Appendix

Appendix A. Execution of the Virtual Global Preference Survey

This virtual Global Preference Survey (Virtual GPS) follows the framework of the original GPS program established by Falk et al. (2018). A summary of the datasets from each model is provided in Table A.1. To ensure comparability across models, we retain only those respondents from whom we successfully obtained responses from all eight LLMs. The merged dataset constructed from these eight models is described in Table A.2.

There are, however, several minor differences between our virtual survey and the original program.

1. **Translation.** We use the Google Translation API to translate personas and survey items from English to other languages. Since this API does not support Dari, we use GPT 4o-mini to translate into Dari.
2. **Sample exclusion.** Morocco is excluded from the virtual GPS because the Google Translate API does not support Moroccan Arabic, and prior work reports that GPT-based models performed unreliably on Moroccan Arabic (Aftiss et al. 2025). For Cambodia, pilot data collected with GPT-3.5-turbo revealed a high rate of non-parsable or off-scale answers (e.g., missing items, inconsistent response formats). To avoid introducing country-specific measurement error, we therefore omit Cambodia from the LLM simulations for all models. These two countries account for only a small fraction of the original GPS sample.
3. **Language selection.** For each country, we predict preferences only for individuals surveyed in the most commonly sampled language. As a result, the total sample size of our virtual GPS is slightly smaller than that of the original GPS program, at approximately 90% of the original number of respondents.

We implement the virtual GPS project in three steps:

1. We extract respondents' demographic and country information (age, gender, subjective cognitive skills, region, and country) from the open-source GPS dataset and embed this information in a system prompt.

2. We present the GPS survey items as user prompts. Both system and user prompts are written in the local survey language rather than English (except for the GPT 4o-mini English-only version, in which all prompts are English). The survey items are listed in Table A.3, and the ordering of the lottery choice and intertemporal choice questions follows the original GPS program.
3. During the simulation, we present only one survey item to an LLM per API request, ensuring that the model answers each question without access to its previous responses. LLMs often produce relatively long answers in various languages, with the chosen option number contained in the text. We therefore use GPT 4o-mini to parse the output and extract the specific option number.

The prompts are as follows:

- **System prompt**:

  You are [age]-year-old [gender], living in [region], [country]. You are [subjective cognitive skills]. You are taking a survey. Please make your own choice according to the question. Please begin with "I will".

- **User prompt to elicit economic preferences**:

  [Survey item], see Table A.3

We collected GPT 4o-mini data from August 7 to August 9, 2024, Gemini 2.0-flash and Deepseek V3 data from March 7 to March 19, 2025, GPT 3.5 turbo, GPT 4.1-mini, GPT 4o-mini (English-only) data from April 28 to May 6, 2025, GPT 4o data from September 12 to 14, 2025, and Gemini 2.5-flash data from September 14 to October 4, 2025. Due to differences in model coverage across countries and languages, the resulting sample sizes vary slightly by model (see Table A.1). During data collection with GPT 4o-mini, Gemini 2.0-flash and DeepSeek V3, we did not collect data for Morocco. For GPT 3.5 turbo, GPT 4.1-mini, GPT 4o, Gemini 2.5-flash, we did not collect data for either Morocco or Cambodia.

During the initial data collection with GPT 4o-mini, a programming error affected the self-assessment items (e.g., "How willing are you to give up something that is beneficial for you today in order to benefit more from that in the future?"). When the model responded with "0", the script incorrectly treated this as a missing-value error and re-

collected the response until a non-zero value was obtained. We corrected this error for all subsequent rounds of data collection.

Table A.1. Sample sizes and country coverage for each LLM in the virtual GPS

Model	Number of observations	Percentage of original GPS observations	Number of countries
GPT 3.5-turbo	69,679	86.73	74
GPT 4o-mini	70,650	87.94	75
GPT 4o-mini (English)	70,335	87.55	75
GPT 4o	69,680	86.73	74
GPT 4.1-mini	69,671	86.72	74
Gemini 2.0-flash	70,239	87.43	75
Gemini 2.5-flash	69,281	86.24	74
DeepSeek V3	69,304	86.27	75
Original GPS	80,337	100	76

Notes: Reported observations and countries refer to respondents retained in the virtual GPS. Percentages indicate coverage relative to the original GPS sample (N = 80,337).

Table A.2. Sample sizes in the virtual GPS and original GPS for each country

Country	Number of LLM observations	Number of GPS observations	Percentage of simulated observations
Afghanistan	620	1000	62.00
Algeria	1020	1022	99.80
Argentina	1000	1000	100.00
Australia	1002	1002	100.00
Austria	1001	1001	100.00
Bangladesh	998	999	99.90
Bolivia	998	998	100.00
Bosnia Herzegovina	522	1004	51.99
Botswana	580	1000	58.00
Brazil	1003	1003	100.00
Cameroon	832	1000	83.20
Canada	784	1001	78.32
Chile	1003	1003	100.00
China	2574	2574	100.00
Colombia	1000	1000	100.00
Costa Rica	1000	1000	100.00
Croatia	992	992	100.00
Czech Republic	1004	1005	99.90

Egypt	1020	1020	100.00
Estonia	683	1004	68.03
Finland	999	1000	99.90
France	1000	1001	99.90
Georgia	950	1000	95.00
Germany	997	997	100.00
Ghana	482	1000	48.20
Greece	1000	1000	100.00
Guatemala	1000	1000	100.00
Haiti	504	504	100.00
Hungary	846	1004	84.26
India	1268	2539	49.94
Indonesia	1000	1000	100.00
Iran	2388	2507	95.25
Iraq	861	1000	86.10
Israel	822	999	82.28
Italy	1004	1004	100.00
Japan	995	1000	99.50
Jordan	999	1000	99.90
Kazakhstan	716	999	71.67
Kenya	510	1000	51.00
Lithuania	999	999	100.00
Malawi	842	1000	84.20
Mexico	1000	1000	100.00
Moldova	771	1000	77.10
Netherlands	1000	1000	100.00
Nicaragua	1000	1000	100.00
Nigeria	658	1000	65.80
Pakistan	1004	1004	100.00
Peru	1000	1000	100.00
Philippines	412	1000	41.20
Poland	998	999	99.90
Portugal	998	998	100.00
Romania	994	994	100.00
Russia	1498	1498	100.00
Rwanda	288	1000	28.80
Saudi Arabia	1035	1035	100.00
Serbia	1023	1023	100.00
South Africa	316	1000	31.60
South Korea	1000	1000	100.00
Spain	1000	1000	100.00
Sri Lanka	710	1000	71.00

Suriname	504	504	100.00
Sweden	999	1000	99.90
Switzerland	659	1000	65.90
Tanzania	968	1000	96.80
Thailand	1000	1000	100.00
Turkey	1000	1000	100.00
Uganda	533	1000	53.30
Ukraine	620	1000	62.00
United Arab Emirates	1000	1000	100.00
United Kingdom	1030	1030	100.00
United States	1068	1072	99.63
Venezuela	998	999	99.90
Vietnam	1000	1000	100.00
Zimbabwe	485	1000	48.50

Notes: *Number of LLM observations* aggregates synthetic respondents across all eight LLMs in the merged dataset. *Percentage of simulated observations* reports coverage relative to the original GPS sample for each country.

73 **Table A.3.** Survey items in the virtual GPS

Preference	Type of item	Survey item
Risk taking	Sequence of five interdependent quantitative questions	*Which option would you prefer: Option (1): A draw with a 50-percent chance of receiving* **[Amount A] [Local currency]** *and the same 50-percent chance of receiving nothing, or Option (2): the amount of* **[Amount B] [Local currency]** *as a sure payment?*
	Self-assessment	In general, how willing are you to take risks, using a scale from 0 to 10, where 0 means you are "completely unwilling to take risks" and 10 means you are "very willing to take risks". You can also use any number between 0 and 10 to indicate where you fall on the scale, using 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10.
Patience	Sequence of five interdependent quantitative questions	*Which option would you prefer: Option (1): receive* **[Amount A] [Local currency]** *today or Option (2): receive* **[Amount B] [Local currency]** *in 12 months? Please assume there is no inflation.*
	Self-assessment	*How willing are you to give up something that is beneficial for you today in order to benefit more from that in the future? Please indicate your answer on a scale from 0 to 10. A 0 means "completely unwilling to do so", and a 10 means "'very willing to do so". You can also use any number between 0 and 10 to indicate where you fall on the scale, using 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10.*
Negative Reciprocity	Willingness to act	*How willing are you to punish someone who treats* ***[you]*** *unfairly, even if there may be costs for you? Please indicate your answer on a scale from 0 to 10. A 0 means "completely unwilling to do so", and a 10 means "very willing to do so". You can also use any number between 0 and 10 to indicate where you fall on the scale, using 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10.*
	Willingness to act	*How willing are you to punish someone who treats* ***[others]*** *unfairly, even if there may be costs for you? Please indicate your answer on a scale from 0 to 10. A 0 means "completely unwilling to do so", and a 10 means "very willing to do so". You can also use any number between 0 and 10 to indicate where you fall on the scale, using 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10.*
	Self-assessment	*How well does each of the following statement describe you? "If I am treated very unjustly, I will take revenge at the first occasion, even if there is a cost to do so". Please indicate your answer on a scale from 0 to 10. A 0 means "does not describe me at all", and a 10 means "describes me perfectly". You can use any number between 0 and 10 to indicate where you fall on the scale, using 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10.*
Positive Reciprocity	Self-assessment	*How well does each of the following statement describe you? "When someone does me a favor, I am willing to*

		return it". Please indicate your answer on a scale from 0 to 10. A 0 means "does not describe me at all", and a 10 means "describes me perfectly". You can use any number between 0 and 10 to indicate where you fall on the scale, using 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10.
	Hypothetical situation	*You are in an area you are not familiar with, and you realize that you have lost your way. You ask a stranger for directions. The stranger offers to take you to your destination. Helping you costs the stranger about* **[Amount A] [Local currency]** *in total. However, the stranger says he or she does not want any money from you. You have six presents with you. The cheapest present costs* **[Min amount] [Local currency]**, *and the most expensive one costs* **[Max amount] [Local currency]**. *Do you give one of the presents to the stranger as a "thank you" gift? Choose between： Option (1) No, would not give present. Option (2) Yes, a present worth* **[Amount B] [Local currency]**. *Option (3) Yes, a present worth* **[Amount C] [Local currency]**. *Option (4) Yes, a present worth* **[Amount D] [Local currency]**. *Option (5) Yes, a present worth* **[Amount E] [Local currency]**. *Option (6) Yes, a present worth* **[Amount F] [Local currency]**. *Option (7) Yes, a present worth* **[Amount G] [Local currency]**.
Altruism	Willingness to act	*How willing are you to give to good causes without expecting anything in return? Please indicate your answer on a scale from 0 to 10. A 0 means "completely unwilling to do so", and a 10 means "very willing to do so". You can also use any number between 0 and 10 to indicate where you fall on the scale, using 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10.*
	Hypothetical situation	*Today you unexpectedly received* **[Amount] [Local currency]**. *How much of this amount would you donate to a good cause? Values between 0 and* **[Amount]** *are allowed.*
Trust	Self-assessment	*How well does each of the following statement describe you? "I assume that people have only the best intentions." Please indicate your answer on a scale from 0 to 10. A 0 means "does not describe me at all," and a 10 means "describes me perfectly." You can use any number between 0 and 10 to indicate where you fall on the scale, using 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10.*

Appendix B. Alignment at the individual and country levels

Here, we provide additional details on the alignment between LLM-predicted and GPS-surveyed preferences at the individual and country levels:

1. **Within-country individual-level correlations**: a robustness check on the individual-level alignment reported in the main text.
2. **Half-split internal consistency**: assessing the reliability of aggregating individual-level LLM-predicted preferences into country-level averages.

B.1 Individual-level Spearman correlations within each country

As a complement to the analysis in the main text, we compute individual-level Spearman correlations between GPS-surveyed and LLM-predicted preferences separately for each country, pooling all individuals within that country. A positive individual-level correlation in a given country indicates that the LLM's predicted preferences are directionally aligned with the GPS responses in that setting. This allows us to examine not only overall individual-level consistency, but also its geographical distribution.

Table B.1 reports, for each model and each preference domain, the number and percentage of countries with positive individual-level correlations. On average, 84.6% of these correlations are positive across all preferences and models (95% CI = [83.4%, 85.8%]; risk-taking: 98.7%; patience: 97.3%; negative reciprocity: 85.3%; positive reciprocity: 83.3%; altruism: 92.5%; trust: 50.7%).

Most models show high consistency for risk-taking and patience, with almost all countries showing positive correlations (around or above 95%). In contrast, the percentage of consistent cases for trust drops sharply (average: 50.7%). In particular, GPT-4o-mini (English), GPT-4o, Gemini 2.0-flash, and Gemini 2.5-flash all fall below 50%. Regardless of statistical significance, the proportion of within-country individual-level correlations that are positive is far above what would be expected by chance (50%; two-sided binomial test, $P < 0.001$) for all models and for all preference domains except trust.

Panels A of Figures B.1-B.8 visualize the geographical distribution of individual-level

consistency (i.e., Spearman correlations for each country). Panels B of Figures B.1-B.8 shows the country-level alignment, measured by the Spearman correlations between GPS-surveyed and LLM-predicted country-level preference averages (sampling-weighted means of individual-level preferences using the GPS ex post weights).

108

Table B.1. Number and percentage of consistent case

	Risk-taking		**Patience**		**Neg. recip.**		**Pos. recip.**		**Altruism**		**Trust**	
	Num. of ρ>0	% of ρ>0	Num. of ρ>0	% of ρ>0	Num. of ρ>0	% of ρ>0	Num. of ρ>0	% of ρ>0	Num. of ρ>0	% of ρ>0	Num. of ρ>0	% of ρ>0
GPT 3.5-turbo	72	97.3	70	94.6	61	82.4	61	82.4	66	89.2	63	85.1
GPT 4o-mini	72	97.3	70	94.6	65	87.8	65	87.8	68	91.9	62	83.8
GPT 4o-mini English	74	100	73	98.6	73	98.6	61	82.4	71	95.9	27	36.5
GPT 4o	74	100	72	97.3	64	86.5	66	89.2	67	90.5	16	21.6
GPT 4.1-mini	74	100	74	100	72	97.3	70	94.6	70	94.6	56	75.7
Gemini 2.0-flash	74	100	73	98.6	57	77.0	64	86.5	71	95.9	18	24.3
Gemini 2.5-flash	72	97.3	71	95.9	49	66.2	46	62.2	69	93.2	18	24.3
DeepSeek V3	72	97.3	73	98.6	64	86.5	60	81.1	65*	89.0	40	54.1
Overall	73.0	98.7	72.0	97.3	63.1	85.3	61.6	83.3	68.4	92.5	37.5	50.7

Notes: The total number of countries is 74. However, because there is no variation in altruism among the DeepSeek-predicted samples for Botswana, we are unable to compute the Spearman correlation for Botswana and therefore exclude it from DeepSeek's altruism column.

