic (WEIRD) populations (Abdurahman et al. 2024; Atari et al. 2023; Crockett and Messeri 2025; Lin 2025; Tao et al. 2024), and other potential risks (Bisbee et al. 2024; Y. Chen et al. 2023; Gao et al. 2025). Prior work has focused largely on political opinions (Argyle et al. 2023; Bisbee et al. 2024; Ceron et al. 2024; Jiang, Wei, and Zhang 2024; Röttger et al. 2024; Rozado 2024; Santurkar et al. 2023), and social and moral attitudes (Abdurahman et al. 2024; Atari et al. 2023; Durmus et al. 2024; Li et al. 2024; Sinacola, Pachot, and Petit 2025; Tao et al. 2024), while paying much less attention to economic preferences, even though these preferences play a central role in linking individual behavior to market outcomes and social welfare, and constitute one of the most compelling use cases for LLMs in efforts to improve individual and societal well-being.

As a central component of economic theory, preference heterogeneity helps explain individual behaviors (Åkerlund et al. 2016; Della Vigna and Malmendier 2006; Genesove

and Mayer 2001), socioeconomic status (Epper et al. 2020; Kosse and Tincani 2020), and macroeconomic outcomes (Guiso, Sapienza, and Zingales 2009; Knack and Keefer 1997; Sunde et al. 2022), and it is indispensable for policy design and evaluation (Bierbrauer et al. 2017; Dupas and Robinson 2013; Kremer, Rao, and Schilbach 2019). Understanding group-level heterogeneity in preferences is especially important for development economics and policy evaluation, which seek to capture how policies affect different populations rather than only the average effect. If LLMs can be shown to align closely with human economic preferences, they could be used to generate counterfactual data for policy analysis, partially alleviating the fundamental problem of causal inference, namely that the counterfactual is unobserved (Holland 1986). However, the credibility of such applications depends on the models' ability to reproduce well-documented patterns of heterogeneity. Given the pronounced demographic variation (Bishai 2004; Burks et al. 2009; Croson and Gneezy 2009; Dohmen et al. 2010; Falk et al. 2018) and culturally specific patterns (Enke 2019; Falk et al. 2018; Galor and Özak 2016; Henrich et al. 2001; Luttmer and Singhal 2011) in economic preferences, the study of economic preferences offers a natural testbed for a global, multi-level audit of LLM-generated synthetic data across individuals, demographic groups, countries, and cultures.

Unlike prior studies (Gao et al. 2025; Goli and Singh 2024; Horton 2023; Leng and Yuan 2024; Mei et al. 2024), we implement a one-to-one prompting design to build "digital twin" panels of a large cross-country economic preference survey. That is, for each human respondent, we generate a matched LLM-respondent facing the same questionnaire. A second key feature is that we use localized-language prompts rather than English-only prompts, in order to better reflect local context (AlKhamissi et al. 2024; Lu, Song, and Zhang 2025). We also include an English-only prompting condition as a robustness check.

Figure 1 illustrates our study design. We use the Global Preference Survey (GPS) as the human benchmark (Falk et al. 2018). The original GPS includes responses from more than 80,000 individuals across 76 countries and elicits six canonical economic preferences – risk taking, patience, negative reciprocity, positive reciprocity, altruism, and trust – that have been validated against incentivized experimental tasks (Falk et al. 2023;

see Methods for details).

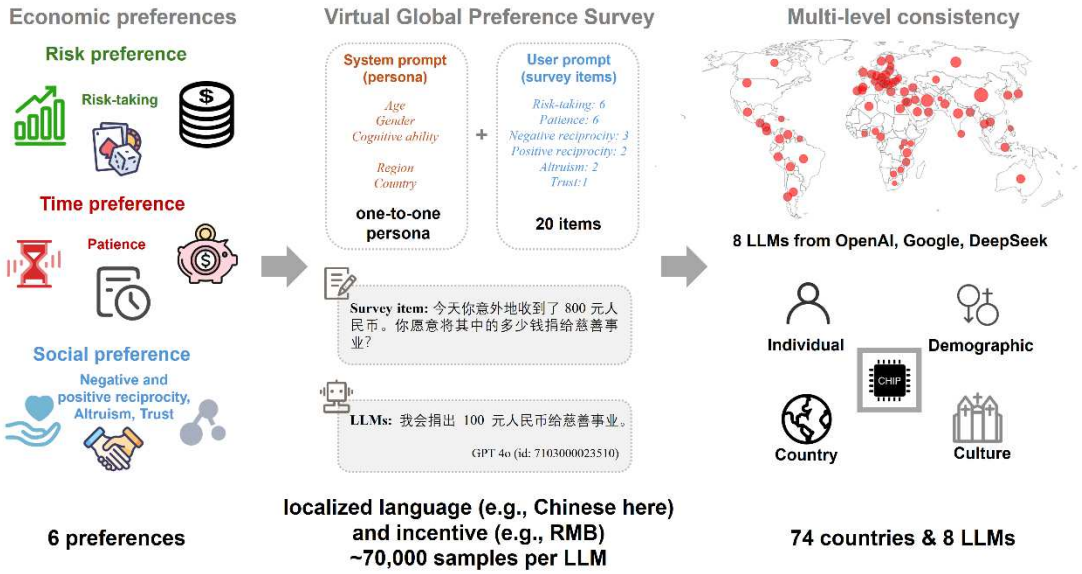


Figure 1. Study Design. Following the Global Preference Survey (GPS), we focus on six economic preferences: risk-taking, patience, negative reciprocity, positive reciprocity, altruism, and trust. For each item, respondent demographics and locations are encoded in a system prompt, the GPS question text is provided as a user prompt to large language models (LLMs), and the model's output is recorded as its predicted response. In the prompts, the language and incentives are localized, as in the example where we use Chinese and RMB (the English translation of this survey item: "Today you unexpectedly received 800 RMB. How much of this amount would you donate to a good cause?" The translation of LLM's answer: "I would donate 100 RMB to charity."). We assess alignment between LLM-predicted and GPS-surveyed (human) preferences at the individual, demographic, country, and cultural levels. Eight LLMs from OpenAI, Google, and DeepSeek generate predicted responses for GPS respondents in 74 countries. Filled circles on the world map mark sample countries and the circle diameter shows the number of observations within each country.

Building on the GPS, we conducted a "virtual GPS" in which eight LLM-prompting conditions – seven leading multilingual models (GPT 3.5-turbo, GPT 4o-mini, GPT 4o, GPT 4.1-mini, Gemini 2.0 Flash, Gemini 2.5 Flash, DeepSeek V3) plus an English-only GPT 4o-mini condition – complete the original GPS module for about 70,000 personas. Our one-to-one prompting design assigns one prompt per persona, mirroring each original GPS respondent. Each persona embeds age, gender, cognitive ability, and country/region, and prompts are delivered in the respondents' original survey language (see Methods and

Appendix A for details). In total, this protocol yields about 11.2 million LLM-generated answers and roughly 560,000 synthetic participants (70,000 personas × 8 LLMs) that are directly comparable to the human GPS benchmark (see Appendix A, Tables A.1 and A.2 for dataset description). Together with the GPS benchmark, our design allows for a more rigorous and fine-grained comparison of LLM-generated responses with human data at multiple levels.

Our study makes both theoretical and practical contributions. First, we provide a global, multi-level audit of LLM-as-respondent validity using a benchmark that is theoretically grounded, experimentally validated, and widely used in cross-country economic research (Becker, Enke, and Falk 2020; Hanushek et al. 2022; Kanagaretnam et al. 2022; Nieminen and Kuziemska-Pawlak 2024; Pickard, Dohmen, and Van Landeghem 2024; Sunde et al. 2022; Tannenbaum et al. 2025). Second, we clarify the conditions under which LLM-generated responses are informative, by assessing whether synthetic preference data can plausibly support policy-relevant analyses and the study of group-level heterogeneity across demographic and national contexts. Third, we relate cross-country variation in LLM performance to macro-level characteristics such as economic development and digital infrastructure, providing a framework for diagnosing where synthetic respondents are more versus less reliable and for thinking about how more inclusive models might be developed. Fourth, we introduce a transparent, reusable protocol based on persona-level, language-matched prompting that can be used to track model updates over time, compare architectures, and integrate LLM audits into ongoing economic and policy research. Taken together, these contributions provide a systematic, large-scale assessment of the conditions under which LLMs can be used responsibly for policy-relevant measurement and analysis in global economic research.

