

Simulated responses from large language models are weak proxies for human economic preferences across 74 countries

In the format provided by the
authors and unedited

Supplementary Information

Section 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 [Supplementary Table A.1](#). To ensure comparability across models, we retain only those respondents for whom we successfully obtained responses from all twelve LLMs. The merged dataset constructed from these twelve models is described in [Supplementary Table A.2](#).

There are, however, several minor differences between the virtual GPS and the original GPS.

1. **Translation.** We use the Google Translate API to translate personas and survey items from English to the local languages. Since this API does not support Dari, we instead 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 ². 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 widely sampled language for that country. As a result, the total sample size of the virtual GPS is slightly smaller than that of the original GPS, at approximately 86–88% of the original number of respondents.

We implement the virtual GPS in three steps:

1. We extract each 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 English-only conditions of GPT-4o-mini, GPT-5.4-mini, and GPT-5.4, in which all prompts are written in English). The survey items are listed in [Supplementary 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 the 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 option number.

The prompts are as follows:

- **System prompt:**

You are a [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 [Supplementary 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) 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. We then collected GPT-5.4 and GPT-5.4-mini data (including the robustness experiments: ablations, paraphrasing, and response-constraint experiments) from March 23 to April 9, 2026, GPT-5.4-mini (high-temperature experiment) data from May 8 to May 9, 2026. Due to differences in model coverage across countries and languages, the resulting sample sizes vary slightly by model (see [Supplementary Table A.1](#)). During data collection with GPT-4o-mini, GPT-4o-mini (English), 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, GPT-5.4, GPT-5.4(English), GPT-5.4-mini and GPT-5.4-mini (English), 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 waves of data collection.

In our analysis, we retained the data affected by the programming error. Because the error only triggered re-collection when a model returned “0”, its impact is bounded by how often zero responses occur. [Supplementary Table A.4](#) reports, for each model and each survey item where a zero response is possible, the proportion of zero responses observed in the corrected data. Across all eleven unaffected conditions, the overall rate of zero responses is 0.4%. Only GPT-3.5-turbo exhibits a rate exceeding 1%. Notably, GPT-4o-mini (English), which shares the same base model as the affected GPT-4o-mini condition, never returned a zero response on these items. These patterns indicate that zero responses are rare in general and essentially absent for the GPT-4o-mini base model, implying that only a negligible fraction of observations in the affected condition could have been altered by the re-collection procedure.

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
Panel A. Main models			
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
GPT-5.4-mini	69,681	86.74	74
GPT-5.4-mini (English)	69,671	86.72	74
GPT-5.4	69,681	86.74	74
GPT-5.4 (English)	69,681	86.74	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
Panel B. Robustness experiments (GPT-5.4-mini; ablation, paraphrasing, response-constraint, and high-temperature experiments)			
Demographic-persona	69,666	86.72	74
Country-persona	69,620	86.66	74
Partial-persona (w/o gender)	67,940	84.57	74
Partial-persona (w/o age)	67,022	83.43	74
Partial-persona (w/o cognitive ability)	69,681	86.74	74
Paraphrasing (English)	69,681	86.74	74
Response-constraint	69,681	86.74	74
High-temperature	69,681	86.74	74
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 ($n = 80,337$). Panel A reports the twelve LLM configurations used for cross-model comparison, together with the original GPS benchmark. Panel B reports robustness experiments based on GPT-5.4-mini: rows 1–5 vary the persona specification, row 6 paraphrases the survey items, row 7 enforces a constrained response format, and row 8 increases the decoding temperature.

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 GPS sample covered
Afghanistan	310	1000	31.00
Algeria	1015	1022	99.32
Argentina	960	1000	96.00
Australia	1001	1002	99.90
Austria	998	1001	99.70
Bangladesh	997	999	99.80
Bolivia	977	998	97.90
Bosnia Herzegovina	513	1004	51.10
Botswana	580	1000	58.00
Brazil	1002	1003	99.90
Cameroon	827	1000	82.70
Canada	784	1001	78.32
Chile	993	1003	99.00
China	2573	2574	99.96
Colombia	988	1000	98.80
Costa Rica	996	1000	99.60
Croatia	987	992	99.50
Czech Republic	985	1005	98.01
Egypt	1016	1020	99.61
Estonia	678	1004	67.53
Finland	999	1000	99.90
France	994	1001	99.30
Georgia	949	1000	94.90
Germany	994	997	99.70
Ghana	479	1000	47.90
Greece	904	1000	90.40
Guatemala	986	1000	98.60
Haiti	501	504	99.40
Hungary	834	1004	83.07
India	1265	2539	49.82
Indonesia	999	1000	99.90
Iran	2012	2507	80.26
Iraq	859	1000	85.90
Israel	796	999	79.68
Italy	1000	1004	99.60
Japan	992	1000	99.20
Jordan	995	1000	99.50
Kazakhstan	714	999	71.47

Kenya	502	1000	50.20
Lithuania	994	999	99.50
Malawi	840	1000	84.00
Mexico	987	1000	98.70
Moldova	770	1000	77.00
Netherlands	999	1000	99.90
Nicaragua	978	1000	97.80
Nigeria	656	1000	65.60
Pakistan	1003	1004	99.90
Peru	989	1000	98.90
Philippines	410	1000	41.00
Poland	998	999	99.90
Portugal	996	998	99.80
Romania	991	994	99.70
Russia	1487	1498	99.27
Rwanda	279	1000	27.90
Saudi Arabia	1031	1035	99.61
Serbia	1011	1023	98.83
South Africa	314	1000	31.40
South Korea	997	1000	99.70
Spain	998	1000	99.80
Sri Lanka	704	1000	70.40
Suriname	504	504	100.00
Sweden	994	1000	99.40
Switzerland	658	1000	65.80
Tanzania	955	1000	95.50
Thailand	998	1000	99.80
Turkey	989	1000	98.90
Uganda	533	1000	53.30
Ukraine	614	1000	61.40
United Arab Emirates	999	1000	99.90
United Kingdom	1030	1030	100.00
United States	1064	1072	99.25
Venezuela	992	999	99.30
Vietnam	1000	1000	100.00
Zimbabwe	485	1000	48.50

Notes: Number of LLM observations matched LLM-simulated observations across all twelve LLMs in the merged dataset. *Percentage of GPS sample covered* reports coverage relative to the original GPS sample for each country. There are no virtual samples from Cambodia and Morocco.

Table A.3. Survey items in the virtual GPS

Preference	Type of item	Survey item
Risk-taking	Sequence of five interdependent quantitative questions	<i>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?</i>
	Self-assessment	<i>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.</i>
Patience	Sequence of five interdependent quantitative questions	<i>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.</i>
	Self-assessment	<i>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.</i>
Negative Reciprocity	Willingness to act	<i>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.</i>
	Willingness to act	<i>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.</i>
	Self-assessment	<i>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.</i>
Positive Reciprocity	Self-assessment	<i>How well does each of the following statement describe you? “When someone does me a favor, I am willing to</i>

		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.

Table A.4. Proportion of responses with output 0 by model and survey item

Survey item	GPT-3.5-turbo	GPT-4o-mini (English)	GPT-4.1-mini	GPT-4o	GPT-5.4-mini	GPT-5.4-mini (English)	GPT-5.4	GPT-5.4 (English)	DeepSeek-V3	Gemini-2.0 Flash	Gemini-2.5 Flash	Average
Risk-taking: Self-assessment	0.005	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.011	0.002
Patience: Self-assessment	0.003	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000
Negative reciprocity: Self-assessment	0.094	0.000	0.003	0.005	0.063	0.000	0.019	0.000	0.009	0.004	0.021	0.02
Negative reciprocity: Willingness to act (yourself)	0.015	0.000	0.001	0.001	0.012	0.000	0.004	0.000	0.004	0.002	0.007	0.004
Negative reciprocity: Willingness to act (others)	0.010	0.000	0.001	0.000	0.001	0.000	0.000	0.000	0.003	0.000	0.003	0.002
Positive reciprocity: Self-assessment	0.004	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.001	0.001
Altruism: Hypothetical situation	0.012	0.000	0.001	0.000	0.001	0.000	0.000	0.000	0.000	0.002	0.001	0.002
Altruism: Willingness to act	0.003	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000
Trust	0.019	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.01	0.003
Average	0.018	0.000	0.001	0.001	0.009	0.000	0.003	0.000	0.002	0.001	0.006	0.004

Notes: Two LLM configurations use the GPT-4o-mini model: GPT-4o-mini (in localized languages) and GPT-4o-mini (English). The bug had already been fixed when we collected data for GPT-4o-mini (English).

Section 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. **Individual-level and country-level correlations:** using Spearman’s correlation coefficients as an alternative way to measure alignment.
2. **Within-country individual-level correlations:** a robustness check on the individual-level alignment reported in the main text.
3. **Split-half internal consistency:** assessing the reliability of aggregating individual-level LLM-predicted preferences into country-level averages.

B.1 Individual-level and country-level correlations and explanatory power

We first examine individual-level alignment by computing Spearman correlations between LLM-predicted and GPS-surveyed preferences (see [Methods](#) in main text for details on the construction of individual-level preferences), pooling all respondents from 74 countries. Panel A of [Figure B.1](#) reports the 72 correlations (6 preference domains $\times$ 12 LLMs). Of these, 63 are significantly positive (87.5%, based on 95% bootstrapped CI; 63 of 72 based on Benjamini-Hochberg FDR-adjusted P values, $P < 0.05$); the remaining nine non-positive estimates fall in the negative reciprocity and trust domains. This high proportion of positive correlations is significantly above chance (two-sided binomial test; proportion = 63/72, $P < 0.001$, 95% CI = [77.6%, 94.1%]). Repeating the analysis within each country yields a similarly high proportion (see Section B.2 below).

Despite the high statistical significance with the large pooled sample, the magnitudes of the correlations remain modest. 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 ($\rho = 0.252$), which indicates only a modest association. Turning to model comparisons, we find no evidence that GPT-5.4 and GPT-5.4-mini outperform earlier OpenAI models (i.e., GPT-3.5-turbo, GPT-4o-mini, GPT-4.1-mini, and GPT-4o; two-sided unpaired t -test on the individual-level correlations across the six preference domains: $t(42.56) = -0.669$, Benjamini-Hochberg FDR corrected $P = 0.658$,

Cohen’s $d = -0.189$, 95% CI = [-0.06, 0.03]) or that Gemini 2.5 Flash outperforms Gemini 2.0 Flash (two-sided unpaired t -test: $t(9.81) = -0.553$; Benjamini-Hochberg FDR corrected $P = 0.658$, Cohen’s $d = -0.319$, 95% CI = [-0.13, 0.08]). We also do not find that models with more parameters (i.e., GPT-4o, GPT-5.4 and GPT-5.4 Eng) outperform their counterparts with fewer parameters (i.e., GPT-4o-mini, GPT-5.4-mini and GPT-5.4-mini English; two-sided unpaired t -test: $t(30.21) = -0.447$, Benjamini-Hochberg FDR corrected $P = 0.658$, Cohen’s $d = -0.149$, 95% CI = [-0.07, 0.04]). Thus, we find no evidence that newer models achieve stronger individual-level alignment on economic preferences.

In Panel B of [Figure B.1](#), 58 of 72 (80.6%) country-level correlations between LLM-predicted and GPS-surveyed preferences are positive, but only 22 of 72 have 95% CIs that exclude zero (16 of 72 based on Benjamini-Hochberg FDR-adjusted P values, $P < 0.05$). Compared with their individual-level counterparts, the country-level correlations are on average slightly larger ($M_{\text{country}} = 0.137$ vs $M_{\text{individual}} = 0.094$, two-sided paired t -test: $t(71) = -2.956$, $P = 0.004$, Cohen’s $d = -0.341$, 95% CI = [-0.07, -0.01]), but they remain modest in absolute magnitude, 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.

Spearman’s correlation captures the recoverability of preference rankings but ignores the magnitude of agreement. As a complementary measure, we compute the explanatory power of LLM-predicted preferences using regressions. We regress GPS-surveyed preference on LLM-predicted preference for each domain without other covariates, and report the R^2 computed at the individual and country levels in [Table B.1](#) and [Table B.2](#). These results indicate that LLM-predicted preferences explain only 1.3% of the individual-level variance (risk-taking: 4.7%; patience: 0.9%; negative reciprocity: 0.6%; positive reciprocity: 0.6%; altruism: 0.7%; trust: 0.2%) and 4.2% of the country-level variance (risk-taking: 8.1%; patience: 5.5%; negative reciprocity: 1.4%; positive reciprocity: 4.2%; altruism: 3.1%; trust: 2.6%) in human preferences. These R^2 estimates

reinforce the conclusion from the Spearman's correlation analysis: alignment is positive but small in magnitude across both individual and country levels.

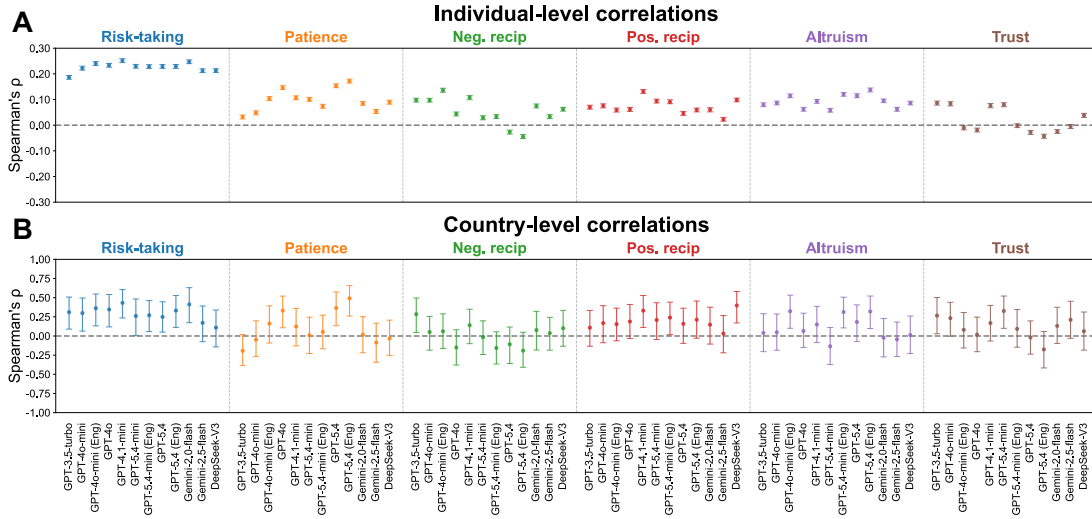


Figure B.1. Alignment between LLM-predicted and GPS-surveyed preferences. (A) Individual-level alignment. Each point represents the Spearman correlation coefficient between GPS-surveyed and LLM-predicted preferences, pooling all individual observations across 74 countries ($n = 67,201$; the number of observations varies by preference domain due to missing values in the original GPS). The predominance of positive coefficients (63 of 72, 87.5%) indicates systematic, though weak, alignment between LLM-predicted and GPS-surveyed preferences at the individual level. **(B) Country-level alignment.** 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 GPS ex-post sampling weights. Of the 72 coefficients, 58 (80.6%) are positive and 22 (30.6%) 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.

B.2 Individual-level Spearman correlations within each country

As a complement to the analysis in the main text and [Section B.1](#), we compute individual-level Spearman correlations between GPS-surveyed and LLM-predicted preferences separately for each country, pooling all individuals from 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 country. This allows us to examine not only overall individual-level alignment, but also its geographical

distribution.

[Table B.3](#) reports the number of positive within-country correlations and, among these, the number that are significantly positive after Benjamini-Hochberg multiple comparison correction.

Most models show high alignment for risk-taking and patience, with almost all countries showing positive correlations (around or above 95%). In contrast, the percentage of cases with positive correlations for trust drops sharply (average: 47.1%). In particular, GPT-4o-mini (English), GPT-4o, Gemini 2.0 Flash, Gemini 2.5 Flash, GPT-5.4 and GPT-5.4 (English) all fall below 50% in the trust domain. 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, $P < 0.05$) for all models, and for all preference domains except trust and negative reciprocity.

[Panels A of Figures B.2-B.13](#) report the distribution of individual-level alignment (i.e., Spearman correlations for each country). [Panels B of Figures B.2-B.13](#) show 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).

Table B.1. Individual-level explanatory power of LLM-predicted preferences

Model	Risk-taking	Patience	Negative reciprocity	Positive reciprocity	Altruism	Trust	Average
GPT-3.5-turbo	0.031	0.001	0.008	0.003	0.007	0.007	0.009
GPT-4o-mini	0.044	0.001	0.010	0.007	0.003	0.005	0.012
GPT-4o-mini (English)	0.039	0.009	0.020	0.004	0.014	0.000	0.014
GPT-4o	0.054	0.018	0.003	0.004	0.001	0.000	0.013
GPT-4.1-mini	0.058	0.009	0.012	0.018	0.008	0.005	0.018
GPT-5.4-mini	0.047	0.007	0.000	0.010	0.001	0.005	0.012
GPT-5.4-mini (English)	0.046	0.004	0.001	0.008	0.012	0.000	0.012
GPT-5.4	0.052	0.021	0.001	0.002	0.005	0.002	0.014
GPT-5.4 (English)	0.054	0.025	0.002	0.003	0.011	0.002	0.016
Gemini 2.0 Flash	0.057	0.003	0.006	0.002	0.007	0.000	0.013
Gemini 2.5 Flash	0.039	0.002	0.002	0.000	0.003	0.000	0.008
DeepSeek-V3	0.045	0.007	0.002	0.012	0.009	0.001	0.013
Average	0.047	0.009	0.006	0.006	0.007	0.002	0.013

Notes: The value in each cell indicates the proportion of variance in GPS-surveyed preferences explained by LLM-predicted preferences (i.e., the R^2). We estimate this by running individual-level OLS regressions of GPS-surveyed preferences on LLM-predicted preferences.

Table B.2. Country-level explanatory power of LLM-predicted preferences

Model	Risk-taking	Patience	Negative reciprocity	Positive reciprocity	Altruism	Trust	Average
GPT-3.5-turbo	0.079	0.030	0.030	0.002	0.019	0.068	0.038
GPT-4o-mini	0.062	0.001	0.010	0.044	0.004	0.020	0.024
GPT-4o-mini (English)	0.109	0.032	0.000	0.006	0.077	0.001	0.038
GPT-4o	0.168	0.129	0.019	0.019	0.000	0.002	0.056
GPT-4.1-mini	0.177	0.006	0.021	0.103	0.008	0.029	0.057
GPT-5.4-mini	0.034	0.013	0.007	0.061	0.014	0.068	0.033
GPT-5.4-mini (English)	0.052	0.006	0.011	0.041	0.093	0.014	0.036
GPT-5.4	0.049	0.136	0.028	0.020	0.018	0.001	0.042
GPT-5.4 (English)	0.050	0.292	0.031	0.018	0.087	0.023	0.084
Gemini 2.0 Flash	0.157	0.001	0.011	0.012	0.003	0.041	0.038
Gemini 2.5 Flash	0.029	0.014	0.000	0.007	0.000	0.041	0.015
DeepSeek-V3	0.010	0.000	0.000	0.173	0.052	0.001	0.039
Average	0.081	0.055	0.014	0.042	0.031	0.026	0.042

Notes: The value in each cell indicates the proportion of variance in GPS-surveyed preferences explained by LLM-predicted preferences (i.e., the R^2). We estimate this by running country-level OLS regressions of GPS-surveyed preferences on LLM-predicted preferences.

Table B.3. Number of cases with positive correlation and number of cases with significantly positive correlation.

	Risk-taking	Patience	Neg. recip	Pos. recip	Altruism	Trust
GPT-3.5-turbo	61/72	44/70	30/61	24/61	35/66	32/63
GPT-4o-mini	69/72	37/70	44/65	29/65	41/68	23/62
GPT-4o-mini (Eng)	69/74	42/73	59/73	27/61	44/71	2/27
GPT-4o	69/74	53/72	39/64	26/66	29/67	3/16
GPT-4.1-mini	68/74	52/74	50/72	42/70	43/70	20/56
GPT-5.4-mini	67/74	44/69	35/63	28/61	42/69	7/45
GPT-5.4-mini (Eng)	66/74	46/71	21/68	31/69	45/69	4/47
GPT-5.4	68/74	51/72	7/41	15/63	35/66	2/12
GPT-5.4 (Eng)	67/74	44/73	3/29	12/65	40/68	1/14
Gemini-2.0-flash	70/74	43/73	39/57	18/64	48/71	1/18
Gemini-2.5-flash	66/72	47/71	23/49	8/46	43/69	4/18
DeepSeek-V3	67/72	49/73	34/64	19/60	30/65	7/40

Notes: In each cell, the first number represents the number of cases with significantly positive correlation after Benjamini-Hochberg FDR correction, and the second number represents the number of cases with positive correlation. 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.

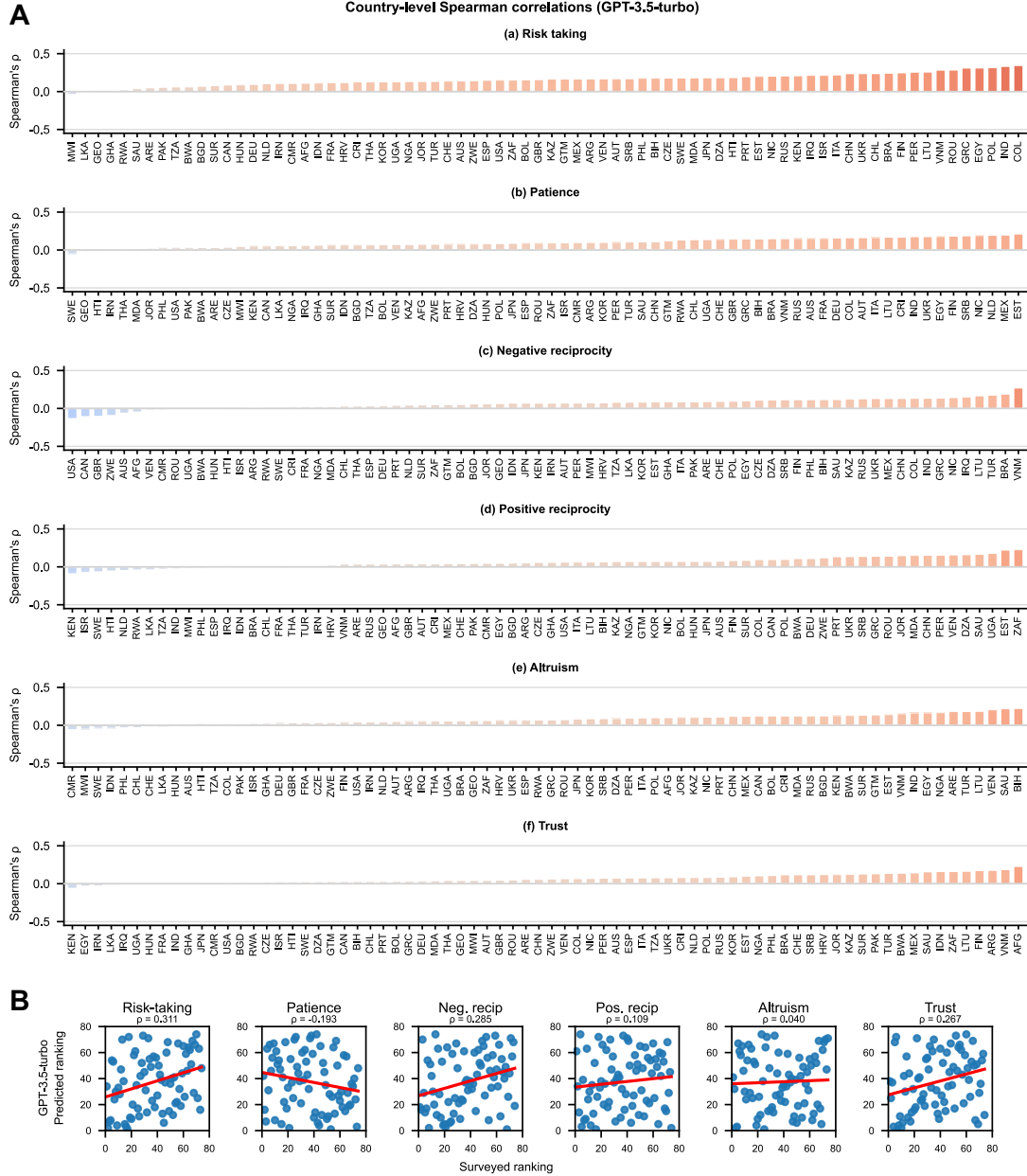


Figure B.2. Alignment across countries (GPT-3.5-turbo). (A) **Individual-level alignment.** For each country and preference domain, we compute the Spearman correlation between GPS-surveyed and LLM-predicted preferences across individuals. Countries are shaded from blue (negative correlations) to red (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 alignment.** 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.