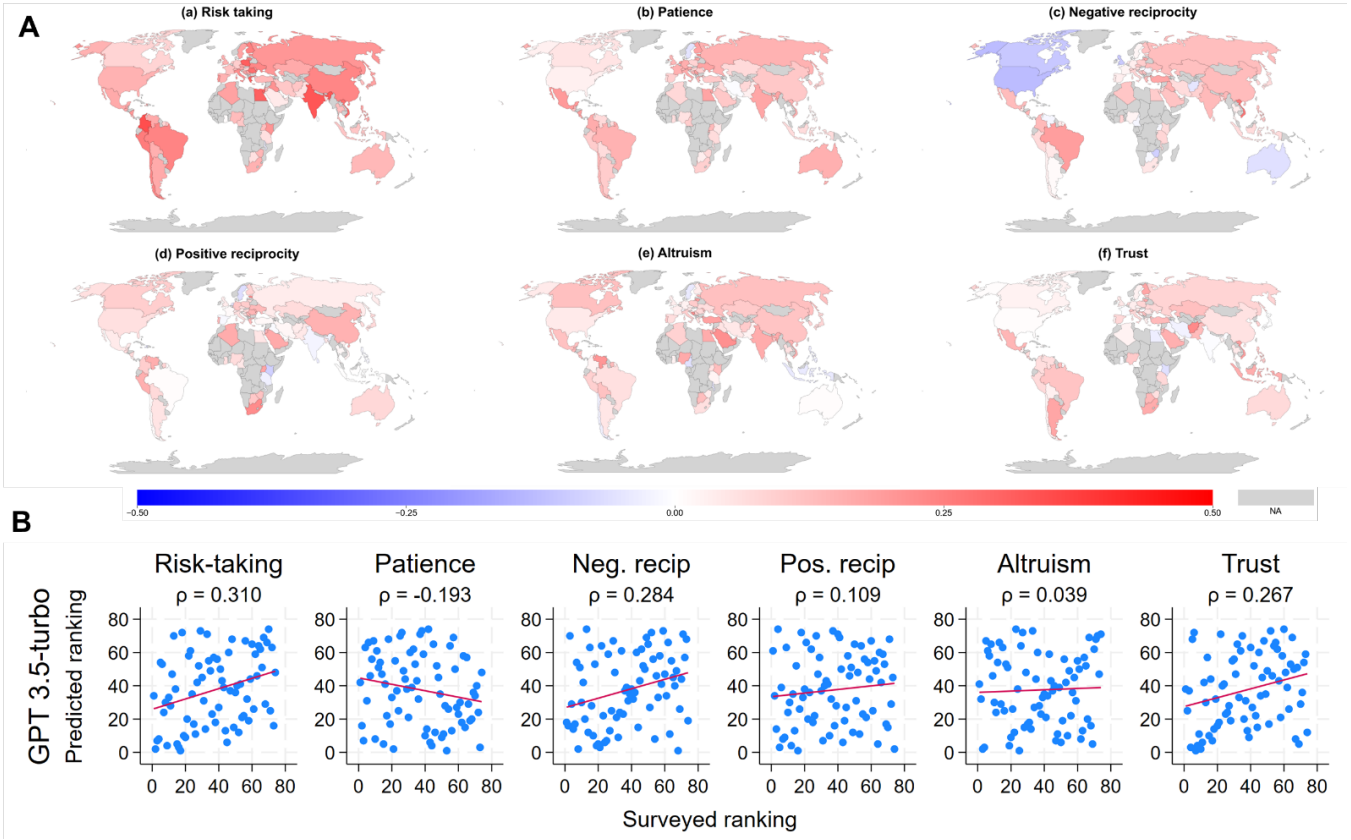

Figure B.1. Consistency across countries (GPT 3.5-turbo). (A) Individual-level consistency. For each country and preference domain, we compute the Spearman correlation between GPS-surveyed and LLM-predicted preferences across individuals. Countries are shaded from blue (negative correlations) to red (positive correlations). Positive correlations indicate that higher human scores tend to be associated with higher LLM scores, whereas negative correlations indicate the opposite pattern. **(B) Country-level consistency.** For each preference domain, we compute the Spearman correlation between country-level average GPS-surveyed preferences and country-level average LLM-predicted preferences. In the scatterplots, each point is a country, the x-axis shows the ranking based on GPS-surveyed country means, and the y-axis shows the ranking based on LLM-predicted country means. Positive correlations indicate that countries ranked higher in the survey also tend to be ranked higher by the LLM, whereas negative correlations indicate the reverse ordering.

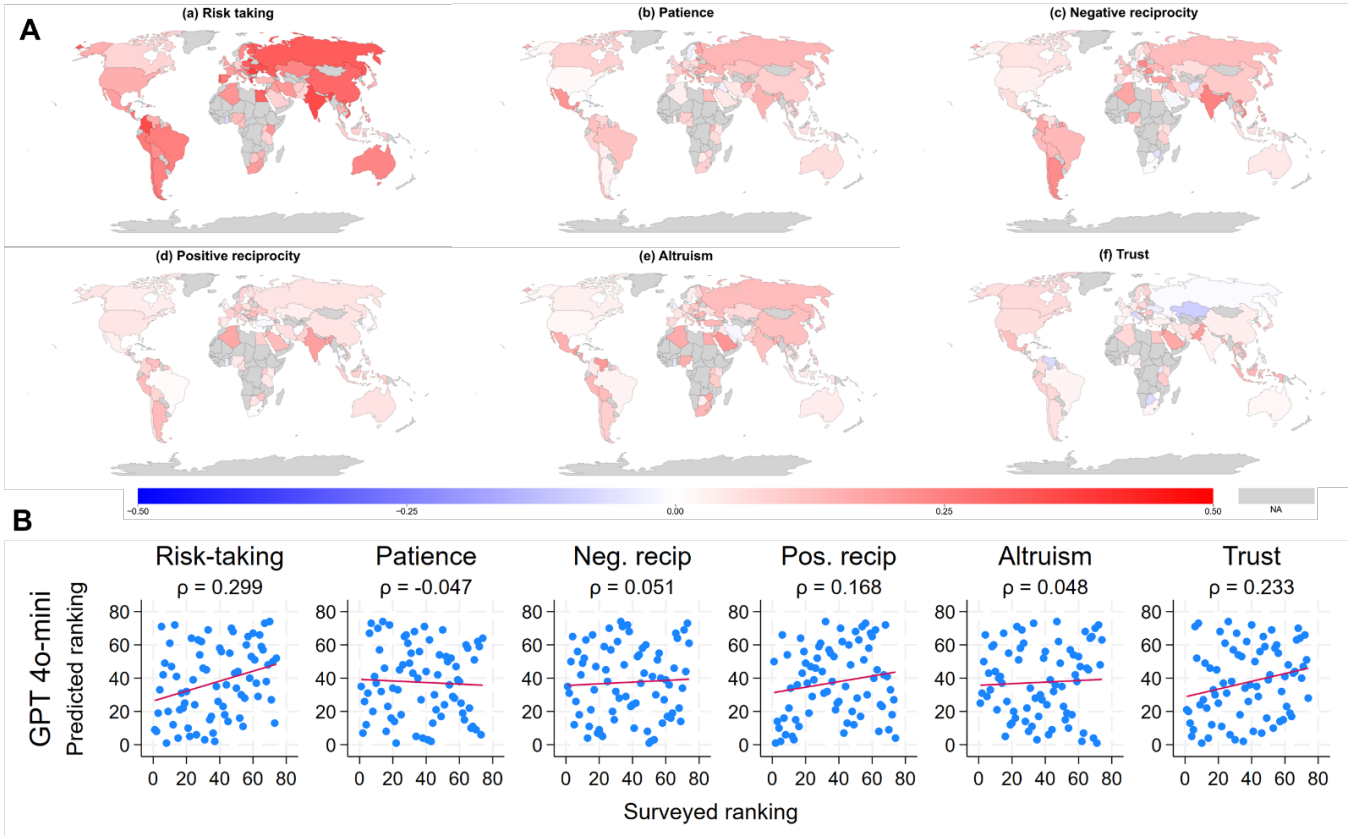

Figure B.2 Consistency across countries (GPT 4o-mini). (A) Individual-level consistency. For each country and preference domain, we compute the Spearman correlation between GPS-surveyed and LLM-predicted preferences across individuals. Countries are shaded from blue (negative correlations) to red (positive correlations). Positive correlations indicate that higher human scores tend to be associated with higher LLM scores, whereas negative correlations indicate the opposite pattern. **(B) Country-level consistency.** For each preference domain, we compute the Spearman correlation between country-level average GPS-surveyed preferences and country-level average LLM-predicted preferences. In the scatterplots, each point is a country, the x-axis shows the ranking based on GPS-surveyed country means, and the y-axis shows the ranking based on LLM-predicted country means. Positive correlations indicate that countries ranked higher in the survey also tend to be ranked higher by the LLM, whereas negative correlations indicate the reverse ordering.

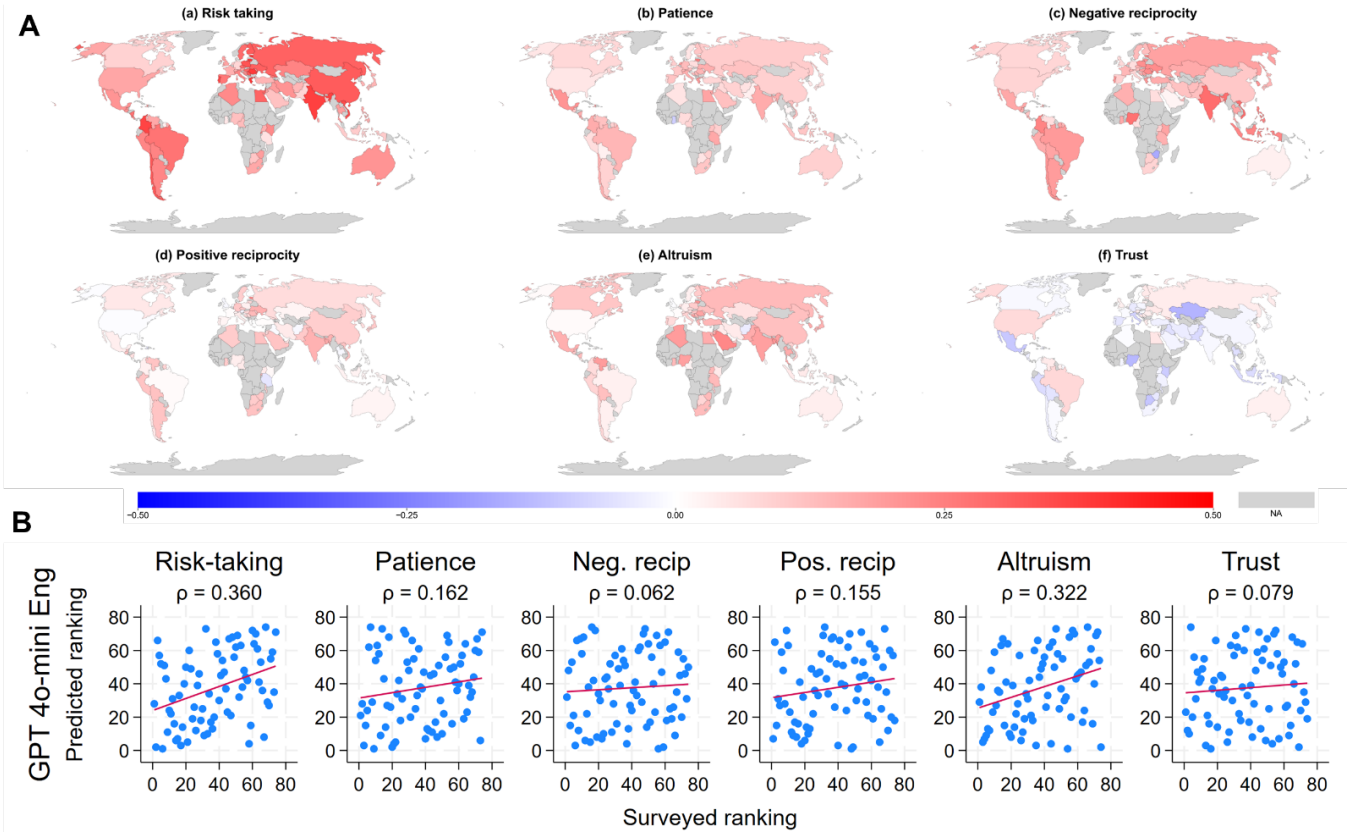

Figure B.3. Consistency across countries (GPT 4o-mini English). (A) Individual-level consistency. For each country and preference domain, we compute the Spearman correlation between GPS-surveyed and LLM-predicted preferences across individuals. Countries are shaded from blue (negative correlations) to red (positive correlations). Positive correlations indicate that higher human scores tend to be associated with higher LLM scores, whereas negative correlations indicate the opposite pattern. **(B) Country-level consistency.** For each preference domain, we compute the Spearman correlation between country-level average GPS-surveyed preferences and country-level average LLM-predicted preferences. In the scatterplots, each point is a country, the x-axis shows the ranking based on GPS-surveyed country means, and the y-axis shows the ranking based on LLM-predicted country means. Positive correlations indicate that countries ranked higher in the survey also tend to be ranked higher by the LLM, whereas negative correlations indicate the reverse ordering.