Results

LLM-predicted preferences align positively but weakly with GPS-surveyed preferences at the individual and country levels

We first examine individual-level alignment by computing Spearman correlations

between LLM-predicted and GPS-surveyed preferences (see Methods for details on the construction of individual-level preferences), pooling all respondents from 74 countries. Figure 2A reports the 48 correlations (6 preference domains × 8 LLMs). Of these, 44 are significantly positive (91.7%), the only non-positive estimates are trust in Gemini 2.0 Flash, Gemini 2.5-flash, GPT 4o, and GPT 4o-mini (English-only prompts). This high proportion of positive correlations is significantly above chance (95% CI = [80.0%, 97.7%]). Repeating the analysis within each country yields a similarly high proportion. On average, 84.6% of the within-country correlations are positive across models and preference domains (95% CI = [83.4%, 85.8%], see Table B.1 in Appendix B.1 and Figures B.2-B.8 in Appendix B.1). These patterns indicate systematic individual-level alignment between LLM-predicted and GPS-surveyed preferences.

Although most individual-level correlations are statistically significant, their magnitudes are small. Across preference domains, correlations are generally highest for risk-taking and lowest for trust. For instance, the largest individual-level estimate is for GPT 4.1-mini on risk-taking (ρ = 0.252), which merely reflects a modest association. Turning to model comparisons, we find no evidence that GPT 4o outperforms earlier OpenAI models (two-sided t-test, P = 0.541) or that Gemini 2.5-flash outperforms Gemini 2.0-flash (two-sided t-test, P = 0.592). Thus, advanced models do not appear to achieve stronger individual-level alignment on economic preferences.

We next examine country-level alignment by aggregating GPS-surveyed and LLM-predicted preferences to country means using the GPS ex-post sampling weights (see Methods for details on the construction of country-level preferences). We first compute half-split internal consistency for country-level averages. The half-split internal consistency of LLM-predicted preferences is comparable to, and slightly higher than, that of the GPS-surveyed data (see Appendix B.2 and Figure B.9 for details). This supports the feasibility of constructing country-level average preference indices from LLM-predicted preferences. In Figure 2B, 41 of 48 (85.4%) country-level correlations between LLM-predicted and GPS-surveyed preferences are positive, but only 12 of 48 have 95% CIs that exclude zero. Compared with their individual-level counterparts, the country-level

correlations are on average slightly larger ($M_{country}$ = 0.137 vs $M_{individual}$ =0.096, two-sided paired t-test, P = 0.019), yet they remain small, and many model-preference pairs do not reach conventional significance thresholds. Together with the individual-level results above, these findings indicate that LLM-predicted preferences exhibit only weak positive alignment with GPS-surveyed preferences at both the individual and country levels.

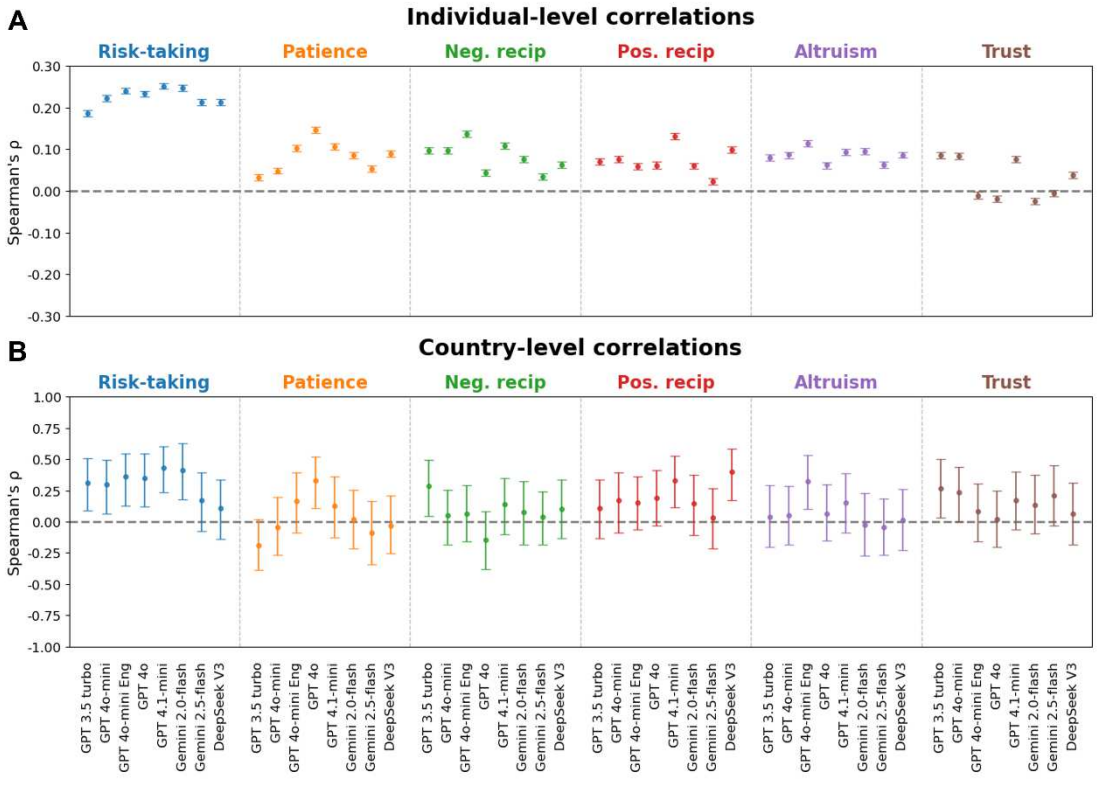


Figure 2. Consistency between LLM-predicted and GPS-surveyed preferences. (A) Individual-level consistency. Each point displays the Spearman correlation coefficient between GPS-surveyed and LLM-predicted preferences, pooling all individual observations across 74 countries (N>65,000, due to missing values in the original GPS, the number of observations varies by preference domain). The predominance of positive coefficients (44 of 48, 91.7%) indicates systematic, though weak, alignment between LLM-predicted and GPS-surveyed preferences at the individual level. **(B) Country-level consistency**. Each point represents the Spearman correlation coefficient between country-level average GPS-surveyed and LLM-predicted preferences (N=74). Country-level averages are weighted by the sampling weights provided in the GPS dataset. Of the 48 coefficients, 41 (85.4%) are positive and 12 (25.0%) have 95% CIs that exclude zero, reflecting limited power at the country level despite generally positive correlations. Error bars indicate 95% bootstrapped CIs in both panels.

Next, we assess whether these patterns are specific to particular models or instead reflect a more general regularity. Figure 3 shows the between-model correlation matrices for LLM-predicted preferences. At the individual level (Figure 3A), LLM-predicted preferences are generally and positively correlated across models (162 of 168 correlations are positive; mean correlation across all preferences: 0.359; risk-taking: 0.662; patience: 0.546; negative reciprocity: 0.233; positive reciprocity: 0.229; altruism: 0.403; trust: 0.079). The country-level between-model correlations in Figure 3B show a similar pattern (159 of 168 are positive; mean correlation across all preferences: 0.265; risk-taking: 0.249; patience: 0.247; negative reciprocity: 0.233; positive reciprocity: 0.334; altruism: 0.364; trust: 0.166).

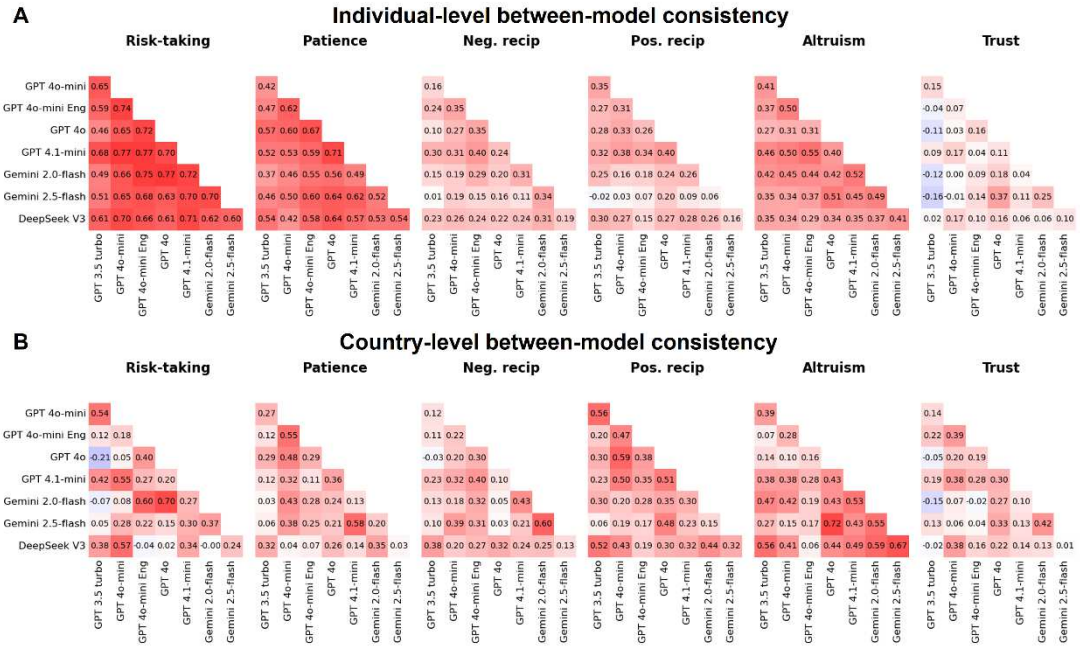


Figure 3. Between-model consistency of LLM-predicted preferences. (A) Individual-level between-model consistency. Each cell reports the Spearman correlation between two LLMs, computed by pooling all individuals from all countries. **(B) Country-level between-model consistency.** Each cell reports the Spearman correlation between two LLMs, computed using country-level averages of LLM-predicted preferences and GPS sampling weights.