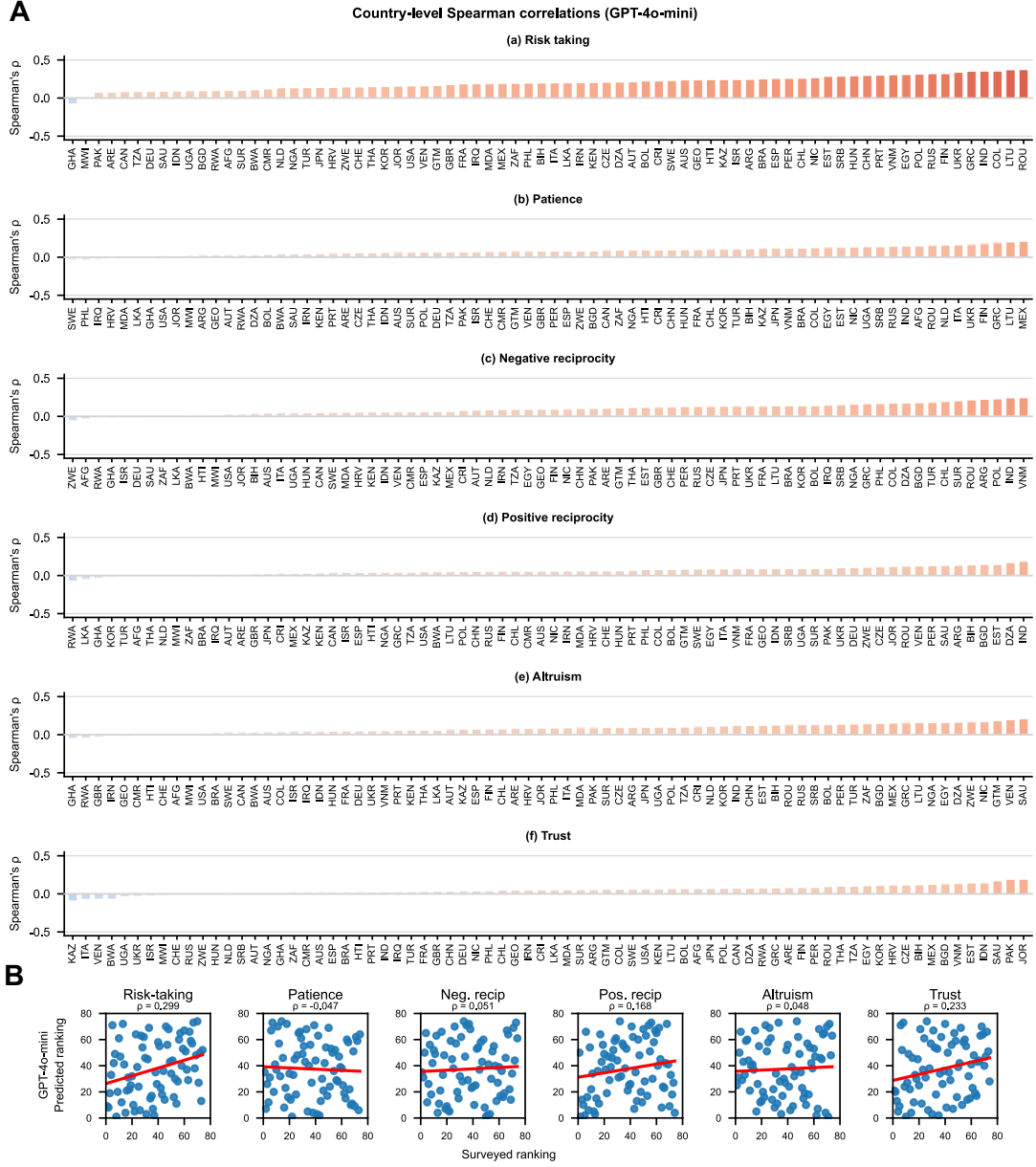


Figure B.3. Alignment across countries (GPT-4o-mini). (A) **Individual-level alignment.** For each country and preference domain, we compute the Spearman correlation between GPS-surveyed and LLM-predicted preferences across individuals. Countries are shaded from blue (negative correlations) to red (positive 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 alignment.** 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.

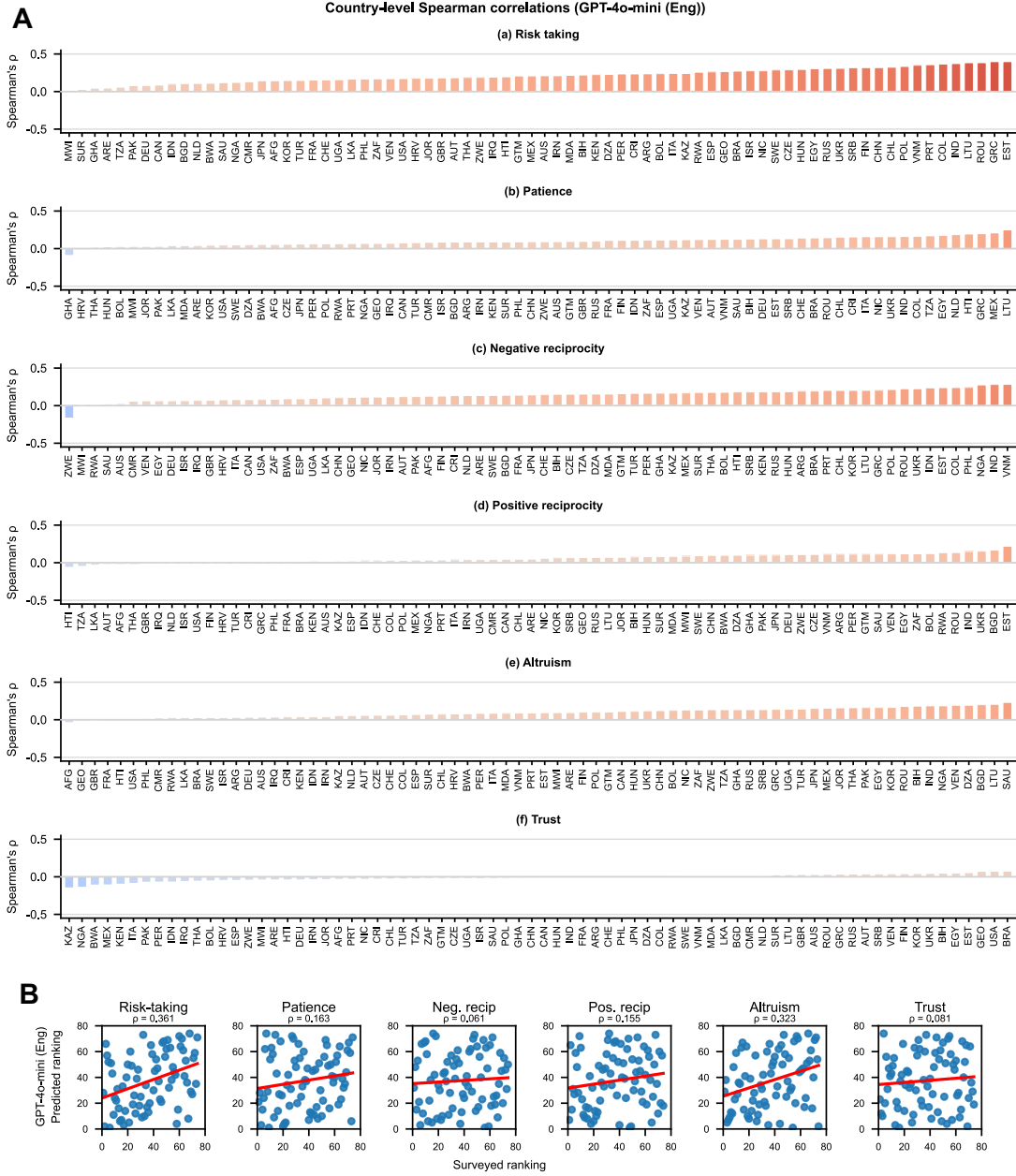


Figure B.4. Alignment across countries (GPT-4o-mini English). **(A) Individual-level alignment.** For each country and preference domain, we compute the Spearman correlation between GPS-surveyed and LLM-predicted preferences across individuals. Countries are shaded from blue (negative correlations) to 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 alignment.** 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.

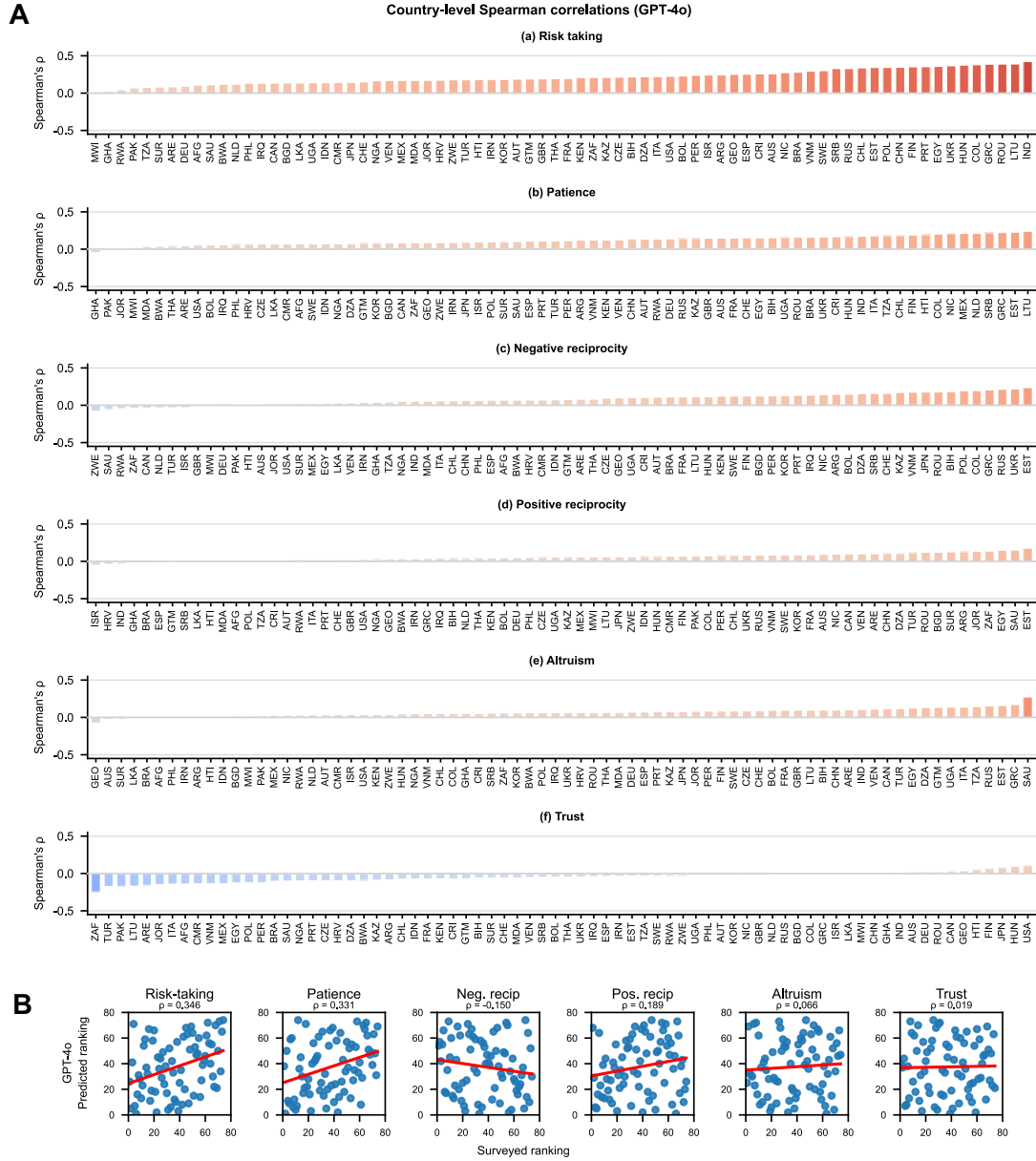


Figure B.5. Alignment across countries (GPT-4o). **(A) Individual-level alignment.** For each country and preference domain, we compute the Spearman correlation between GPS-surveyed and LLM-predicted preferences across individuals. Countries are shaded from blue (negative correlations) to red (positive 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 alignment.** 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.

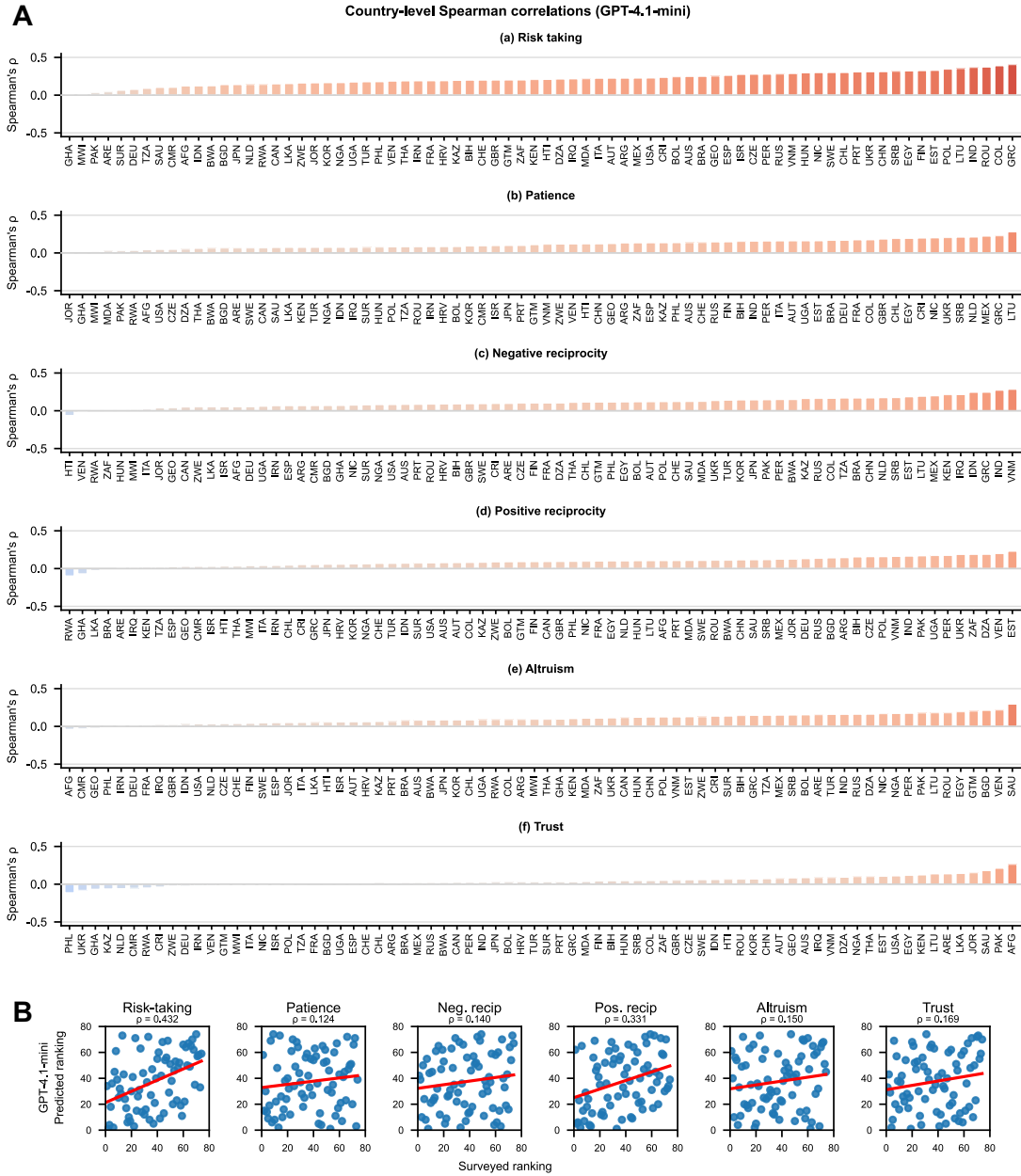


Figure B.6. Alignment across countries (GPT-4.1-mini). (A) **Individual-level alignment.** For each country and preference domain, we compute the Spearman correlation between GPS-surveyed and LLM-predicted preferences across individuals. Countries are shaded from blue (negative correlations) to red (positive 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 alignment.** 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.

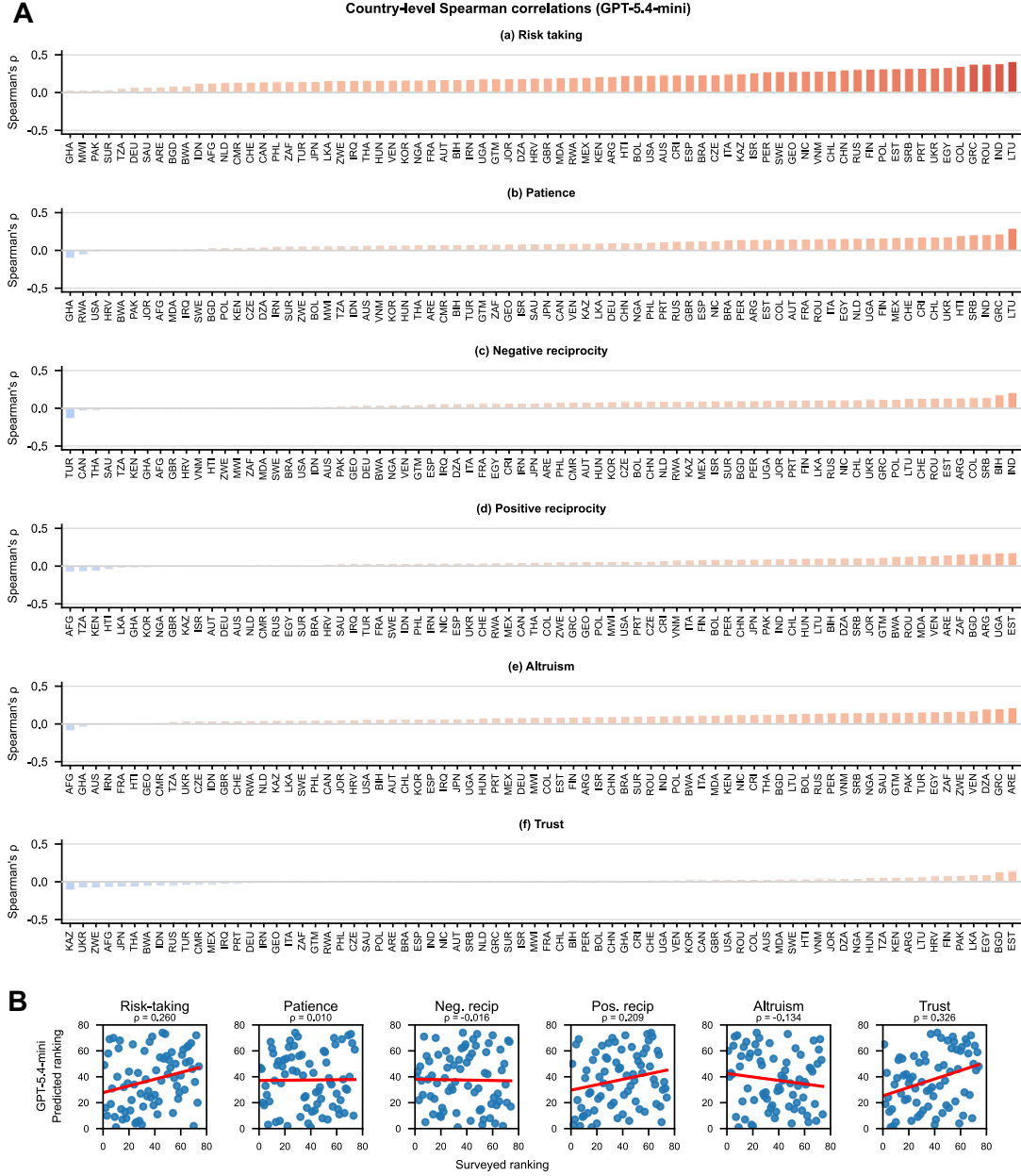


Figure B.7. Alignment across countries (GPT-5.4-mini). (A) **Individual-level alignment.** For each country and preference domain, we compute the Spearman correlation between GPS-surveyed and LLM-predicted preferences across individuals. Countries are shaded from blue (negative correlations) to red (positive 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 alignment.** 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.

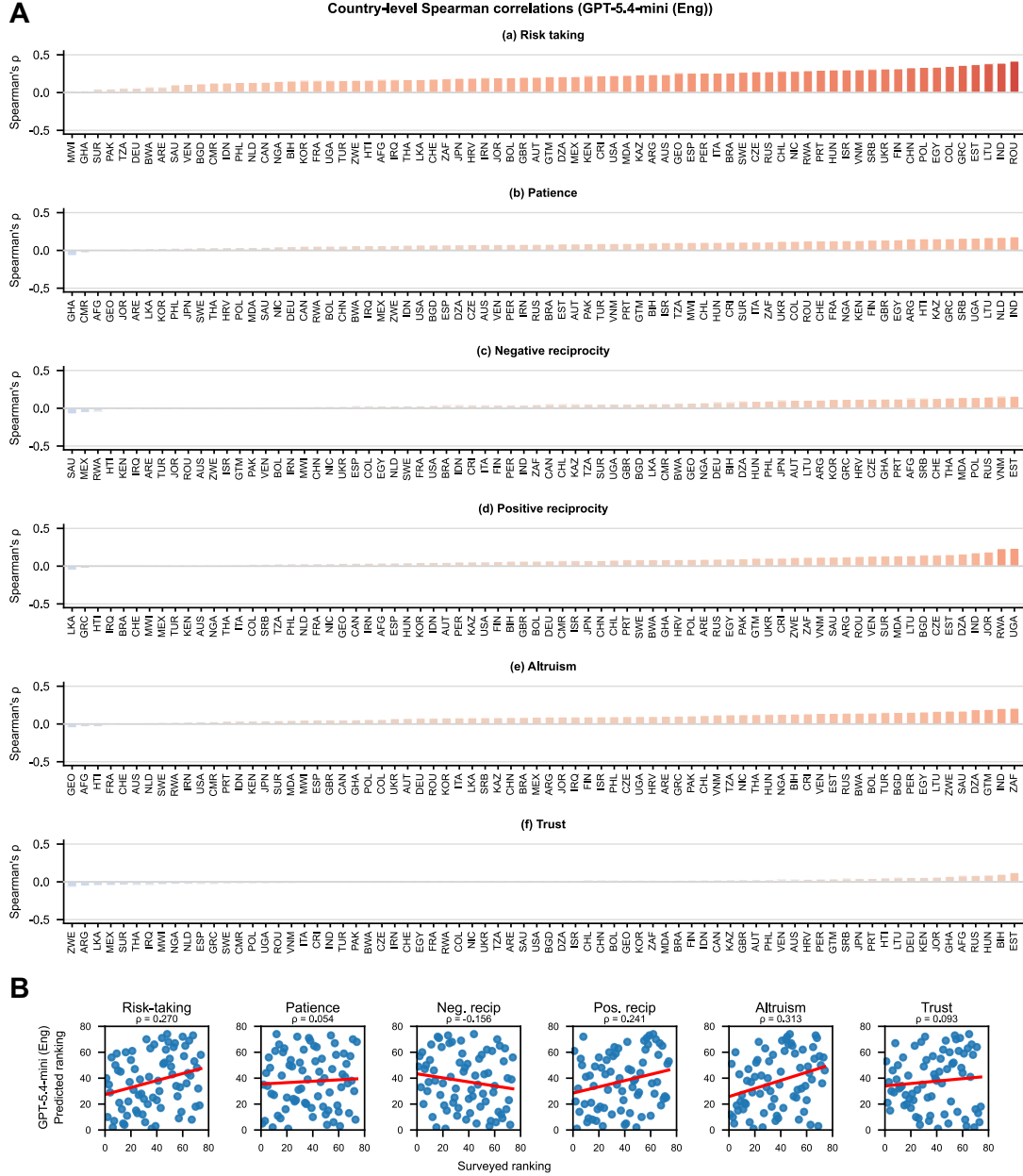


Figure B.8. Alignment across countries (GPT-5.4-mini English). **(A) Individual-level alignment.** For each country and preference domain, we compute the Spearman correlation between GPS-surveyed and LLM-predicted preferences across individuals. Countries are shaded from blue (negative correlations) 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 alignment.** 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.

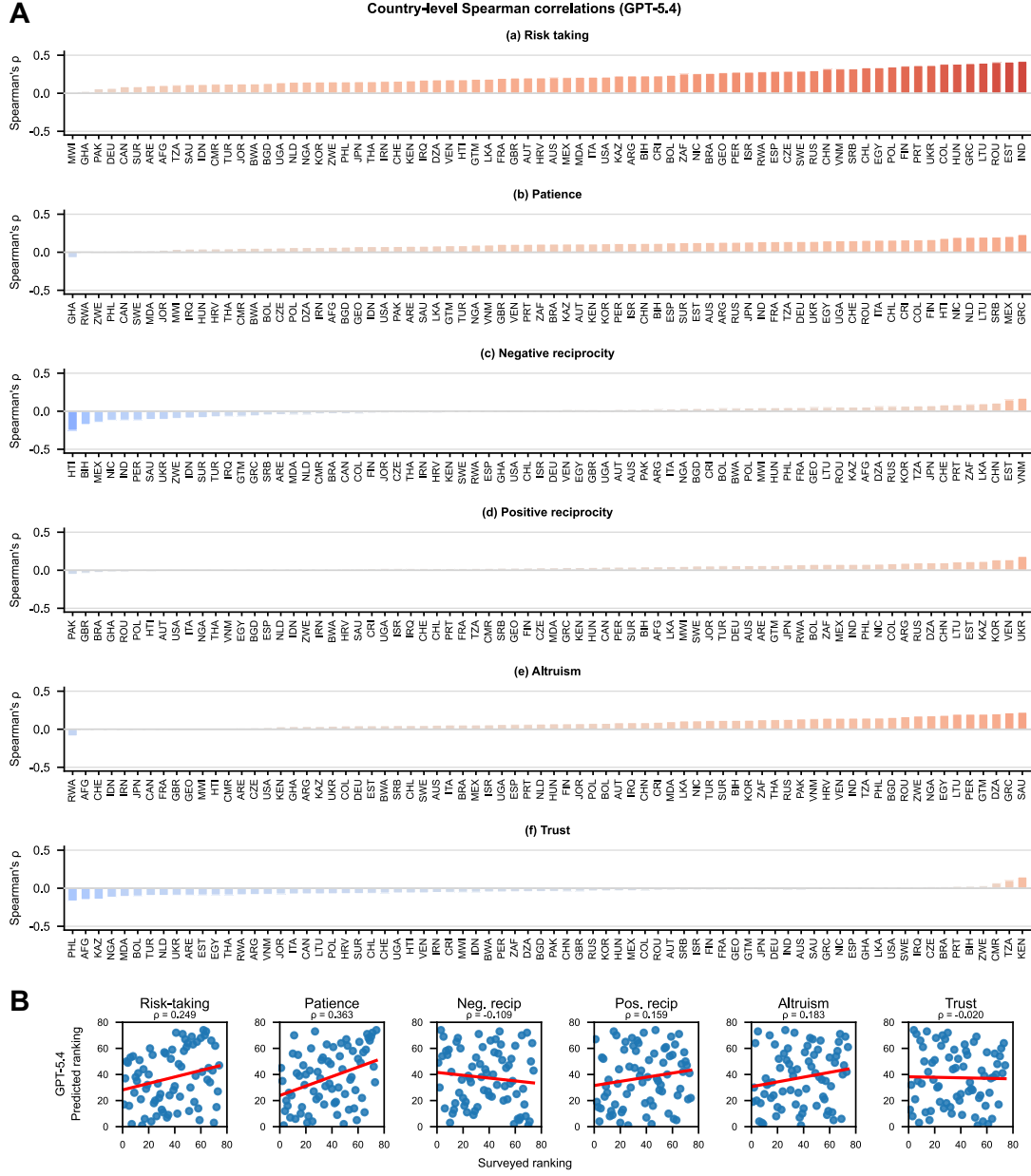


Figure B.9. Alignment across countries (GPT-5.4). (A) **Individual-level alignment.** For each country and preference domain, we compute the Spearman correlation between GPS-surveyed and LLM-predicted preferences across individuals. Countries are shaded from blue (negative correlations) to red (positive 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 alignment.** 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.