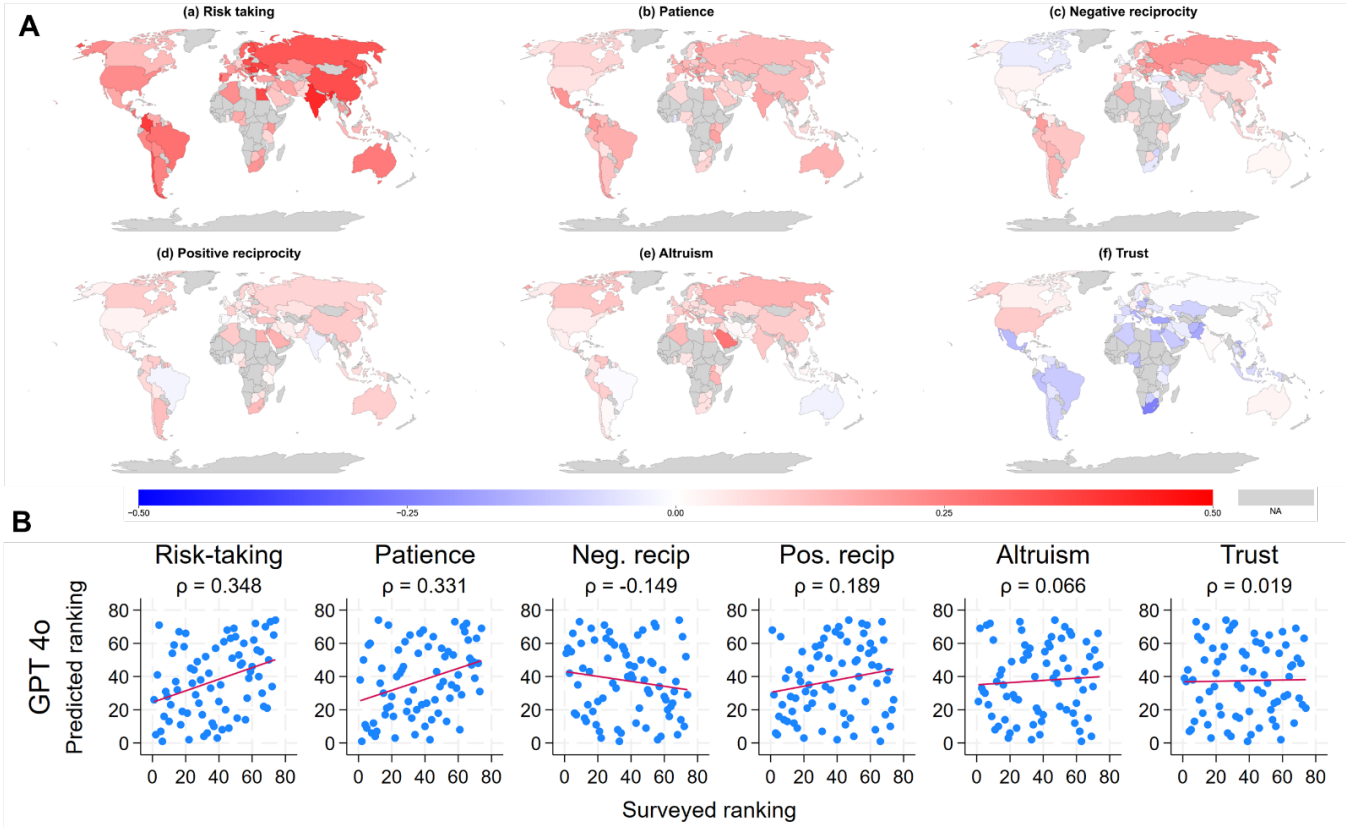

Figure B.4. Consistency across countries (GPT 4o). (A) Individual-level consistency. For each country and preference domain, we compute the Spearman correlation between GPS-surveyed and LLM-predicted preferences across individuals. Countries are shaded from blue (negative correlations) to red (positive correlations). Positive correlations indicate that higher human scores tend to be associated with higher LLM scores, whereas negative correlations indicate the opposite pattern. **(B) Country-level consistency.** For each preference domain, we compute the Spearman correlation between country-level average GPS-surveyed preferences and country-level average LLM-predicted preferences. In the scatterplots, each point is a country, the x-axis shows the ranking based on GPS-surveyed country means, and the y-axis shows the ranking based on LLM-predicted country means. Positive correlations indicate that countries ranked higher in the survey also tend to be ranked higher by the LLM, whereas negative correlations indicate the reverse ordering.

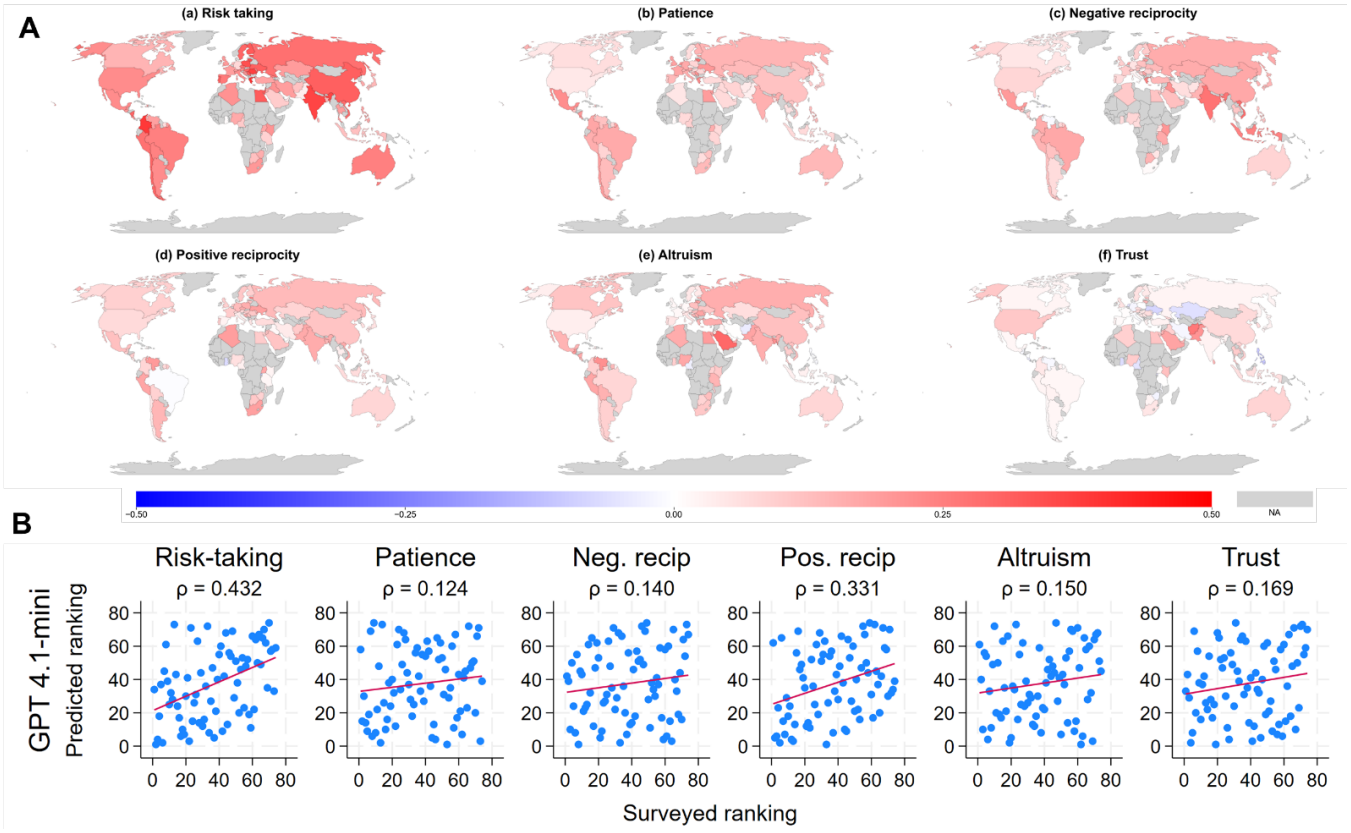

Figure B.5. Consistency across countries (GPT 4.1-mini). (A) Individual-level consistency. For each country and preference domain, we compute the Spearman correlation between GPS-surveyed and LLM-predicted preferences across individuals. Countries are shaded from blue (negative correlations) to red (positive correlations). Positive correlations indicate that higher human scores tend to be associated with higher LLM scores, whereas negative correlations indicate the opposite pattern. **(B) Country-level consistency across.** For each preference domain, we compute the Spearman correlation between country-level average GPS-surveyed preferences and country-level average LLM-predicted preferences. In the scatterplots, each point is a country, the x-axis shows the ranking based on GPS-surveyed country means, and the y-axis shows the ranking based on LLM-predicted country means. Positive correlations indicate that countries ranked higher in the survey also tend to be ranked higher by the LLM, whereas negative correlations indicate the reverse ordering.

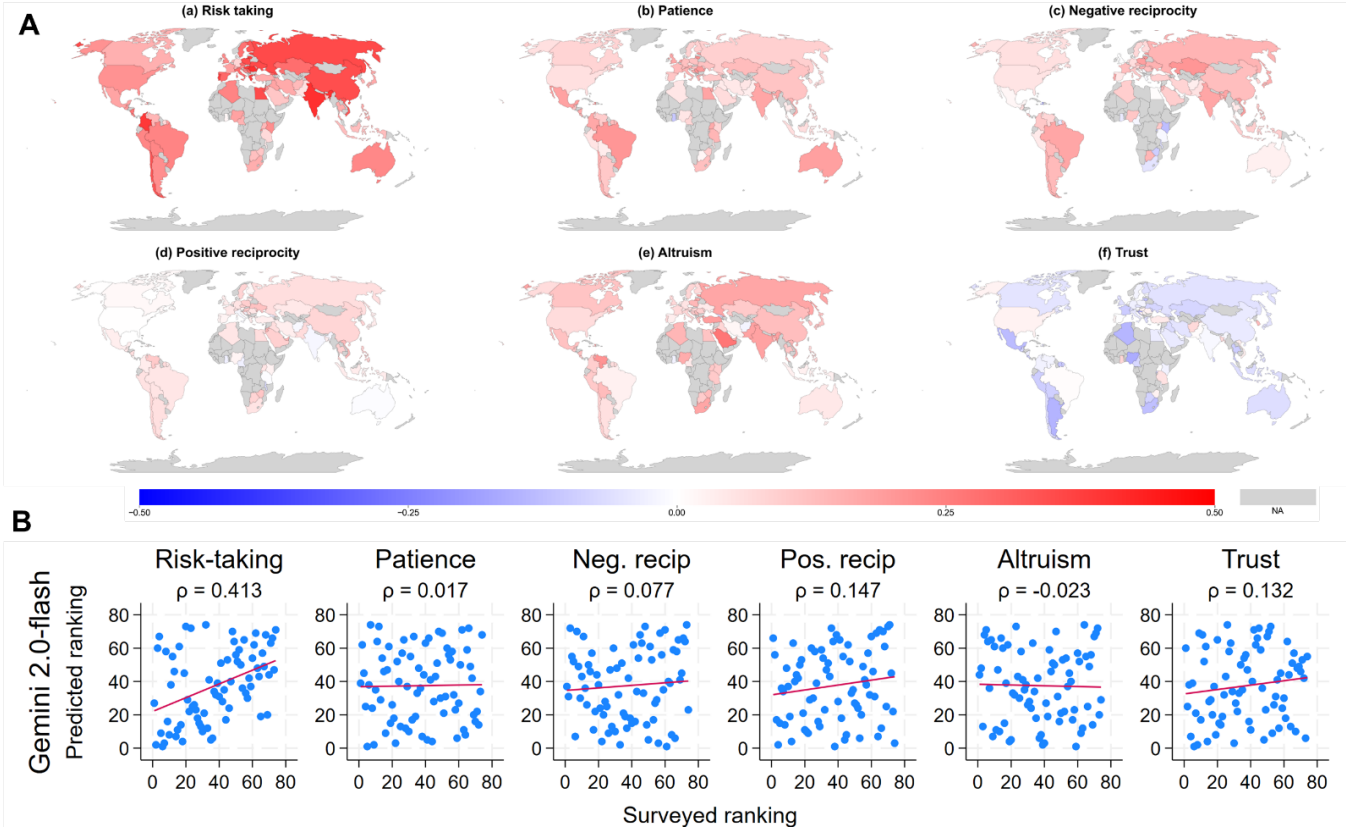

Figure B.6. Consistency across countries (Gemini 2.0-flash). (A) Individual-level consistency. For each country and preference domain, we compute the Spearman correlation between GPS-surveyed and LLM-predicted preferences across individuals. Countries are shaded from blue (negative correlations) to red (positive correlations). Positive correlations indicate that higher human scores tend to be associated with higher LLM scores, whereas negative correlations indicate the opposite pattern. **(B) Country-level consistency.** For each preference domain, we compute the Spearman correlation between country-level average GPS-surveyed preferences and country-level average LLM-predicted preferences. In the scatterplots, each point is a country, the x-axis shows the ranking based on GPS-surveyed country means, and the y-axis shows the ranking based on LLM-predicted country means. Positive correlations indicate that countries ranked higher in the survey also tend to be ranked higher by the LLM, whereas negative correlations indicate the reverse ordering.