Regardless of statistical significance, the proportion of between-model correlation coefficients that are positive is far above what would be expected by chance (50%; two-sided binomial test, $P < 0.001$ for each preference at both individual and country levels).

These results suggest that LLMs make broadly similar predictions to one another at both levels, but these shared predictions align only weakly with human economic preferences, indicating a common but imperfect underlying mechanism for how current LLMs represent and extrapolate such preferences.

It is worth noting that across preference domains, alignment is not uniform. Individual- and country-level correlations in Figure 2 are consistently higher for risk-taking and lower for trust (mean individual-level correlations: 0.226 vs. 0.028; country-level: 0.305 vs. 0.146), mirroring the between-model consistency patterns in Figure 3 (mean individual-level consistency: 0.662 vs. 0.079; country-level: 0.249 vs. 0.166). This convergence between human–model alignment and model–model consistency suggests that LLMs capture some preference domains more readily than others, with trust standing out as a domain that models both struggle to approximate and fail to represent consistently across architectures.

LLM-predicted preferences align with GPS-surveyed preferences at the demographic level, but not at the cultural level

We next investigate whether LLM-predicted preferences reproduce the demographic gradients with respect to gender, age, and cognitive ability observed in the GPS data. For each LLM and preference domain, we compute effect sizes (Cohen's d) for gender (male vs. female), age (younger vs. older), and cognitive ability (high vs. low; see Methods for details). We then construct a consistency ratio, defined as the proportion of LLM–preference pairs whose estimated effects are directionally concordant with the GPS benchmark based on 95% CIs: either both CIs include zero, or both exclude zero and the effects have the same sign (see Methods) (see Methods). A higher consistency ratio indicates a greater ability to recover observed demographic (or cultural) patterns.

Figure 4 presents the demographic-level effect sizes (Cohen's d) across gender, age, and cognitive ability groups. Overall, the consistency ratio is 77.1% (95% CI = [69.3%, 83.7%]; 79.2% (38/48) for gender, 66.7% (32/48) for age, and 85.4% (41/48) for cognitive ability). These relatively high consistency ratios indicate that LLM-predicted

data reproduce most of the demographic patterns documented in the GPS benchmark. Across models, demographic effects for risk-taking, patience and negative reciprocity are particularly well aligned, whereas LLMs frequently fail to reproduce the gender and age differences observed for positive reciprocity.

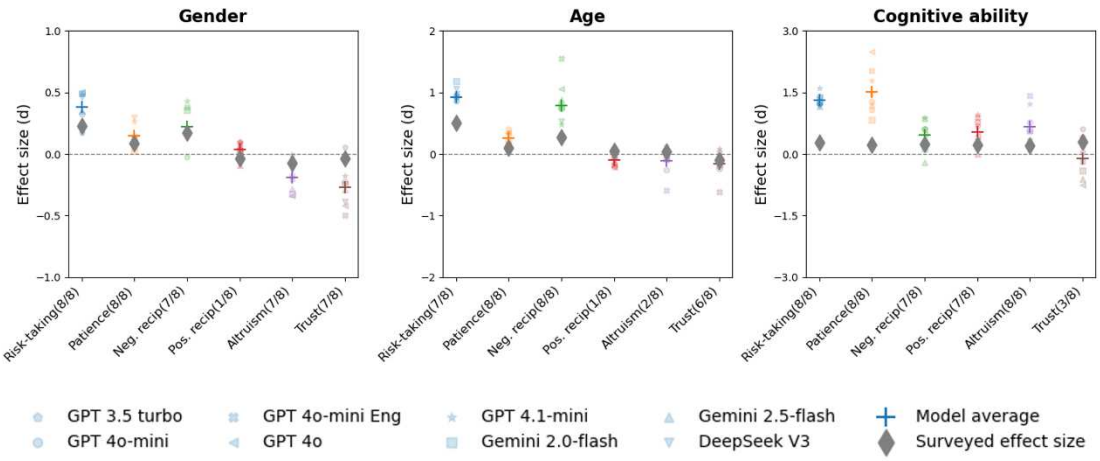


Figure 4. Demographic-level consistency and gradients in GPS-surveyed and LLM-predicted preferences. Each panel plots effect sizes (Cohen's *d*) across preference domains for gender (left), age (middle), and cognitive ability (right). Demographic variables are encoded as binary indicators for female, older (≥ 60), and lower cognitive ability (< 5). Positive effect sizes indicate higher preference levels for males (vs females), younger (vs older), and higher cognitive ability (vs lower). Colored symbols show effects sizes predicted by LLMs, crosses denote model-averaged effect sizes, and gray diamonds denote effect sizes estimated from the GPS data. Numbers in parentheses on the x-axis indicate, for each preference domain, the number of models (out of eight) whose effect is "consistent" with the GPS effect. Based on the consistency ratio, 79.2% (38/48) of model-preference pairs match the GPS gender effect, 66.7% (32/48) match the age effect, and 85.4% (41/48) match the cognitive ability effect.

There is also evidence that LLMs tend to amplify demographic effects, i.e., model-averaged effect sizes are often larger than their survey-based counterparts. Gender and age differences in risk-taking, patience and negative reciprocity, as well as cognitive ability differences for all preferences except trust, are systematically more pronounced in LLM-predicted preferences than in the GPS data (Figure 4). This amplification is consistent with concerns that LLMs, when used as stand-ins for human participants, may misrepresent identity groups from out-group perspective, thereby potentially overstating group

differences (Exley et al. 2025; Wang, Morgenstern, and Dickerson 2025).

Demographic differences, however, capture only one source of variation in economic preferences. A large literature shows that preferences also vary systematically across cultural contexts (Cao et al. 2021; M. K. Chen 2013; Enke 2019, 2022; Galor and Özak 2016; Guiso, Sapienza, and Zingales 2006). We therefore examine whether LLM-predicted preferences recover cross-cultural differences along cultural dimensions from the World Values Survey (WVS, i.e. Inglehart–Welzel cultural map), Hofstede's, and Schwartz's cultural theories(Hofstede, Hofstede, and Minkov 2010; Inglehart and Welzel 2005; Schwartz 2006). For each dimension, we classify countries into "high" and "low" groups using a median split on that dimension, assign individuals to these groups based on their country, and estimate the corresponding difference in preferences between individuals from high- and low-group countries (see Methods for details).

We again summarize models' performance using the consistency ratio. Figure 5 shows the cultural-level effect sizes estimated from LLM-predicted and GPS-surveyed preferences. The overall cultural-level consistency ratio is 49.6% (95% CI = [45.9%, 53.3%]), substantially lower than the demographic level consistency ratio of 77.1% (see Table C.1 in Appendix C for details). The highest cultural-level consistency ratio, 62.5%, is obtained for WVS traditional values and Schwartz's hierarchy dimensions, yet even this does not reach the average or the lowest demographic-level consistency ratio (age, 66.7%). Performance is particularly weak for some widely studied cultural dimensions. For example, only one model (out of eight) recovers the relationship between Hofstede's long-term orientation and patience in GPS, and likewise only one model recovers the relationship between Hofstede's uncertainty avoidance and risk-taking in GPS.