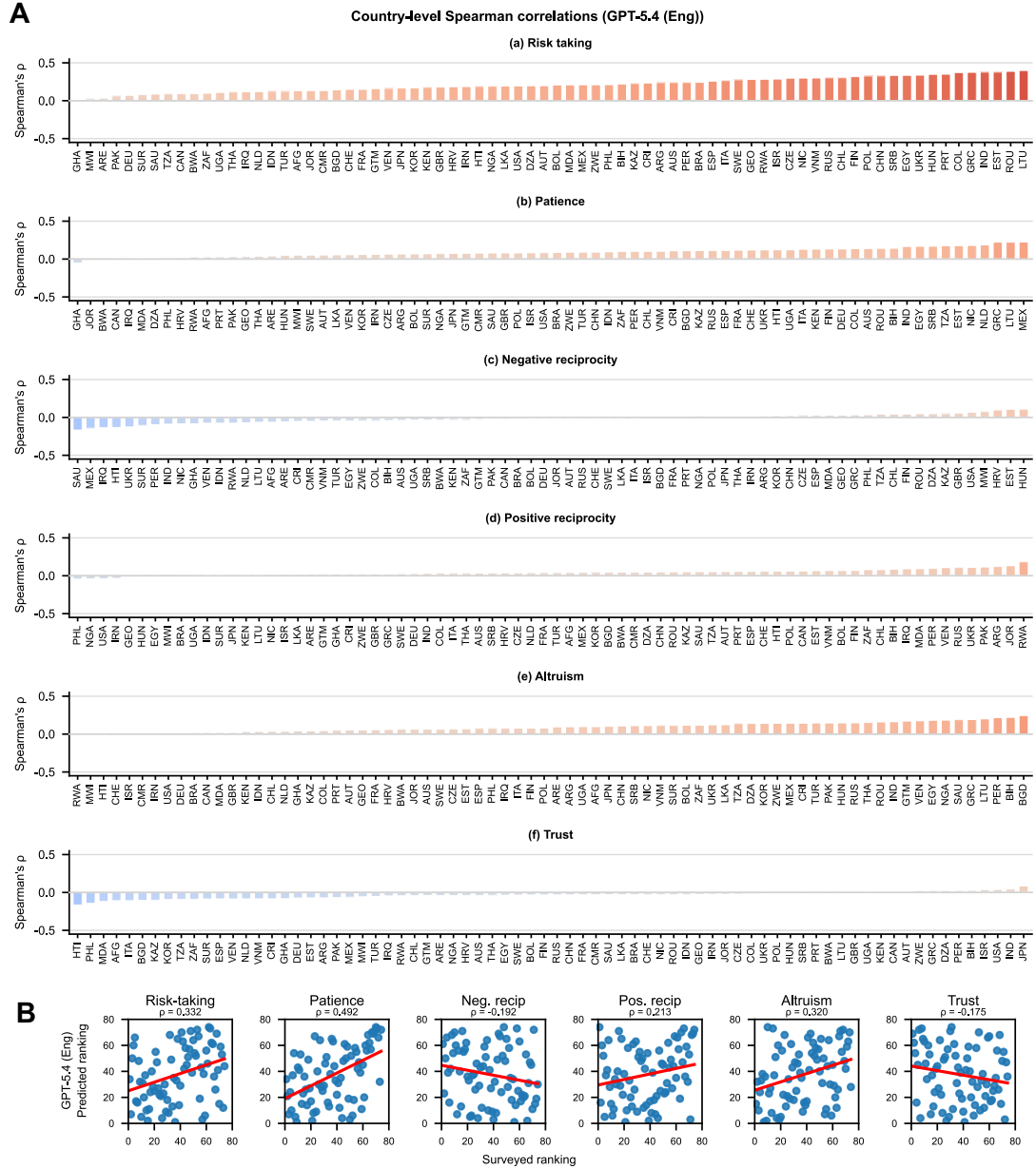


Figure B.10. Alignment across countries (GPT-5.4 English). **(A) Individual-level alignment.** For each country and preference domain, we compute the Spearman correlation between GPS-surveyed and LLM-predicted preferences across individuals. Countries are shaded from blue (negative correlations) to red (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 alignment.** 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.

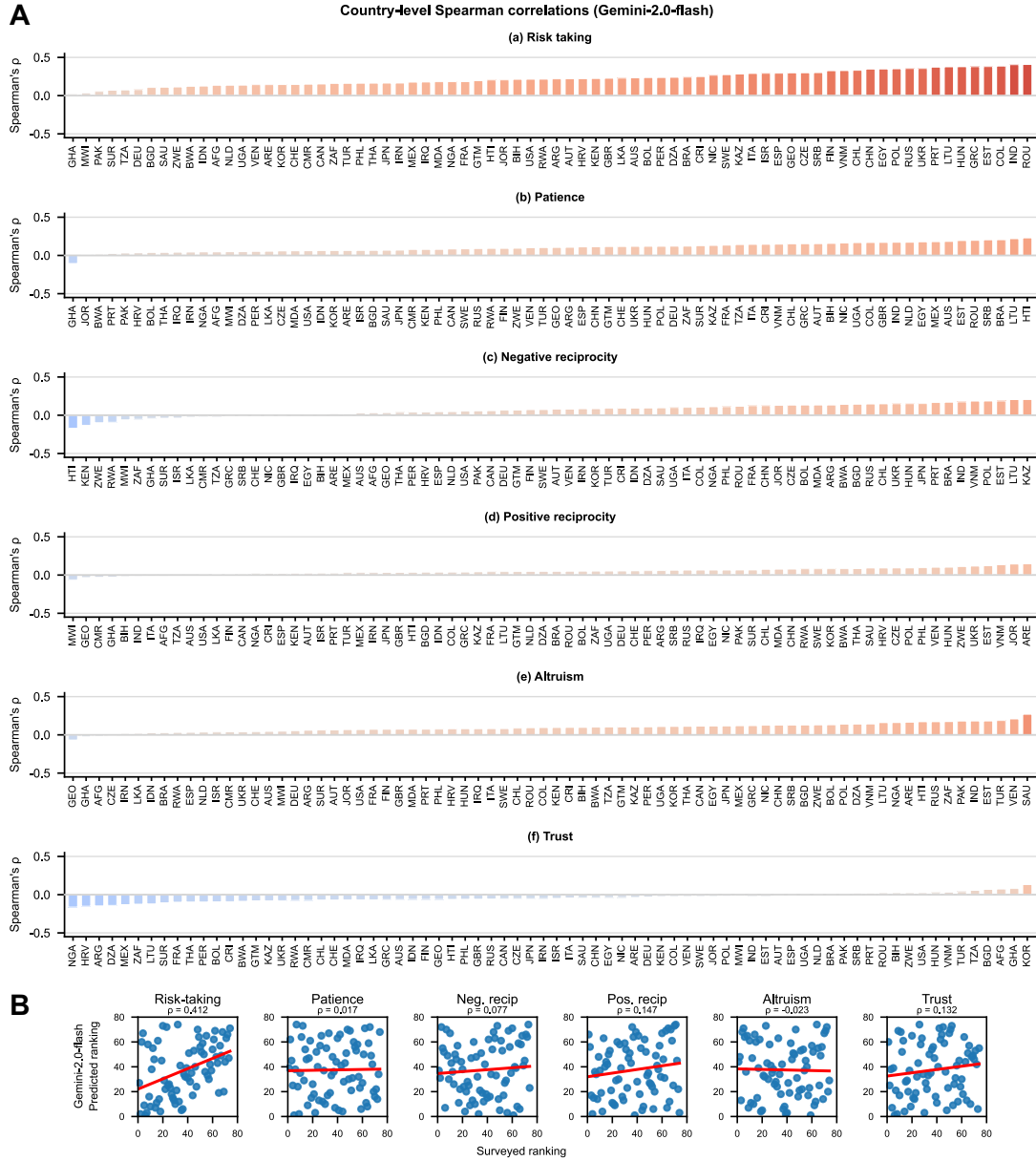


Figure B.11. Alignment across countries (Gemini 2.0 Flash). **(A) Individual-level alignment.** For each country and preference domain, we compute the Spearman correlation between GPS-surveyed and LLM-predicted preferences across individuals. Countries are shaded from blue (negative correlations) to 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 alignment.** 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.

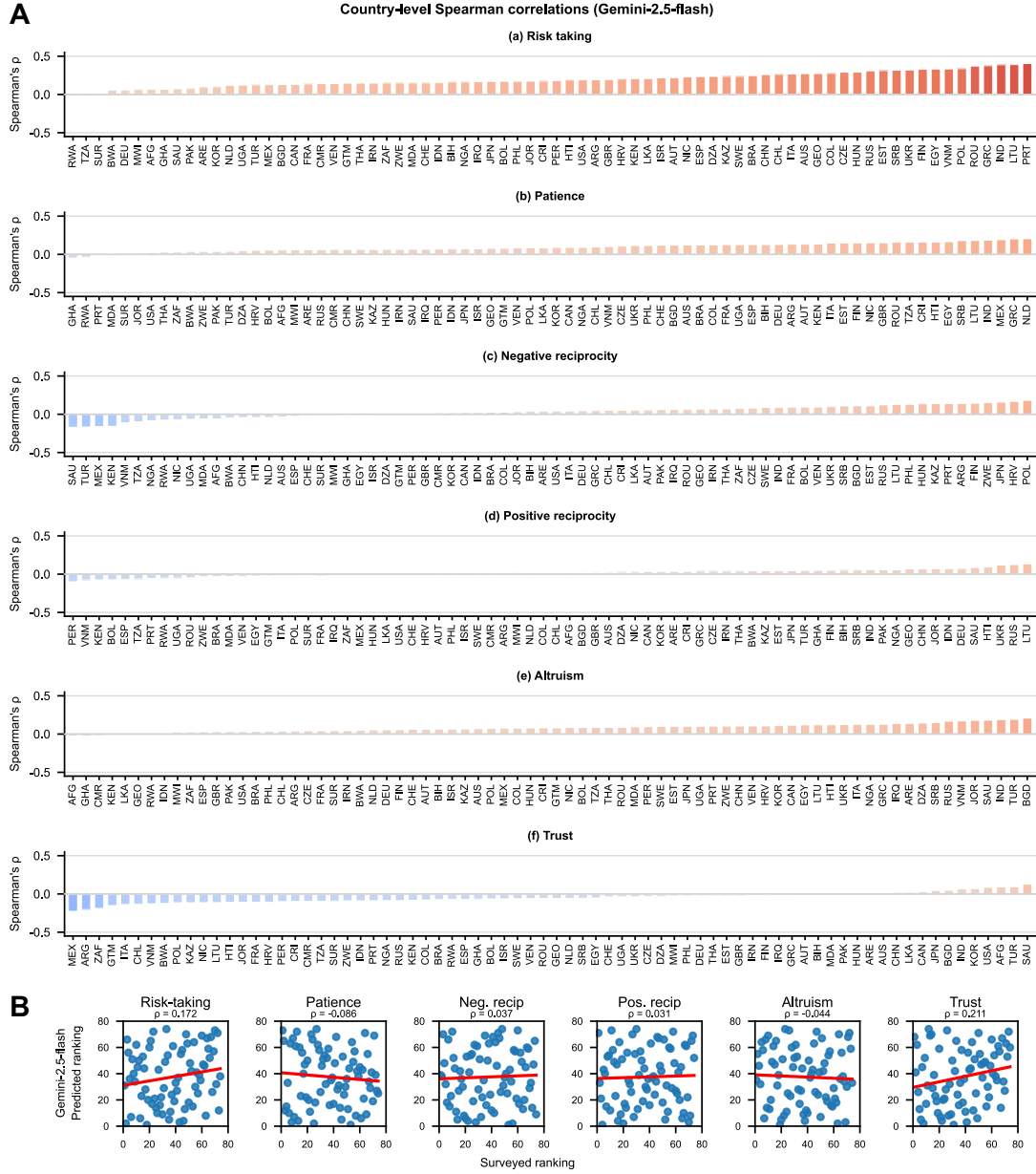


Figure B.12. Alignment across countries (Gemini 2.5 Flash). **(A) Individual-level alignment.** For each country and preference domain, we compute the Spearman correlation between GPS-surveyed and LLM-predicted preferences across individuals. Countries are shaded from blue (negative correlations) to 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 alignment.** 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.

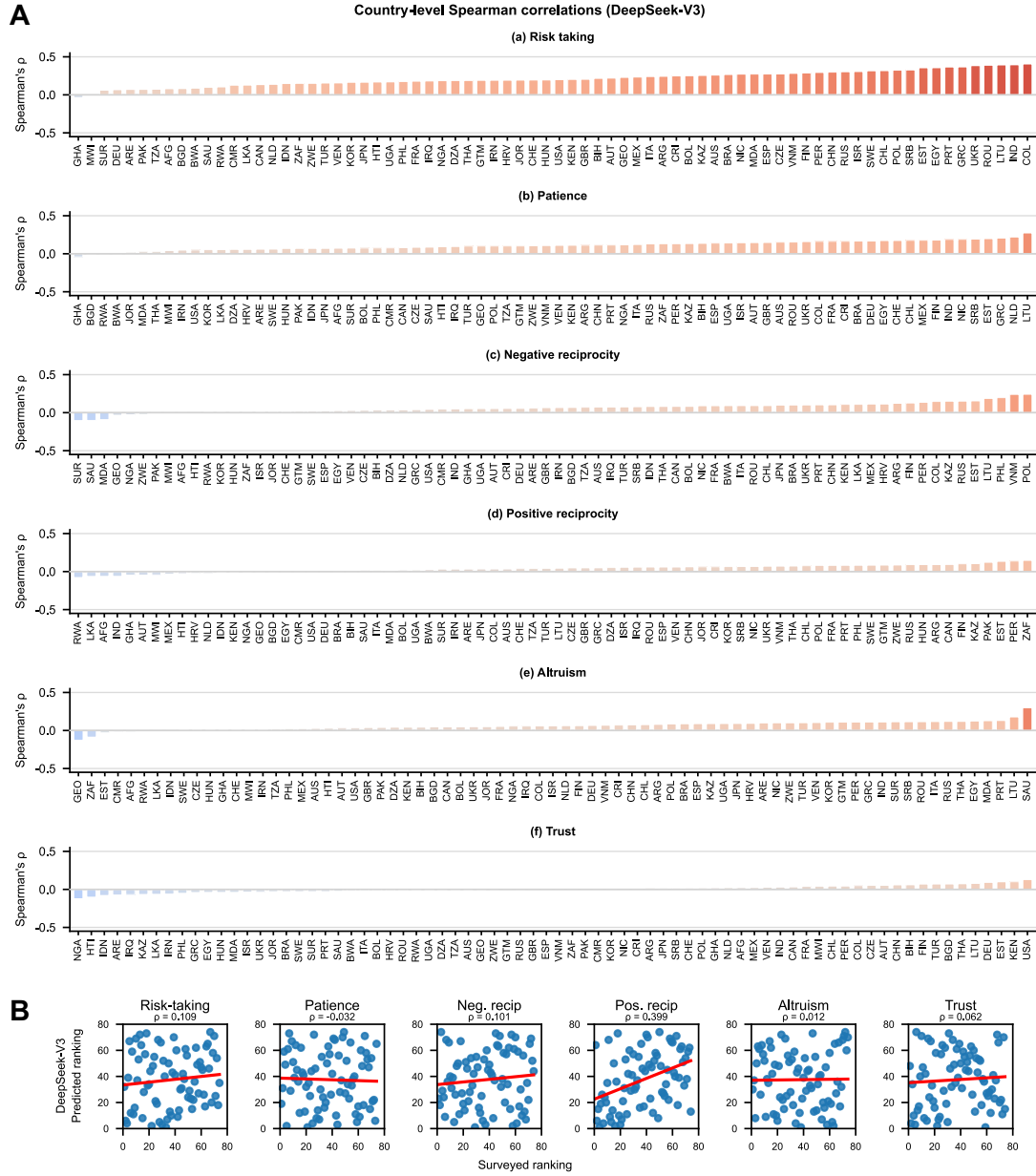


Figure B.13. Alignment across countries (DeepSeek-V3). (A) **Individual-level alignment.** For each country and preference domain, we compute the Spearman correlation between GPS-surveyed and LLM-predicted preferences across individuals. Countries are shaded from blue (negative correlations) to red (positive 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 alignment.** 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.3 Split-half internal consistency

We estimate the internal consistency of GPS-surveyed and LLM-predicted preferences using a permutation-based split-half 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.14 shows the split-half internal consistency for country-level averages.

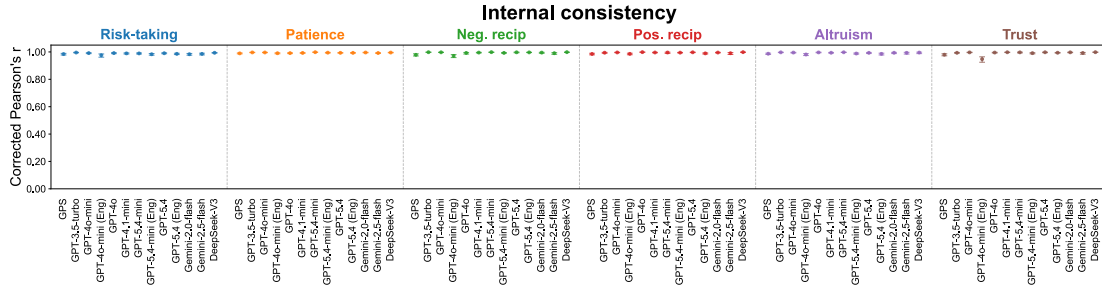


Figure B.14. Within-model internal consistency (split-half reliability). For each model, we randomly assign individuals within each country into two halves (1,000 permutations), compute country-level means for each half, and report the Spearman–Brown corrected Pearson’s r averaged across permutations. Error bars indicate the 2.5th and 97.5th percentiles of the Spearman–Brown-corrected correlations across 1,000 random splits.

The split-half internal consistency of LLM-predicted preferences is slightly but significantly higher than that of the GPS-surveyed data (mean_{GPS} = 0.984 vs mean_{LLMs} = 0.992, two-sided Welch’s t -test: $t(8.00) = 4.353$, $P = 0.002$, Cohen’s $d = 1.142$, 95% CI = [0.00, 0.01]). This supports the feasibility of aggregating individual-level LLM-predicted preferences into country-level average preference indices.

Table B.4. Individual-level Spearman’s correlations and MSEs

		Risk-taking	Patience	Neg. recip	Pos. recip	Altruism	Trust
GPT-3.5-turbo	Spearman’s ρ	0.186	0.032	0.098	0.07	0.08	0.087
	Adjusted P -value	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
	95% CI of ρ	[0.179, 0.194]	[0.024, 0.040]	[0.090, 0.105]	[0.063, 0.078]	[0.072, 0.087]	[0.078, 0.094]
	MSE	1.651	1.95	1.818	1.89	1.838	1.838
GPT-4o-mini	95% CI of MSE	[1.633, 1.669]	[1.929, 1.970]	[1.799, 1.838]	[1.866, 1.915]	[1.816, 1.861]	[1.815, 1.864]
	Spearman’s ρ	0.222	0.048	0.097	0.076	0.087	0.084
	Adjusted P -value	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
	95% CI of ρ	[0.215, 0.230]	[0.041, 0.056]	[0.090, 0.105]	[0.068, 0.083]	[0.079, 0.094]	[0.075, 0.092]
GPT-4o-mini (English)	MSE	1.582	1.923	1.803	1.839	1.895	1.857
	95% CI of MSE	[1.564, 1.600]	[1.903, 1.944]	[1.780, 1.827]	[1.817, 1.861]	[1.864, 1.929]	[1.833, 1.882]
	Spearman’s ρ	0.24	0.103	0.136	0.059	0.114	-0.011
	Adjusted P -value	<0.001	<0.001	<0.001	<0.001	<0.001	0.007
GPT-4o	95% CI of ρ	[0.233, 0.248]	[0.096, 0.111]	[0.129, 0.144]	[0.051, 0.067]	[0.107, 0.122]	[-0.018, -0.003]
	MSE	1.606	1.812	1.714	1.873	1.763	2.021
	95% CI of MSE	[1.556, 1.657]	[1.792, 1.831]	[1.695, 1.733]	[1.853, 1.892]	[1.744, 1.783]	[1.997, 2.046]
	Spearman’s ρ	0.233	0.147	0.044	0.062	0.062	-0.019
GPT-4.1-mini	Adjusted P -value	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
	95% CI of ρ	[0.226, 0.241]	[0.139, 0.154]	[0.036, 0.051]	[0.054, 0.070]	[0.055, 0.069]	[-0.026, -0.011]
	MSE	1.536	1.733	1.892	1.881	1.936	2.013
	95% CI of MSE	[1.520, 1.552]	[1.715, 1.750]	[1.867, 1.921]	[1.858, 1.903]	[1.913, 1.962]	[1.988, 2.035]
GPT-5.4-mini	Spearman’s ρ	0.252	0.106	0.108	0.131	0.093	0.076
	Adjusted P -value	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
	95% CI of ρ	[0.244, 0.259]	[0.099, 0.114]	[0.099, 0.115]	[0.124, 0.139]	[0.086, 0.100]	[0.068, 0.085]
	MSE	1.518	1.815	1.784	1.729	1.826	1.854
GPT-5.4-mini (English)	95% CI of MSE	[1.501, 1.535]	[1.796, 1.834]	[1.765, 1.806]	[1.708, 1.750]	[1.803, 1.849]	[1.823, 1.887]
	Spearman’s ρ	0.23	0.1	0.029	0.094	0.058	0.08
	Adjusted P -value	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
	95% CI of ρ	[0.222, 0.237]	[0.093, 0.108]	[0.021, 0.037]	[0.086, 0.102]	[0.051, 0.066]	[0.072, 0.088]
GPT-5.4	MSE	1.568	1.832	1.989	1.802	1.935	1.864
	95% CI of MSE	[1.552, 1.585]	[1.813, 1.849]	[1.959, 2.020]	[1.780, 1.824]	[1.908, 1.959]	[1.834, 1.897]
	Spearman’s ρ	0.228	0.073	0.033	0.091	0.12	-0.001
	Adjusted P -value	<0.001	<0.001	<0.001	<0.001	<0.001	0.707
GPT-5.4 (English)	95% CI of ρ	[0.221, 0.236]	[0.066, 0.081]	[0.026, 0.041]	[0.084, 0.099]	[0.113, 0.127]	[-0.009, 0.006]
	MSE	1.569	1.881	1.926	1.824	1.781	2.009
	95% CI of MSE	[1.552, 1.588]	[1.857, 1.903]	[1.906, 1.947]	[1.802, 1.843]	[1.762, 1.800]	[1.983, 2.036]
	Spearman’s ρ	0.229	0.153	-0.027	0.046	0.115	-0.028
GPT-5.4 (Flash)	Adjusted P -value	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
	95% CI of ρ	[0.221, 0.237]	[0.146, 0.161]	[-0.035, -0.019]	[0.038, 0.054]	[0.107, 0.122]	[-0.036, -0.020]
	MSE	1.542	1.711	2.076	1.903	1.859	2.08
	95% CI of MSE	[1.526, 1.558]	[1.693, 1.728]	[2.051, 2.103]	[1.882, 1.924]	[1.833, 1.885]	[2.054, 2.109]
Gemini 2.0 Flash	Spearman’s ρ	0.229	0.172	-0.044	0.06	0.138	-0.043
	Adjusted P -value	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
	95% CI of ρ	[0.222, 0.237]	[0.164, 0.179]	[-0.052, -0.036]	[0.052, 0.067]	[0.130, 0.145]	[-0.051, -0.035]
	MSE	1.534	1.681	2.081	1.898	1.792	2.078
Gemini 2.5 Flash	95% CI of MSE	[1.518, 1.550]	[1.662, 1.700]	[2.059, 2.102]	[1.873, 1.923]	[1.760, 1.825]	[2.055, 2.099]
	Spearman’s ρ	0.247	0.085	0.075	0.06	0.095	-0.025
	Adjusted P -value	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
	95% CI of ρ	[0.240, 0.255]	[0.078, 0.092]	[0.067, 0.083]	[0.052, 0.068]	[0.088, 0.102]	[-0.032, -0.017]
DeepSeek-V3	MSE	1.521	1.891	1.84	1.902	1.827	2.008
	95% CI of MSE	[1.504, 1.536]	[1.869, 1.912]	[1.817, 1.865]	[1.879, 1.926]	[1.802, 1.850]	[1.981, 2.035]
	Spearman’s ρ	0.212	0.054	0.033	0.023	0.062	-0.005
	Adjusted P -value	<0.001	<0.001	<0.001	<0.001	<0.001	0.199
DeepSeek-V3 (Flash)	95% CI of ρ	[0.205, 0.220]	[0.046, 0.061]	[0.026, 0.041]	[0.016, 0.030]	[0.054, 0.070]	[-0.013, 0.003]
	MSE	1.603	1.914	1.919	1.988	1.898	1.996
	95% CI of MSE	[1.585, 1.621]	[1.894, 1.935]	[1.893, 1.948]	[1.933, 2.038]	[1.869, 1.928]	[1.970, 2.021]
	Spearman’s ρ	0.213	0.09	0.062	0.098	0.086	0.038
DeepSeek-V3 (Flash)	Adjusted P -value	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
	95% CI of ρ	[0.206, 0.220]	[0.082, 0.097]	[0.055, 0.070]	[0.091, 0.106]	[0.079, 0.094]	[0.030, 0.046]
	MSE	1.574	1.836	1.918	1.785	1.811	1.935
	95% CI of MSE	[1.558, 1.590]	[1.818, 1.856]	[1.889, 1.946]	[1.762, 1.810]	[1.784, 1.842]	[1.910, 1.960]

Notes: The individual-level correlations and MSEs between LLM-predicted and GPS-surveyed preferences are computed by pooling all individual observations across 74 countries ($n = 67,201$). The P -values are adjusted by Benjamini–Hochberg FDR method and 95% CIs are bootstrapped.