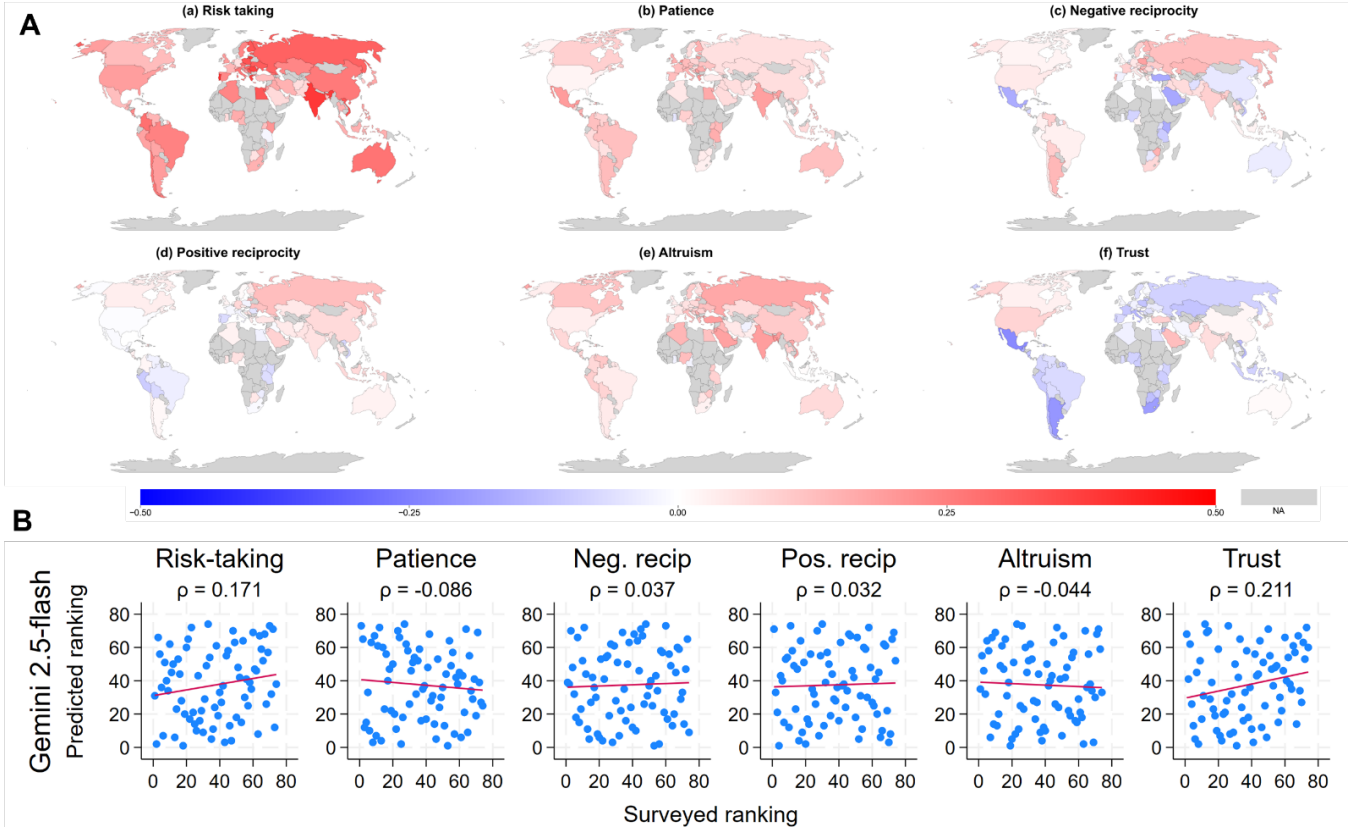

Figure B.7. Consistency across countries (Gemini 2.5-flash). (A) Individual-level consistency. For each country and preference domain, we compute the Spearman correlation between GPS-surveyed and LLM-predicted preferences across individuals. Countries are shaded from blue (negative correlations) to red (positive correlations). Positive correlations indicate that higher human scores tend to be associated with higher LLM scores, whereas negative correlations indicate the opposite pattern. **(B) Country-level consistency.** For each preference domain, we compute the Spearman correlation between country-level average GPS-surveyed preferences and country-level average LLM-predicted preferences. In the scatterplots, each point is a country, the x-axis shows the ranking based on GPS-surveyed country means, and the y-axis shows the ranking based on LLM-predicted country means. Positive correlations indicate that countries ranked higher in the survey also tend to be ranked higher by the LLM, whereas negative correlations indicate the reverse ordering.

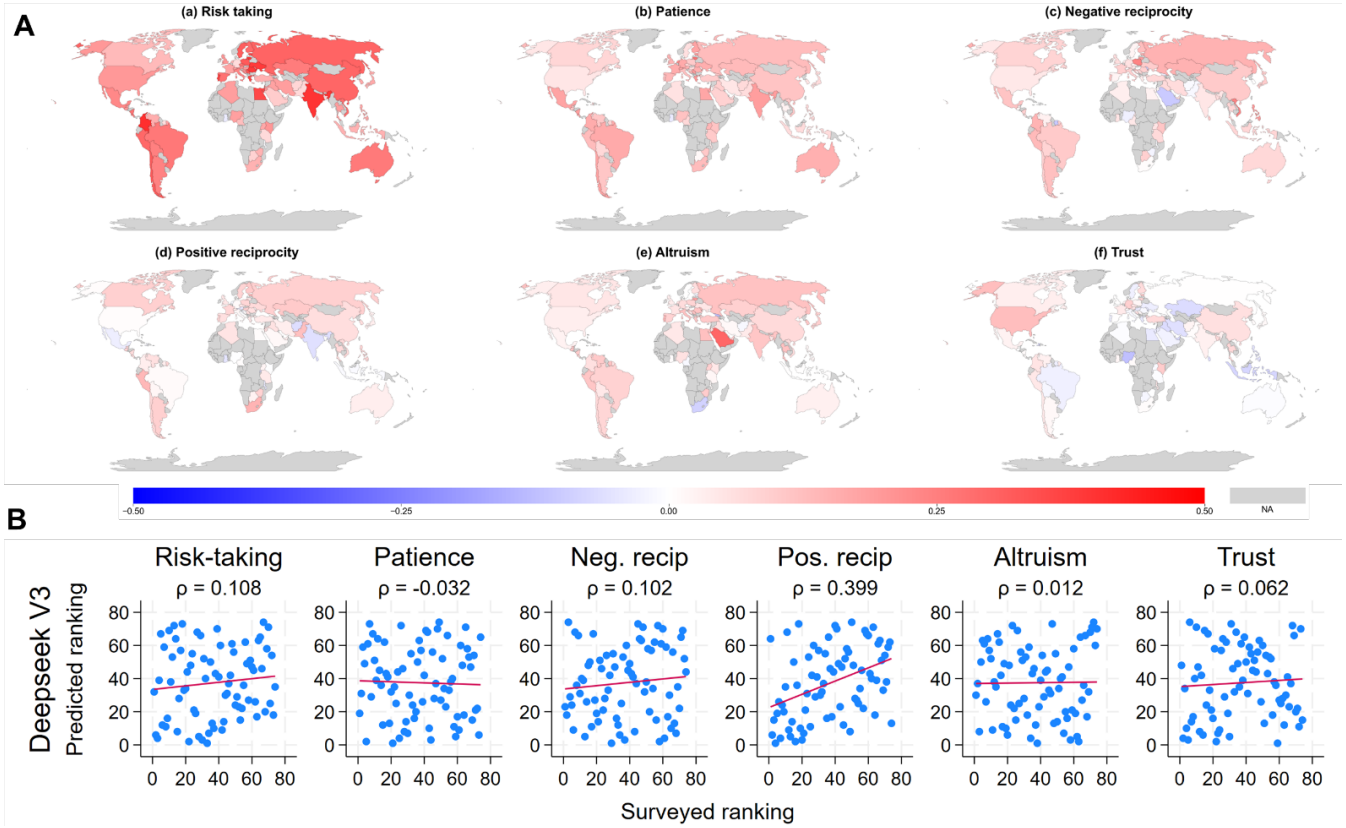

Figure B.8. Consistency across countries (DeepSeek V3). (A) Individual-level consistency. For each country and preference domain, we compute the Spearman correlation between GPS-surveyed and LLM-predicted preferences across individuals. Countries are shaded from blue (negative correlations) to red (positive correlations). Positive correlations indicate that higher human scores tend to be associated with higher LLM scores, whereas negative correlations indicate the opposite pattern. **(B) Country-level consistency.** For each preference domain, we compute the Spearman correlation between country-level average GPS-surveyed preferences and country-level average LLM-predicted preferences. In the scatterplots, each point is a country, the x-axis shows the ranking based on GPS-surveyed country means, and the y-axis shows the ranking based on LLM-predicted country means. Positive correlations indicate that countries ranked higher in the survey also tend to be ranked higher by the LLM, whereas negative correlations indicate the reverse ordering.

B.2 Half-split internal consistency

We estimate the internal consistency of GPS-surveyed and LLM-predicted preferences using a permutation-based half-split procedure. For each country, all individual observations are randomly divided into two halves of equal size. Country-level average preferences are computed for each half, and internal consistency is measured using the Spearman–Brown corrected Pearson's *r* between the two halves. This procedure is repeated 1,000 times. Figure B.9 shows the half-split internal consistency for country-level averages.

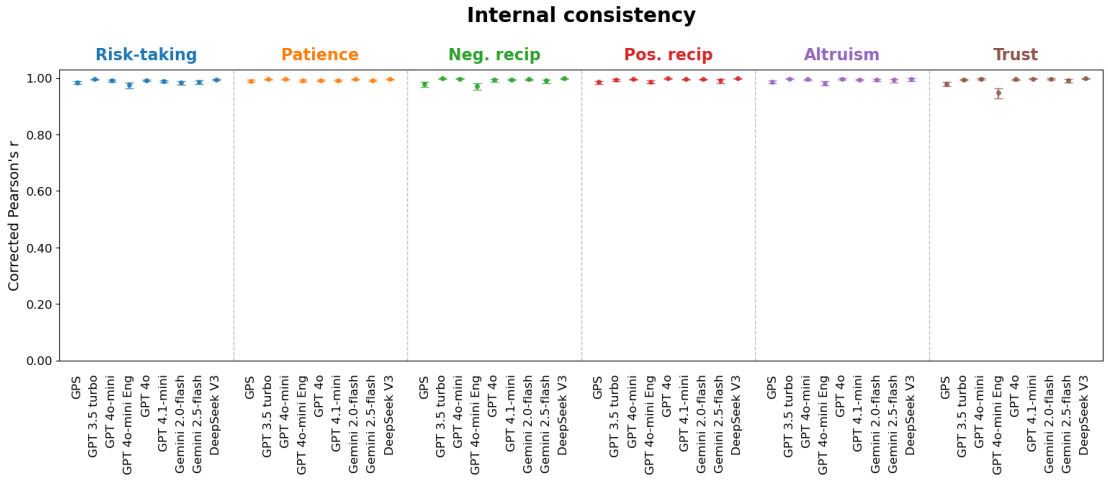

Figure B.9. Within-model internal consistency (split-half reliability). For each model, we randomly assign individuals within each country into two halves (1,000 permutations), compute country-level means for each half, and report the Spearman–Brown corrected Pearson's *r* averaged across permutations. Error bars denote bootstrapped 95% CI.

The half-split internal consistency of LLM-predicted preferences is comparable to, and slightly higher than, that of the GPS-surveyed data (two-sided t-test, $\text{mean}_{\text{GPS}} = 0.984$ vs $\text{mean}_{\text{LLMs}} = 0.992$, $P = 0.033$). This supports the feasibility of aggregating individual-level LLM-predicted preferences into country-level average preference indices.

Appendix C. Alignment at the demographic and cultural levels

Here, we provide additional details on the alignment between LLM-predicted and GPS-surveyed preferences at the demographic and cultural levels, including the definition of the consistency ratio, robustness checks, and between-model comparisons.

C.1 Consistency ratios for demographic and cultural alignment

As described in the Methods, we quantify alignment between LLM-predicted and GPS-surveyed preferences using a consistency ratio. For a given binary grouping variable (gender, high- vs. low-split on age, cognitive ability, or on a cultural dimension) and preference domain, a model-dimension case is counted as "consistent" if either (i) the 95% CIs of the effect sizes (Cohen's d) for GPS and LLM both include zero, or (ii) both 95% CIs exclude zero and the two effect sizes have the same sign. The consistent ratio is the share of such consistent cases, so higher values indicate stronger alignment in the pattern of demographic and cultural differences.

We examine 15 cultural dimensions from the World Values Survey (WVS, i.e. Inglehart–Welzel cultural map), Hofstede's, and Schwartz's cultural theories (Inglehart and Welzel 2005, Schwartz 2006, Hofstede et al. 2010). Table C.1 reports the consistency ratios across demographic and cultural dimensions. The overall demographic-level 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). The overall cultural-level consistency ratio is 49.6% (95% CI = [45.9%, 53.3%]), which is significantly lower than the demographic-level consistency ratio (chi-square test, $\chi^2(1) = 36.55$, $P < 0.001$).

We also examine whether effect sizes estimated from GPS-surveyed data are correlated with those estimated from LLM-predicted data. Table C.2 reports the Spearman rank correlations. At the demographic level, correlations are large and positive for gender ($\rho = 0.878$, $P < 0.001$) and age ($\rho = 0.854$, $P < 0.001$), whereas the correlation for cognitive ability is small and not significantly different from zero ($\rho = -0.134$, $P = 0.364$). For the cultural dimensions, correlations range from slightly negative to moderate positive. Several dimensions show statistically significant positive associations, including WVS traditional

values, WVS survival values and etc. These results indicate that survey-based effect sizes are echoed to some extent in the LLM-predicted data, particularly for gender and age, and for a subset of cultural dimensions, while the correspondence is much weaker for cognitive ability and many other cultural dimensions.

Table C.1. The number of consistent cases and consistency ratios for demographic and cultural gradients in LLM-predicted and GPS-surveyed preferences

	Risk-taking	Patience	Negative reciprocity	Positive reciprocity	Altruism	Trust	Total consistent cases	Consistency ratio/ %
Demographic level								
Gender	8	8	7	1	7	7	38	79.2
Cognitive ability	8	8	7	7	8	3	41	85.4
Age	7	8	8	1	2	6	32	66.7
Cultural level								
WVS Tradition values	8	4	2	4	8	4	30	62.5
WVS Survival values	7	6	0	4	7	3	27	56.3
Schwartz Harmony	5	3	0	3	6	4	21	43.8
Schwartz Embeddedness	8	6	1	6	0	6	27	56.3
Schwartz Hierarchy	6	7	6	2	4	5	30	62.5
Schwartz Mastery	3	1	6	1	2	2	15	31.3
Schwartz Affective Autonomy	7	6	0	4	7	5	29	60.4
Schwartz Intellectual Autonomy	6	7	1	0	6	2	22	45.8
Schwartz Egalitarianism	3	4	5	0	4	0	16	33.3
Hofstede PDI	7	7	3	2	5	5	29	60.4
Hofstede IDV	0	4	2	5	7	0	18	37.5
Hofstede MAS	7	3	7	0	2	6	25	52.1
Hofstede UAI	1	5	0	5	4	6	21	43.8
Hofstede LTO	8	1	2	3	8	4	26	54.2
Hofstede IVR	5	5	1	1	3	6	21	43.8

Notes: For each demographic or cultural dimension, we construct a binary grouping variable and compute effect sizes (Cohen's *d*) on each preference across each LLM with GPS-surveyed effect sizes as benchmark. A case is classified as "consistent" if either (a) the 95% CIs of the two effect sizes both include zero, or (b) both CIs exclude zero and the effect sizes have the same sign. Each row is based on 48 model–preference combinations (6 preferences × 8 models). The last two columns report, for each dimension, the total number of consistent cases and the corresponding consistency ratio (in %).