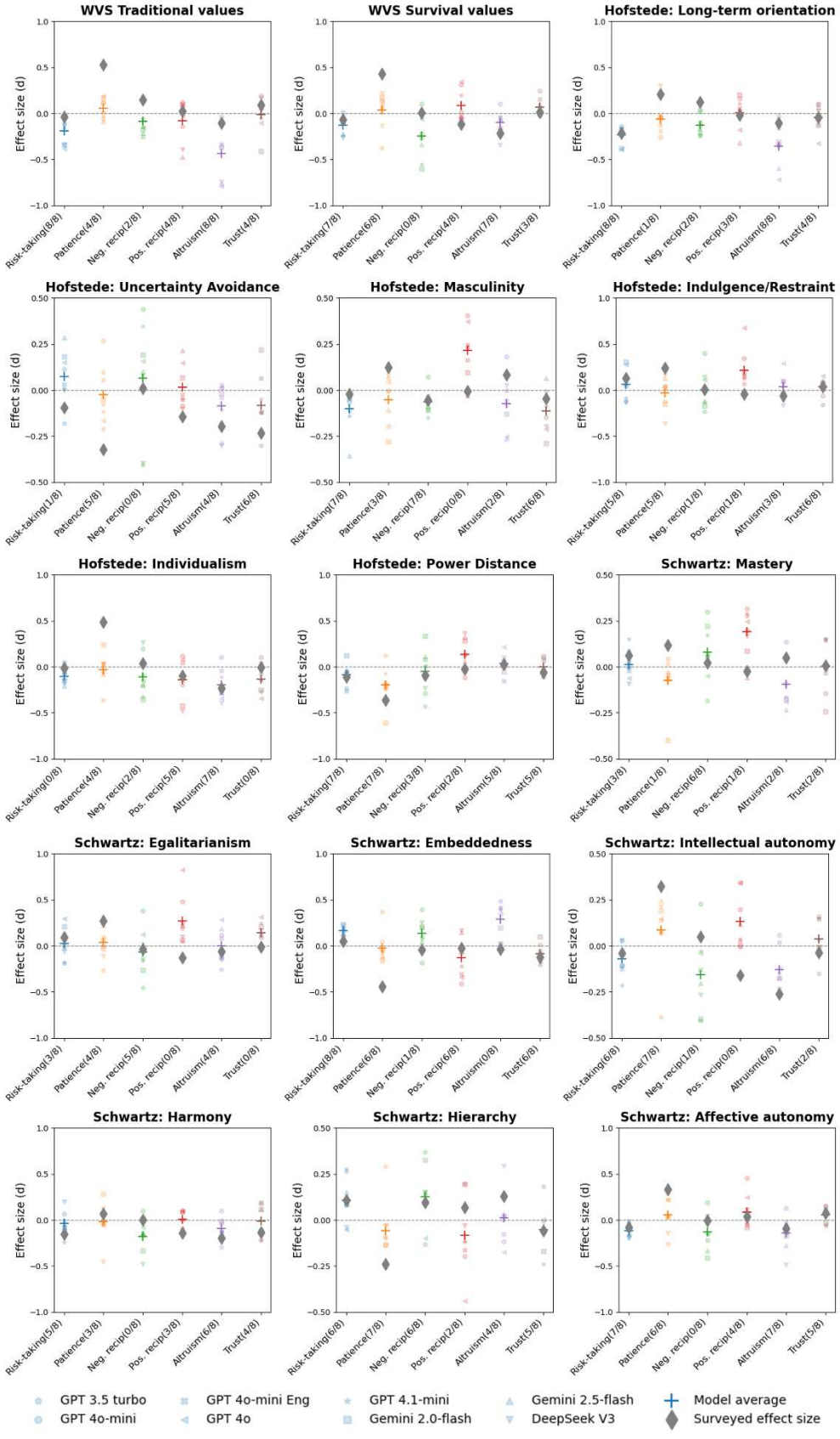


Figure 5. Cultural-level consistency and gradients in GPS-surveyed and LLM-predicted preferences. Each panel plots effect sizes (Cohen's *d*) across preference domains for one

cultural dimension from the World Values Survey (WVS), Hofstede's, and Schwartz's cultural theories. For each dimension, countries are classified into "high" and "low" groups using a median split on that dimension, and individuals are assigned accordingly. We quantify the difference between respondents in high- versus low-group countries using Cohen's *d*, a unit-free standardized mean difference. Positive values indicate higher preference levels among respondents in high-group countries relative to low-group countries. Colored symbols show effect sizes for individual models, crosses denote model-averaged effect sizes, and gray diamonds denote effect sizes estimated from the GPS data. Numbers in parentheses on the x-axis indicate, for each preference domain and cultural dimension, the number of models (out of eight) whose effect size is "consistent" with the GPS benchmark. The overall cultural-level consistency ratio is 49.6%, substantially lower than the demographic-level consistency ratio (77.1%), indicating that LLMs only partially recover cross-cultural heterogeneity in economic preferences.

These results indicate that, while LLMs mirror many demographic gradients in economic preferences, they only partially capture cultural variation. As a robustness check, we also construct a sign-consistency measure that records the proportion of cases in which the sign of the effect size based on LLM-predicted data matches that based on GPS-surveyed data, irrespective of statistical significance. This complementary indicator yields a similar pattern (see also Appendix C.2). Effect sizes for demographic and cultural gradients are also positively correlated across models (see Appendix C.3 for details), indicating that different LLMs tend to make similar directional predictions even when they diverge from the survey estimates.

Sources of misalignment between LLM-predicted and GPS-surveyed preferences

We next investigate potential sources of misalignment that might explain why LLMs yield only weak correlations with GPS-surveyed preferences at the individual and country levels and struggle to reproduce distinctive cultural patterns. One possibility is that country information in the prompts plays only a limited role, so that models do not meaningfully condition on it. To diagnose this, we estimate a sequence of regressions in which GPS-surveyed preferences are regressed on LLM-predicted preferences. In the baseline specification, we include no covariates. We then add demographic characteristics, country indicators, and both sets of controls in turn. If the coefficient on LLM-predicted

preferences shrinks after adding a given set of controls, this suggests that the corresponding variables account for part of the observed alignment (see Linear mixed-effects regressions in Methods).

Figure 6A shows that adding country indicators has essentially no effect on the regression coefficients, whereas adding demographic controls substantially attenuates them. A linear mixed-effects regression across model × preference combinations confirms this pattern (relative to the baseline: b_{country} = 0.001, P = 0.864; b_{Demo} = -0.070, P < 0.001; $b_{\text{Demo+Country}}$ = -0.076, P < 0.001; see also Appendix D.1). These results indicate that demographic information in the prompts accounts for much of the individual-level alignment, while country labels contribute little.

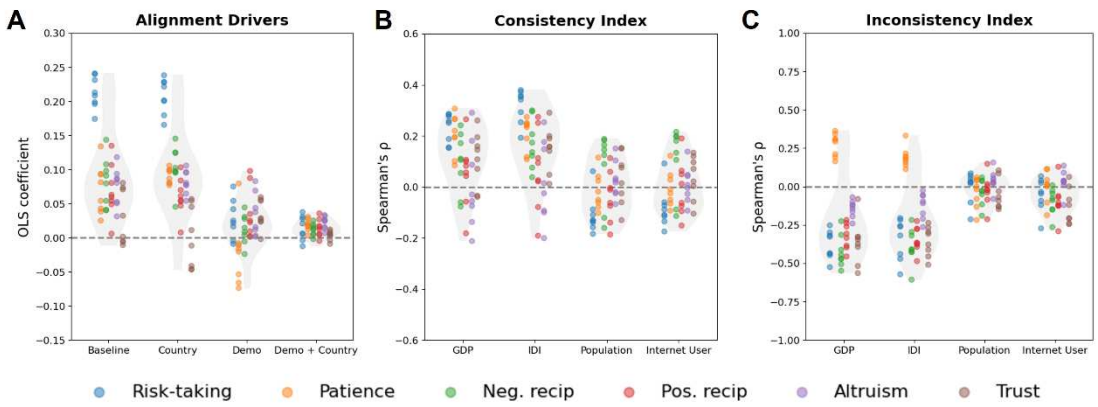


Figure 6. Potential sources of variation in LLM-human consistency in economic preferences. (A) Regression coefficients from models regressing GPS-surveyed preferences on LLM-predicted preferences. Each point represents an OLS coefficient estimated on pooled individual-level data for a given model × preference combination. Covariates (country indicators, demographic characteristics, and both) are sequentially added as indicated on the *x*-axis. **(B)** Spearman correlations between the consistency index (the correlation between LLM-predicted and GPS-surveyed preferences within each country) and four country-level characteristics: log GDP per capita, the Information and Communications Technology Development Index (IDI), log population, and log number of internet users. Each point represents a Spearman correlation coefficient computed across 74 countries. **(C)** Spearman correlations between the inconsistency index (the absolute country-level prediction error) and the same four characteristics. Most correlations with GDP and IDI are positive in panel (B), and negative in panel (C), indicating that economic development and digital infrastructure are more strongly associated with LLM-human consistency than population size or the number of internet users.

To better understand why alignment deteriorates at the country and cultural levels, we construct two country-level measures. First, a consistency index, defined as the Spearman correlation between GPS-surveyed and LLM-predicted preferences within each country. Second, an inconsistency index, defined as the absolute prediction error, i.e., the absolute difference between country-level LLM-predicted and GPS-surveyed preferences (see Methods). According to the Scaling Law (Kaplan et al. 2020), we relate these indices to four country characteristics that plausibly track LLMs' training conditions. Higher-income countries may attract greater optimization effort from LLM developers, as they tend to host larger user bases (Appel et al. 2025; Chatterji et al. 2025), increasing the commercial returns to improving model performance for these contexts. We use log GDP per capita and the Information and Communications Technology Development Index (IDI), as proxies for the amount of compute used for training, and log population and log number of internet users, as proxies for corpus size (Villalobos et al. 2024). Figures 6B and 6C report the spearman correlations between these characteristics and the two country-level indices.