Table B.5. Country-level Spearman’s correlations and MSEs

		Risk-taking	Patience	Neg. recip	Pos. recip	Altruism	Trust
GPT-3.5-turbo	Spearman’s ρ	0.311	-0.193	0.285	0.109	0.04	0.267
	Adjusted P -value	0.034	0.262	0.059	0.535	0.854	0.082
	95% CI of ρ	[0.090, 0.508]	[-0.386, 0.019]	[0.045, 0.495]	[-0.134, 0.334]	[-0.204, 0.291]	[0.029, 0.502]
	MSE	1.417	2.317	1.63	1.877	1.7	1.459
GPT-4o-mini	95% CI of MSE	[0.932, 1.984]	[1.753, 2.994]	[1.012, 2.445]	[1.195, 2.705]	[1.085, 2.454]	[0.951, 2.034]
	Spearman’s ρ	0.299	-0.047	0.051	0.168	0.048	0.233
	Adjusted P -value	0.044	0.829	0.824	0.324	0.829	0.143
	95% CI of ρ	[0.064, 0.496]	[-0.267, 0.198]	[-0.185, 0.256]	[-0.088, 0.396]	[-0.186, 0.286]	[-0.003, 0.440]
GPT-4o-mini (English)	MSE	1.481	2.036	1.774	1.557	1.841	1.697
	95% CI of MSE	[0.940, 2.130]	[1.470, 2.647]	[1.213, 2.537]	[1.031, 2.155]	[1.104, 2.678]	[0.982, 2.629]
	Spearman’s ρ	0.361	0.163	0.061	0.155	0.323	0.081
	Adjusted P -value	0.019	0.342	0.775	0.357	0.03	0.681
GPT-4o	95% CI of ρ	[0.132, 0.548]	[-0.092, 0.394]	[-0.162, 0.289]	[-0.062, 0.363]	[0.101, 0.533]	[-0.157, 0.307]
	MSE	1.321	1.618	1.934	1.814	1.424	1.904
	95% CI of MSE	[0.846, 1.838]	[1.093, 2.193]	[1.320, 2.599]	[1.197, 2.540]	[0.927, 2.105]	[1.323, 2.578]
	Spearman’s ρ	0.346	0.331	-0.15	0.189	0.066	0.019
GPT-4.1-mini	Adjusted P -value	0.026	0.029	0.366	0.266	0.771	0.916
	95% CI of ρ	[0.118, 0.542]	[0.107, 0.521]	[-0.379, 0.081]	[-0.034, 0.411]	[-0.149, 0.298]	[-0.205, 0.246]
	MSE	1.164	1.266	2.244	1.698	1.945	1.896
	95% CI of MSE	[0.811, 1.629]	[0.942, 1.626]	[1.628, 2.919]	[1.071, 2.376]	[1.333, 2.590]	[1.345, 2.479]
GPT-5.4-mini	Spearman’s ρ	0.432	0.124	0.14	0.331	0.15	0.169
	Adjusted P -value	0.004	0.468	0.402	0.029	0.366	0.324
	95% CI of ρ	[0.235, 0.606]	[-0.128, 0.360]	[-0.100, 0.349]	[0.112, 0.528]	[-0.086, 0.387]	[-0.066, 0.399]
	MSE	1.143	1.822	1.685	1.341	1.796	1.637
GPT-5.4-mini (English)	95% CI of MSE	[0.761, 1.575]	[1.257, 2.548]	[1.221, 2.215]	[0.874, 1.917]	[1.115, 2.513]	[1.040, 2.415]
	Spearman’s ρ	0.26	0.01	-0.016	0.209	-0.134	0.326
	Adjusted P -value	0.09	0.934	0.916	0.205	0.424	0.03
	95% CI of ρ	[0.004, 0.483]	[-0.229, 0.245]	[-0.241, 0.198]	[-0.045, 0.434]	[-0.373, 0.111]	[0.100, 0.522]
GPT-5.4	MSE	1.608	1.75	2.139	1.486	2.21	1.46
	95% CI of MSE	[1.027, 2.244]	[1.318, 2.260]	[1.354, 3.220]	[0.994, 2.019]	[1.325, 3.179]	[1.000, 1.950]
	Spearman’s ρ	0.27	0.054	-0.156	0.241	0.313	0.093
	Adjusted P -value	0.08	0.816	0.357	0.127	0.034	0.618
GPT-5.4 (English)	95% CI of ρ	[0.058, 0.462]	[-0.166, 0.271]	[-0.366, 0.055]	[0.018, 0.442]	[0.105, 0.507]	[-0.146, 0.346]
	MSE	1.523	1.82	2.175	1.575	1.371	2.204
	95% CI of MSE	[1.096, 1.980]	[1.341, 2.360]	[1.612, 2.785]	[1.062, 2.118]	[0.954, 1.822]	[1.275, 3.219]
	Spearman’s ρ	0.249	0.363	-0.109	0.159	0.183	-0.02
GPT-5.4 (English)	Adjusted P -value	0.111	0.019	0.535	0.351	0.285	0.916
	95% CI of ρ	[0.051, 0.447]	[0.135, 0.574]	[-0.359, 0.117]	[-0.096, 0.363]	[-0.073, 0.406]	[-0.237, 0.197]
	MSE	1.536	1.247	2.305	1.696	1.708	2.037
	95% CI of MSE	[1.062, 2.032]	[0.882, 1.668]	[1.414, 3.631]	[1.182, 2.351]	[1.112, 2.417]	[1.389, 2.806]
GPT-5.4 (English)	Spearman’s ρ	0.332	0.492	-0.192	0.213	0.32	-0.175
	Adjusted P -value	0.029	<0.001	0.262	0.204	0.031	0.314
	95% CI of ρ	[0.112, 0.528]	[0.261, 0.657]	[-0.408, 0.049]	[-0.029, 0.457]	[0.097, 0.522]	[-0.417, 0.058]
	MSE	1.533	0.906	2.321	1.708	1.391	2.271
Gemini 2.0 Flash	95% CI of MSE	[0.988, 2.067]	[0.638, 1.220]	[1.732, 2.943]	[1.042, 2.549]	[0.941, 1.888]	[1.664, 2.956]
	Spearman’s ρ	0.412	0.017	0.077	0.147	-0.023	0.132
	Adjusted P -value	0.006	0.916	0.699	0.369	0.916	0.43
	95% CI of ρ	[0.174, 0.630]	[-0.218, 0.253]	[-0.183, 0.321]	[-0.105, 0.375]	[-0.273, 0.228]	[-0.097, 0.374]
Gemini 2.5 Flash	MSE	1.191	1.909	1.762	1.755	1.867	1.575
	95% CI of MSE	[0.810, 1.635]	[1.270, 2.652]	[1.181, 2.456]	[1.087, 2.618]	[1.277, 2.580]	[1.107, 2.116]
	Spearman’s ρ	0.172	-0.086	0.037	0.031	-0.044	0.211
	Adjusted P -value	0.323	0.658	0.861	0.877	0.837	0.204
DeepSeek-V3	95% CI of ρ	[-0.074, 0.391]	[-0.343, 0.167]	[-0.186, 0.243]	[-0.220, 0.267]	[-0.269, 0.181]	[-0.030, 0.453]
	MSE	1.637	2.205	1.976	1.807	1.984	1.573
	95% CI of MSE	[1.017, 2.276]	[1.520, 2.941]	[1.452, 2.605]	[1.169, 2.525]	[1.386, 2.643]	[1.055, 2.115]
	Spearman’s ρ	0.109	-0.032	0.101	0.399	0.012	0.062
DeepSeek-V3	Adjusted P -value	0.535	0.877	0.578	0.008	0.93	0.775
	95% CI of ρ	[-0.142, 0.339]	[-0.253, 0.207]	[-0.134, 0.335]	[0.170, 0.581]	[-0.229, 0.261]	[-0.184, 0.313]
	MSE	1.774	1.98	2.007	1.152	1.525	1.914
	95% CI of MSE	[1.197, 2.460]	[1.434, 2.563]	[1.156, 3.079]	[0.761, 1.664]	[1.056, 2.033]	[1.132, 3.028]

Notes: Country-level averages are weighted by the GPS ex-post sampling weights. Before computing the MSEs, both the individual- and country-level preference indices are normalized to mean 0 and standard deviation 1. The country-level correlations and MSEs between LLM-predicted and GPS-surveyed preferences are computed from 74 countries ($n = 74$). The P -values are adjusted by Benjamini–Hochberg FDR method and 95% CIs are bootstrapped.

Section C. Between-model consistency on LLM-predicted preferences

We assess whether different LLM configurations produce similar predictions using two complementary metrics: Spearman's rank correlations and mean squared errors (MSEs). For each pair of models, we compute the correlation (or MSE) between their predicted preferences, separately at the individual and country levels.

[Figure C.1](#) shows the between-model correlation matrices for LLM-predicted preferences at the individual level. At the individual level, LLM-predicted preferences are positively and substantially correlated across models (384 of 396 correlations are positive; mean correlation across all preferences: 0.346; risk-taking: 0.676; patience: 0.510; negative reciprocity: 0.179; positive reciprocity: 0.219; altruism: 0.403; trust: 0.092). The country-level between-model correlations in [Figure C.2](#) show a similar pattern (365 of 396 are positive; mean correlation across all preferences: 0.262; risk-taking: 0.293; patience: 0.228; negative reciprocity: 0.218; positive reciprocity: 0.269; altruism: 0.363; trust: 0.201). The complete pairwise correlation matrices, together with 95% CIs and Benjamini–Hochberg FDR-adjusted P values, are provided in [online Supplementary Data 2](#). 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). The between-model correlations are higher than GPS-LLMs correlations (two-sided unpaired t -test; individual-level: 0.346 vs 0.094, $t(326.38) = 17.726$, $P < 0.001$, Cohen's $d = 1.224$, 95% CI = [0.22, 0.28]; country-level: 0.262 vs 0.137, $t(105.74) = 5.822$, $P < 0.001$, Cohen's $d = 0.692$, 95% CI = [0.08, 0.17]).

We then complement the correlation analysis with mean squared error (MSE) comparisons. [Figure 3](#) in main text reports the individual-level between-model MSEs, with a mean of 1.377. Across preference domains, MSEs are generally lowest for risk-taking (average MSEs across all models: risk-taking: 0.770; patience: 1.049; negative reciprocity: 1.679; positive reciprocity: 1.579; altruism: 1.329; trust: 1.856). [Figure C.3](#) reports the country-level between-model MSEs, with a mean of 1.491 (risk-taking: 1.423; patience:

1.500; negative reciprocity: 1.628; positive reciprocity: 1.485; altruism: 1.199; trust: 1.709). The between-model MSEs are lower than GPS-LLM MSEs (two-sided unpaired t -test; individual-level: 1.377 vs 1.827, $t(313.31) = -16.569$, $P < 0.001$, Cohen's $d = -1.158$, 95% CI = [-0.50, -0.40]; country-level: 1.491 vs 1.722, $t(115.62) = -5.468$, $P < 0.001$, Cohen's $d = -0.600$, 95% CI = [-0.32, -0.15]). Together with the correlation results, this indicates that LLMs make broadly similar predictions to one another at both levels.

Both individual-level and country-level comparison between GPS-LLMs correlations and between-model correlations for each preference are reported in [Table C.1](#), and the comparison between MSEs is reported in [Table C.2](#). These tables show a consistent pattern: at the individual level, between-model correlations are larger (and MSEs smaller) than GPS-LLM correlations (MSEs) for every preference domain, while at the country level the differences are not significant for some domains.

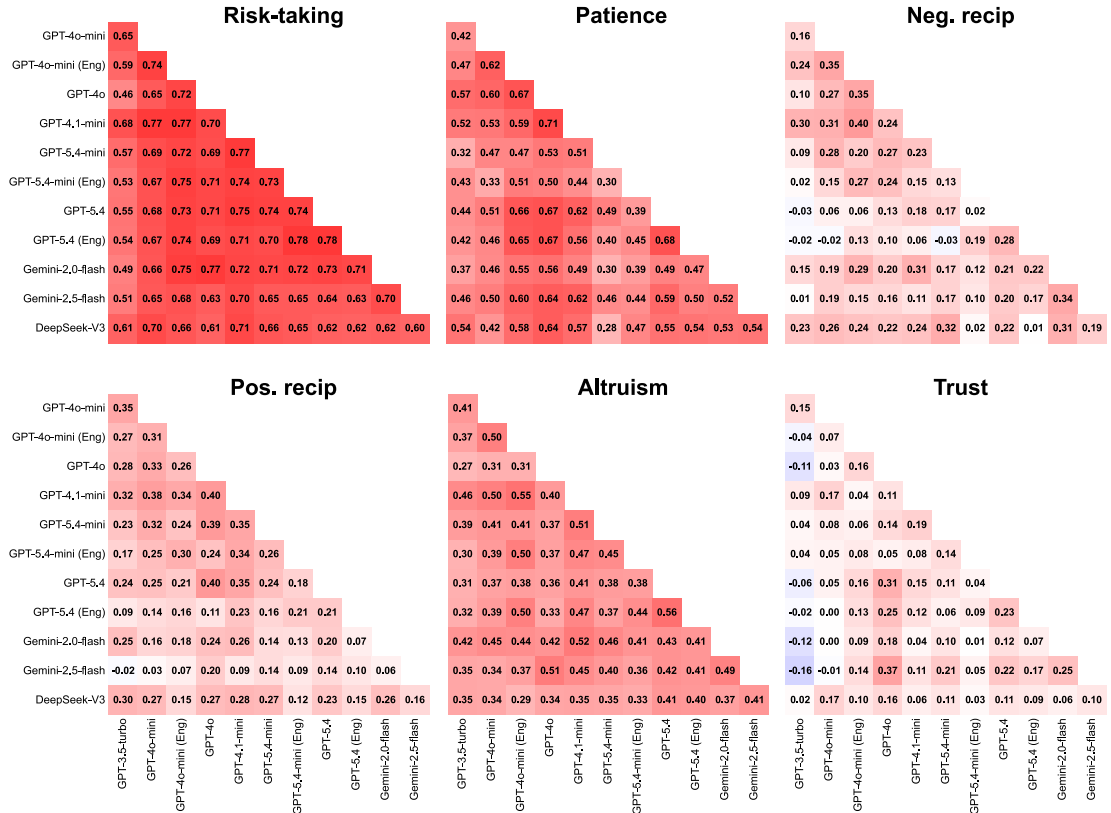


Figure C.1. Individual-level between-model Spearman's correlation coefficients. Each cell reports the Spearman correlation between two LLMs, computed by pooling all individuals from all countries. LLM-predicted preferences are broadly and positively correlated across models (risk-

taking: 66; patience: 66; negative reciprocity: 62; positive reciprocity: 65; altruism: 66; trust: 59; out of 66 pairs of LLMs). In the heatmap, blue indicates negative values and red indicates positive values, with darker shades representing larger absolute values.

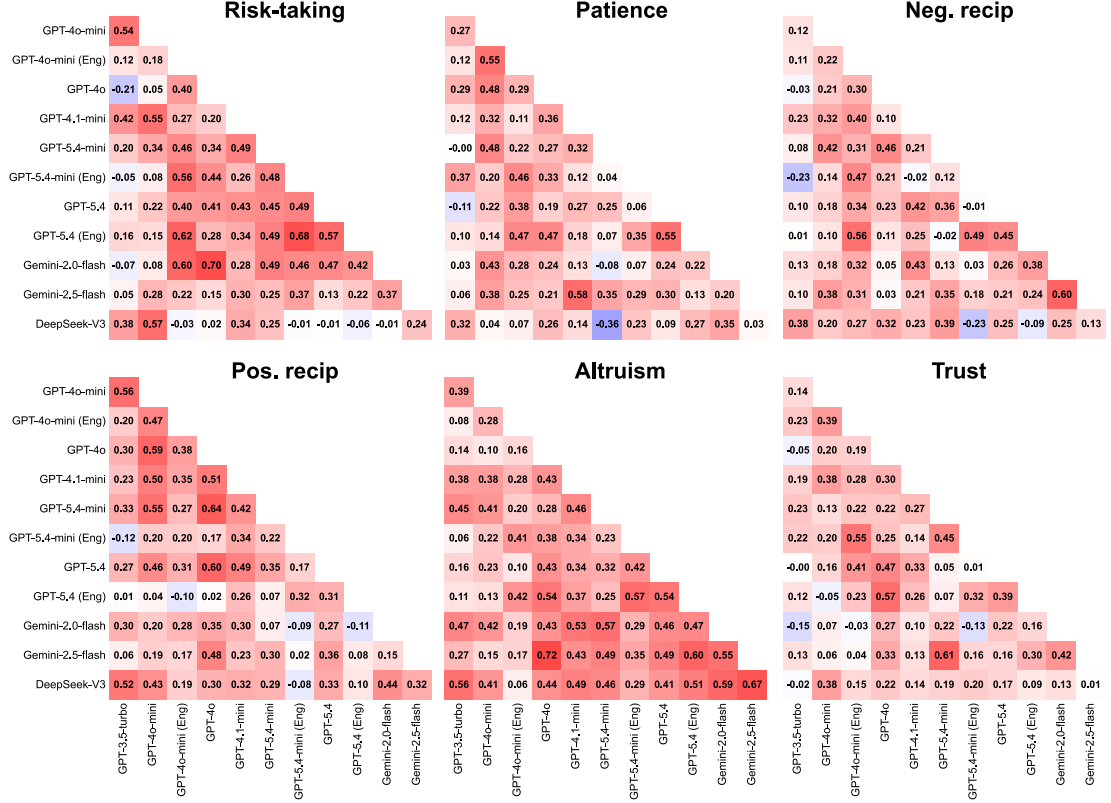


Figure C.2. Country-level between-model Spearman's correlation coefficients. Each cell reports the Spearman correlation between two LLMs, computed using country-level averages of LLM-predicted preferences (weighted by GPS ex-post sampling weights). LLM-predicted preferences are broadly and positively correlated across models (risk-taking: 58; patience: 62; negative reciprocity: 59; positive reciprocity: 61; altruism: 66; trust: 59). In the heatmap, blue indicates negative values and red indicates positive values, with darker shades representing larger absolute values.

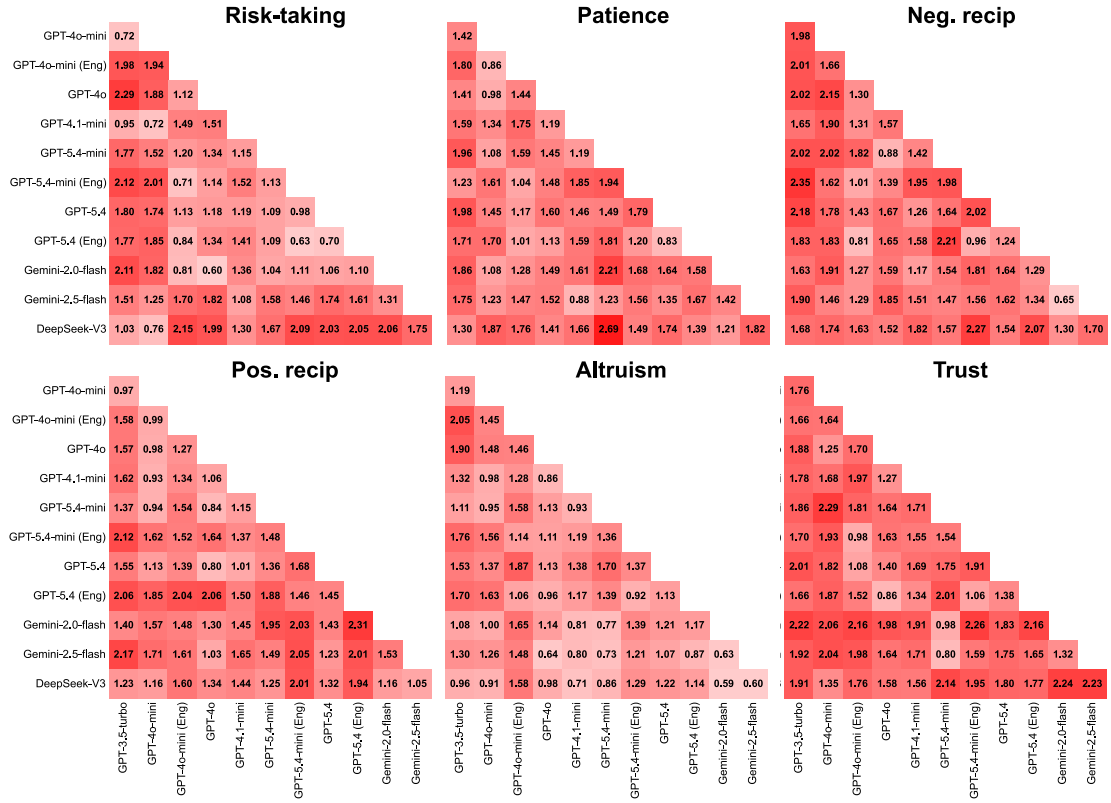


Figure C.3. Country-level between-model MSEs. Each cell reports the MSE between two LLM configurations, computed using country-level averages of LLM-predicted preferences (weighted by GPS ex-post sampling weights). Country-level preference indices are standardized to mean 0 and standard deviation 1 prior to MSE computation. In the heatmap, darker shades represent larger MSEs, with lower values indicating greater alignment between models.

Table C.1. Comparison between GPS-LLM and between-model Spearman correlations

Preference	GPS-LLM correlations	Between-model correlations	<i>t</i> -value	Degree of freedom	P-value	Cohen's d	95% CI
Individual-level							
Risk-taking	0.227	0.676	43.869	70.597	<0.001	6.674	[0.43, 0.47]
Patience	0.097	0.510	23.593	36.667	<0.001	4.412	[0.38, 0.45]
Negative reciprocity	0.054	0.179	6.208	28.837	<0.001	1.269	[0.08, 0.17]
Positive reciprocity	0.072	0.219	10.201	60.564	<0.001	1.637	[0.12, 0.17]
Altruism	0.093	0.403	28.674	42.503	<0.001	5.114	[0.29, 0.33]
Trust	0.019	0.092	3.934	27.725	0.001	0.818	[0.03, 0.11]
Country-level							
Risk-taking	0.296	0.293	-0.086	35.087	0.932	-0.016	[-0.08, 0.07]
Patience	0.100	0.228	2.056	13.902	0.059	0.734	[-0.01, 0.26]
Negative reciprocity	0.011	0.218	4.519	17.469	<0.001	1.239	[0.11, 0.30]
Positive reciprocity	0.196	0.269	2.030	28.204	0.052	0.419	[-0.00, 0.15]
Altruism	0.104	0.363	5.347	15.645	<0.001	1.633	[0.16, 0.36]
Trust	0.116	0.201	1.895	16.682	0.076	0.542	[-0.01, 0.18]

Notes: For each preference domain, the correlations for each pair of GPS-LLM and correlations for each pair of models are computed separately. The P-values are calculated from an unpaired *t*-test by pooling all correlations across 12 models.

Table C.2. Comparison between GPS-LLM and between-model MSEs

Preference	GPS-LLM MSEs	Between-model MSEs	t-value	Degree of freedom	P-value	Cohen's d	95% CI
Individual-level							
Risk-taking	1.567	0.770	-34.526	70.178	<0.001	-5.266	[-0.84, -0.75]
Patience	1.832	1.049	-20.744	44.693	<0.001	-3.643	[-0.86, -0.71]
Negative reciprocity	1.897	1.679	-5.241	27.005	<0.001	-1.102	[-0.30, -0.13]
Positive reciprocity	1.859	1.579	-9.160	48.575	<0.001	-1.57	[-0.34, -0.22]
Altruism	1.847	1.329	-20.524	43.881	<0.001	-3.625	[-0.57, -0.47]
Trust	1.963	1.856	-3.276	27.317	0.003	-0.686	[-0.17, -0.04]
Country-level							
Risk-taking	1.444	1.423	-0.264	35.350	0.794	-0.050	[-0.18, 0.14]
Patience	1.740	1.500	-1.887	13.642	0.081	-0.695	[-0.51, 0.03]
Negative reciprocity	1.996	1.628	-4.441	20.610	<0.001	-1.078	[-0.54, -0.20]
Positive reciprocity	1.622	1.485	-1.811	24.528	0.082	-0.398	[-0.29, 0.02]
Altruism	1.730	1.199	-6.166	18.120	<0.001	-1.641	[-0.71, -0.35]
Trust	1.802	1.709	-1.025	17.811	0.319	-0.277	[-0.28, 0.10]

Notes: For each dimension, the MSEs for each pair of GPS-LLM and MSEs for each pair of models are computed separately. All MSEs are computed after standardizing each preference index to mean 0 and SD 1 within the relevant level (individual or country). The P-values are calculated from an unpaired *t*-test by pooling all MSEs across 12 models.

Section D. Alignment at the demographic and cultural levels

This section provides additional details on the alignment between LLM-predicted and GPS-surveyed preferences at the demographic and cultural levels, including the consistency-ratio results reported in the main text, a sign-based match ratio as a robustness check, and between-model comparisons.

D.1 Consistency ratios for demographic and cultural alignment

Our primary alignment metric, reported in the main text, is the consistency ratio (see [Methods](#) for the definition). 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.

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.

[Table D.1](#) reports the consistency ratios across demographic and cultural dimensions. The overall demographic-level consistency ratio is 77.8% (two-sided binomial test; proportion = 168/216, $P < 0.001$, 95% CI = [71.6%, 83.1%]; 81.9% (59/72) for gender, 66.7% (48/72) for age, and 84.7% (61/72) for cognitive ability). The overall cultural-level consistency ratio is 48.1% (two-sided binomial test: proportion = 520/1080, $P = 0.235$, 95% CI = [45.1%, 51.2%]), which is significantly lower than the demographic-level consistency ratio (chi-square test: $\chi^2(1) = 62.27$, $P < 0.001$, Cramer’s $V = 0.22$, 95% CI = [-35.5%, -22.9%]).

D.2 Sign-based match ratios as a robustness check

As a robustness check for the consistency-ratio results in D.1, we construct a sign-based

match ratio using 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 D.2](#) reports the distribution of effect sizes across the four quadrants.

The overall demographic-level match ratio (Quadrants I and III combined) is 78.7% (two-sided binomial test; proportion = 170/216, $P < 0.001$, 95% CI = [72.6%, 84.0%]; gender: 83.3%; age: 68.1%; cognitive ability: 84.7%), whereas the cultural-level match ratio is 56.6% (two-sided binomial test; proportion = 611/1080, $P < 0.001$, 95% CI = [53.6%, 59.6%]), ranging from 37.5% to 70.8%. The cultural-level match ratio is significantly lower than the demographic-level ratio (chi-square test: $\chi^2(1) = 35.89$, $P < 0.001$, Cramer's $V = 0.17$, 95% CI = [-27.9%, -15.5%]).