228

Table C.2. Spearman correlations between GPS- and LLM-based effect sizes across demographic and cultural dimensions

Dimension	Spearman's ρ	P-value
Demographic level		
Gender	0.878	<0.001
Cognitive ability	-0.134	0.364
Age	0.854	<0.001
Cultural level		
WVS Tradition values	0.526	<0.001
WVS Survival values	0.248	0.089
Schwartz Harmony	0.027	0.854
Schwartz Embeddedness	0.306	0.035
Schwartz Hierarchy	0.322	0.025
Schwartz Mastery	-0.456	0.001
Schwartz Affective Autonomy	0.496	<0.001
Schwartz Intellectual Autonomy	0.167	0.256
Schwartz Egalitarianism	-0.153	0.298
Hofstede PDI	0.496	<0.001
Hofstede IDV	0.234	0.109
Hofstede MAS	0.209	0.155
Hofstede UAI	0.336	0.019
Hofstede LTO	0.390	<0.001
Hofstede IVR	-0.175	0.234

Notes: For each demographic or cultural dimension, we construct a binary grouping variable and compute effect sizes (Cohen's *d*) on each preference across each LLM with GPS-surveyed effect sizes as benchmark. We then calculate the Spearman correlation between the vector of GPS-based effect sizes and the corresponding vector of LLM-based effect sizes. For each dimension, all effect sizes across the six preference domains and eight LLMs (48 combinations) are pooled.

229

C.2 Robustness check based on sign consistency

As a robustness check, we construct a sign-consistency measure based on a quadrant classification of effect sizes, irrespective of statistical significance. For each effect-size pair (GPS vs. LLM), we record whether it falls in Quadrant I ($x > 0$ and $y > 0$), Quadrant II ($x < 0$ and $y > 0$), Quadrant III ($x < 0$ and $y < 0$), or Quadrant IV ($x > 0$ and $y < 0$), where x is the GPS effect size and y is the LLM effect size. We then compute, for each dimension, the percentage of cases in Quadrants I and III—that is, the proportion of cases in which the sign of the effect size based on GPS data matches the sign of the effect size based on LLM data. Table C.3 reports the distribution of effect sizes across the four quadrants.

The overall demographic-level match ratio (Quadrants I and III combined) is 78.5% (95% CI = [70.9%, 84.9%]), whereas the cultural-level match ratio is 57.9% (95% CI = [54.2%, 61.6%]). The cultural-level match ratio is significantly lower than the demographic-level ratio (chi-square test, $\chi^2(1) = 21.38$, $P < 0.001$). Although this sign-based match ratio differs slightly from the CI-based consistency ratio used in the main text, both methods lead to the same conclusion, i.e., alignment at the cultural-level is markedly weaker than at the demographic-level.

Table C.3. Quadrant distribution and sign consistency of GPS- and LLM-based effect sizes across demographic and cultural dimensions

	Quadrant I	Quadrant II	Quadrant III	Quadrant IV	# in quadrant I and III	% in quadrant I and III
			Demographic level			
Gender	23	8	16	1	39	81.2
Cognitive ability	41	0	0	7	41	85.4
Age	27	2	6	13	33	68.8
			Cultural level			
WVS Tradition values	16	0	16	16	32	66.7
WVS Survival values	13	6	18	11	31	64.6
Schwartz Harmony	3	12	28	5	31	64.6
Schwartz Embeddedness	8	21	19	0	27	56.2
Schwartz Hierarchy	19	3	13	13	32	66.7
Schwartz Mastery	18	7	1	22	19	39.6
Schwartz Affective Autonomy	15	3	21	9	36	75
Schwartz Intellectual Autonomy	8	15	17	8	25	52.1
Schwartz Egalitarianism	8	22	10	8	18	37.5
Hofstede PDI	5	16	24	3	29	60.4
Hofstede IDV	6	7	25	10	31	64.6
Hofstede MAS	5	9	23	11	28	58.3
Hofstede UAI	6	17	23	2	29	60.4
Hofstede LTO	3	9	23	13	26	54.2
Hofstede IVR	19	12	4	13	23	47.9

Notes: For each demographic or cultural dimension, we construct a binary grouping variable (e.g., male vs. female; above- vs. below-median on a cultural index) and compute effect sizes (Cohen's d) for each LLM–preference combination in both the GPS and LLM data. Each GPS-based effect size (x) and the corresponding LLM-based effect size (y) form a pair that is classified into Quadrant I ($x > 0, y > 0$), Quadrant II ($x < 0, y > 0$), Quadrant III ($x < 0, y < 0$), or Quadrant IV ($x > 0, y < 0$). For each dimension, all effect sizes across the six preference domains and eight LLMs are pooled. The last two columns report, for each dimension, the number and percentage of cases in Quadrants I and III combined (i.e., with matching signs).

C.3 Between-model consistency in demographic and cultural effects

Here we assess between-model consistency in demographic- and cultural-level effects. For each pair of LLMs, we compute Spearman's ρ between their effect sizes (Cohen's d), pooling across all demographic (or cultural) dimensions and all preference domains. Figure C.1 summarizes the resulting correlation matrices.

All 28 demographic-level correlations are positive, with an average correlation of 0.815. At the cultural level, 26 of 28 correlations are positive, and the average correlation is 0.300. Regardless of statistical significance, the proportion of between-model correlations that are positive is far above what would be expected by chance (50%; two-sided binomial test, $P < 0.001$ at both the demographic and cultural levels). These results suggest that LLMs make broadly similar predictions to one another at both levels, even when these shared patterns deviate from the survey-based estimates.

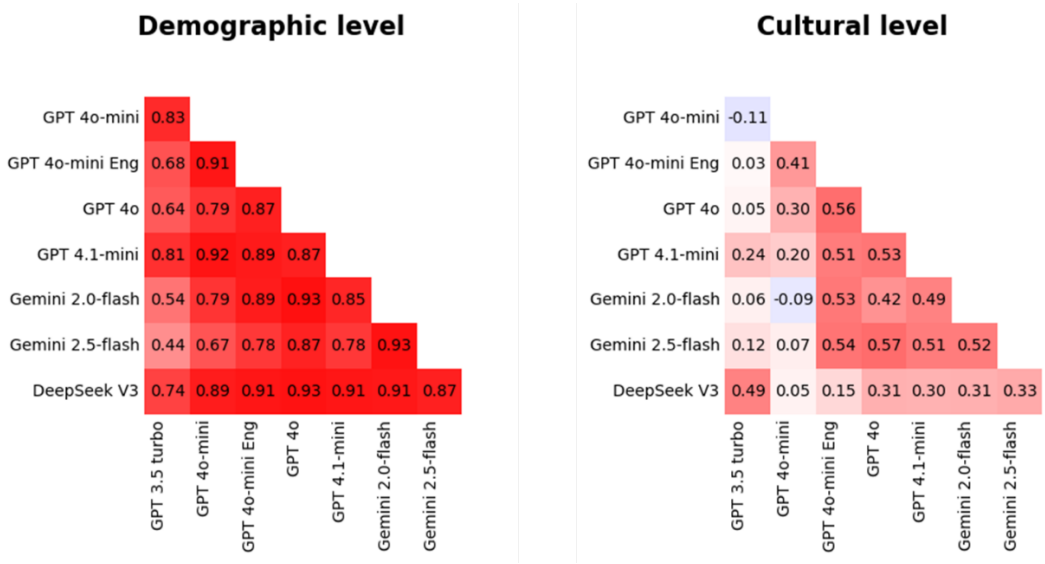

Figure C.1. Between-model consistency in demographic- and cultural-level effects. (A) Demographic-level between-model consistency. Each cell reports Spearman's ρ between effect sizes (Cohen's d) estimated from a pair of LLMs, pooling all demographic dimensions and all preference domains. **(B) Cultural-level between-model consistency.** Each cell reports Spearman's ρ between effect sizes estimated from a pair of LLMs, pooling across all cultural dimensions and all preference domains.

Appendix D. Country-level sources of consistency and inconsistency

We first examine whether demographic information or country information plays a more important role in shaping alignment between LLM-predicted and GPS-surveyed preferences. We then examine the extent of cross-country heterogeneity in alignment and investigate which country-level characteristics explain this heterogeneity. We construct two country-specific indices to measure how closely LLM-predicted preferences align with GPS-surveyed preferences. The consistency index is defined as the Spearman correlation between LLM-predicted and GPS-surveyed preferences, computed by pooling all individual observations within a given country. The inconsistency index is defined as the sampling-weighted average of individual-level absolute differences between LLM-predicted and GPS-surveyed preferences within that country. Higher consistency and lower inconsistency indicate better country-level alignment between model-generated and human preferences.

D.1 Decomposing individual-level alignment with demographics and country factors

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. The mean OLS coefficients across eight LLMs for different preference domains under each specification are summarized in Table D.1. Specifically, the average coefficient changes little from 0.0901 in the baseline specification to 0.0915 with country controls only, then drops to 0.0205 when only demographic controls are included, and further to 0.0144 when both demographics and country fixed effects are added.

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 (see Table D.2).

Relative to the baseline specification, adding country controls has no significant effect (b = 0.001, P = 0.864), whereas including demographic controls leads to a significant reduction in the OLS coefficient (b = -0.070, P < 0.001). A post hoc comparison shows no statistically significant difference between the demographic-only, and demographic and country specifications (P = 0.454).

On average, the coefficients drop from 0.090 in the baseline specification to 0.020 when demographic controls are included. This suggests that demographic variables absorb most of the baseline association between LLM-predicted and GPS-surveyed preferences (approximately 78%, i.e., 0.070/0.090), whereas adding country fixed effects contributes little additional explanatory power.

D.2 Consistency index

Motivated by scaling-law results for LLMs (Kaplan et al. 2020), we estimate a series of OLS regressions to examine whether measures of economic development (log GDP per capita and the Information and Communications Technology Development Index, IDI, used as proxies for compute intensity) or measures of population (log population and log number of internet users, used as proxies for corpus size, Villalobos et al. 2022) are more strongly associated with alignment between LLM-predicted and GPS-surveyed preferences.

Columns (1)–(4) of Table D.3 report regressions of the country-level consistency index on these economic-development and population measures, including model fixed effects, preference fixed effects, and their interaction terms. Across specifications, the coefficients on log GDP per capita range from 0.008 to 0.009 (P<0.001), and those on IDI from 0.080 to 0.087 (P<0.001), indicating that higher levels of economic development are systematically associated with stronger LLM–human alignment. By contrast, the coefficients on log population and log number of internet users are small in magnitude (at most 0.002 in absolute value) and not robustly different from zero across specifications.

Because these four country-level variables are highly collinear, we do not include them

jointly in a single specification. Instead, we conduct Principal Component Analysis (PCA) separately for the two sets of variables and extract the first principal component for economic development and for population. Column (5) of Table D.3 reports regressions including both principal components. The first principal component of economic development is positively and significantly associated with the consistency index (b=0.010, P<0.001), whereas the first principal component of the population measures is essentially zero and not statistically significant.

Replacing Spearman's ρ with Pearson's r as the dependent variable yields similar patterns (Table D.4). Coefficients on log GDP per capita (0.009–0.009) and IDI (0.085–0.090) remain positive and highly significant (P<0.001), while the population measures again show no robust relationship with the consistency index. These regressions indicate that LLMs generate better-aligned predictions for countries with higher levels of economic development, rather than for countries with larger populations or more internet users.

D.3 Inconsistency index

We next repeat the analysis in Section D.2 using the inconsistency index as the dependent variable (Table D.5). In all specifications, higher economic development is associated with lower inconsistency. The coefficients on log GDP per capita range from −0.050 to −0.055 (P<0.001), and those on IDI from −0.343 to −0.389 (P<0.001). Population-related measures are also negatively associated with the inconsistency index, but their coefficients are smaller in magnitude: between −0.010 and −0.007 for log population and between −0.008 and −0.007 for log number of internet users. In the PCA specification in column (5), the economic-development score has a coefficient of −0.053 (P<0.001), compared with −0.007 (P = 0.015) for the population score.

As a robustness check, we redefine the inconsistency index as the absolute difference between sampling-weighted country-level average preferences in the LLM and GPS data. Table D.6. reports regression results using this alternative definition. Under this specification, economic-development variables remain strongly and negatively associated with the inconsistency index. Coefficients on log GDP per capita range from −0.044 to −0.051 (P<0.001),

and those on IDI from −0.294 to −0.368 (P<0.001). In contrast, the coefficients on log population and log number of internet users are small and statistically insignificant in columns (3) and (4). In the PCA specification, the economic-development score remains significantly negative (b=−0.049, P<0.001), whereas the population score is close to zero and not significant.

Taken together, these results show that LLMs produce higher level alignment (i.e., smaller absolute prediction errors) for countries with higher levels of economic development and larger deviations for countries with lower levels of development, while population size and the number of internet users play at most a secondary role.

Table D.1 OLS coefficients on LLM-predicted preferences across four conditions

Preference	Baseline	Country	Demo	Demo + Country
Risk-taking	0.2132	0.2082	0.0279	0.0144
Patience	0.0705	0.0910	-0.0208	0.0188
Neg. recip	0.0822	0.1044	0.0129	0.0123
Pos. recip	0.0696	0.0606	0.0353	0.0145
Altruism	0.0752	0.0817	0.0335	0.0225
Trust	0.0299	0.0030	0.0341	0.0039
Average	0.0901	0.0915	0.0205	0.0144

Notes: We estimate OLS regressions for each model × preference combination, regressing GPS-surveyed preferences on LLM-predicted preferences using (i) no covariates (baseline), (ii) country fixed effects only, (iii) demographic controls only (gender, age, cognitive ability), and (iv) both demographic controls and country fixed effects. This table reports the mean OLS coefficients pooled across all models, grouped by preference domain and regression condition.

Table D.2 Linear mixed-effects decomposition of individual-level alignment coefficients

	DV: OLS coefficient (GPS on LLM)
Condition	
Only country dummies	0.001
	(0.008)
Only demographic controls	-0.070***
	(0.008)
Demographic controls + country dummies	-0.076***
	(0.008)
Preference	
Risk-taking	0.063***
	(0.010)
Patience	-0.013
	(0.010)
Neg. recip	-0.000
	(0.010)
Pos. recip	-0.008
	(0.010)
Trust	-0.036***
	(0.010)
Observations	192
Number of groups	8

Notes: The dependent variable is the OLS coefficient from regressions of GPS-surveyed preferences on LLM-predicted preferences estimated under four specifications (baseline, + country dummies, + demographic controls, and + both). We fit a linear mixed-effects model with indicators for specification (reported under "Condition") and preference-domain dummies (reported under "Preference"), and include model-level random intercepts. The reference categories are the baseline specification and altruism. Standard errors are in parentheses. * $P < 0.1$, ** $P < 0.05$, *** $P < 0.01$.