Most correlations between the consistency index and GDP or IDI are positive (GDP: 37/48, 77.1%; IDI: 42/48, 87.5%), whereas most correlations between the inconsistency index and GDP or IDI are negative (GDP: 40/48, 83.3%; IDI: 40/48, 83.3%). Thus, higher levels of economic development and digital infrastructure development are associated with stronger alignment and smaller absolute prediction errors in LLM-predicted preferences. In contrast, population size and the number of internet users show little systematic relation to either index (consistency index: population 22/48 positive, 45.8%; internet users 24/48 positive, 50.0%; inconsistency index: population 23/48 negative, 47.9%; internet user: 28/48 negative, 58.3%; see Appendix D.2 for regression evidence and robustness check about the consistency index; Appendix D.3 for details about the inconsistency index).

An exception arises for patience in Figure 6C. For this domain, all models show greater inconsistency in countries with higher GDP and IDI than in those with lower development. This inversion is consistent with prior findings that patience is itself

strongly correlated with economic development. That is, patience explains nearly 40% of the cross-country variation in log GDP per capita, whereas other preference domains do not exhibit such systematic associations with GDP (Falk et al. 2018; Sunde et al. 2022).

Discussion

How to use LLMs responsibly in empirical research is a pressing question. We contribute to this literature by evaluating the plausibility of using LLMs as synthetic respondents in a global survey setting. Specifically, we compare LLM-generated responses with human data from the Global Preference Survey (GPS) across multiple levels, from individuals to demographic groups, cultures and countries, and thereby clarify the boundary conditions of their scientific utility.

Our findings reveal two key limitations of using LLM-generated preference data. First, although LLMs are consistent with one another, their correlations with human data are weak at both the individual and country levels. Individual-level correlations are statistically significant largely because of the very large sample size, and country-level correlations are only slightly larger in magnitude but often fail to reach conventional significance thresholds. LLMs reproduce a substantial share of demographic gradients but perform poorly at the cultural level, suggesting a micro-to-macro gap in alignment. Second, prediction errors of LLMs systematically covary with economic development (as proxied by GDP per capita and the IDI) but show much weaker and less robust associations with population size or the number of internet users. Errors are thus geographically and culturally patterned in ways that coincide with economic development and digital infrastructure. When such uneven errors are aggregated, they accumulate into macro-level misalignment.

These limitations directly constrain applications of synthetic data in research, policy-design, and applied practice. The most basic promise of synthetic data is to approximate broad, average tendencies. Our results suggest that such averages can be misleading when they depend on macro-level structure, i.e., models can agree with one another while still poorly approximating human preferences. We attempt to replicate published GPS findings

(Falk et al. 2018; Falk and Hermle 2018; Sunde et al. 2022) using LLM-predicted preferences. For GPT 4o, 13 of 35 (37.1%) country-level findings are reproduced in the expected direction, and other models show even weaker performance (see Appendix E). These exercises indicate that the feasibility of using LLMs to generate unobserved counterfactual data for policy analysis cannot be taken for granted, and requires more careful, domain-specific validation.

Our results also raise the question of why errors track economic development more closely than population size. Two non-exclusive mechanisms are consistent with our evidence and with recent work on geographical biases in LLMs(Atari et al. 2023; Tao et al. 2024). First, even with language-matched prompts, languages and content from more developed countries may receive disproportionate emphasis in LLMs development and training, yielding lower errors there (Joshi et al. 2020; Ruprechter, Burghardt, and Helic 2023). Second, developed countries are typically more digitized, providing more high-quality, structured data per capita that can improve model performance (Graham, Ojanperä, and Dittus 2019). Together, these mechanisms align with the observed associations between model error and GDP/IDI, though we cannot make any causal claims.

As stated in the Universal Declaration on Cultural Diversity, "Cultural diversity is as necessary for humankind…, is the common heritage of humanity." Our results, however, expose a practical dilemma in training and developing LLMs. Although models perform relatively poorly in lower-income countries, developers have limited commercial incentives to optimize them for these contexts. The equity risks are nontrivial. Downstream, developed countries may be able to leverage AI more effectively in research, productivity, and education, while users in developing countries receive answers that are less adapted to local contexts. Over time, if LLM performance systematically favors developed countries, cultural knowledge from developing countries may remain under-represented, encouraging homogenization and centralization of the global cultural ecosystem, eroding cross-cultural diversity (Abdurahman et al. 2024; Atari et al. 2023), and widening digital inequity (Capraro et al. 2024; Pataranutaporn et al. 2025). These risks underscore the need to align LLMs not only with universal ethical principles but also

with diverse cultural, linguistic, and social contexts.

Beyond alignment, we identify two additional practical constraints. One is gender bias (see Appendix F). Across countries, LLMs tend to overpredict women's prosocial preferences relative to the GPS benchmark, and the resulting gender-gap misalignment is larger in countries with higher gender-stereotype indices. This pattern suggests that model outputs are sensitive not only to demographic persona inputs but also to country-level gender norms embedded in the models' training corpora. The other is limited response diversity (see Appendix G). Compared with human responses in GPS, LLM-generated responses are more concentrated, and this reduction in response diversity is more pronounced in high-income countries. Model outputs for respondents in higher-development contexts are more deterministic, whereas responses for less developed countries display greater variability.

At the same time, our individual-and demographic-level alignment results highlight several promising scientific uses of LLMs as synthetic respondents. First, as shown in prior work, LLMs can be used to propose and probe latent motivational structures underlying human behavior in WEIRD populations, complementing formal decision theory (Bhatia 2024; Bhatia et al. 2025; Binz et al. 2025; Xie et al. 2025; Zhu, Yan, and Griffiths 2025). Second, LLMs can accelerate early-stage hypothesis generation and study-design diagnostics, helping researchers identify promising interventions or measurement strategies before turning to human samples. Our results suggest using them as tools for simulation and diagnostics, particularly when research questions concern relative subgroup differences (e.g., by gender or age). Third, LLMs can serve as a fast, low-cost environment for follow-up experiments and robustness checks on main effects of interest (Guilbeault, Delecourt, and Desikan 2025). Finally, they offer a safe testbed for high-risk but policy-relevant counterfactual manipulations (e.g., changes in partisan animosity) that would be difficult or unethical to implement with human subjects.

Beyond its implications for economic and policy analysis, our study also offers a framework for constructing multi-level benchmarks of alignment of LLMs. By pairing a large, nationally representative survey with one-to-one synthetic sampling, we obtain

comparable human-model measures at the individual, demographic, country, and cultural levels. The resulting metrics can be used to track how alignment evolves across successive model generations. More broadly, as LLM evaluation increasingly relies on standardized test suites for coding, reasoning, and other well-defined objective tasks, there is a parallel need for benchmarks that assess predictive accuracy and cross-cultural alignment in high-impact, policy-relevant and subjective domains. Using existing large-scale cross-national surveys as behavioral benchmarks, as we do here with the GPS, provides one practical route and can lower the barrier for researchers to build more culturally diverse tests for evaluating LLMs across different contexts and subjects (Ni et al. 2025).

Several limitations qualify our conclusions. Because GPS was conducted in 2012, portions of its content may appear in LLM training data. Successful predictions could therefore partially reflect memorization, making our alignment estimates an upper bound. Future large-scale surveys and intervention experiments could construct similar digital twin panels in domains where training data contamination is less likely, helping to disentangle memorization from genuine reasoning. Moreover, although we find systematic individual-level alignment, the mechanisms by which models reproduce these patterns remain opaque. Future work should combine audits with targeted interventions, such as data augmentation in under-represented contexts, cross-lingual grounding, and domain-specific alignment, to test whether the micro-to-macro gap can be narrowed. It will also be important to develop diagnostics that warn researchers when synthetic respondents are unsuitable for aggregate inference, particularly in high-stakes development and policy applications. Ultimately, realizing the promise of LLMs for global economic research will require sustained efforts to improve cultural and developmental coverage. We believe that global audits such as ours are essential for constructing and validating these tools, and the framework developed here can serve as a foundation for future efforts to build more reliable and inclusive LLM-based approaches to economic and policy analysis.

Methods

Global preference survey

We use the Global Preference Survey (GPS) as the human benchmark. Conducted as part of the 2012 Gallup World Poll, the GPS includes more than 80,000 respondents from 76 countries that together account for about 90% of the world's population and global GDP. It is one of the most comprehensive cross-country datasets for systematically measuring economic preferences and provides globally comparable preference measures (Falk et al. 2018).