We also examine whether effect sizes estimated from GPS-surveyed data are correlated with those estimated from LLM-predicted data. [Table D.3](#) reports the Spearman rank correlations, corresponding to [Figure 4](#) and [Figure 5](#) in the main text. At the demographic level, the effect sizes estimated from LLM-predicted preferences increase as those estimated from GPS (i.e., positive Spearman correlations), except for trust. At the cultural level, the coefficients are less than those at the demographic level (0.169 vs 0.501), indicating that LLMs are more sensitive in demographic dimensions, compared to the cultural dimensions.

We further check, within each dimension, the correlation between effect sizes estimated from GPS-surveyed and LLM-predicted data. As [Table D.4](#) reports, at the demographic level, correlations are large and positive for gender ($\rho = 0.898$, $P < 0.001$) and age ($\rho = 0.830$, $P < 0.001$), whereas the correlation for cognitive ability is weakly negative and not significantly different from zero ($\rho = -0.165$, $P = 0.166$). For the cultural dimensions, correlations range from -0.351 (Schwartz mastery) to 0.458 (WVS Tradition

values). Several dimensions show statistically significant positive associations, including WVS Tradition values, Schwartz Embeddedness, Schwartz Hierarchy, Schwartz Affective Autonomy, Hofstede PDI, UAI, and LTO (all $P < 0.05$).

Beyond the dimension-level Spearman correlations reported above, [Table D.5](#) disaggregates the match ratios introduced in [Table D.2](#) by preference domain, providing a finer-grained view across the six preference types.

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. Although the sign-based match ratio differs from the CI-based consistency ratio reported in the main text, both metrics lead to the same conclusion: alignment at the cultural level is markedly weaker than at the demographic level.

Although LLM-simulated responses recover the direction of most demographic differences, they do so with systematically inflated magnitudes ³ (i.e., the effect sizes estimated from LLM-predicted preferences are larger than their survey-based counterparts). This is especially the case for risk-taking and patience (Figure 4, above the 45° line; risk-taking: 34 out of 36; patience: 30 out of 36). This amplification is consistent with concerns that LLMs over-rely on stereotyped associations from their training data, producing overstating rather than calibrated representations of demographic groups ^{4,5}.

Table D.1. Consistency ratios for demographic and cultural gradients between LLM-predicted and GPS-surveyed preferences

	Risk-taking	Patience	Negative reciprocity	Positive reciprocity	Altruism	Trust	Total consistent cases	Consistency ratio (%)
Demographic level								
Gender	12	12	11	2	11	11	59	81.9
Cognitive ability	12	12	9	11	12	5	61	84.7
Age	12	12	12	1	3	8	48	66.7
Cultural level								
WVS Tradition values	12	6	3	6	12	6	45	62.5
WVS Survival values	9	9	0	4	8	3	33	45.8
Schwartz Harmony	9	5	0	5	8	4	31	43.1
Schwartz Embeddedness	12	9	2	9	0	8	40	55.6
Schwartz Hierarchy	10	10	9	2	5	7	43	59.7
Schwartz Mastery	6	3	8	1	3	2	23	31.9
Schwartz Affective Autonomy	9	9	0	7	9	7	41	56.9
Schwartz Intellectual Autonomy	8	9	2	1	8	4	32	44.4
Schwartz Egalitarianism	5	7	7	1	4	0	24	33.3
Hofstede PDI	10	10	4	6	7	7	44	61.1
Hofstede IDV	0	7	3	5	9	0	24	33.3
Hofstede MAS	10	6	8	0	4	8	36	50.0
Hofstede UAI	1	8	1	8	7	8	33	45.8
Hofstede LTO	12	3	2	5	12	6	40	55.6
Hofstede IVR	8	9	1	1	3	9	31	43.1

Notes: For each demographic or cultural dimension, we construct a binary grouping variable and compute effect sizes (Cohen’s d) for each LLM–preference combination, separately for the GPS data and for each LLM. 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 have the same sign. Each row pools 72 LLM–preference combinations (6 preferences $\times$ 12 LLMs). The last two columns report, for each dimension, the total number of consistent cases and the corresponding consistency ratio (in %).

Table D.2. 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 Quadrants I and III
Demographic level						
Gender	35	11	25	1	60	83.3
Cognitive ability	61	0	0	11	61	84.7
Age	40	3	9	20	49	68.1
Cultural level						
WVS Tradition values	23	0	24	25	47	65.3
WVS Survival values	19	13	23	17	42	58.3
Schwartz Harmony	5	20	40	7	45	62.5
Schwartz Embeddedness	12	31	29	0	41	56.9
Schwartz Hierarchy	27	6	18	21	45	62.5
Schwartz Mastery	28	11	1	32	29	40.3
Schwartz Affective Autonomy	23	8	28	13	51	70.8
Schwartz Intellectual Autonomy	12	24	24	12	36	50.0
Schwartz Egalitarianism	13	34	14	11	27	37.5
Hofstede PDI	7	23	37	5	44	61.1
Hofstede IDV	10	13	35	14	45	62.5
Hofstede MAS	10	18	30	14	40	55.6
Hofstede UAI	9	25	35	3	44	61.1
Hofstede LTO	5	13	35	19	40	55.6
Hofstede IVR	31	20	4	17	35	48.6

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 twelve 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).

Table D.3. Spearman correlations between GPS- and LLM-based effect sizes, computed within each preference domain (pooling dimensions and LLMs)

Preference	Spearman's ρ	P -value
Demographic level		
Risk-taking	0.426	0.010
Patience	0.812	<0.001
Negative reciprocity	0.580	<0.001
Positive reciprocity	0.432	0.008
Altruism	0.835	<0.001
Trust	-0.079	0.649
Cultural level		
Risk-taking	0.446	<0.001
Patience	0.380	<0.001
Negative reciprocity	-0.129	0.083
Positive reciprocity	-0.095	0.202
Altruism	0.270	<0.001
Trust	0.141	0.058

Notes: For each demographic or cultural dimension, we construct a binary grouping variable and compute effect sizes (Cohen's d) for each LLM–preference combination, separately for the GPS data and for each LLM. For each preference at the demographic level (top six rows), we then pool effect sizes across the 3 demographic dimensions and 12 LLMs (36 combinations) and compute Spearman's ρ between the vector of GPS-based and the corresponding vector of LLM-based effect sizes. For each preference at the cultural level (bottom six rows), we pool effect sizes across the 15 cultural dimensions and 12 LLMs (180 combinations).

Table D.4. Spearman correlations between GPS- and LLM-based effect sizes, computed within each demographic/cultural dimension (pooling preferences and LLMs)

Dimension	Spearman's ρ	P-value
Demographic level		
Gender	0.898	<0.001
Cognitive ability	-0.165	0.166
Age	0.830	<0.001
Cultural level		
WVS Tradition values	0.458	<0.001
WVS Survival values	0.190	0.110
Schwartz Harmony	0.005	0.966
Schwartz Embeddedness	0.286	0.015
Schwartz Hierarchy	0.261	0.027
Schwartz Mastery	-0.351	0.003
Schwartz Affective Autonomy	0.427	<0.001
Schwartz Intellectual Autonomy	0.119	0.320
Schwartz Egalitarianism	-0.113	0.346
Hofstede PDI	0.370	0.001
Hofstede IDV	0.165	0.167
Hofstede MAS	0.174	0.143
Hofstede UAI	0.386	0.001
Hofstede LTO	0.321	0.006
Hofstede IVR	-0.140	0.241

Notes: For each demographic or cultural dimension, we construct a binary grouping variable and compute effect sizes (Cohen's d) for each LLM-preference combination, separately for the GPS data and for each LLM. For each dimension, we then pool effect sizes across the 6 preference domains and 12 LLMs (72 combinations) and compute Spearman's ρ between the vector of GPS-based and the corresponding vector of LLM-based effect sizes.

Table D.5. Match ratios for demographic and cultural gradients between LLM-predicted and GPS-surveyed preferences

	Risk-taking	Patience	Negative reciprocity	Positive reciprocity	Altruism	Trust	Total matched cases	Match ratio (%)
Demographic level								
Gender	12	12	11	2	12	11	60	83.3
Cognitive ability	12	12	9	11	12	5	61	84.7
Age	12	12	12	1	3	9	49	68.1
Cultural level								
WVS Tradition values	12	7	3	6	12	7	47	65.3
WVS Survival values	10	9	2	4	9	8	42	58.3
Schwartz Harmony	10	5	10	6	9	5	45	62.5
Schwartz Embeddedness	12	9	2	9	1	8	41	56.9
Schwartz Hierarchy	10	10	9	2	6	8	45	62.5
Schwartz Mastery	7	4	8	1	3	6	29	40.3
Schwartz Affective Autonomy	10	9	9	7	9	7	51	70.8
Schwartz Intellectual Autonomy	8	10	2	2	8	6	36	50.0
Schwartz Egalitarianism	6	7	7	1	4	2	27	37.5
Hofstede PDI	10	10	4	6	7	7	44	61.1
Hofstede IDV	11	7	3	6	10	8	45	62.5
Hofstede MAS	11	6	8	1	4	10	40	55.6
Hofstede UAI	3	8	9	8	8	8	44	61.1
Hofstede LTO	12	3	2	5	12	6	40	55.6
Hofstede IVR	8	9	5	1	3	9	35	48.6

Notes: For each demographic or cultural dimension, we construct a binary grouping variable and compute effect sizes (Cohen's d) for each LLM–preference combination, separately for the GPS data and for each LLM. A case is classified as “match” if the GPS- and LLM-based effect sizes have the same sign. Each row pools 72 LLM–preference combinations (6 preferences $\times$ 12 LLMs). The last two columns report, for each dimension, the total number of matched cases and the corresponding match ratio (in %).

D.3 Between-model alignment in demographic and cultural effects

Here we assess between-model alignment 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 D.1](#) summarizes the resulting correlation matrices. The full set of between-model Spearman correlations of effect sizes, together with 95% CIs and Benjamini–Hochberg FDR-adjusted P values, is reported in [Table D.6](#).

All 66 demographic-level correlations are positive, with an average correlation of 0.821. At the cultural level, 59 of 66 correlations are positive, and the average correlation is 0.298. 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.

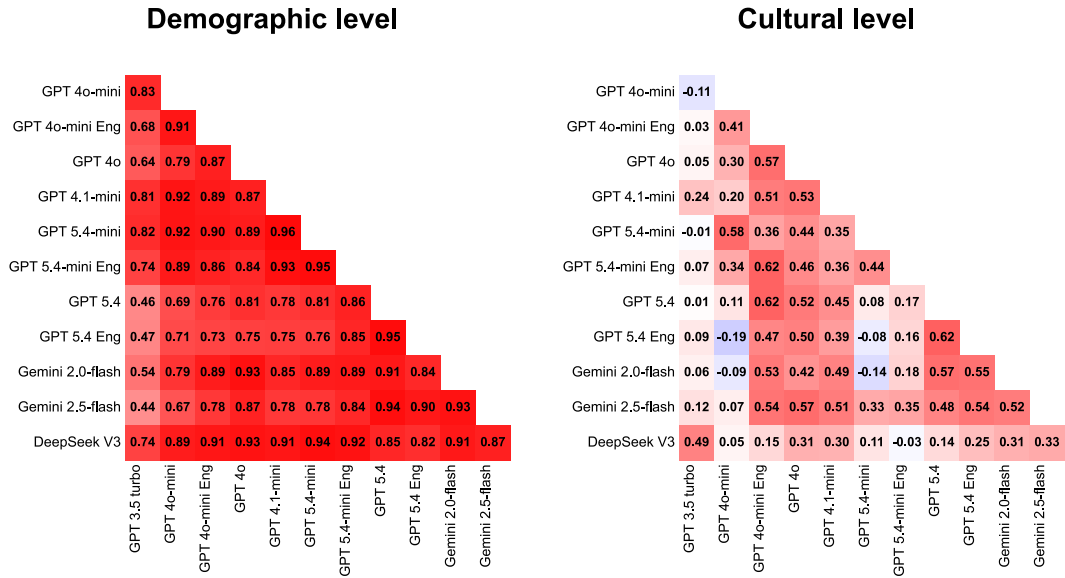


Figure D.1. Between-model alignment in demographic- and cultural-level effects. (A) Demographic-level between-model alignment. Each cell reports Spearman’s ρ between effect sizes (Cohen’s d) estimated from a pair of LLMs, pooling all demographic dimensions and all preference domains. **(B) Cultural-level between-model alignment.** Each cell reports Spearman’s ρ between effect sizes estimated from a pair of LLMs, pooling across all cultural dimensions and all preference domains. In the heatmap, blue indicates negative values and

red indicates positive values, with darker shades representing larger absolute values. Details are reported in [Table D.6](#).

Table D.6. Between-model alignment in demographic- and cultural-level effects.

Model A	Model B	Cultural level			Demographic level		
		rho	P	Adj. P	rho	P	Adj. P
GPT-3.5-turbo	GPT-4o-mini	-0.106	0.320	0.398	0.835	<0.001	<0.001
GPT-3.5-turbo	GPT-4o-mini (Eng)	0.027	0.799	0.824	0.676	0.002	0.002
GPT-3.5-turbo	GPT-4o	0.046	0.664	0.707	0.641	0.004	0.004
GPT-3.5-turbo	GPT-4.1-mini	0.240	0.023	0.036	0.808	<0.001	<0.001
GPT-3.5-turbo	GPT-5.4-mini	-0.015	0.891	0.904	0.822	<0.001	<0.001
GPT-3.5-turbo	GPT-5.4-mini (Eng)	0.070	0.514	0.585	0.738	<0.001	<0.001
GPT-3.5-turbo	GPT-5.4	0.009	0.935	0.935	0.461	0.054	0.055
GPT-3.5-turbo	GPT-5.4 (Eng)	0.088	0.409	0.490	0.470	0.049	0.051
GPT-3.5-turbo	Gemini-2.0-flash	0.059	0.581	0.639	0.544	0.020	0.021
GPT-3.5-turbo	Gemini-2.5-flash	0.119	0.265	0.350	0.441	0.067	0.067
GPT-3.5-turbo	DeepSeek-V3	0.488	<0.001	<0.001	0.736	<0.001	<0.001
GPT-4o-mini	GPT-4o-mini (Eng)	0.405	<0.001	<0.001	0.907	<0.001	<0.001
GPT-4o-mini	GPT-4o	0.297	0.004	0.007	0.789	<0.001	<0.001
GPT-4o-mini	GPT-4.1-mini	0.197	0.062	0.098	0.915	<0.001	<0.001
GPT-4o-mini	GPT-5.4-mini	0.584	<0.001	<0.001	0.920	<0.001	<0.001
GPT-4o-mini	GPT-5.4-mini (Eng)	0.341	0.001	0.002	0.893	<0.001	<0.001
GPT-4o-mini	GPT-5.4	0.109	0.305	0.390	0.692	0.001	0.002
GPT-4o-mini	GPT-5.4 (Eng)	-0.195	0.066	0.101	0.707	0.001	0.001
GPT-4o-mini	Gemini-2.0-flash	-0.090	0.400	0.489	0.789	<0.001	<0.001
GPT-4o-mini	Gemini-2.5-flash	0.066	0.539	0.603	0.668	0.002	0.003
GPT-4o-mini	DeepSeek-V3	0.047	0.662	0.707	0.886	<0.001	<0.001
GPT-4o-mini (Eng)	GPT-4o	0.565	<0.001	<0.001	0.868	<0.001	<0.001
GPT-4o-mini (Eng)	GPT-4.1-mini	0.509	<0.001	<0.001	0.895	<0.001	<0.001
GPT-4o-mini (Eng)	GPT-5.4-mini	0.362	<0.001	0.001	0.899	<0.001	<0.001
GPT-4o-mini (Eng)	GPT-5.4-mini (Eng)	0.616	<0.001	<0.001	0.864	<0.001	<0.001
GPT-4o-mini (Eng)	GPT-5.4	0.617	<0.001	<0.001	0.763	<0.001	<0.001
GPT-4o-mini (Eng)	GPT-5.4 (Eng)	0.473	<0.001	<0.001	0.734	<0.001	<0.001
GPT-4o-mini (Eng)	Gemini-2.0-flash	0.530	<0.001	<0.001	0.886	<0.001	<0.001
GPT-4o-mini (Eng)	Gemini-2.5-flash	0.542	<0.001	<0.001	0.777	<0.001	<0.001
GPT-4o-mini (Eng)	DeepSeek-V3	0.153	0.151	0.212	0.907	<0.001	<0.001
GPT-4o	GPT-4.1-mini	0.535	<0.001	<0.001	0.872	<0.001	<0.001
GPT-4o	GPT-5.4-mini	0.440	<0.001	<0.001	0.886	<0.001	<0.001
GPT-4o	GPT-5.4-mini (Eng)	0.458	<0.001	<0.001	0.841	<0.001	<0.001
GPT-4o	GPT-5.4	0.517	<0.001	<0.001	0.812	<0.001	<0.001
GPT-4o	GPT-5.4 (Eng)	0.499	<0.001	<0.001	0.748	<0.001	<0.001
GPT-4o	Gemini-2.0-flash	0.424	<0.001	<0.001	0.932	<0.001	<0.001
GPT-4o	Gemini-2.5-flash	0.572	<0.001	<0.001	0.866	<0.001	<0.001
GPT-4o	DeepSeek-V3	0.306	0.003	0.006	0.928	<0.001	<0.001
GPT-4.1-mini	GPT-5.4-mini	0.351	<0.001	0.001	0.959	<0.001	<0.001
GPT-4.1-mini	GPT-5.4-mini (Eng)	0.361	<0.001	0.001	0.930	<0.001	<0.001
GPT-4.1-mini	GPT-5.4	0.445	<0.001	<0.001	0.781	<0.001	<0.001
GPT-4.1-mini	GPT-5.4 (Eng)	0.386	<0.001	<0.001	0.754	<0.001	<0.001
GPT-4.1-mini	Gemini-2.0-flash	0.490	<0.001	<0.001	0.851	<0.001	<0.001
GPT-4.1-mini	Gemini-2.5-flash	0.511	<0.001	<0.001	0.777	<0.001	<0.001
GPT-4.1-mini	DeepSeek-V3	0.297	0.004	0.007	0.911	<0.001	<0.001
GPT-5.4-mini	GPT-5.4-mini (Eng)	0.444	<0.001	<0.001	0.946	<0.001	<0.001
GPT-5.4-mini	GPT-5.4	0.081	0.449	0.530	0.808	<0.001	<0.001
GPT-5.4-mini	GPT-5.4 (Eng)	-0.076	0.479	0.554	0.761	<0.001	<0.001
GPT-5.4-mini	Gemini-2.0-flash	-0.138	0.196	0.264	0.886	<0.001	<0.001
GPT-5.4-mini	Gemini-2.5-flash	0.329	0.002	0.003	0.783	<0.001	<0.001
GPT-5.4-mini	DeepSeek-V3	0.109	0.308	0.390	0.944	<0.001	<0.001
GPT-5.4-mini (Eng)	GPT-5.4	0.172	0.104	0.153	0.860	<0.001	<0.001
GPT-5.4-mini (Eng)	GPT-5.4 (Eng)	0.160	0.131	0.188	0.847	<0.001	<0.001
GPT-5.4-mini (Eng)	Gemini-2.0-flash	0.178	0.093	0.139	0.893	<0.001	<0.001
GPT-5.4-mini (Eng)	Gemini-2.5-flash	0.352	<0.001	0.001	0.843	<0.001	<0.001
GPT-5.4-mini (Eng)	DeepSeek-V3	-0.028	0.796	0.824	0.924	<0.001	<0.001
GPT-5.4	GPT-5.4 (Eng)	0.623	<0.001	<0.001	0.953	<0.001	<0.001
GPT-5.4	Gemini-2.0-flash	0.569	<0.001	<0.001	0.905	<0.001	<0.001
GPT-5.4	Gemini-2.5-flash	0.482	<0.001	<0.001	0.944	<0.001	<0.001
GPT-5.4	DeepSeek-V3	0.144	0.176	0.243	0.845	<0.001	<0.001
GPT-5.4 (Eng)	Gemini-2.0-flash	0.549	<0.001	<0.001	0.837	<0.001	<0.001
GPT-5.4 (Eng)	Gemini-2.5-flash	0.536	<0.001	<0.001	0.901	<0.001	<0.001
GPT-5.4 (Eng)	DeepSeek-V3	0.254	0.016	0.026	0.818	<0.001	<0.001
Gemini-2.0-flash	Gemini-2.5-flash	0.515	<0.001	<0.001	0.928	<0.001	<0.001
Gemini-2.0-flash	DeepSeek-V3	0.312	0.003	0.005	0.905	<0.001	<0.001
Gemini-2.5-flash	DeepSeek-V3	0.330	0.002	0.003	0.868	<0.001	<0.001

Notes: The Spearman’s ρ between effect sizes (Cohen’s d) estimated from a pair of LLMs, pooling all demographic (or cultural) dimensions and all preference domains. The adj. P columns report the P -value adjusted by Benjamini–Hochberg FDR method.

Section E. 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 construct two country-specific indices to measure how closely LLM-predicted preferences align with GPS-surveyed preferences within each country, and investigate which country-level characteristics explain cross-country heterogeneity in these indices. 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 MSEs between LLM-predicted and GPS-surveyed preferences within that country. Higher consistency and lower inconsistency indicate better country-level alignment between LLM-predicted and GPS-surveyed preferences. At the end, we run a series of diagnostic regressions to decompose the prediction error of LLMs.

E.1 Decomposing individual-level alignment with demographics and country controls

To assess how much demographic and country information contribute to individual-level alignment, we first estimate four OLS regressions for each model $\times$ 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 $\times$ preference combination. The mean OLS coefficients across 12 LLMs for different preference domains under each specification are summarized in [Table E.1](#). The average coefficient is essentially unchanged from 0.087 (baseline) to 0.087 (country dummies only), but falls to 0.020 with demographic controls only, and further to 0.013 when both demographic controls and country fixed effects are included.

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 E.2](#)).

Relative to the baseline specification, adding country controls yields a coefficient that is not statistically significant ($b = 0.000$, $P = 0.961$), whereas including demographic controls leads to a significant reduction in the OLS coefficient ($b = -0.066$, $P < 0.001$). A post hoc comparison shows no statistically significant difference between the demographic-only specification, and the demographic and country specification (Tukey’s HSD-adjusted $P = 0.883$).

These results indicate that demographic variables absorb most of the baseline association between LLM-predicted and GPS-surveyed preferences (approximately 77%, i.e., $0.067/0.087$), whereas adding country fixed effects contributes little additional explanatory power.

E.2 Consistency index

Motivated by scaling-law results for LLMs (Kaplan et al., 2020), we use two sets of country-level proxies. The first set proxies for compute intensity and consists of log GDP per capita and the Information and Communications Technology Development Index (IDI). The second set proxies for corpus size and consists of log population and log number of internet users (Villalobos et al., 2022). We then estimate a series of OLS regressions to examine which set is more strongly associated with country-level alignment.

Columns (1)–(4) of [Table E.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 0.007 to 0.008 ($P < 0.001$), and those on IDI from 0.070 to 0.077 ($P < 0.001$), indicating that higher levels of economic development are systematically associated with stronger alignment between LLM-predicted and GPS-surveyed preferences. By contrast, the coefficients on log population and log number of

internet users are small in magnitude (at most 0.001 in absolute value) and the statistical significance is not robust across specifications.

Because these four country-level variables are highly collinear, we cannot 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 E.3](#) reports regressions including both principal components. The first principal component of economic development is positively and significantly associated with the consistency index ($b = 0.009$, $P < 0.001$), whereas the first principal component of the population measures is not statistically significant. Column (6) of [Table E.3](#) reports this specification on the English-only LLM configurations where all the input and output are in English, eliminating potential noise from machine translation. The first principal component of economic development remains positively and significantly associated with the consistency index ($b = 0.006$, $P < 0.001$), and the first principal component of population is small and negative ($b = -0.003$, $P = 0.038$).

Replacing Spearman's ρ with Pearson's r as the dependent variable yields similar patterns ([Table E.4](#)). Coefficients on log GDP per capita (0.008–0.009) and IDI (0.077–0.083) remain positive and significant ($P < 0.001$), while the significant association between population measures and consistency index is not robust across the two identifications. 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.

E.3 Inconsistency index

We next repeat the analysis in Section E.2 using the inconsistency index as the dependent variable ([Table E.5](#)). In all specifications, higher economic development is associated with lower inconsistency. The coefficients on log GDP per capita range from -0.194 to -0.205 ($P < 0.001$), and those on IDI from -1.343 to -1.494 ($P < 0.001$). Population-related measures are also negatively associated with the inconsistency index, but their

coefficients are small in magnitude: between -0.020 and -0.006 for log population and between -0.011 and -0.006 for log number of internet users. In the PCA specification in column (5), the economic-development score has a coefficient of -0.201 ($P < 0.001$), compared with -0.007 ($P = 0.448$) for the population score. Column (6) of [Table E.5 restricts the sample to the English-only experiments](#). The economic-development score remains significantly negative ($b = -0.058$, $P < 0.001$), whereas the population score is not statistically significant ($b = 0.015$, $P = 0.251$).

As a robustness check, we redefine the inconsistency index as the mean squared error between sampling-weighted country-level average preferences in the LLM and GPS data. [Table E.6](#) reports regression results using this alternative definition. Under this alternative definition, economic-development variables remain strongly and negatively associated with the inconsistency index. Coefficients on log GDP per capita range from -0.244 to -0.260 ($P < 0.001$), and those on IDI from -1.885 to -2.141 ($P < 0.001$). In contrast, the coefficients on log population and log number of internet users are small in magnitude and not robustly significant across column (1) to column (4). In both the PCA (column 5) and English-only (column 6) specifications, the economic-development score remains significantly negative, whereas the population score is small and only marginally significant.

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

E.4 Prediction error decomposition

The above results establish two main points: (1) demographic information accounts for more of the individual-level alignment than country information does; and (2) compared with population-related variables, economic-related variables play a more important role in shaping prediction errors. However, it remains unclear which specific variables explain the prediction errors best.

We decompose the prediction error by regressing the MSE of each model on demographic variables, country characteristics, and country fixed effects, and use R^2 to measure the explanatory power of each variable. In each regression, the dependent variable is the MSEs and the independent variable is indicated by the name of rows in [Figures E.1–E.6](#).

[Figures E.1–E.6](#) report the decomposition results for each model and for each preference. At most, 6% of the prediction errors can be explained by demographic variables and country characteristics; this maximum is attained for trust predicted by DeepSeek-V3, with GDP and IDI as the explanatory variables. In contrast, country fixed effects have the greatest explanatory power, although this may partly reflect the large number of dummy variables (73 dummies) included. We also find that the explanatory power of GDP and IDI (i.e., economic-related variables) is substantially larger than that of population-related variables (0.020 vs 0.002; two-sided t -test: $t(23.92) = 4.351$, $P < 0.001$, Cohen's $d = 1.256$, 95% CI= [0.01, 0.03]).