Table D.3. OLS regressions of the country-level consistency index (Spearman's ρ) on country-level variables

	(1)	(2)	(3)	(4)	(5)
log [GDP p/c]	0.008***	0.009***			
	(0.001)	(0.001)			
IDI			0.080***	0.087***	
			(0.006)	(0.007)	
log population	0.002**		0.000		
	(0.001)		(0.001)		
log internet user		0.000		-0.000	
		(0.001)		(0.001)	
Score for economic development					0.010***
					(0.001)
Score for population					-0.000
					(0.001)
Observations	3503	3263	3455	3263	3263
R^2	0.507	0.513	0.518	0.524	0.519

Notes: The dependent variable is the country-level consistency index (Spearman's ρ between LLM-predicted and GPS-surveyed preferences within each country). Reported coefficients are OLS estimates, with robust standard errors in parentheses. All regressions include preference fixed effects, model fixed effects, and their interaction terms. * $P < 0.1$, ** $P < 0.05$, *** $P < 0.01$.

Table D.4. OLS regressions of the country-level consistency index (Pearson's r) on country-level variables

	(1)	(2)	(3)	(4)	(5)
log [GDP p/c]	0.009***	0.009***			
	(0.001)	(0.001)			
ICT Development Index			0.085***	0.090***	
			(0.006)	(0.007)	
log population	0.002**		-0.000		
	(0.001)		(0.001)		
log internet user		0.000		-0.000	
		(0.001)		(0.001)	
Score for economic development					0.011***
					(0.001)
Score for population					-0.000
					(0.001)
Observations	3503	3263	3455	3263	3263
R^2	0.506	0.514	0.521	0.525	0.520

Notes: The dependent variable is the country-level consistency index (Pearson's r between LLM-predicted and GPS-surveyed preferences within each country). Reported coefficients are OLS estimates, with robust standard errors in parentheses. All regressions include preference fixed effects, model fixed effects, and their interaction terms. * $P < 0.1$, ** $P < 0.05$, *** $P < 0.01$.

374

Table D.5. OLS regressions of the country-level inconsistency index (individual-level absolute errors, then averaged) on country-level variables

	(1)	(2)	(3)	(4)	(5)
log [GDP p/c]	-0.050***	-0.055***			
	(0.004)	(0.005)			
ICT Development Index			-0.343***	-0.389***	
			(0.034)	(0.040)	
log population	-0.010***		-0.007**		
	(0.003)		(0.003)		
log internet user		-0.008***		-0.007**	
		(0.003)		(0.003)	
Score for economic development					-0.053***
					(0.005)
Score for population					-0.007**
					(0.003)
Observations	3504	3264	3456	3264	3264
R^2	0.082	0.085	0.075	0.081	0.086

Notes: The dependent variable is the country-level inconsistency index, defined as the sampling-weighted average of individual-level absolute differences between GPS-surveyed and LLM-predicted preferences. Reported coefficients are OLS estimates, with robust standard errors in parentheses. All regressions include preference fixed effects, model fixed effects, and their interaction terms. *$P < 0.10$, **$P < 0.05$, ***$P < 0.01$.

375

376

377

Table D.6. OLS regressions of the country-level inconsistency index (absolute differences between country-level averages) on country-level variables

	(1)	(2)	(3)	(4)	(5)
log [GDP p/c]	-0.044***	-0.051***			
	(0.006)	(0.007)			
ICT Development Index			-0.294***	-0.368***	
			(0.045)	(0.051)	
log population	-0.004		0.001		
	(0.004)		(0.004)		
log internet user		-0.001		0.000	
		(0.005)		(0.004)	
Score for economic development					-0.049***
					(0.007)
Score for population					-0.000
					(0.005)
Observations	3504	3264	3456	3264	3264
R^2	0.075	0.078	0.070	0.077	0.079

Notes: The dependent variable is the country-level inconsistency index, defined as the absolute difference between GPS-surveyed and LLM-predicted sampling-weighted country-level average preferences. Reported coefficients are OLS estimates, with robust standard errors in parentheses. All regressions include preference fixed effects, model fixed effects, and their interaction terms. *$P < 0.10$, **$P < 0.05$, ***$P < 0.01$.

378

379

Appendix E. Replicating GPS-based findings using LLM-predicted preferences

In this section, we assess the extent to which LLM-predicted preferences can reproduce previously published GPS-based findings. These findings come from articles in leading journals such as *Science* (Falk and Hermle 2018), *Quarterly Journal of Economics* (Falk et al. 2018) and *Review of Economic Studies* (Sunde et al. 2022). For exposition, we distinguish between individual-level findings, which rely on microdata (e.g., gender, age), and country-level findings, which rely on aggregated measures (e.g., country-level average patience, GDP).

We focus on three core studies using the GPS (Falk et al. 2018; Falk and Hermle 2018; Sunde et al. 2022). In designing the replication exercise, we follow three principles. First, in terms of scope, we aim to reproduce as many of the key results in Falk et al. (2018) as possible, since this first GPS article reports both individual- and country-level analyses. In addition, we target the main headline results in Falk and Hermle (2018), and Sunde et al. (2022). Second, for specification choice, when replicating published regression results, we use the baseline specifications reported in the original articles, rather than extended models with additional controls. Third, for the definition of successful replication, we focus primarily on positive (statistically significant) findings in the original articles and assess whether the replication matches the sign of the reported effect at the 10% significance level.

Our estimation strategy mirrors the original GPS analyses. For country-level findings, we compute Pearson correlation coefficients between the two variables of interest, matching the way these results are typically reported. For individual-level findings, all target results come from Table V in Falk et al. (2018), so we apply exactly the same regression specification as in the table.

In total, we include country-level results reported in Tables II, V, VI, IX, and X of Falk et al. (2018), four key country-level results from Sunde et al. (2022), and two key country-level findings from Falk and Hermle (2018). Individual-level replication is based on Table V in Falk et al. (2018). Since we do not have access to the full 2012 Gallup World Poll dataset, we cannot replicate results that require merging the GPS with additional Gallup World Poll variables (e.g.

Table VII in Falk et al. 2018). Overall, the replication set comprises 35 country-level findings and 18 individual-level findings.

For the country-level findings, Pearson correlation coefficients are reported in Table E.1. The first two columns list the two variables being correlated. The third column reports the correlations based on the original GPS-surveyed preferences, and the remaining columns report correlations based on LLM-predicted preferences. Among the LLMs, GPT 4o performs best, successfully reproducing 13 of 35 (37.1%) country-level findings in the expected direction. The weakest performance is obtained for GPT-4o-mini, which reproduces only a single finding.

One subtle issue arises for the correlation between positive reciprocity and the share of Protestants. In the full original GPS sample, this correlation is statistically significant ($P < 0.1$). In our censored GPS sample, where we exclude minor-language respondents and omit Cambodia and Morocco, the corresponding correlation is no longer significant ($P > 0.10$). When judging replication success, we nonetheless treat this case according to the significance pattern reported in the full original GPS dataset.

Individual-level replication results are reported in Table E.2. Following the GPS program, we regress each preference index on gender, subjective math skills, age, and age squared, including country fixed effects, thereby reproducing the specification in Table V of Falk et al. (2018). For age, we jointly evaluate the coefficients on age and age squared. The age effect is labeled as successfully replicated only if both coefficients satisfy our sign-agreement and significance rule. Under this criterion, the best-performing models at the individual level are GPT 4.1-mini and GPT 4o-mini, each reproducing 14 of 18 (77.8%) findings, whereas GPT 3.5-turbo performs worst, reproducing 9 of 18 (50.0%) findings.

Table E.1. Replication of GPS-based country-level results using LLM-predicted preferences

Variable 1	Variable 2	GPS	GPT 3.5-turbo	GPT 4o-mini	GPT 4o-mini Eng	GPT 4o	GPT 4.1-mini	Gemini 2.0-flash	Gemini 2.5-flash	DeepSeek V3	***Successful replications***
Findings from *Patience and comparative development*											
Patience	Economic development: log GDP per capita	0.63***	-0.29**	-0.03	**0.25****	**0.29****	-0.07	0.06	-0.17	0.04	**2**
	Physical capital: Log [capital stock p/c]	0.56***	-0.16	-0.04	0.19	**0.29****	-0.03	0.07	-0.11	0.17	**1**
	Human capital: Skilled fraction over 25-year-old	0.56***	-0.28**	0.06	**0.29****	**0.27****	-0.07	0.14	-0.14	0.02	**2**
	Productivity: TFP	0.23*	-0.34***	-0.10	-0.15	0.17	0.08	-0.15	-0.15	-0.20	**0**
Findings from *Relationship of gender differences in preferences to economic development and gender equality*											
Gender difference in preference	Gender equality	0.55***	**0.22***	-0.23*	**0.62*****	0.06	-0.02	-0.19	-0.02	**0.20***	**3**
	log GDP per capita	0.64***	-0.00	0.01	**0.54*****	0.15	**0.22***	0.08	0.06	**0.23****	**3**
Findings from *Global evidence on economic preferences*											
Patience	Hofstede long-term orientation	0.33**	-0.21	-0.19	0.07	-0.07	-0.24*	0.06	-0.27**	0.21	**0**
	Geographic conditions	0.43***	-0.16	-0.28**	-0.05	**0.31****	-0.22	0.09	-0.36**	**0.42*****	**2**
	Absolute latitude	0.47***	-0.08	-0.06	0.08	**0.34*****	-0.07	0.17	-0.16	**0.38*****	**2**
	Biology conditions	0.38***	-0.19	-0.30**	-0.01	0.22	-0.26*	0.07	-0.36***	**0.42*****	**1**
	Weak future time reference	-0.33***	0.22*	0.23*	-0.01	0.11	0.10	-0.06	0.22*	-0.07	**0**
	Pronoun drop	0.57***	-0.09	0.13	0.07	**0.28****	0.05	-0.02	0.05	-0.02	**1**
	Share of protestants	0.45***	-0.12	0.14	**0.22***	0.15	0.18	**0.25****	**0.20***	-0.09	**3**
	Individualism	0.67***	-0.28**	0.01	0.13	**0.27****	0.05	0.10	0.03	0.01	**1**
	Family ties	-0.54***	0.17	-0.01	-0.10	-0.19	-0.09	**-0.42*****	0.09	**-0.26***	**2**
Risk-taking	Hofstede uncertainty avoidance	-0.32**	-0.20	-0.08	-0.08	-0.01	**-0.23***	0.03	0.22*	-0.03	**1**
	Biology conditions	-0.36**	-0.03	-0.09	-0.22	**-0.35****	-0.13	**-0.48*****	-0.20	-0.08	**2**
	Geographic conditions	-0.30**	-0.05	-0.15	-0.23	**-0.35****	-0.16	**-0.50*****	-0.10	-0.15	**2**
	Family ties	0.34**	0.07	0.07	**0.30****	**0.45*****	**0.27***	**0.56*****	0.16	0.14	**4**
Pos. reciprocity	Share of protestants	-0.15	-0.01	0.16	0.13	-0.00	-0.01	**-0.22***	-0.18	**-0.26****	**2**
	Biology conditions	0.28**	-0.12	0.08	0.23	-0.03	-0.09	-0.11	-0.02	0.03	**0**

Neg reciprocity	Geographic conditions	0.34**	0.05	-0.15	-0.34**	-0.15	-0.09	-0.04	-0.15	0.12	**0**
	Absolut latitude	0.23**	-0.07	-0.28**	-0.36***	-0.12	0.01	-0.12	-0.32***	0.05	**0**
	Biology conditions	0.31**	0.12	-0.22	-0.33**	-0.23	-0.07	-0.06	-0.21	0.00	**0**
	Armed conflicts	0.33***	**0.20***	-0.15	-0.21*	-0.06	-0.16	-0.04	-0.30***	-0.01	**1**
Altruism	Agricultural suitability	-0.21*	0.01	0.16	-0.16	-0.20	-0.17	-0.00	**-0.25****	-0.03	**1**
	Crop suitability	-0.21*	-0.09	0.11	-0.05	**-0.32*****	**-0.23***	-0.07	**-0.37*****	-0.13	**3**
	Family ties	0.27*	**0.42*****	0.07	0.17	**0.57*****	**0.38*****	**0.37****	**0.49*****	**0.39*****	**6**
Trust	WVS Trust	0.51***	**0.29****	0.13	0.16	-0.00	0.01	0.04	0.21	0.12	**1**
	Geographic conditions	0.41***	0.14	0.18	0.22	-0.42***	-0.01	0.15	0.04	-0.08	**0**
	Absolut latitude	0.28**	-0.00	0.12	-0.02	-0.52***	-0.24**	0.10	-0.09	-0.03	**0**
	Agricultural suitability	-0.47***	0.04	0.01	0.21*	-0.10	-0.16	**-0.26****	-0.14	0.13	**1**
	Crop suitability	-0.36***	-0.00	0.01	0.17	**-0.26****	**-0.26****	**-0.25****	-0.16	0.14	**3**
	Biology conditions	0.46***	0.02	0.15	0.23	-0.50***	-0.09	0.16	0.06	-0.09	**0**
	Weak future time reference	-0.23*	-0.10	**-0.20***	**-0.27****	0.01	0.17	-0.04	**-0.23****	-0.09	**3**
Successful replications			**4**	**1**	**7**	**13**	**6**	**9**	**5**	**8**	**58**

Notes: Entries are two-tailed Pearson correlation coefficients. * $P < 0.10$, ** $P < 0.05$, *** $P < 0.01$; Bold coefficients indicate successful replications under our sign- and significance-based criterion. A replication is classified as successful if the LLM-based correlation has the same sign as the GPS-based correlation and both are significant at the 10% level. When replicating the results in Falk and Hermle (2018), we conduct principal component analysis separately for each model's data.