The GPS elicits six economic preference domains: risk-taking, patience, positive reciprocity, negative reciprocity, trust, and altruism. Participants answer 8 self-assessment items, 2 hypothetical choice-based quantitative items, 5 staircase intertemporal-choice items, and 5 staircase lottery-choice items. Each preference index is constructed by combining one abstract qualitative item with one hypothetical choice-based quantitative item into a single preference variable. Trust is measured solely with a qualitative item (see Table A.3 in Appendix A for details).

LLM-simulated survey data

To obtain LLM-predicted preferences, we conduct a large-scale one-to-one virtual survey. For each GPS respondent, we extract demographic information (gender, age, and cognitive ability) and country information, and encode them in the system prompt so that the LLM "adopts" that respondent's perspective. The corresponding GPS question is then presented as a user prompt. To minimize stochastic variability, we set the temperature to 0, making outputs as deterministic as possible.

Following the GPS design, all payoff amounts are denominated in local currencies and calibrated accordingly. For each country, we simulate responses only in the most widely sampled survey language whenever possible. Translations from English into other languages are generated using Google Translate API. On average, we obtain about 70,000 valid observations per model, for a total of 558,837 simulated respondents across the eight LLMs (see Appendix A for details).

Construction of individual- and country-level preference indices

To maintain comparability with the original GPS dataset, we follow its scoring and aggregation procedure for the LLM outputs. For each survey item, we first compute individual-level z-scores for each item. We then construct the six preference indices as weighted sums of item-based z-scores using the experimentally validated GPS weights.

To ensure comparability across models, we retain only those respondents observed for whom predictions are available from all 8 LLMs and re-standardize the resulting preference indices. Country-level preferences are computed as sampling-weighted means of the individual indices using the ex post sampling weights provided in the GPS.

Binary encoding of demographic and cultural variables

Age, cognitive ability, and cultural dimensions are encoded as binary indicator variables. Older respondents are those aged 60 years or above. Cognitive ability in the GPS is measured on an 11-point Likert scale (0–10) based on self-assessed agreement with the statement "I am good at math". Scores below 5 are coded as low cognitive ability, and scores of 5 or above as high.

For cultural dimensions and other country-level variables, we construct binary indicators using country-level medians. For each cultural dimension (e.g., WVS traditional values), countries at or below the median are coded as low, and those above the median as high. Individual respondents are then assigned to the corresponding high/low group based on their country's classification.

Consistency ratio

We use a consistency ratio to quantify how well the LLM-predicted preferences reproduce demographic and cultural gradients observed in the GPS. For each binary grouping variable (gender, age, cognitive ability, or high vs. low on a given cultural dimension) and each preference domain, we compute effect sizes (Cohen's d) and their 95% CIs separately for the GPS data and for each LLM. For example, the effect of cognitive ability on risk-taking is summarized by comparing the mean standardized risk-taking score of respondents with higher versus lower cognitive ability and dividing this difference by the pooled standard deviation. In general, these effect sizes provide a unit-free measure of the average difference between the two groups, with positive effect sizes indicating higher preference levels in the

reference group (e.g., males, younger individuals, higher ability, or high on the cultural dimension).

An effect is counted as "consistent" if either (a) the 95% CIs for both GPS-based and LLM-based effect sizes include zero (a concordant null effect), or (b) the 95% CIs for both effect sizes exclude zero and the two effect sizes have the same sign. The consistency ratio is defined as the proportion of such consistent cases across all model-preference combinations. Higher consistency ratios indicate stronger alignment between LLM-predicted and GPS-surveyed patterns of heterogeneity.

Linear mixed-effects regressions

To assess how much demographic and country information contribute to individual-level alignment, we first estimate four OLS regressions for each model × preference combination, regressing GPS-surveyed preferences on LLM-predicted preferences using (i) no covariates (baseline), (ii) country dummies only, (iii) demographic controls only (gender, age, cognitive ability), and (iv) both demographic controls and country dummies. This yields four OLS coefficients per model × preference combination.

We then estimate a linear mixed-effects regression using these OLS coefficients as the dependent variable and a set of dummy variables indicating the regression specification (baseline vs. + country vs. + demographics vs. + demographics and country) as predictors. Preference-domain dummies (for the six preference types) are included as additional controls to account for systematic differences across domains. To capture unobserved heterogeneity across LLMs, we include random intercepts at the model level.

Consistency and inconsistency indices

The consistency and inconsistency indices summarize, for each country, how closely LLM-predicted preferences associate GPS-surveyed preferences. For each country, the consistency index is defined as the Spearman correlation between LLM-predicted and GPS-surveyed preferences, computed by pooling all individual observations within that country. The inconsistency index is defined as the sampling-weighted country-level mean of the absolute individual prediction errors between LLM-predicted and GPS-surveyed preferences. That is, we first compute absolute differences between GPS-surveyed and LLM-predicted preferences

583 at the individual level and then take their sampling-weighted average within each country.
584 Higher consistency indices and lower inconsistency indices indicate better country-level
585 alignment between model-generated and human preferences.
586

Acknowledgements

Fadong Chen gratefully acknowledges support from the National Natural Science Foundation of China (Grant Nos. 72322009, 72173113), and Zhejiang Provincial Philosophy and Social Sciences Planning Project (23SYS04ZD). Qijun Wang gratefully acknowledges support from Zhejiang University. Lei Wang gratefully acknowledges support from the Ministry of Science and Technology [STI 2030—Major Projects 2021ZD0200409] and the National Natural Science Foundation of China (Grant Nos. 72371226). Sudeep Bhatia gratefully acknowledges support from the NSF (Grant SES-1847794).

Data availability

The data generated in this study have been deposited to the Open Science Framework (https://osf.io/pt2n4/overview?view_only=b705f153817542249d6e04bd511882f7).

Code availability

The code for the analyses presented in this article have been deposited to the Open Science Framework (https://osf.io/pt2n4/overview?view_only=b705f153817542249d6e04bd511882f7).

References

Abdurahman, Suhaib, Mohammad Atari, Farzan Karimi-Malekabadi, Mona J Xue, Jackson Trager, Peter S Park, Preni Golazizian, Ali Omrani, and Morteza Dehghani. 2024. "Perils and Opportunities in Using Large Language Models in Psychological Research." *PNAS Nexus* 3(7): pgae245. doi:10.1093/pnasnexus/pgae245.

Åkerlund, David, Bart H. H. Golsteyn, Hans Grönqvist, and Lena Lindahl. 2016. "Time Discounting and Criminal Behavior." *Proceedings of the National Academy of Sciences* 113(22): 6160–65. doi:10.1073/pnas.1522445113.

AlKhamissi, Badr, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. "Investigating Cultural Alignment of Large Language Models." In Bangkok, Thailand: Association for Computational Linguistics, 12404–22. doi:10.18653/v1/2024.acl-long.671.

Appel, Ruth, Peter McCrory, Alex Tamkin, Miles McCain, Tyler Neylon, and Michael Stern. 2025. "Anthropic Economic Index Report: Uneven Geographic and Enterprise AI Adoption." doi:10.48550/arXiv.2511.15080.

Argyle, Lisa P., 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): 337–51. doi:10.1017/pan.2023.2.

Atari, Mohammad, Mona J. Xue, Peter S. Park, Damián Ezequiel Blasi, and Joseph Henrich. 2023. "Which Humans?" doi:10.31234/osf.io/5b26t.

Becker, Anke, Benjamin Enke, and Armin Falk. 2020. "Ancient Origins of the Global Variation in Economic Preferences." *AEA Papers and Proceedings* 110: 319–23. doi:10.1257/pandp.20201071.

Bhatia, Sudeep. 2024. "Exploring Variability in Risk Taking with Large Language Models." *Journal of Experimental Psychology: General* 153(7): 1838–60. doi:10.1037/xge0001607.

Bhatia, Sudeep, Simon T. Van Baal, Feiyi Wang, and Lukasz Walasek. 2025. "Computational Analysis of 100 K Choice Dilemmas: Decision Attributes, Trade-off Structures, and Model-Based Prediction." *Proceedings of the National Academy of Sciences* 122(17): e2406489122. doi:10.1073/pnas.2406489122.

Bierbrauer, Felix, Axel Ockenfels, Andreas Pollak, and Désirée Rückert. 2017. "Robust Mechanism Design and Social Preferences." *Journal of Public Economics* 149: 59–80. doi:10.1016/j.jpubeco.2017.03.003.

Binz, Marcel, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, et al. 2025. "A Foundation Model to Predict and Capture Human Cognition." *Nature* 644(8078): 1002–9. doi:10.1038/s41586-025-09215-4.

Bisbee, James, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson. 2024. "Synthetic Replacements for Human Survey Data? The Perils of Large Language Models." *Political Analysis* 32(4): 401–16. doi:10.1017/pan.2024.5.