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

Preference	Baseline	Country	Demo	Demo + Country
Risk-taking	0.217	0.212	0.023	0.018
Patience	0.084	0.090	0.008	0.018
Negative reciprocity	0.052	0.079	-0.005	0.006
Positive reciprocity	0.070	0.060	0.035	0.015
Altruism	0.077	0.081	0.033	0.021
Trust	0.019	-0.003	0.026	0.002
Average	0.087	0.087	0.020	0.013

Notes: We estimate OLS regressions for each model $\times$ 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 E.2 Linear mixed-effects decomposition of individual-level alignment coefficients

DV: OLS coefficient (GPS on LLM)	
Condition	
Only country fixed effects	0.000 (0.961) [-0.014, 0.015]
Only demographic controls	-0.066*** (<0.001) [-0.081, -0.052]
Demographic controls + fixed effects	-0.073*** (<0.001) [-0.088, -0.059]
Preference	
Risk-taking	0.065*** (<0.001) [0.047, 0.082]
Patience	-0.003 (0.773) [-0.020, 0.015]
Negative reciprocity	-0.020* (0.030) [-0.037, -0.002]
Positive reciprocity	-0.008 (0.404) [-0.025, 0.010]
Trust	-0.042*** (<0.001) [-0.060, -0.024]
Observations	288
Number of groups	12

Notes: The dependent variable is the OLS coefficient from regressions of GPS-surveyed preferences on LLM-predicted preferences estimated under four specifications (baseline, + country fixed effects, + 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. Unadjusted P -values are in parentheses. 95% confidence intervals are in brackets. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.

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

	(1)	(2)	(3)	(4)	(5)	(6)
log GDP per capita	0.007*** (<0.001) [0.006,0.008]	0.008*** (<0.001) [0.006,0.009]				
ICT Development Index			0.070*** (<0.001) [0.060,0.080]	0.077*** (<0.001) [0.065,0.088]		
log population	0.001 (0.181) [-0.000,0.002]		-0.001 (0.154) [-0.002,0.000]			
log number of internet users		-0.001 (0.262) [-0.002,0.001]		-0.001 (0.084) [-0.002,0.000]		
Score for economic development					0.009*** (<0.001) [0.007,0.010]	0.006*** (<0.001) [0.003,0.009]
Score for population					-0.001 (0.112) [-0.002,0.000]	-0.003* (0.038) [-0.006,-0.000]
Observations	5255	4895	5183	4895	4895	1224
R-squared	0.531	0.536	0.540	0.544	0.540	0.584

Notes: The dependent variable is the country-level consistency index, defined as the Spearman's ρ between GPS-surveyed and LLM-predicted preferences within each country. Reported coefficients are OLS estimates, with P -value in parentheses and 95% confidence interval in brackets. All regressions include preference fixed effects, model fixed effects, and their interaction terms. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.

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

	(1)	(2)	(3)	(4)	(5)	(6)
log GDP per capita	0.008*** (<0.001) [0.007,0.009]	0.009*** (<0.001) [0.007,0.010]				
ICT Development Index			0.077*** (<0.001) [0.067,0.087]	0.083*** (<0.001) [0.071,0.094]		
log population	0.001 (0.298) [-0.001,0.002]		-0.001 (0.055) [-0.002,0.000]			
log number of internet users		-0.001 (0.100) [-0.002,0.000]		-0.001* (0.023) [-0.003,-0.000]		
Score for economic development					0.010*** (<0.001) [0.008,0.011]	0.007*** (<0.001) [0.004,0.010]
Score for population					-0.001* (0.037) [-0.003,-0.000]	-0.004** (0.009) [-0.006,-0.001]
Observations	5255	4895	5183	4895	4895	1224
R-squared	0.534	0.541	0.548	0.550	0.546	0.587

Notes: The dependent variable is the country-level consistency index, defined as the Pearson's r between GPS-surveyed and LLM-predicted preferences within each country. Reported coefficients are OLS estimates, with P -value in parentheses and 95% confidence interval in brackets. All regressions include preference fixed effects, model fixed effects, and their interaction terms. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.

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

	(1)	(2)	(3)	(4)	(5)	(6)
log GDP per capita	-0.194*** (<0.001) [-0.228,-0.161]	-0.205*** (<0.001) [-0.243,-0.167]				
ICT Development Index			-1.343*** (<0.001) [-1.599,-1.087]	-1.494*** (<0.001) [-1.788,-1.200]		
log population	-0.020* (0.026) [-0.038,-0.002]		-0.006 (0.550) [-0.024,0.013]			
log number of internet users		-0.011 (0.287) [-0.030,0.009]		-0.006 (0.549) [-0.025,0.013]		
Score for economic development					-0.201*** (<0.001) [-0.239,-0.163]	-0.058*** (<0.001) [-0.091,-0.025]
Score for population					-0.007 (0.448) [-0.027,0.012]	0.015 (0.251) [-0.011,0.041]
Observations	5256	4896	5184	4896	4896	1224
R-squared	0.061	0.059	0.058	0.058	0.061	0.055

Notes: The dependent variable is the country-level inconsistency index, defined as the sampling-weighted average of individual-level mean squared error between GPS-surveyed and LLM-predicted preferences within each country. Before computing the MSEs, the individual-level preference index is normalized to mean 0 and standard deviation 1. Reported coefficients are OLS estimates, with P -value in parentheses and 95% confidence interval in brackets. All regressions include preference fixed effects, model fixed effects, and their interaction terms. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.

Table E.6. OLS regressions of the country-level inconsistency index (country-level MSE) on country-level variables.

	(1)	(2)	(3)	(4)	(5)	(6)
log GDP per capita	-0.244*** (<0.001) [-0.315,-0.174]	-0.260*** (<0.001) [-0.336,-0.184]				
ICT Development Index			-1.885*** (<0.001) [-2.415,-1.355]	-2.141*** (<0.001) [-2.721,-1.561]		
log population	0.018 (0.477) [-0.031,0.066]		0.048 (0.072) [-0.004,0.100]			
log number of internet users		0.042 (0.122) [-0.011,0.095]		0.051 (0.059) [-0.002,0.103]		
Score for economic development					-0.264*** (<0.001) [-0.339,-0.189]	-0.172** (0.003) [-0.284,-0.059]
Score for population					0.048 (0.075) [-0.005,0.102]	0.104 (0.053) [-0.001,0.209]
Observations	5256	4896	5184	4896	4896	1224
R-squared	0.025	0.025	0.028	0.028	0.027	0.026

Notes: The dependent variable is the country-level inconsistency index, defined as country-level mean squared error between GPS-surveyed and LLM-predicted preferences. Before computing the MSEs, the country-level preference index is normalized to mean 0 and standard deviation 1. Reported coefficients are OLS estimates, with P -value in parentheses and 95% confidence interval in brackets. All regressions include preference fixed effects, model fixed effects, and their interaction terms. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.

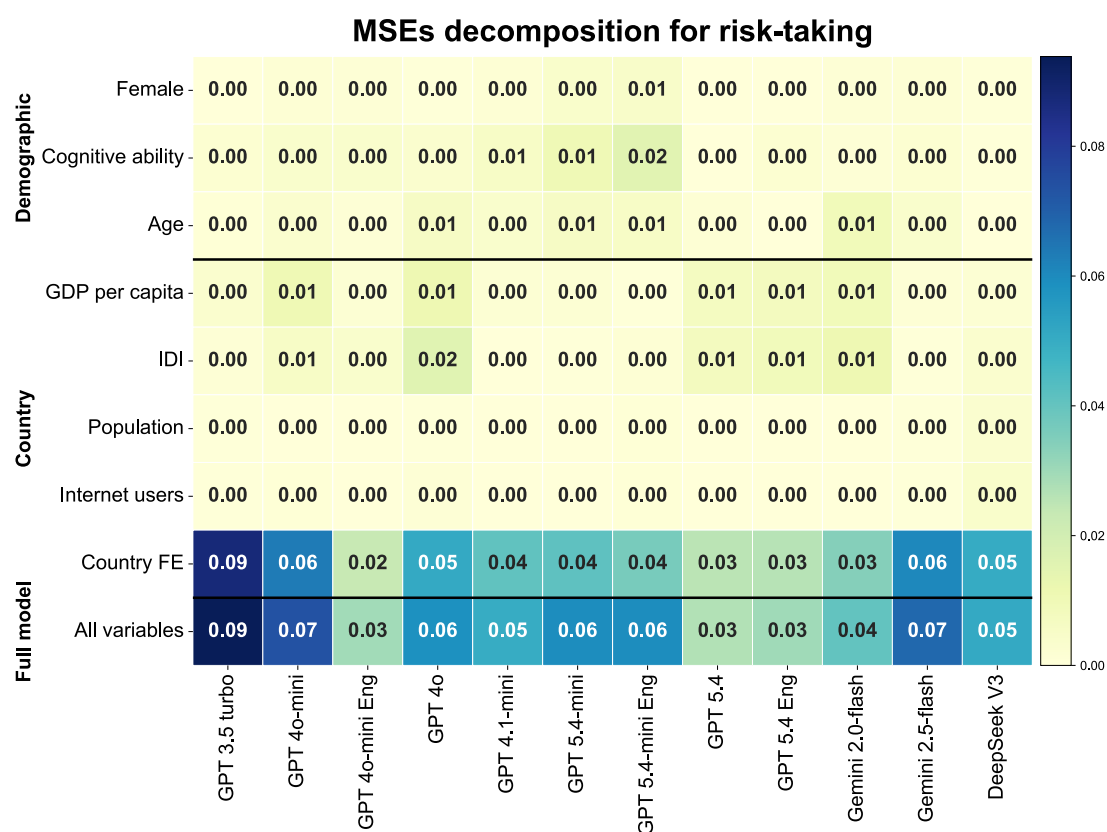


Figure E.1. Error decomposition of LLM-predicted preferences (risk-taking). The number in each cell represents the R^2 estimated from a regression where the individual-level MSE for risk-taking is the dependent variable and corresponding variables indicated by the row label are the independent variables. Columns correspond to the 12 LLM configurations.

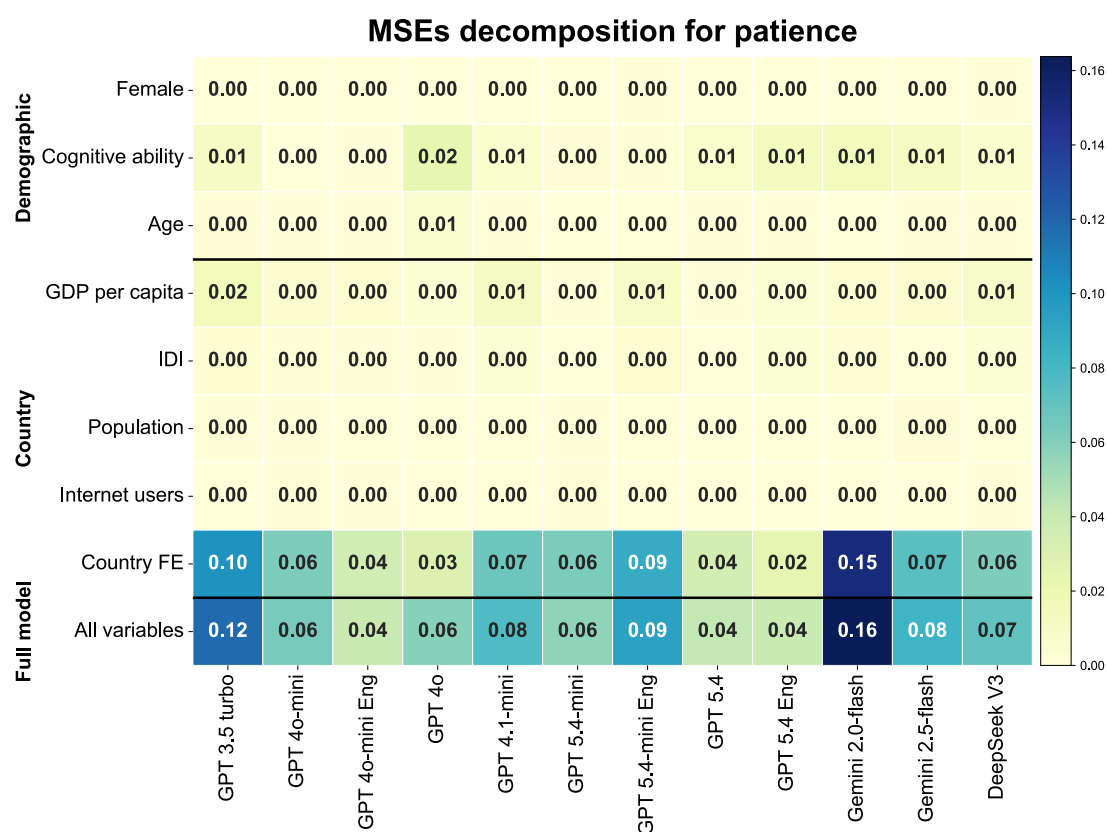


Figure E.2. Error decomposition of LLM-predicted preferences (patience). The number in each cell represents the R^2 estimated from a regression where the individual-level MSE for patience is the dependent variable and corresponding variables indicated by the row label are the independent variables. Columns correspond to the 12 LLM configurations.

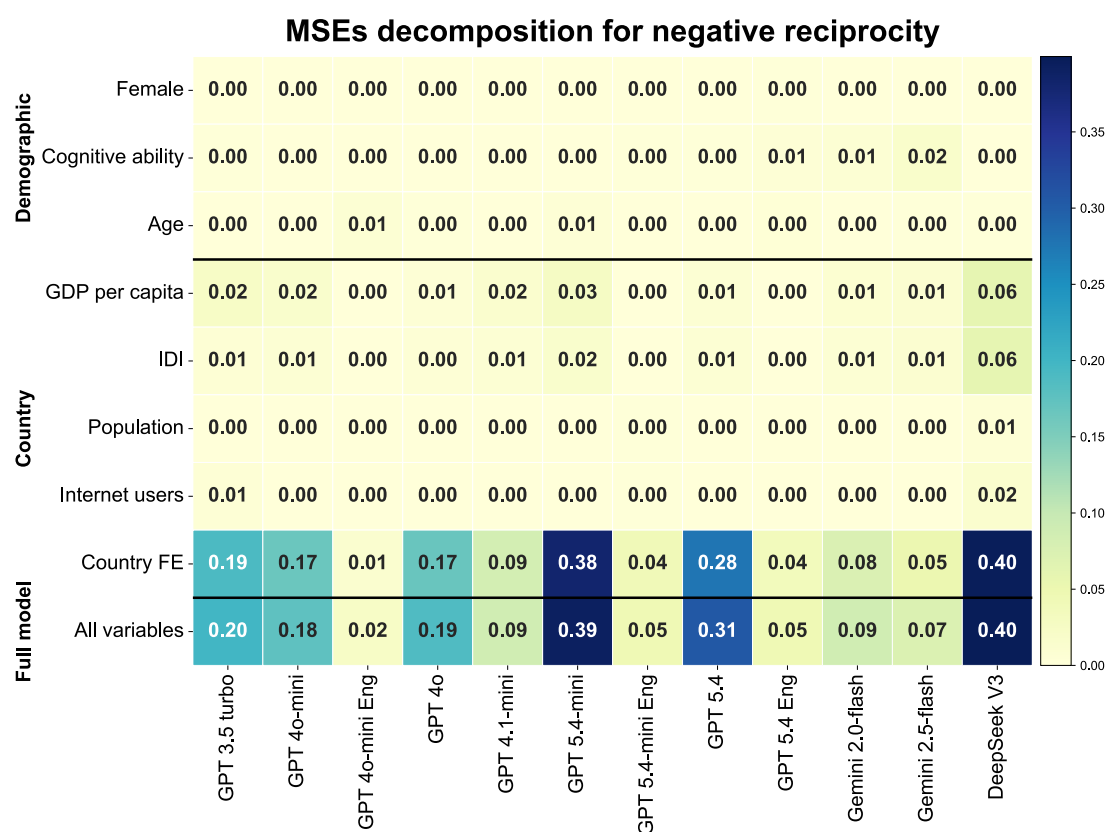


Figure E.3. Error decomposition of LLM-predicted preferences (negative reciprocity).
The number in each cell represents the R^2 estimated from a regression where the individual-level MSE for negative reciprocity is the dependent variable and corresponding variables indicated by the row label are the independent variables. Columns correspond to the 12 LLM configurations.

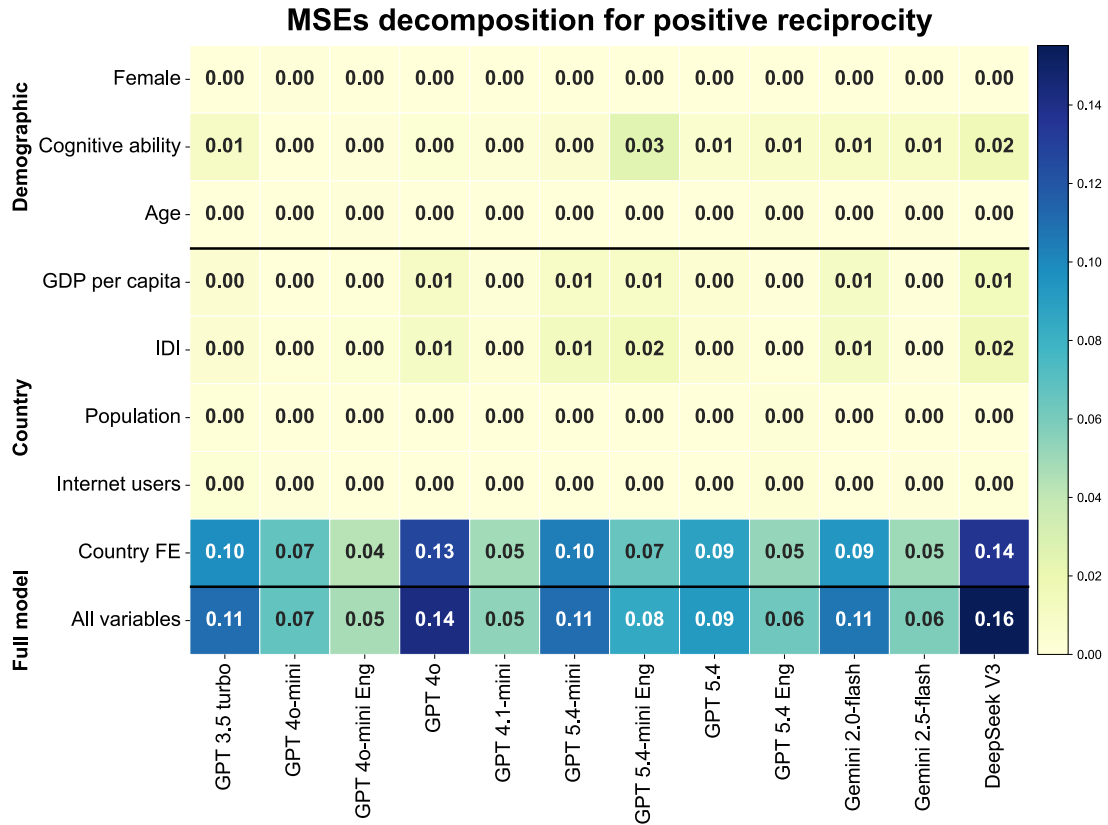


Figure E.4. Error decomposition of LLM-predicted preferences (positive reciprocity).
The number in each cell represents the R^2 estimated from a regression where the individual-level MSE for positive reciprocity is the dependent variable and corresponding variables indicated by the row label are the independent variables. Columns correspond to the 12 LLM configurations.

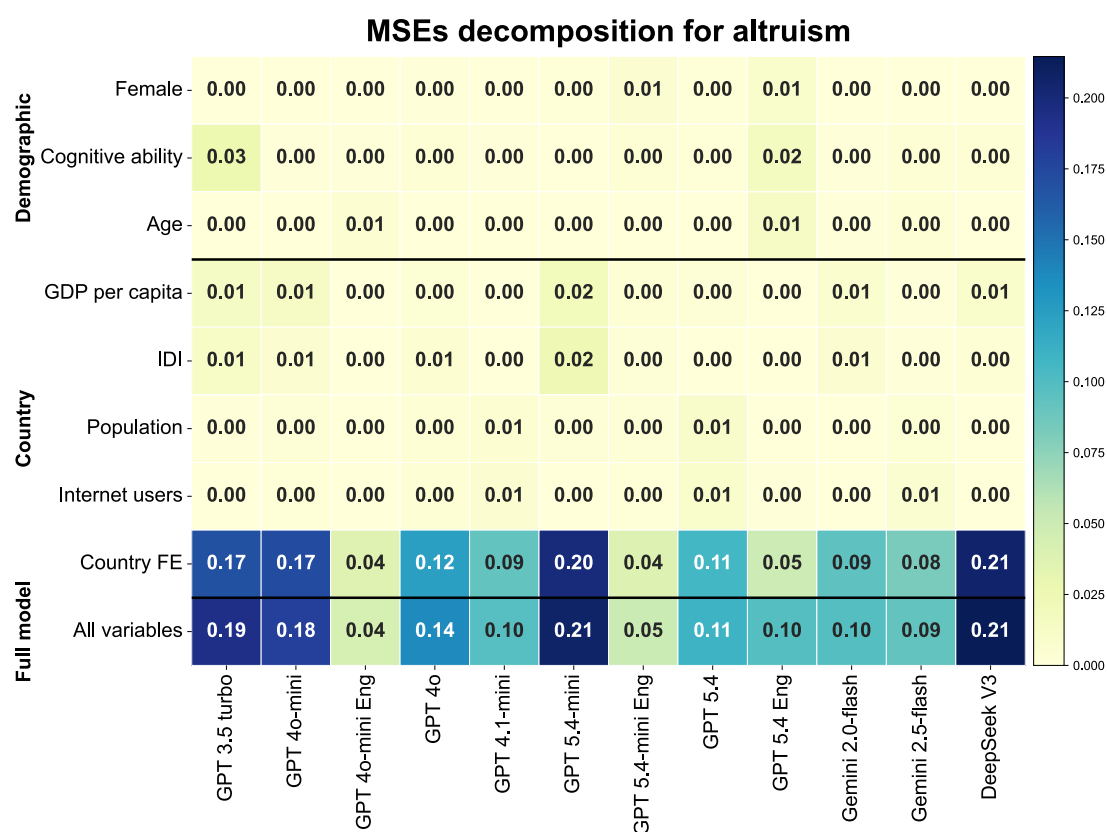


Figure E.5. Error decomposition of LLM-predicted preferences (altruism). The number in each cell represents the R^2 estimated from a regression where the individual-level MSE for altruism is the dependent variable and corresponding variables indicated by the row label are the independent variables. Columns correspond to the 12 LLM configurations.

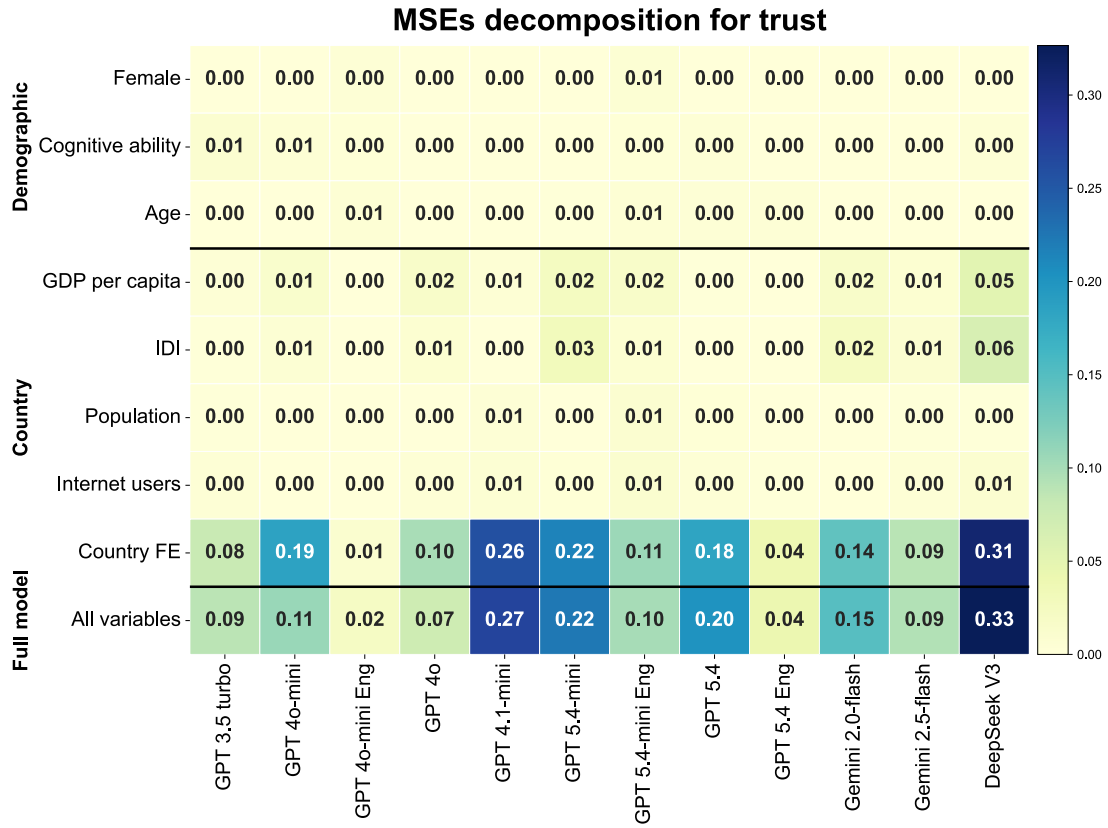


Figure E.6. Error decomposition of LLM-predicted preferences (trust). The number in each cell represents the R^2 estimated from a regression where the individual-level MSE for trust is the dependent variable and corresponding variables indicated by the row label are the independent variables. Columns correspond to the 12 LLM configurations.

Section F. Ablation experiments

This section presents the prompt-component ablation experiments. [Section F.1](#) specifies the persona conditions and the corresponding system prompt templates. [Sections F.2](#) and [F.3](#) then evaluate, respectively, how alignment and demographic gradients change across these conditions.