Table E.2. Replication of GPS-based individual-level results using LLM-predicted preferences

Dependent variables	Independent variables	GPS	GPT 3.5-turbo	GPT 4o-mini	GPT 4o-mini Eng	GPT 4o	GPT 4.1-mini	Gemini 2.0-flash	Gemini 2.5-flash	DeepSeek V3	**Successful replications**
Risk-taking	Female	-0.18***	-0.00	**-0.16*** **	**-0.10*** **	**-0.35*** **	**-0.30*** **	**-0.34*** **	**-0.23*** **	**-0.31*** **	**7**
	Sub. Math skill	0.05***	**0.22*** **	**0.22*** **	**0.20*** **	**0.20*** **	**0.24*** **	**0.19*** **	**0.18*** **	**0.20*** **	**8**
	Age	-0.19	**-0.19**	0.79***	**0.06**	-2.82***	**-0.19**	-4.88***	-2.93***	0.36**	
	Age square	-1.11***	**-0.70*** **	-2.83***	**-2.32*** **	0.30	**-1.90*** **	1.80***	0.38*	-2.90***	**3**
Patience	Female	-0.07***	0.04**	**-0.04*** **	0.05***	0.02*	**-0.11*** **	0.07***	0.06***	**-0.14*** **	**3**
	Sub. Math skill	0.03***	**0.20*** **	**0.20*** **	**0.29*** **	**0.27*** **	**0.25*** **	**0.17*** **	**0.23*** **	**0.23*** **	**8**
	Age	0.77***	-0.04	**1.94*** **	**2.65*** **	**1.73*** **	**1.67*** **	**1.77*** **	**1.05*** **	**1.09*** **	
	Age square	-1.51***	-0.47**	**-2.22*** **	**-2.58*** **	**-2.59*** **	**-2.45*** **	**-2.31*** **	**-1.29*** **	**-1.73*** **	**7**
Neg. reciprocity	Female	-0.14***	0.13***	**-0.10*** **	**-0.26*** **	**-0.13*** **	**-0.31*** **	**-0.30*** **	**-0.17*** **	**-0.08*** **	**7**
	Sub. Math skill	0.04***	**0.12*** **	**0.09*** **	**0.15*** **	**0.05*** **	**0.16*** **	**0.05*** **	-0.04***	**0.04*** **	**7**
	Age	-0.43**	-0.35	1.71***	2.65***	4.07***	3.69***	0.66***	3.31***	2.66***	
	Age square	-0.40**	-0.10	-3.64***	-6.11***	-6.88***	-4.57***	-2.46***	-5.48***	-3.89***	**0**
Pos. reciprocity	Female	0.05***	-0.01	**0.02** **	**0.05*** **	**0.03***	0.02	0.01	**0.10*** **	-0.04**	**4**
	Sub. Math skill	0.04***	**0.15*** **	**0.15*** **	**0.18*** **	**0.10*** **	**0.19*** **	**0.08*** **	0.01	**0.07*** **	**7**
	Age	1.14***	**0.92*** **	**1.77*** **	0.86***	**2.83*** **	**1.75*** **	**1.43*** **	**1.40*** **	**1.01*** **	
	Age square	-1.28***	**-1.01*** **	**-1.45*** **	-0.29	**-2.24*** **	**-1.66*** **	**-1.05*** **	**-0.75*** **	**-0.49** **	**7**
Altruism	Female	0.10***	**0.09*** **	**0.28*** **	**0.36*** **	**0.42*** **	**0.24*** **	**0.45*** **	**0.38*** **	**0.24*** **	**8**
	Sub. Math skill	0.04***	**0.13*** **	**0.17*** **	**0.25*** **	**0.08*** **	**0.23*** **	**0.17*** **	**0.12*** **	**0.09*** **	**8**
	Age	0.03	-0.54**	0.54**	0.67***	1.84***	-0.62***	0.17	0.74***	-1.08***	
	Age square	-0.01	0.75***	0.40*	1.36***	-0.79**	0.72***	0.52**	0.36	1.81***	**0**
Trust	Female	0.07***	0.03	**0.26*** **	**0.44*** **	**0.35*** **	**0.25*** **	**0.19*** **	**0.24*** **	**0.32*** **	**7**
	Sub. Math skill	0.05***	**0.12*** **	**0.08*** **	-0.05***	-0.12***	**0.04*** **	-0.08***	-0.12***	-0.01	**3**
	Age	0.36	0.79***	-0.71***	-4.59***	-1.51***	-1.09***	-1.28***	**-0.07**	-0.86***	
	Age square	0.05	-0.72***	1.37***	5.96***	2.07***	1.26***	1.65***	**0.25**	1.26***	**0(1)**
Successful replications			**9**	**14**	**12**	**12**	**14**	**11**	**10 (11)**	**12**	**94 (95)**

Notes: For each preference domain, we regress the preference index on gender, subjective math skills, age, and age squared, including country fixed effects, following Table V in (Falk et al. 2018). Reported coefficients are OLS estimates, with standard errors clustered at the country level. * $P < 0.10$, ** $P < 0.05$, *** $P < 0.01$; Bold coefficients indicate successful replications under our sign- and significance-based criterion. A replication is classified as successful if the replication estimate and the original estimate have the same sign and both are significant at the 10% level, or if both are statistically insignificant. For the age effect, both the age and age-squared coefficients must satisfy this criterion. Under this definition, the (jointly insignificant) age and age-squared effects on trust for Gemini 2.5-flash are also treated as a successful replication. The numbers in parentheses in the last row include this case.

Appendix F. Gender-related biases in Large Language Models

Given the importance of gender in economic and social life, a growing literature examines gender-related biases in LLMs (Kotek et al. 2023, Zack et al. 2024, Armstrong et al. 2024, Fulgu and Capraro 2024). Building on this work, we investigate whether gender differences in economic preferences are amplified by LLMs. We follow the estimation approach of Falk and Hermle (2018), standardizing the six preference indices within a country and estimate separate OLS regressions for each country and preference domain. In these regressions, gender differences are captured by the coefficients on a female indicator, controlling for age, age squared, and cognitive ability.

Table F.1 reports paired t tests comparing country-specific gender gaps in LLM-predicted and GPS-surveyed preferences. Relative to the GPS-surveyed gender gaps, LLMs tend to amplify gender gaps. On average, they predict that women are more altruistic (8 out of 8 models, $P < 0.05$, two-tailed paired t tests) and display higher trust (7 out of 8 models, $P < 0.05$, two-tailed paired t tests). This pattern is consistent with earlier studies showing that LLMs prompted with demographic identities tend to reflect societal beliefs or stereotypes held by out-group members about that group (Wang et al. 2025), in this case beliefs about women's prosocial orientations (Exley et al. 2025).

To examine how this amplification varies across cultural contexts, we relate gender gaps in preferences to a country-level index of gender stereotypes. We first draw on gender-stereotype items from the World Values Survey (WVS) and use principal component analysis (PCA) to extract the first principal component as a WVS-based gender-stereotype index (see Appendix H for variable definitions). We then combine this WVS index with three gender-equality indicators, the Global Gender Gap Index, the UNDP Gender Inequality Index, and the UNDP Gender Development Index, by extracting the first principal component across these four measures to obtain a composite gender-stereotype index.

We first investigate whether LLMs remain globally neutral with respect to women. That is, whether gender gaps in LLM-predicted preferences do not systematically covary with the intensity of gender stereotypes across countries. Panel A of Figure F.1 plots, for each model

and preference domain, the Spearman correlations between country-specific gender gaps in LLM-predicted preferences and the composite gender-stereotype index. GPT 4o-mini (English) exhibits negative correlations for risk-taking, patience, and negative reciprocity, and positive correlations for positive reciprocity and altruism. For the other models, no clear and consistent pattern emerges between LLM-based gender gaps and the stereotype index.

We then benchmark LLM-based gender gaps against those observed in the GPS data. For each country, preference, and model, we define gender bias as the difference between the LLM-based and GPS-based gender gaps in that preference. Panel B of Figure F.1 shows the Spearman correlations between these gender-bias measures and the composite gender-stereotype index. All correlations for risk-taking (8/8), patience (8/8), and negative reciprocity (8/8) are negative, while most correlations for positive reciprocity (7/8), altruism (8/8), and trust (7/8) are positive. For GPT-4o-mini (English), the 95% CIs for these correlations do not include zero in any preference domain. This indicates that, in countries with stronger gender stereotypes, LLMs predict that women are less willing to take risks, less patient, and less negatively reciprocal, but more positively reciprocal, more altruistic, and more trusting relative to the survey benchmark.

Taken together, most LLMs tend to amplify gender differences in altruism and trust, systematically portraying women as more prosocial than suggested by the survey benchmark. Moreover, LLMs, and GPT 4o-mini (English) in particular, appear to encode country-specific gender norms that covary with the intensity of local gender stereotypes, rather than maintaining a neutral treatment of gender across countries.

Table F.1. Differences between LLM-predicted and GPS-surveyed gender gaps in preferences (paired t tests across countries)

	Difference in gender gap (LLM − GPS)					
	Risk-taking	Patience	Neg. recip	Pos. recip	Altruism	Trust
GPT 3.5-turbo	0.19***	0.12***	0.41***	-0.03	0.06**	-0.04
GPT 4o-mini	0.01	0.03	0.03	-0.00	0.28***	0.29***
GPT 4o-mini English	0.08***	0.12***	-0.13***	0.02	0.28***	0.37***
GPT 4o	-0.23***	0.09***	-0.01	0.01	0.43***	0.35***
GPT 4.1-mini	-0.14***	-0.05*	-0.27***	-0.04	0.18***	0.29***
Gemini 2.0-flash	-0.19***	0.15***	-0.22***	-0.03	0.45***	0.18***
Gemini 2.5-flash	-0.08***	0.13***	-0.05	0.13***	0.39***	0.23***
DeepSeek V3	-0.17***	-0.07***	0.02	-0.09***	0.24***	0.44***

Notes: * $P < 0.10$, ** $P < 0.05$, *** $P < 0.01$. The number of countries is 74, except for altruism in DeepSeek V3, where $N = 73$. Each cell reports the mean country-level difference in gender gaps between LLM-predicted and GPS-surveyed preferences, defined as (gender gap estimated from LLM data − gender gap estimated from GPS data). Positive values indicate that LLMs predict women to have higher preference levels than those observed in the GPS survey data.

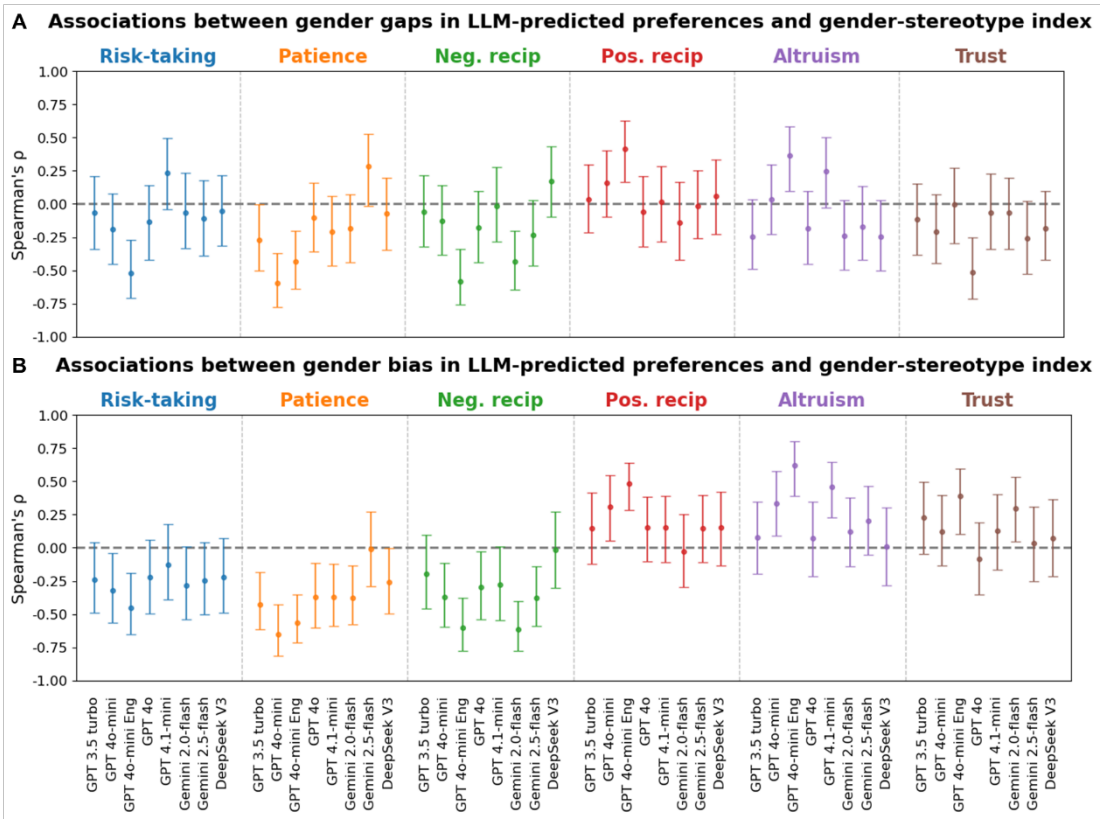

Figure F.1. Associations between gender gaps, gender biases, and gender-stereotype index. **(A) Gender gaps in LLM-predicted preferences and the gender-stereotype index.** For each model and preference domain, we estimate country-specific gender gaps in preferences using the regression specification of Falk and Hermle (2018). The gender gap is captured by the female indicator, controlling for age, age squared, and cognitive ability. Each point reports the Spearman correlation between these country-level gender gaps and the composite gender-stereotype index. **(B) Gender bias in LLM-predicted preferences and gender-stereotype index.** Gender bias is defined, for each country, preference domain, and model, as the difference between the LLM-predicted and GPS-surveyed gender gaps in that preference (LLM gap minus GPS gap). Each point in Panel B reports the Spearman correlation between these gender-bias measures and the composite gender-stereotype index. Error bars indicate 95% bootstrapped CIs in both panels.

Appendix G. Response Diversity in Large Language Models

One argument in favor of replacing or supplementing human participants with LLMs is that models might generate a wider range of responses than conventional survey methods, "possibly giving them a higher degree of freedom to generate diverse responses than that of conventional human participant methods" (Grossmann et al. 2023). Using the GPS as a benchmark, we directly assess this claim by comparing the diversity of LLM-generated responses with that of human responses.

For each country, preference domain, and model, we count the number of distinct values observed in the GPS data and in the corresponding LLM-generated data across each preference domain and each country (Wang et al. 2025). Because each preference index (except trust) aggregates multiple survey items and the raw item-level GPS data are not available, we use the number of distinct values as our measure of response diversity. We then define a diversity ratio as the number of distinct LLM-generated values divided by the number of distinct values in the GPS data. If LLMs produced more diverse responses than humans, this ratio would be at least 1.

Averaged across countries and models, the diversity ratio is 0.252 (risk-taking: 0.214; patience: 0.416; negative reciprocity: 0.142; positive reciprocity: 0.163; altruism: 0.099; trust: 0.252). Figure G.1 reports diversity ratios for each of the six preference domains and eight models. All 48 model-preference combinations (6 domains × 8 LLMs) have mean ratios significantly below 1 (two-sided t-test $P < 0.001$), indicating that LLMs never reach human levels of response diversity. For five of the six domains (all except trust), most ratios are below 0.5 (39 of 40 combinations), and for altruism all eight ratios are below 0.2. Thus, LLM-generated responses are substantially less diverse than human responses.

Because the GPS imputes missing values whenever a preference index is constructed from multiple items and does not flag which observations are imputed, trust provides the cleanest comparison, as it is based on a single survey question. Even in this domain, the largest diversity ratio (for GPT-3.5-turbo) is 0.647, which implies that the model covers, on average, about 7 of the 11 available response categories. Combined with earlier work, our one-to-one setup

strengthens the conclusion that LLMs approximate "average-human" behavior while generating markedly less diverse responses (Mei et al. 2024, Wang et al. 2025).