Bishai, David M. 2004. "Does Time Preference Change with Age?" *Journal of Population Economics* 17(4): 583–602. doi:10.1007/s00148-004-0187-0.

Burks, Stephen V., Jeffrey P. Carpenter, Lorenz Goette, and Aldo Rustichini. 2009. "Cognitive Skills Affect Economic Preferences, Strategic Behavior, and Job Attachment." *Proceedings of the National Academy of Sciences* 106(19): 7745–50.

649 doi:10.1073/pnas.0812360106.
650 Cao, Yiming, Benjamin Enke, Armin Falk, Paola Giuliano, and Nathan Nunn. 2021. “Herding,
651 Armed Conflict, and a Culture of Honor: Global Evidence.” NBER working paper.
652 doi:10.3386/w29250.
653 Capraro, Valerio, Austin Lentsch, Daron Acemoglu, Selin Akgun, Aisel Akhmedova, Ennio
654 Bilancini, Jean-François Bonnefon, et al. 2024. “The Impact of Generative Artificial
655 Intelligence on Socioeconomic Inequalities and Policy Making.” *PNAS Nexus* 3(6):
656 pgae191. doi:10.1093/pnasnexus/pgae191.
657 Ceron, Tanise, Neele Falk, Ana Barić, Dmitry Nikolaev, and Sebastian Padó. 2024. “Beyond
658 Prompt Brittleness: Evaluating the Reliability and Consistency of Political
659 Worldviews in LLMs.” *Transactions of the Association for Computational Linguistics*
660 12: 1378–1400. doi:10.1162/tacl_a_00710.
661 Chatterji, Aaron, Thomas Cunningham, David J. Deming, Zoe Hitzig, Christopher Ong, Carl Yan
662 Shan, and Kevin Wadman. 2025. “How People Use ChatGPT.” NBER working paper.
663 doi:10.3386/w34255.
664 Chen, M. Keith. 2013. “The Effect of Language on Economic Behavior: Evidence from Savings
665 Rates, Health Behaviors, and Retirement Assets.” *American Economic Review* 103(2):
666 690–731. doi:10.1257/aer.103.2.690.
667 Chen, Yiting, Tracy Xiao Liu, You Shan, and Songfa Zhong. 2023. “The Emergence of Economic
668 Rationality of GPT.” *Proceedings of the National Academy of Sciences* 120(51):
669 e2316205120. doi:10.1073/pnas.2316205120.
670 Crockett, M.J., and Lisa Messeri. 2025. “AI Surrogates and Illusions of Generalizability in
671 Cognitive Science.” *Trends in Cognitive Sciences*: S1364661325002517.
672 doi:10.1016/j.tics.2025.09.012.
673 Croson, Rachel, and Uri Gneezy. 2009. “Gender Differences in Preferences.” *Journal of*
674 *Economic Literature* 47(2): 448–74. doi:10.1257/jel.47.2.448.
675 Cui, Ziyan, Ning Li, and Huaikang Zhou. 2025. “A Large-Scale Replication of Scenario-Based
676 Experiments in Psychology and Management Using Large Language Models.” *Nature*
677 *Computational Science* 5(8): 627–34. doi:10.1038/s43588-025-00840-7.
678 Della Vigna, Stefano, and Ulrike Malmendier. 2006. “Paying Not to Go to the Gym.” *American*
679 *Economic Review* 96(3): 694–719. doi:10.1257/aer.96.3.694.
680 Dillion, Danica, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. “Can AI Language Models
681 Replace Human Participants?” *Trends in Cognitive Sciences* 27(7): 597–600.
682 doi:10.1016/j.tics.2023.04.008.
683 Dohmen, Thomas, Armin Falk, David Huffman, and Uwe Sunde. 2010. “Are Risk Aversion and
684 Impatience Related to Cognitive Ability?” *American Economic Review* 100(3): 1238–
685 60. doi:10.1257/aer.100.3.1238.
686 Dupas, Pascaline, and Jonathan Robinson. 2013. “Why Don’t the Poor Save More? Evidence
687 from Health Savings Experiments.” *American Economic Review* 103(4): 1138–71.
688 doi:10.1257/aer.103.4.1138.
689 Durmus, Esin, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton
690 Bakhtin, Carol Chen, et al. 2024. “Towards Measuring the Representation of
691 Subjective Global Opinions in Language Models.” doi:10.48550/arXiv.2306.16388.
692 Enke, Benjamin. 2019. “Kinship, Cooperation, and the Evolution of Moral Systems.” *The*

Quarterly Journal of Economics 134(2): 953–1019. doi:10.1093/qje/qjz001.
Enke, Benjamin. 2022. "Market Exposure and Human Morality." *Nature Human Behaviour* 7(1): 134–41. doi:10.1038/s41562-022-01480-x.
Epper, Thomas, Ernst Fehr, Helga Fehr-Duda, Claus Thustrup Kreiner, David Dreyer Lassen, Søren Leth-Petersen, and Gregers Nytoft Rasmussen. 2020. "Time Discounting and Wealth Inequality." *American Economic Review* 110(4): 1177–1205. doi:10.1257/aer.20181096.
Exley, Christine L, Oliver P Hauser, Molly Moore, and John-Henry Pezzuto. 2025. "Believed Gender Differences in Social Preferences." *The Quarterly Journal of Economics* 140(1): 403–58. doi:10.1093/qje/qjae030.
Falk, Armin, Anke Becker, Thomas Dohmen, Benjamin Enke, David Huffman, and Uwe Sunde. 2018. "Global Evidence on Economic Preferences." *The Quarterly Journal of Economics* 133(4): 1645–92. doi:10.1093/qje/qjy013.
Falk, Armin, Anke Becker, Thomas Dohmen, David Huffman, and Uwe Sunde. 2023. "The Preference Survey Module: A Validated Instrument for Measuring Risk, Time, and Social Preferences." *Management Science* 69(4): 1935–50. doi:10.1287/mnsc.2022.4455.
Falk, Armin, and Johannes Hermle. 2018. "Relationship of Gender Differences in Preferences to Economic Development and Gender Equality." *Science* 362(6412): eaas9899. doi:10.1126/science.aas9899.
Galor, Oded, and Ömer Özak. 2016. "The Agricultural Origins of Time Preference." *American Economic Review* 106(10): 3064–3103. doi:10.1257/aer.20150020.
Gao, Yuan, Dokyun Lee, Gordon Burtch, and Sina Fazelpour. 2025. "Take Caution in Using LLMs as Human Surrogates." *Proceedings of the National Academy of Sciences* 122(24): e2501660122. doi:10.1073/pnas.2501660122.
Genesove, David, and Christopher Mayer. 2001. "Loss Aversion and Seller Behavior: Evidence from the Housing Market." *The Quarterly Journal of Economics* 116(4): 1233–60. doi:10.1162/003355301753265561.
Goli, Ali, and Amandeep Singh. 2024. "Frontiers: Can Large Language Models Capture Human Preferences?" *Marketing Science* 43(4): 709–22. doi:10.1287/mksc.2023.0306.
Graham, Mark, Sanna Ojanperä, and Martin Dittus. 2019. "Internet Geographies: Data Shadows and Digital Divisions of Labor." In *Society and the Internet*, Oxford University PressOxford, 58–79. doi:10.1093/oso/9780198843498.003.0004.
Grossmann, Igor, Matthew Feinberg, Dawn C. Parker, Nicholas A. Christakis, Philip E. Tetlock, and William A. Cunningham. 2023. "AI and the Transformation of Social Science Research." *Science* 380(6650): 1108–9. doi:10.1126/science.adi1778.
Guilbeault, Douglas, Solène Delecourt, and Bhargav Srinivasa Desikan. 2025. "Age and Gender Distortion in Online Media and Large Language Models." *Nature* 646(8087): 1129–37. doi:10.1038/s41586-025-09581-z.
Guiso, Luigi, Paola Sapienza, and Luigi Zingales. 2006. "Does Culture Affect Economic Outcomes?" *Journal of Economic Perspectives* 20(2): 23–48. doi:10.1257/jep.20.2.23.
Guiso, Luigi, Paola Sapienza, and Luigi Zingales. 2009. "Cultural Biases in Economic Exchange?" *The Quarterly Journal of Economics* 124(3): 1095–1131. doi:10.1162/qjec.2009.124.3.1095.

Hanushek, Eric A, Lavinia Kinne, Philipp Lergetporer, and Ludger Woessmann. 2022. "Patience, Risk-Taking, and Human Capital Investment across Countries." *The Economic Journal* 132(646): 2290–2307. doi:10.1093/ej/ueab105.

Henrich, Joseph, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, and Richard McElreath. 2001. "In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies." *American Economic Review* 91(2): 73–78. doi:10.1257/aer.91.2.73.