F.1. System prompt of ablation experiments

We designed and conducted a series of prompt-component ablation experiments in which we remove some components of the respondents' information embedded in the system prompts. The ablation comprises four types of conditions (six in total): full-persona, partial-persona, demographic-persona and country-persona. The full-persona condition replicates the original experiment with all information included. The partial-persona conditions consist of three variants, each omitting one demographic attribute (gender, age, or cognitive ability) while retaining the others and country. The country-persona condition retains only country-level information, while the demographic-persona condition retains only demographic information.

The detailed system prompt templates are shown in [Table F.1](#). All ablation experiments are conducted in local languages by GPT-5.4-mini, consistent with our original experiments.

F.2. Alignment across ablation experiments

If alignment relies primarily on demographic information, the alignment should decrease in the country-persona condition. [Figure F.1](#) reports the individual-level and country-level Spearman correlations and [Figure F.2](#) reports the MSEs.

We find that the individual-level correlations decrease from 0.096 (full-persona) to 0.030 (country-persona), and MSEs increase for all preferences. However, we do not find such a decrease from full-persona to demographic-persona, in fact, they slightly increase (from 0.096 to 0.098). This result is in line with our claim in [Supplementary Information E.1](#), where we find that demographic variables absorb approximately 77% explanatory

power of alignment.

Table F.1. System prompts of ablation experiments

Condition	Template
Full-persona	You are <i>[age]</i> -year-old <i>[gender]</i> , living in <i>[region]</i> , <i>[country]</i> . You are <i>[subjective cognitive skills]</i> . You are taking a survey. Please make your own choice according to the question. Please begin with I will.
Partial-persona (without cognitive ability)	You are <i>[age]</i> -year-old <i>[gender]</i> , living in <i>[region]</i> , <i>[country]</i> . You are taking a survey. Please make your own choice according to the question. Please begin with I will.
Partial-persona (without gender)	You are <i>[age]</i> years old, living in <i>[region]</i> , <i>[country]</i> . You are <i>[subjective cognitive skills]</i> . You are taking a survey. Please make your own choice according to the question. Please begin with I will.
Partial-persona (without age)	You are <i>[gender]</i> , living in <i>[region]</i> , <i>[country]</i> . You are <i>[subjective cognitive skills]</i> . You are taking a survey. Please make your own choice according to the question. Please begin with I will.
Country-persona	You are living in <i>[region]</i> , <i>[country]</i> . You are taking a survey. Please make your own choice according to the question. Please begin with I will.
Demographic-persona	You are <i>[age]</i> -year-old <i>[gender]</i> . You are <i>[subjective cognitive skills]</i> . You are taking a survey. Please make your own choice according to the question. Please begin with I will.

Turning to MSEs, the individual-level MSEs increase from 1.838 (full-persona) to 1.961 (country-persona), an increase of approximately 7% ($0.123/1.838$), mirroring the correlation-based results. Adding country information on top of demographics to the prompt barely changes the prediction error (an increase of 0.003, or 0.16%; two-sided paired t -test, $t(5) = 0.239$, Benjamini-Hochberg FDR corrected $P = 0.820$, Cohen’s $d = 0.02$, 95% CI = $[-0.03, 0.04]$, $BF_{10} = 0.38$; two-sided Wilcoxon signed-rank test, Benjamini-Hochberg FDR corrected $P = 1.000$), whereas adding demographic information on top of country reduces prediction error by 0.123 (about 6.7% relative to full-persona; two-sided paired t -test, $t(5) = -3.35$, Benjamini-Hochberg FDR corrected $P = 0.040$, Cohen’s $d = -0.99$, 95% CI = $[-0.22, -0.03]$, $BF_{10} = 4.07$; two-sided Wilcoxon signed-rank test, Benjamini-Hochberg FDR corrected $P = 0.062$).

As reported in [Table F.2](#), a linear model with preference-domain fixed effects confirms this pattern (relative to the country-persona condition: $b_{\text{partial-persona}} = -0.097$, $P = 0.001$, 95% CI = $[-0.149, -0.046]$; $b_{\text{demographic-persona}} = -0.126$, $P < 0.001$, 95% CI = $[-0.189, -0.064]$; $b_{\text{full-persona}} = -0.123$, $P < 0.001$, 95% CI = $[-0.186, -0.061]$). In [Table F.3](#), we report the

detailed correlation and MSE values for each preference.

Although correlations in the partial-persona conditions tend to decrease, the overall evidence is mixed. For the cognitive ability dimension, the individual-level correlations decrease and the MSEs increase, compared to the full-persona, indicating that cognitive ability plays an important role in shaping LLMs' predicted preferences. However, for gender and age dimensions, the direction of correlations and MSEs is not consistent. We attribute this to the potentially distorted linkage between age and gender, as shown in prior work ⁸.

Turning to the country-level alignment, the country-level MSEs are similar across the three main conditions ($M_{\text{full-persona}} = 1.779$ vs $M_{\text{demographic-persona}}=1.747$ vs $M_{\text{country-persona}}=1.743$), and the corresponding Spearman correlations are also close (0.109, 0.123, and 0.138, respectively; see [Table F.3](#) for preference-level breakdowns). Most of the 95% confidence intervals overlap ([Figures F.1B](#) and [F.2B](#)), so we cannot conclude that they differ statistically. One possible explanation for this is that even in the country-persona condition, LLMs still cannot adequately adjust their predictions according to country information. Hence, we should expect that the country-level factors contribute the largest explanatory power of prediction errors. Using the same techniques as in the main text, we decompose the MSEs for each ablation experiment independently, and we find that the country fixed effects still explain most of the prediction errors, as shown in [Figure F.3](#) (see [Figures F.4–F.9](#) for per-preference error decomposition).

The recoverability in demographic and cultural dimensions fluctuates as well, but none of the changes is statistically significant. Removing demographic information from full-persona reduces the overall demographic consistency ratio from 88.9% to 61.1%, but the reduction is not statistically significant (full-persona vs country-persona; two-sided McNemar's test, $\chi^2(1) = 2.29$, Benjamini-Hochberg FDR corrected $P=0.360$; match ratio: 88.9% to 72.2%, two-sided McNemar's test, $\chi^2(1) = 0.80$, Benjamini-Hochberg FDR corrected $P=0.500$). Removing country information increases the cultural consistency ratio from 46.7% to 47.8%, but the increase is not statistically significant (full-persona vs demo-persona; two-sided McNemar's test, $\chi^2(1) = 0.00$, Benjamini-Hochberg FDR

corrected $P=1.000$; match ratio: 56.7% to 62.2%, two-sided McNemar's test, $\chi^2(1) = 1.78$, Benjamini-Hochberg FDR corrected $P=0.360$).

Table F.2. Linear regression of individual-level MSEs on ablation conditions

DV: Individual-level MSEs	
Condition	
Full-persona	-0.123*** (<0.001)
Demographic-persona	[-0.186, -0.061] -0.126*** (<0.001)
Partial-persona	[-0.189, -0.064] -0.097** (0.001)
Preference	[-0.149, -0.046]
Risk-taking	-0.350*** (<0.001)
Patience	[-0.413, -0.287] -0.124*** (<0.001)
Negative reciprocity	[-0.186, -0.061] 0.011 (0.715)
Positive reciprocity	[-0.051, 0.074] -0.151*** (<0.001)
Trust	[-0.214, -0.088] -0.103** (0.002)
	[-0.166, -0.040]
R^2	0.884
Observations	36

Notes: The dependent variable is the individual-level MSE computed from each ablation condition. Reported coefficients are OLS estimates from a model including fixed effects for persona condition (Full-persona, Demographic-persona, Country-persona, Partial-persona) and preference domain. The reference categories are the country-persona condition and altruism. P -values are in parentheses and 95% CIs are in brackets. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.

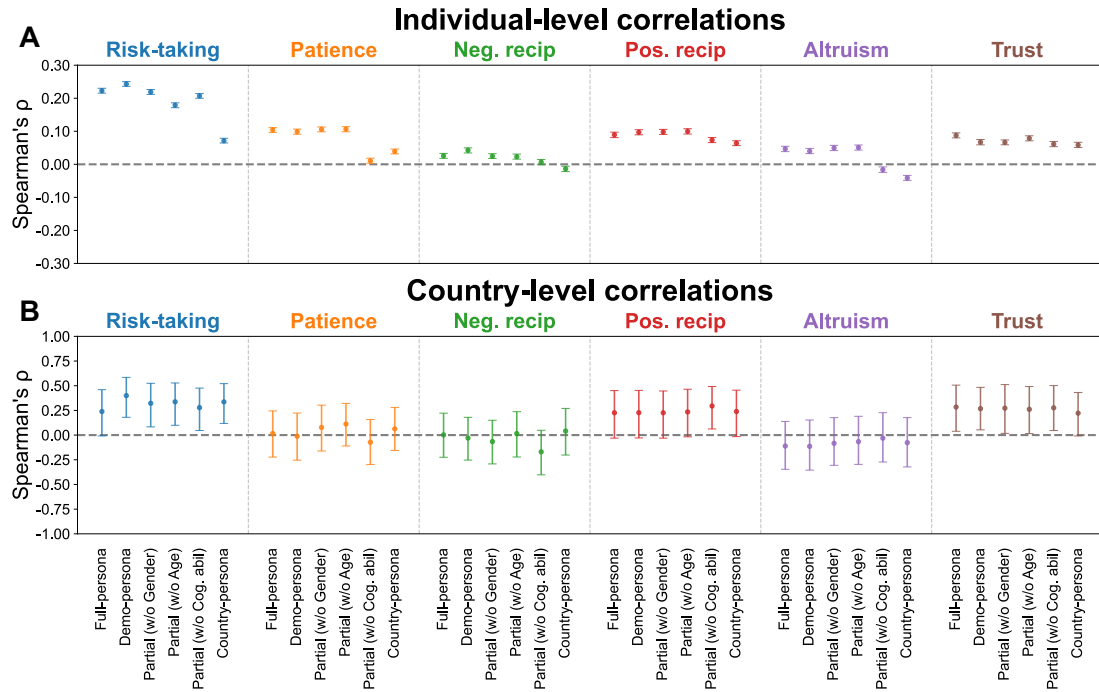


Figure F.1. Alignment between LLM-predicted and GPS-surveyed preferences. (A) Individual-level alignment. Each point displays the Spearman correlation coefficient between GPS-surveyed and LLM-predicted preferences, pooling all individual observations across 74 countries ($n = 65,212$, due to missing values in the original GPS, the number of observations varies by preference domain). **(B) Country-level alignment.** Each point represents the Spearman correlation coefficient between country-level average GPS-surveyed and LLM-predicted preferences ($n = 74$). Error bars indicate 95% bootstrapped CIs in both panels.

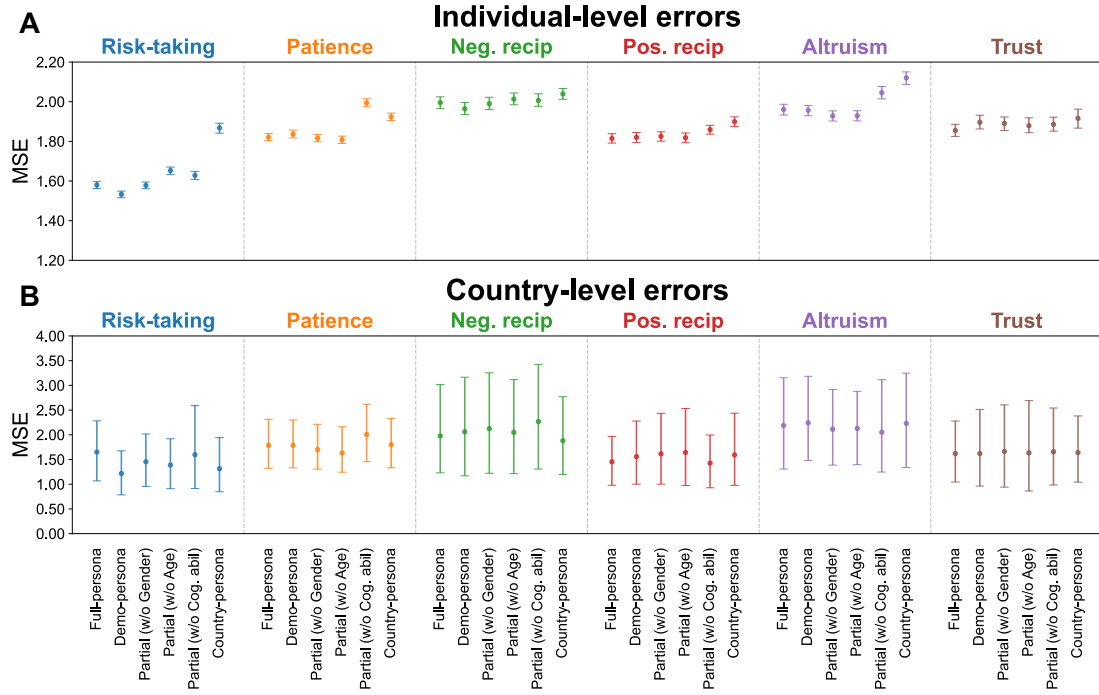


Figure F.2. Mean squared errors (MSEs) between LLM-predicted and GPS-surveyed preferences. (A) Individual-level MSEs. Each point displays the MSE between GPS-surveyed and LLM-predicted preferences, pooling all individual observations across 74 countries ($n = 65,212$, due to missing values in the original GPS, the number of observations varies by preference domain). **(B) Country-level MSEs.** Each point represents the MSE 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. Before computing the MSEs, the country-level preference index is normalized to mean 0 and standard deviation 1. Error bars indicate 95% bootstrapped CIs in both panels.

Table F.3. Individual- and country-level Spearman correlations and MSEs in ablation experiments.

	Risk-taking	Patience	Negative reciprocity	Positive reciprocity	Altruism	Trust	Average
Individual-level Spearman correlations							
Full-persona	0.223	0.104	0.026	0.089	0.047	0.087	0.096
Demo-persona	0.243	0.099	0.043	0.097	0.040	0.067	0.098
Partial (w/o Age)	0.179	0.107	0.023	0.100	0.051	0.079	0.090
Partial (w/o Cognition)	0.207	0.011	0.007	0.073	-0.016	0.061	0.057
Partial (w/o Gender)	0.219	0.106	0.025	0.098	0.049	0.066	0.094
Country-persona	0.071	0.039	-0.014	0.064	-0.041	0.059	0.030
Country-level Spearman correlations							
Full-persona	0.239	0.016	0.002	0.226	-0.111	0.284	0.109
Demo-persona	0.400	-0.013	-0.030	0.227	-0.114	0.268	0.123
Partial (w/o Age)	0.337	0.112	0.016	0.235	-0.066	0.261	0.149
Partial (w/o Cognition)	0.278	-0.072	-0.169	0.295	-0.031	0.276	0.096
Partial (w/o Gender)	0.322	0.077	-0.066	0.226	-0.083	0.273	0.125
Country-persona	0.337	0.063	0.041	0.239	-0.077	0.222	0.138
Individual-level MSEs							
Full-persona	1.580	1.821	1.996	1.814	1.961	1.855	1.838
Demo-persona	1.533	1.837	1.965	1.820	1.957	1.896	1.835
Partial (w/o Age)	1.653	1.808	2.013	1.818	1.929	1.879	1.850
Partial (w/o Cognition)	1.628	1.994	2.006	1.859	2.046	1.886	1.903
Partial (w/o Gender)	1.578	1.816	1.990	1.825	1.929	1.890	1.838
Country-persona	1.868	1.923	2.039	1.899	2.121	1.916	1.961
Country-level MSEs							
Full-persona	1.649	1.787	1.977	1.454	2.188	1.619	1.779
Demo-persona	1.217	1.788	2.060	1.558	2.243	1.620	1.747
Partial (w/o Age)	1.387	1.631	2.049	1.643	2.128	1.635	1.746
Partial (w/o Cognition)	1.597	2.005	2.268	1.426	2.053	1.656	1.834
Partial (w/o Gender)	1.455	1.701	2.125	1.615	2.113	1.662	1.778
Country-persona	1.313	1.799	1.881	1.593	2.231	1.640	1.743

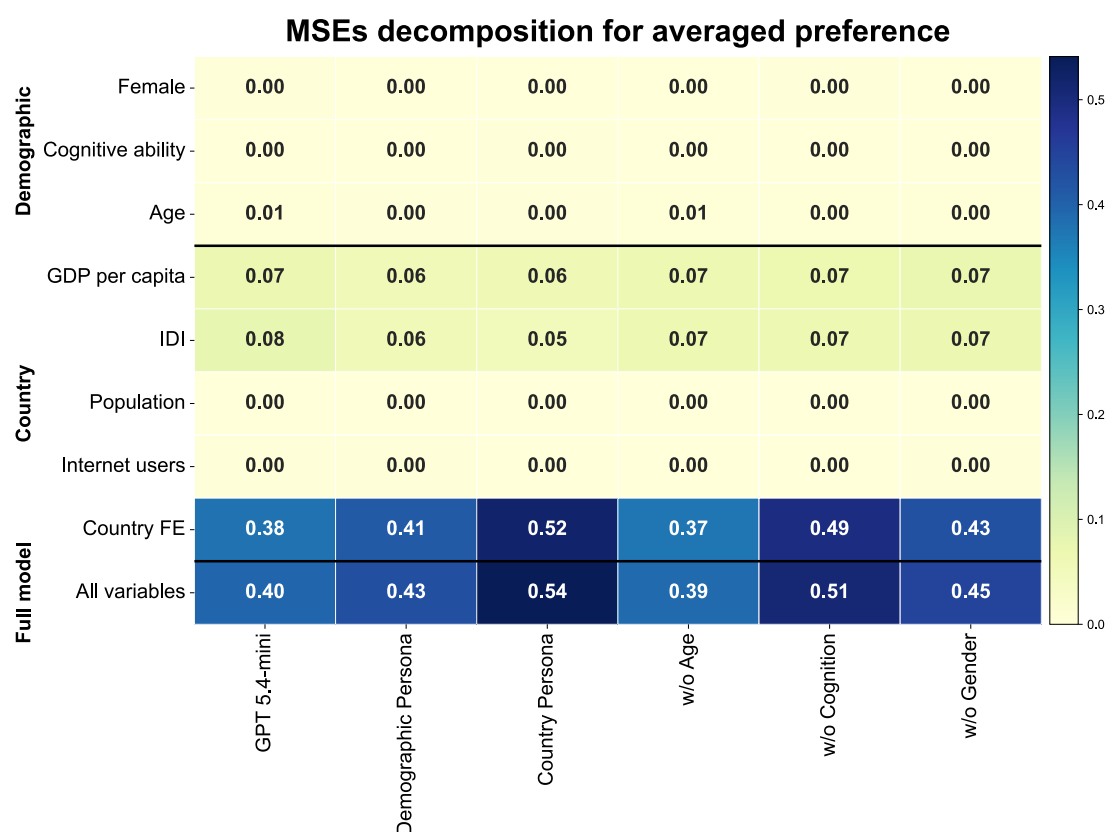


Figure F.3. Error decomposition pooled across LLM-predicted preferences. The number in each cell represents the R^2 estimated from a regression where the individual-level MSE pooled across six preferences is the dependent variable and corresponding variables indicated by the row label are the independent variables.

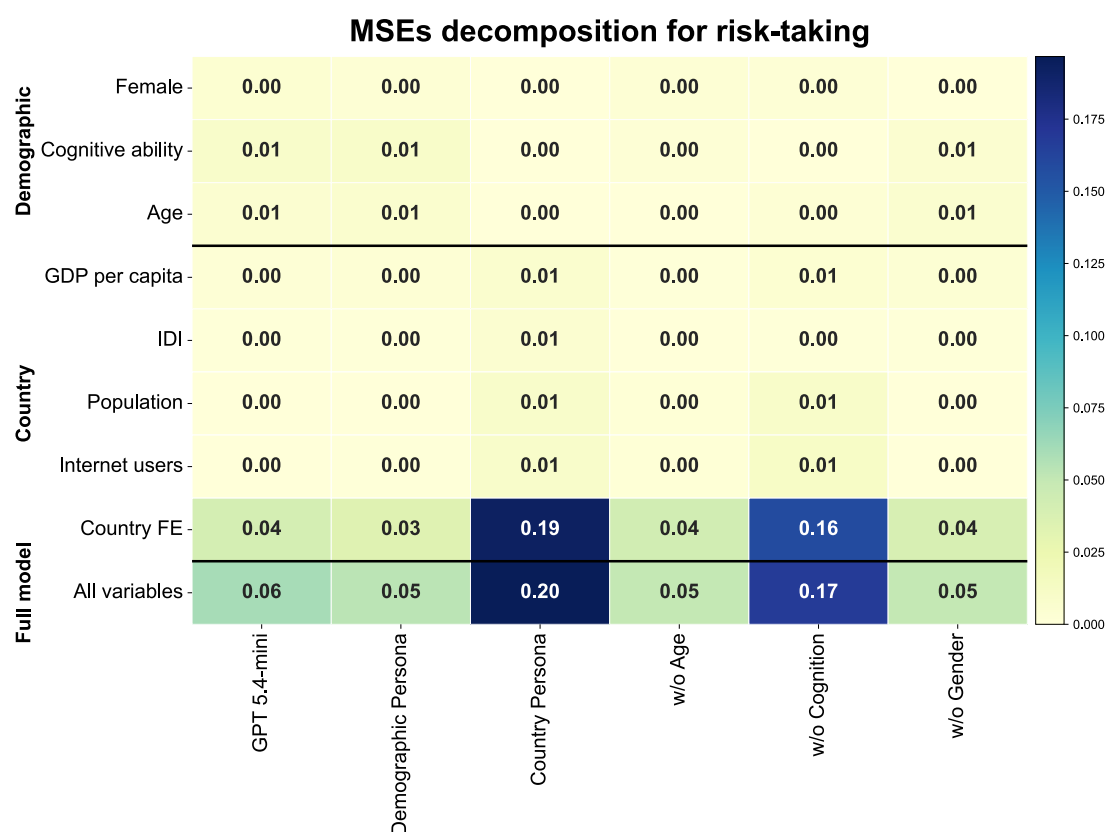


Figure F.4. Error decomposition of LLM-predicted preferences (risk-taking). The number in each cell represents the R^2 estimated from a regression where the individual-level MSE for risk-taking is the dependent variable and corresponding variables indicated by the row label are the independent variables.

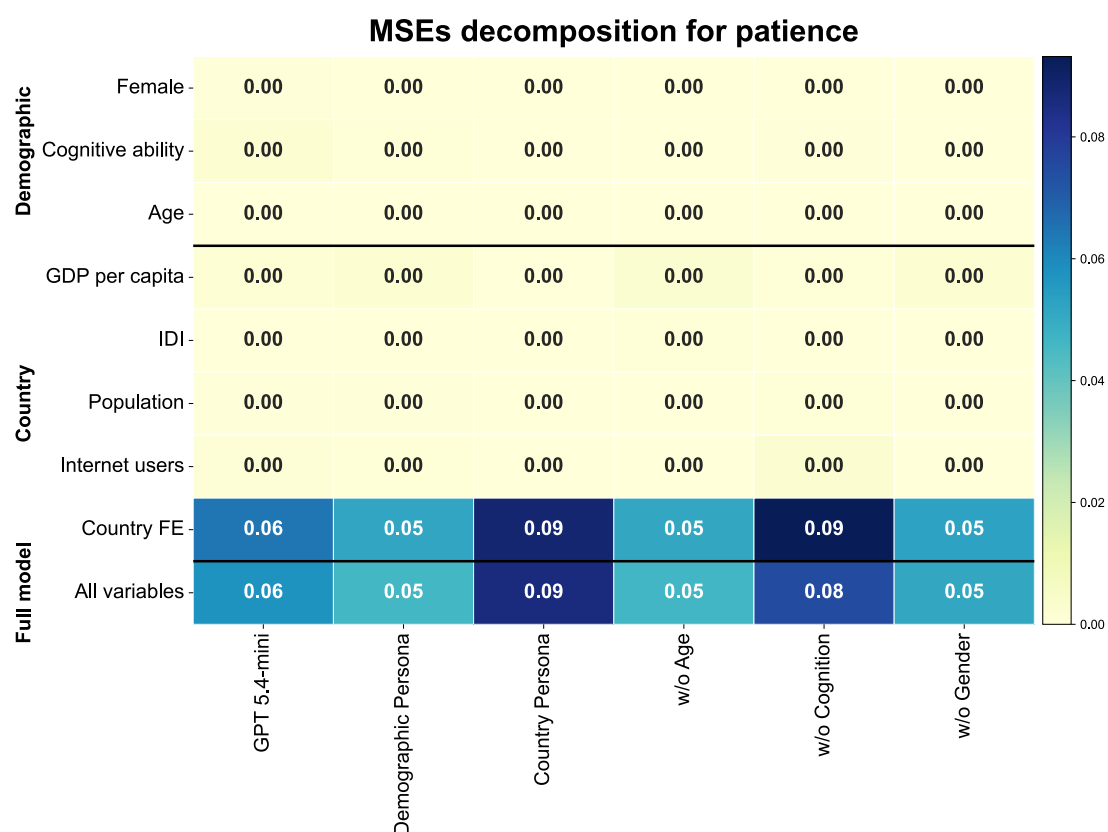


Figure F.5. Error decomposition of LLM-predicted preferences (patience). The number in each cell represents the R^2 estimated from a regression where the individual-level MSE for patience is the dependent variable and corresponding variables indicated by the row label are the independent variables.