We next examine how response diversity varies with country-level variables. Tables G.1 and G.2 report standardized regression coefficients from country-level regressions in which the dependent variable is the diversity index, standardized to have mean 0 and standard deviation 1. The key regressors are standardized measures of economic development (log GDP per capita, IDI) and population indicators (log population, log number of internet users). All regressions control the log of sample size in GPS data and include model fixed effects. In Table G.1, we additionally include preference-domain fixed effects and model-by-preference interaction effects.

When all preference domains are pooled (Table G.1), higher levels of economic development and larger populations are both associated with lower response diversity. A one–standard deviation increase in log GDP per capita is associated with a 0.172-standard-deviation decline in the diversity index ($P < 0.001$; column 1), and a one–standard deviation increase in the IDI is associated with a 0.154-standard-deviation decline ($P < 0.001$; column 3) when controlling the log population. The population measures show smaller but still significant negative associations, with a one–standard-deviation increase in log population associated with a 0.054–standard-deviation decrease in the diversity index ($P < 0.001$; column 1) and a one–standard-deviation increase in log number of internet users associated with a 0.048–standard-deviation decrease ($P < 0.001$; column 2), when controlling the log GDP per capita.

Because these four country-level indicators are highly collinear, we do not include them jointly in a single specification. Instead, we perform principal component analysis (PCA) separately on the two development indicators and on the two population indicators, and extract the first principal component from each set. In Column (5) of Table G.1, the first component of the development indicators is strongly negatively associated with the diversity index (b=−0.182, $P < 0.001$), whereas the first component of the population indicators has a much smaller negative coefficient (b=−0.048, $P < 0.001$). This pattern suggests that LLMs simulating respondents from countries with lower levels of economic development tend to

generate more varied responses, whereas simulations for more economically developed countries are more concentrated and deterministic, with population size playing a secondary but still detectable role.

The trust domain offers the most conservative results, as GPS cannot impute missing values in this domain. Tables G.2 repeats the analysis using only the diversity index for the trust domain. The negative associations with economic development indicators remain robust. The coefficients on standardized log GDP per capita and the IDI range from −0.133 to −0.139 ($P < 0.001$; columns 1 and 2) and from −0.077 to −0.081 ($P < 0.05$, columns 3 and 4), respectively. By contrast, the population indicators show weaker and less stable effects. The coefficients on standardized log population and log number of internet users are −0.057 and −0.050 ($P < 0.1$) when controlling log GDP per capita (columns 1 and 2). However, they are no longer significantly different from zero at the 10% level, when controlling the IDI (columns 3 and 4). In the PCA specification (column 5), the economic-development component is again more strongly associated with lower diversity (−0.118, $P < 0.001$) than the population component (−0.050, $P = 0.076$).

These results indicate that LLMs tend to produce less diverse responses for respondents from more economically developed countries, while generating somewhat more varied responses for less developed contexts. Population size and the number of internet users also correlate negatively with diversity, but their effects are smaller in magnitude and less robust once economic development is taken into account.

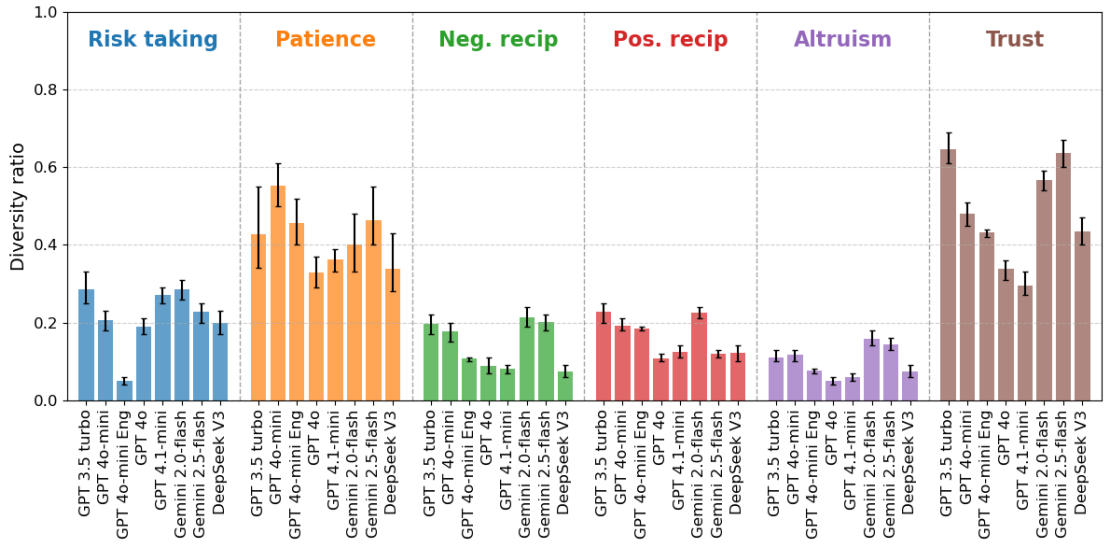

Figure G.1 Diversity ratios across preference domains and models. Bars show the mean diversity ratio across countries for each combination of preference domain and LLM. The diversity ratio is defined as the number of distinct values in the LLM-generated preference index divided by the number of distinct values in the corresponding GPS preference index within a country. Larger ratios indicate greater response diversity in the LLM relative to human respondents. All ratios are significantly below 1 (two-sided t-tests, $P < 0.001$), indicating that none of the models reaches human levels of response diversity. Error bars indicate bootstrapped 95% CIs.

Table G.1. Country-level regressions of the diversity index on development and population indicators (all preference domains pooled)

	(1)	(2)	(3)	(4)	(5)
std. log [GDP p/c]	-0.172***	-0.180***			
	(0.012)	(0.014)			
std. IDI			-0.154***	-0.159***	
			(0.014)	(0.016)	
std. log population	-0.054***		-0.048***		
	(0.010)		(0.011)		
std. log internet user		-0.048***		-0.045***	
		(0.011)		(0.011)	
std. score for economic development					-0.182***
					(0.016)
std. score for population					-0.048***
					(0.011)
Observations	3504	3264	3456	3264	3264
R-squared	0.568	0.567	0.566	0.560	0.565

Notes: The dependent variable is the standardized diversity index (mean 0, standard deviation 1), pooled across all six preference domains. All regressors (log GDP per capita, IDI, log population, log number of internet users, and their principal components) are standardized to have mean 0 and standard deviation 1, so coefficients are directly comparable across and within columns. Reported coefficients are OLS estimates, with robust standard errors in parentheses. Regressions include controls for the log of the GPS sample size in each country, preference-domain fixed effects, model fixed effects, and model × preference interaction effects. * $P < 0.1$, ** $P < 0.05$, *** $P < 0.01$.

Table G.2. Country-level regressions of the diversity index on development and population indicators (trust domain only)

	(1)	(2)	(3)	(4)	(5)
std. log [GDP p/c]	-0.139***	-0.133***			
	(0.030)	(0.030)			
std. IDI			-0.077**	-0.081**	
			(0.034)	(0.034)	
std. log population	-0.057**		-0.042		
	(0.027)		(0.028)		
std. log internet user		-0.050*		-0.045	
		(0.028)		(0.028)	
std. score for economic development					-0.118***
					(0.033)
std. score for population					-0.050*
					(0.028)
Observations	584	544	576	544	544
R-squared	0.475	0.490	0.470	0.477	0.484

Notes: The dependent variable is the standardized diversity index for the trust domain (mean 0, standard deviation 1). All regressors (log GDP per capita, IDI, log population, log number of internet users, and their principal components) are standardized to have mean 0 and standard deviation 1. Reported coefficients are OLS estimates with robust standard errors in parentheses. Regressions include controls for the log of the GPS sample size in each country and model fixed effects. * $P < 0.1$, ** $P < 0.05$, *** $P < 0.01$.

Appendix H. Variables Definitions

Log [GDP per capita]. Natural logarithm of GDP per capita in 2022, taken from the World Bank Development Indicators. Note that in Appendix E, we additionally use historical GDP series aligned with Falk et al. (2018) for comparability with their analyses.

Log [population]. Natural logarithm of total population in 2022, taken from the United Nations Development Programme (UNDP) data center.

Log [internet users]. Natural logarithm of the number of internet users in 2022, taken from the World Bank World Development Indicators.

ICT Development Index (IDI). Country-level ICT Development Index from the International Telecommunication Union website (https://www.itu.int/itu-d/reports/statistics/IDI2023/). The IDI program was relaunched in 2023, so we use the 2023 IDI values.

Gender-stereotype index (WVS). Taken from the World Value Survey website (https://www.worldvaluessurvey.org/WVSEVSjoint2017.jsp). We use the Joint EVS/WVS 2017-2022 dataset (v5.0; Jun 24, 2024). We extract four items related to gender stereotypes:

D059: *Men make better political leaders than women do*;

D060: *University is more important for a boy than for a girl*;

D061: *Pre-school child suffers with working mother*;

D078: *Men make better business executives than women do.*

We perform a principal components analysis (PCA) on these four items and take the first principal component as the WVS gender-stereotype index. The sign of the component is reversed so that higher values indicate stronger endorsement of stereotypical beliefs about gender roles.

Global Gender Gap Index. Country-level Gender Gap Index from Prosperity Data360 (World Bank) (https://prosperitydata360.worldbank.org/en/indicator/WEF+GGR+INDEX), based on the World Economic Forum GGGI series. We compute the average value over 2010–2022 Because higher GGGI values indicate greater gender parity, we use the reciprocal of the index

(1/GGGI) so that larger values correspond to greater gender gaps and standardize it to have mean 0 and standard deviation 1.

UNDP Gender Inequity Index. Country-level Gender Inequality Index taken from the United Nations Development Programme (https://hdr.undp.org/data-center/documentation-and-downloads). We compute the average GII over 2010–2022 and z-standardize it (mean 0, standard deviation 1) to obtain a comparable index.

UNDP Gender Development Index (GDI). Country-level Gender Development Index from the Data center of the United Nations Development Programme website (https://hdr.undp.org/data-center/documentation-and-downloads). We compute the average GDI over 2010–2022, use the reciprocal of the index (1/GDI) and standardize it to have mean 0 and standard deviation 1. This standardized GDI is then used, together with the WVS gender-stereotype index, the GGGI-based measure, and the GII, in the construction of the composite gender-stereotype index described in Appendix F.

Hofstede cultural dimensions: Six cultural dimension (version 2015-12-08) in Hofstede's Cultural Theory taken from Geert Hofstede website (https://geerthofstede.com/research-and-vsm/dimension-data-matrix/).

WVS cultural dimensions: Two WVS cultural dimensions (2023 Version) taken from WVS website (https://www.worldvaluessurvey.org/WVSNewsShow.jsp?ID=467). We use the latest data for each country.

Schwartz's cultural dimensions: Seven Schwartz cultural value orientation scores taken from ResearchGate open-sourced data uploaded by Shalom H Schwartz (https://www.c.net/publication/304715744_The_7_Schwartz_cultural_value_orientation_scores_for_80_countries).

Reference

Aftiss A, Lamsiyah S, Schommer C, El Alaoui SO (2025) Empirical evaluation of pre-trained language models for summarizing moroccan darija news articles. Ezzini S, Alami H, Berrada I, Benlahbib A, El Mahdaouy A, Lamsiyah S, Derrouz H, et al., eds. *Proceedings of the 4th Workshop on Arabic Corpus Linguistics (WACL-4)* (Association for Computational Linguistics, Abu Dhabi, UAE), 77–85.

Armstrong L, Liu A, MacNeil S, Metaxa D (2024) The silicon ceiling: auditing GPT's race and gender biases in hiring. *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* EAAMO '24. (ACM, San Luis Potosi Mexico), 1–18.

Exley CL, Hauser OP, Moore M, Pezzuto JH (2025) Believed gender differences in social preferences. *Q. J. Econ.* 140(1):403–458.

Falk A, Becker A, Dohmen T, Enke B, Huffman D, Sunde U (2018) Global evidence on economic preferences. *Q. J. Econ.* 133(4):1645–1692.

Falk A, Hermle J (2018) Relationship of gender differences in preferences to economic development and gender equality. *Science* 362(6412):eaas9899.

Fulgu RA, Capraro V (2024) Surprising gender biases in GPT. *Comput. Hum. Behav. Rep.* 16:100533.

Grossmann I, Feinberg M, Parker DC, Christakis NA, Tetlock PE, Cunningham WA (2023) AI and the transformation of social science research. *Science* 380(6650):1108–1109.

Hofstede G, Hofstede GJ, Minkov M (2010) *Cultures and organizations: software of the mind: intercultural cooperation and its importance for survival* Revised and expanded third edition. (McGraw-Hill, New York).

Inglehart R, Welzel C (2005) *Modernization, cultural change, and democracy: the human development sequence* (Cambridge University Press).

Kaplan J, McCandlish S, Henighan T, Brown TB, Chess B, Child R, Gray S, Radford A, Wu J, Amodei D (2020) Scaling laws for neural language models. (January 23) http://arxiv.org/abs/2001.08361.

Kotek H, Dockum R, Sun D (2023) Gender bias and stereotypes in large language models. *Proceedings of the ACM Collective Intelligence Conference* CI '23. (ACM, Delft Netherlands), 12–24.

Mei Q, Xie Y, Yuan W, Jackson MO (2024) A turing test of whether AI chatbots are behaviorally similar to humans. *Proc. Natl. Acad. Sci.* 121(9):e2313925121.

Schwartz S (2006) A theory of cultural value orientations: explication and applications. *Comp. Sociol.* 5(2–3):137–182.

Sunde U, Dohmen T, Enke B, Falk A, Huffman D, Meyerheim G (2022) Patience and comparative development. *Rev. Econom. Stud.* 89(5):2806–2840.

Villalobos P, Ho A, Sevilla J, Besiroglu T, Heim L, Hobbhahn M (2022) Will we run out of data? Limits of LLM scaling based on human-generated data. (October 26) https://www.semanticscholar.org/paper/6cf65eb8aa66116e14a97bb8f71552359ff814ba.

Wang A, Morgenstern J, Dickerson JP (2025) Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nat. Mach. Intell.*

7(3):400–411.

Zack T, Lehman E, Suzgun M, Rodriguez JA, Celi LA, Gichoya J, Jurafsky D, et al. (2024) Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. *Lancet Digit. Health* 6(1):e12–e22.