Hofstede, Geert, Gert Jan Hofstede, and Michael Minkov. 2010. *Cultures and Organizations: Software of the Mind: Intercultural Cooperation and Its Importance for Survival*. Revised and expanded third edition. New York: McGraw-Hill.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81(396): 945–60. doi:10.1080/01621459.1986.10478354.

Horton, John J. 2023. "Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?" *ACM Conference on Economics and Computation*. doi:10.1145/3670865.3673513.

Inglehart, Ronald, and Christian Welzel. 2005. *Modernization, Cultural Change, and Democracy: The Human Development Sequence*. Cambridge University Press. doi:10.1017/CBO9780511790881.

Jiang, Shapeng, Lijia Wei, and Chen Zhang. 2024. "Donald Trumps in the Virtual Polls: Simulating and Predicting Public Opinions in Surveys Using Large Language Models." doi:10.48550/ARXIV.2411.01582.

Joshi, Pratik, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. "The State and Fate of Linguistic Diversity and Inclusion in the NLP World." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, 6282–93. doi:10.18653/v1/2020.acl-main.560.

Kanagaretnam, Kiridaran, Jimmy Lee, Chee Yeow Lim, and Gerald J. Lobo. 2022. "Trusting the Stock Market: Further Evidence from IPOs around the World." *Journal of Banking & Finance* 142: 106557. doi:10.1016/j.jbankfin.2022.106557.

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, et al. 2020. "Scaling Laws for Neural Language Models." doi:10.48550/arXiv.2001.08361.

Knack, S., and P. Keefer. 1997. "Does Social Capital Have an Economic Payoff? A Cross-Country Investigation." *The Quarterly Journal of Economics* 112(4): 1251–88. doi:10.1162/003355300555475.

Kosse, Fabian, and Michela M. Tincani. 2020. "Prosociality Predicts Labor Market Success around the World." *Nature Communications* 11(1): 5298. doi:10.1038/s41467-020-19007-1.

Kremer, Michael, Gautam Rao, and Frank Schilbach. 2019. "Behavioral Development Economics." In *Handbook of Behavioral Economics: Applications and Foundations 1*, Handbook of Behavioral Economics - Foundations and Applications 2, North-Holland, 345–458. doi:10.1016/bs.hesbe.2018.12.002.

Leng, Yan, and Yuan Yuan. 2024. "Do LLM Agents Exhibit Social Behavior?" doi:10.48550/arXiv.2312.15198.

Li, Cheng, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024. "CultureLLM: Incorporating Cultural Differences into Large Language Models." *Advances in Neural Information Processing Systems* 37: 84799–838.

Lin, Zhicheng. 2025. "Six Fallacies in Substituting Large Language Models for Human Participants." *Advances in Methods and Practices in Psychological Science* 8(3): 25152459251357566. doi:10.1177/25152459251357566.

Lu, Jackson G., Lesley Luyang Song, and Lu Doris Zhang. 2025. "Cultural Tendencies in Generative AI." *Nature Human Behaviour*: 1–10. doi:10.1038/s41562-025-02242-1.

Luttmer, Erzo F. P., and Monica Singhal. 2011. "Culture, Context, and the Taste for Redistribution." *American Economic Journal: Economic Policy* 3(1): 157–79. doi:10.1257/pol.3.1.157.

Mei, Qiaozhu, Yutong Xie, Walter Yuan, and Matthew O. Jackson. 2024. "A Turing Test of Whether AI Chatbots Are Behaviorally Similar to Humans." *Proceedings of the National Academy of Sciences* 121(9): e2313925121. doi:10.1073/pnas.2313925121.

Meng, Juanjuan. 2024. "AI Emerges as the Frontier in Behavioral Science." *Proceedings of the National Academy of Sciences* 121(10): e2401336121. doi:10.1073/pnas.2401336121.

Morgan, James N. 1975. "Using Survey Data from the University of Michigan's Survey Research Center." *American Economic Review* 65(2): 250–56.

Ni, Shiwen, Guhong Chen, Shuaimin Li, Xuanang Chen, Siyi Li, Bingli Wang, Qiyao Wang, et al. 2025. "A Survey on Large Language Model Benchmarks." doi:10.48550/arXiv.2508.15361.

Nieminen, Mika, and Kamila Kuziemska-Pawlak. 2024. "Cross-Country Variation in Economic Preferences and the Asset Composition of International Investment Positions." *Journal of International Money and Finance* 146: 103130. doi:10.1016/j.jimonfin.2024.103130.

Pataranutaporn, Pat, Nattavudh Powdthavee, Chayapatr Archiwaranguprok, and Pattie Maes. 2025. "Simulating Human Well-Being with Large Language Models: Systematic Validation and Misestimation across 64,000 Individuals from 64 Countries." *Proceedings of the National Academy of Sciences* 122(48): e2519394122. doi:10.1073/pnas.2519394122.

Pickard, Harry, Thomas Dohmen, and Bert Van Landeghem. 2024. "Inequality and Risk Preference." *Journal of Risk and Uncertainty* 69(2): 191–217. doi:10.1007/s11166-024-09440-8.

Qin, Xin, Mingpeng Huang, and Jie Ding. 2024. "AITurk: Using ChatGPT for Social Science Research." doi:10.2139/ssrn.4922861.

Röttger, Paul, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Rose Kirk, Hinrich Schütze, and Dirk Hovy. 2024. "Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in Large Language Models." doi:10.48550/arXiv.2402.16786.

Rozado, David. 2024. "The Political Preferences of LLMs." *PLOS One* 19(7): e0306621. doi:10.1371/journal.pone.0306621.

Ruprechter, Thorsten, Keith Burghardt, and Denis Helic. 2023. "Poor Attention: The Wealth and Regional Gaps in Event Attention and Coverage on Wikipedia." *PLOS One* 18(11):

e0289325. doi:10.1371/journal.pone.0289325.

Santurkar, Shibani, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. "Whose Opinions Do Language Models Reflect?" In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23, Honolulu, Hawaii, USA: JMLR.org, 29971–4.

Schwartz, Shalom. 2006. "A Theory of Cultural Value Orientations: Explication and Applications." *Comparative Sociology* 5(2–3): 137–82. doi:10.1163/156913306778667357.

Sinacola, Enzo, Arnault Pachot, and Thierry Petit. 2025. "LLMs, Virtual Users, and Bias: Predicting Any Survey Question without Human Data." In *Proceedings of 17th International Conference on Machine Learning and Computing*, , 396–407. doi:10.1007/978-3-031-94892-3_29.

Stantcheva, Stefanie. 2023. "How to Run Surveys: A Guide to Creating Your Own Identifying Variation and Revealing the Invisible." *Annual Review of Economics* 15(1): 205–34. doi:10.1146/annurev-economics-091622-010157.

Sunde, Uwe, Thomas Dohmen, Benjamin Enke, Armin Falk, David Huffman, and Gerrit Meyerheim. 2022. "Patience and Comparative Development." *Review of Economic Studies* 89(5): 2806–40. doi:10.1093/restud/rdab084.

Tannenbaum, David, Alain Cohn, Christian Lukas Zünd, and Michel André Maréchal. 2025. "What Do Cross-Country Surveys Tell Us about Social Capital?" *Review of Economics and Statistics* 107(1): 142–51. doi:10.1162/rest_a_01245.

Tao, Yan, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. "Cultural Bias and Cultural Alignment of Large Language Models." *PNAS Nexus* 3(9): pgae346. doi:10.1093/pnasnexus/pgae346.

Villalobos, Pablo, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. 2024. "Position: Will We Run out of Data? Limits of LLM Scaling Based on Human-Generated Data." In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24, Vienna, Austria: JMLR.org, 49523–44.

Wang, Angelina, Jamie Morgenstern, and John P. Dickerson. 2025. "Large Language Models That Replace Human Participants Can Harmfully Misportray and Flatten Identity Groups." *Nature Machine Intelligence* 7(3): 400–411. doi:10.1038/s42256-025-00986-z.

Wu, Jing Cynthia, Jin Xi, and Shihan Xie. 2025. *LLM Survey Framework: Coverage, Reasoning, Dynamics, Identification*. NBER working paper. doi:10.3386/w34308.

Xie, Yutong, Qiaozhu Mei, Walter Yuan, and Matthew O. Jackson. 2025. "Using Large Language Models to Categorize Strategic Situations and Decipher Motivations behind Human Behaviors." *Proceedings of the National Academy of Sciences* 122(35): e2512075122. doi:10.1073/pnas.2512075122.

Zhu, Jian-Qiao, Haijiang Yan, and Thomas L. Griffiths. 2025. "Language Models Trained to Do Arithmetic Predict Human Risky and Intertemporal Choice." In *International Conference on Representation Learning*, 64953–71.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AppendixLargeLanguageModelsasSyntheticRespondentsforEconomicPreferencesAGlobalAuditNumbered.pdf