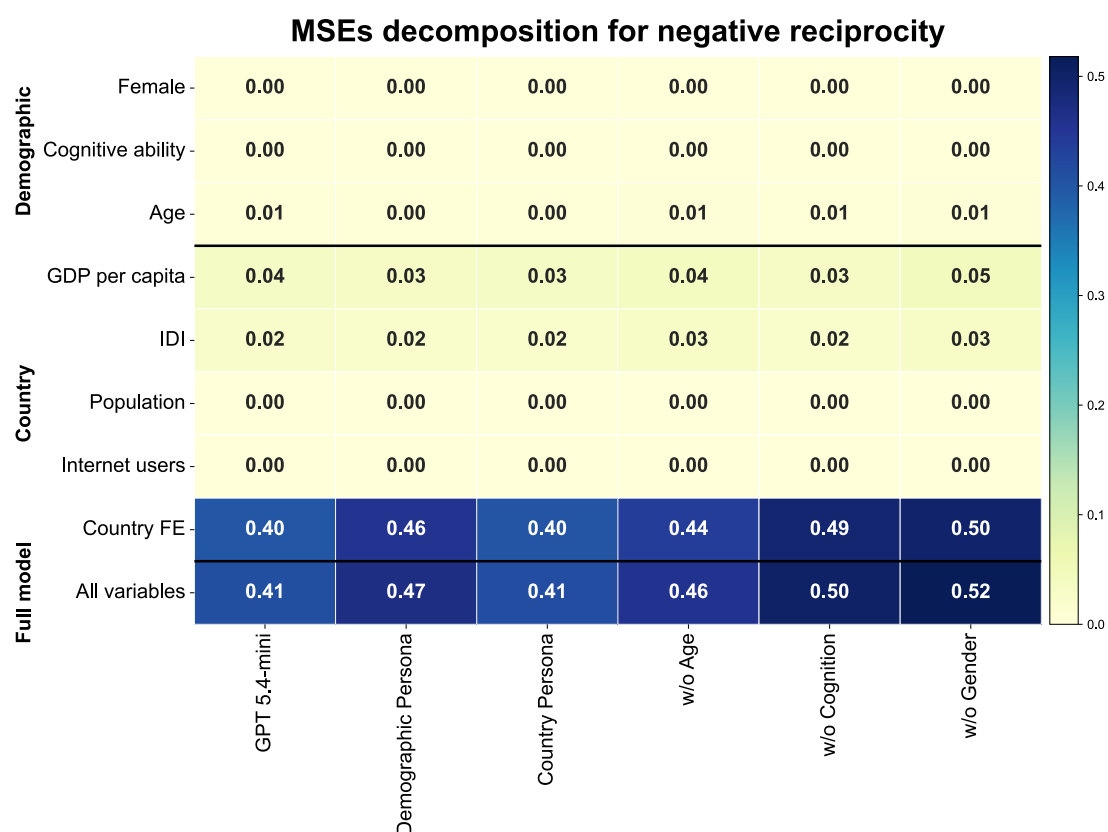


Figure F.6. Error decomposition of LLM-predicted preferences (negative reciprocity). The number in each cell represents the R^2 estimated from a regression where the individual-level MSE for negative reciprocity is the dependent variable and corresponding variables indicated by the row label are the independent variables.

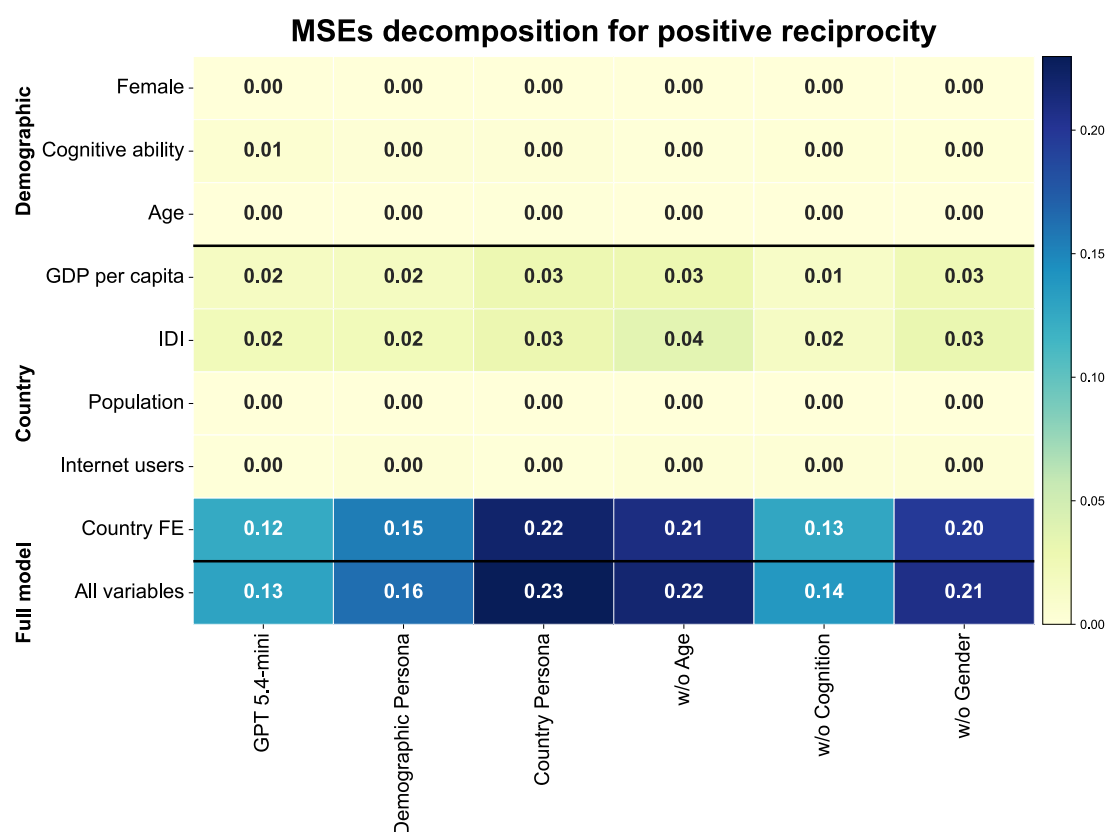


Figure F.7. Error decomposition of LLM-predicted preferences (positive reciprocity). The number in each cell represents the R^2 estimated from a regression where the individual-level MSE for positive reciprocity is the dependent variable and corresponding variables indicated by the row label are the independent variables.

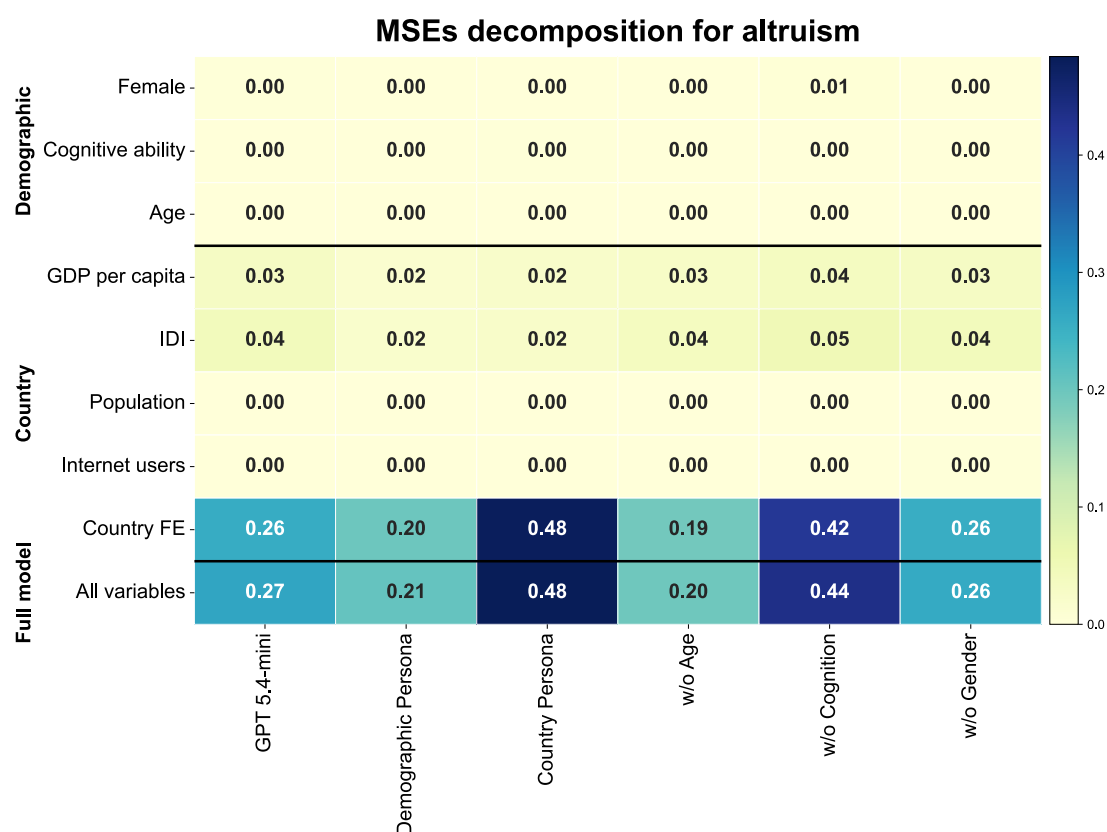


Figure F.8. Error decomposition of LLM-predicted preferences (altruism). The number in each cell represents the R^2 estimated from a regression where the individual-level MSE for altruism is the dependent variable and corresponding variables indicated by the row label are the independent variables.

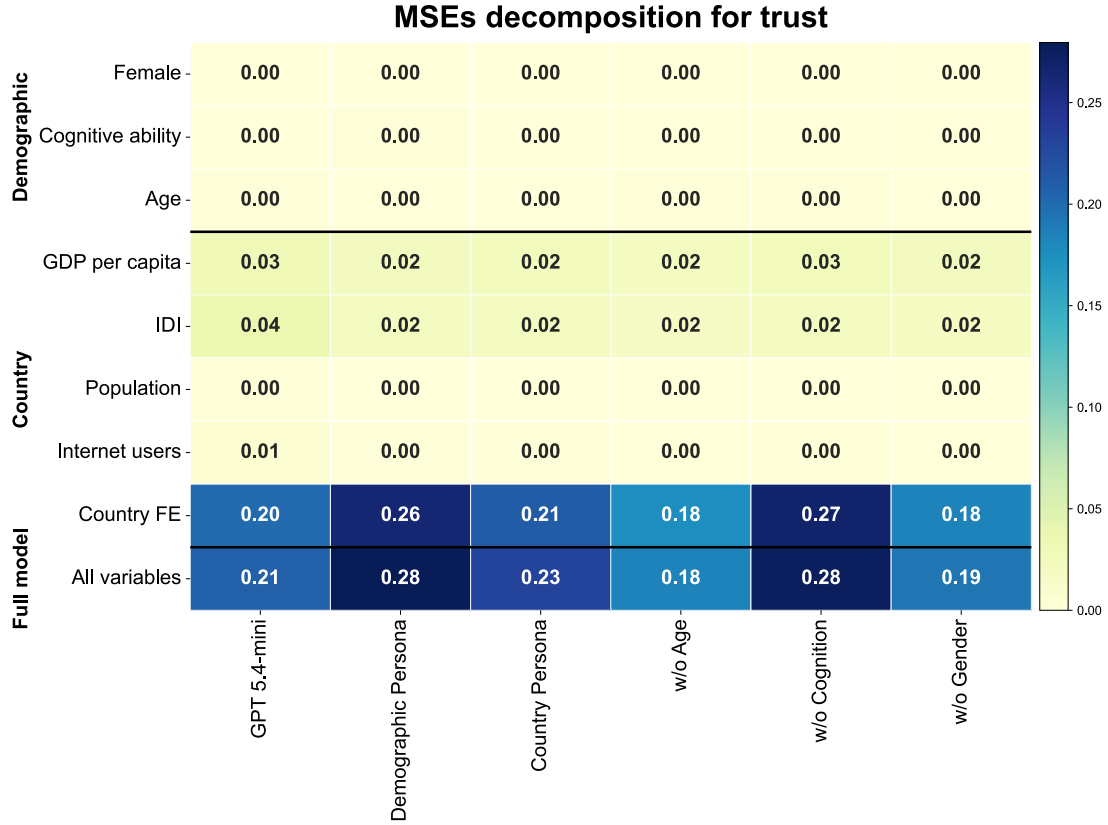


Figure F.9. Error decomposition of LLM-predicted preferences (trust). The number in each cell represents the R^2 estimated from a regression where the individual-level MSE for trust is the dependent variable and corresponding variables indicated by the row label are the independent variables.

F.3. Demographic gradients in ablation experiments

We speculate that demographic information in system prompts produces the demographic gradient, rather than country information or systematic noise. If so, the effect sizes estimated from partial-persona should be attenuated, and those of country-persona could be attenuated more. Hence, we expect that the effect sizes of these conditions satisfy the following pattern: effect sizes (full-persona) $\approx$ effect sizes (demographic-persona) $>$ effect sizes (partial-persona) $\approx$ effect sizes (country-persona).

Here, effect size (partial-persona) specifically refers to the effect size associated with a given dimension when that dimension is removed from the system prompt. For example, this includes the effect size of gender estimated from prompts without gender information, as opposed to the effect size estimated from prompts that explicitly include gender.

We run a linear mixed-effects model by pooling effect-size estimates from the full-persona, demographic-persona, country-persona and partial-persona conditions. As reported in Table F.4, the linear mixed-effect model shows that, compared to the country-persona condition, the coefficient of partial-persona is not significant at 5% significance level, and both coefficients of full-persona and demographic-persona are significant ($b_{\text{full-persona}} = 0.287, P < 0.001$; $b_{\text{demographic-persona}} = 0.287, P < 0.001$). In Table F.5, the detailed effect sizes estimated from each condition are reported.

Table F.4. Linear mixed-effects model of effect sizes (Cohen's d)

	DV: Cohen's d
Condition	
Full-persona	0.287*** (<0.001) [0.132, 0.442]
Demographic-persona	0.287*** (<0.001) [0.132, 0.442]
Partial-persona	0.030 (0.708) [-0.125, 0.184]
Preference	
Risk-taking	0.434*** (<0.001) [0.244, 0.624]
Patience	0.166 (0.087) [-0.024, 0.355]
Negative reciprocity	0.025 (0.795) [-0.164, 0.215]
Positive reciprocity	-0.014 (0.888) [-0.203, 0.176]
Trust	-0.138 (0.153) [-0.328, 0.051]
Observations	72
Number of groups	3

Notes: The dependent variable is Cohen's d effect sizes (gender, age, cognitive ability) computed from each ablation condition. We fit a linear mixed-effects model with indicators for ablation condition (Full-persona, Demographic-persona, Partial-persona) and preference-domain dummies, with random intercepts for the three demographic dimensions (gender, age, cognitive ability). The reference categories are the country-persona condition and altruism. P -values are in parentheses and 95% CIs are in brackets. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.

Table F.5. Effect sizes estimated from ablation experiments

Model	Risk-taking	Patience	Negative reciprocity	Positive reciprocity	Altruism	Trust	Average
Effect sizes for gender							
Full-persona	0.567	0.166	0.091	0.057	-0.147	-0.133	0.100
Demographic-persona	0.540	0.166	0.071	0.038	-0.172	-0.137	0.084
Partial-persona (w/o Gender)	0.164	0.138	0.034	0.050	0.128	0.037	0.092
Country-persona	0.066	-0.018	0.041	-0.029	-0.008	0.003	0.009
Effect sizes for age							
Full-persona	0.912	0.222	0.243	-0.133	0.023	-0.166	0.184
Demographic-persona	0.921	0.294	0.270	-0.116	0.069	-0.113	0.221
Partial-persona (w/o Age)	0.133	0.112	-0.073	-0.085	0.128	-0.056	0.027
Country-persona	0.100	-0.057	0.008	-0.076	0.081	-0.063	-0.001
Effect sizes for cognitive ability							
Full-persona	1.383	0.910	0.277	0.587	0.551	0.041	0.625
Demographic-persona	1.351	1.008	0.298	0.493	0.595	-0.121	0.604
Partial-persona (w/o Cognition)	0.105	0.036	0.054	0.030	-0.129	0.019	0.019
Country-persona	0.022	0.068	0.046	0.079	-0.061	0.088	0.040

Notes: Each row reports the Cohen's d effect size of the indicated demographic dimension (female vs male, older vs younger, lower vs higher cognitive ability) on each preference, estimated separately for each ablation condition. The "Average" column is the mean across the six preferences. Effect sizes are signed; positive values indicate higher predicted preference for the second category.

Section G. Paraphrasing and response-constraint experiments

A fundamental and frequently raised concern about using LLM-simulated responses is whether LLMs genuinely infer or learn people's preferences, rather than simply recalling memorized training data. We use two methods to clarify this point. First, we paraphrase the survey items (as reported in [Table G.1](#)) and run a paraphrasing experiment in English using GPT-5.4-mini. Second, we run a series of experiments in different language settings. Our original experiment is conducted in the local language as the GPS programme (e.g., Mandarin for Chinese respondents), and we run a series of controlled experiments whose only difference is in English (three models: GPT-4o-mini, GPT-5.4-mini, and GPT-5.4).

If the main result we find in the main text is from LLMs' memorization, instead of inference, we should expect that the correlations would decrease and the MSEs would increase sharply in paraphrasing and English experiments. However, as shown in [Figure G.1](#) (individual- and country-level correlations) and [Figure G.2](#) (individual- and country-level MSEs), we find no statistically significant differences between models prompted in local languages and in English in our individual- and country-level metrics (two-sided paired t -test by pooling GPT-4o-mini, GPT-5.4-mini, and GPT-5.4; individual-level correlations: $t(17) = 0.124$, $P = 0.903$, Cohen's $d = 0.014$, 95% CI = [-0.02, 0.02]; country-level correlations: $t(17) = 1.052$, $P = 0.307$, Cohen's $d = 0.232$, 95% CI = [-0.04, 0.12]; individual-level MSEs: $t(17) = -0.709$, $P = 0.488$, Cohen's $d = -0.090$, 95% CI = [-0.06, 0.03]; country-level MSEs: $t(17) = -0.580$, $P = 0.570$, Cohen's $d = -0.138$, 95% CI = [-0.22, 0.12]). The detailed correlation coefficients and MSEs are reported in [Table G.2](#).

Compared to GPT-5.4-mini (English), the individual-level correlations from the paraphrasing experiment are slightly larger for risk-taking, negative reciprocity, altruism, and trust, but smaller for patience and positive reciprocity. The MSE evidence is similarly mixed. That is, the MSEs fluctuate across preferences. Combining the individual- and country-level evidence from two metrics, we do not find statistically significant evidence that paraphrasing systematically changes the alignment (two-sided paired t -test between GPT-5.4-mini and GPT-5.4-mini with paraphrasing; individual-level correlations: $t(5) = -$

0.211 , $P = 0.841$, Cohen’s $d = -0.033$, 95% CI = [-0.04, 0.03]; country-level correlations: $t(5) = 1.707$, $P = 0.148$, Cohen’s $d = 0.230$, 95% CI = [-0.02, 0.11]; individual-level MSEs: $t(5) = 0.454$, $P = 0.669$, Cohen’s $d = 0.066$, 95% CI = [-0.05, 0.07]; country-level MSEs: $t(5) = 0.417$, $P = 0.694$, Cohen’s $d = 0.110$, 95% CI = [-0.21, 0.29]). Taking these two experiments’ results into account, there is no statistically significant evidence showing that memorization (or data contamination) is the primary driver of alignment, either at the individual or at the country level.

Another potential concern is that our main result might be sensitive to the response format. We conduct an additional response-constraint experiment using GPT-5.4-mini, in which the LLM is instructed to respond only with a number. The templates are shown in [Table G.3](#). We do not find statistically significant evidence that the response-constraint instruction systematically changes the alignment (two-sided paired t -test between GPT-5.4-mini and GPT-5.4-mini with response-constraint; individual-level correlations: $t(5) = -0.143$, $P = 0.892$, Cohen’s $d = -0.020$, 95% CI = [-0.03, 0.02]; country-level correlations: $t(5) = 0.228$, $P = 0.828$, Cohen’s $d = 0.060$, 95% CI = [-0.12, 0.15]; individual-level MSEs: $t(5) = 0.250$, $P = 0.812$, Cohen’s $d = 0.026$, 95% CI = [-0.03, 0.04]; country-level MSEs: $t(5) = -0.293$, $P = 0.781$, Cohen’s $d = -0.060$, 95% CI = [-0.20, 0.16]). The detailed correlation coefficients and MSEs are reported in [Table G.4](#).

Beyond the above analysis, we scrutinize the LLMs’ output in an item-by-item way, as reported in [Table G.5](#). Following the GPS programme, 12 raw survey items are collected per respondent. We compare the raw output between the paraphrasing experiment and the original experiment, and compare the raw output between the response-constraint experiment and the original experiment. As we mentioned above, the paraphrasing experiment is conducted in English, so we choose the GPT-5.4-mini (English) condition as the benchmark. We find that the average Spearman correlation across the 12 items for paraphrasing experiment is 0.404 and that for response-constraint experiment is 0.434. All correlations are significantly larger than zero ($P < 0.001$). Both averages exceed the typical between-model correlations (individual-level: 0.346) reported in [Section C](#) and the GPS-LLM correlations (individual-level: 0.094) reported in [Figure B.1](#), indicating that the

inferred preferences are stable across paraphrasing and response-constraint perturbations.

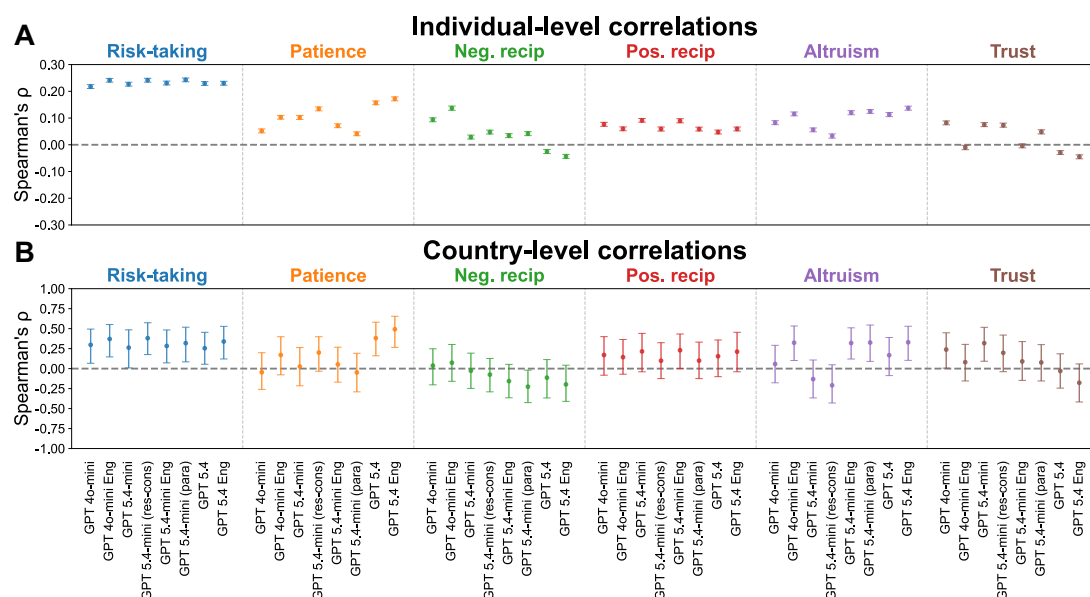


Figure G.1. Spearman correlations between LLM-predicted and GPS-surveyed preferences. **(A) Individual-level correlations.** Each point displays the Spearman's correlation coefficient between GPS-surveyed and LLM-predicted preferences, pooling all individual observations across 74 countries ($n = 69,295$, due to missing values in the original GPS, the number of observations varies by preference domain). **(B) Country-level correlations.** Each point represents the Spearman's 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. Error bars indicate 95% bootstrapped CIs in both panels.

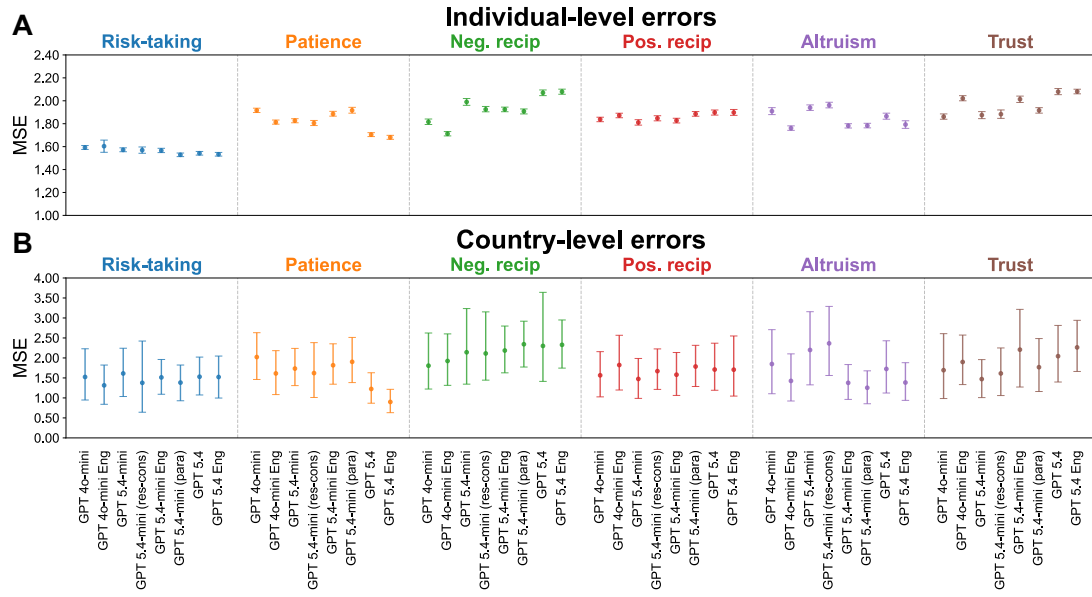


Figure G.2. Mean squared errors (MSEs) between LLM-predicted and GPS-surveyed preferences. (A) Individual-level MSEs. Each point displays the MSE between GPS-surveyed and LLM-predicted preferences, pooling all individual observations across 74 countries ($n = 69,295$, due to missing values in the original GPS, the number of observations varies by preference domain). **(B) Country-level MSEs.** Each point represents the MSE 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. Before computing the MSEs, the country-level preference index is normalized to mean 0 and standard deviation 1. Error bars indicate 95% bootstrapped CIs in both panels.

Table G.1. Paraphrasing experiment design (conducted in English)

Preference	Type of item	Original survey item	Paraphrased 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?	Please choose between the following options: Option (1): A lottery with a 50% chance of receiving [Amount A] [Local currency] and a 50% chance of receiving zero, Option (2): A certain payment of [Amount B] [Local currency].
	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.	Please indicate your general willingness to take risks on a scale from 0 to 10, where 0 represents “completely unwilling to take risks” and 10 represents “very willing to take risks.” You may select any integer between 0 and 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.	Please choose between the following options: Option (1): [Amount A] [Local currency] today, or Option (2): [Amount B] [Local currency] in 12 months. Please assume 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.	Please indicate your willingness to sacrifice current benefits for larger future gains on a scale from 0 to 10, where 0 represents “completely unwilling to do so” and 10 represents “very willing to do so.” You may select any integer between 0 and 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	Please indicate your willingness to incur a personal cost to punish someone who treats [you] unfairly on a scale from 0 to 10, where 0 represents “completely unwilling to do so” and 10 represents “very

		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.	willing to do so.” You may select any integer between 0 and 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.	Please indicate your willingness to incur a personal cost to punish someone who treats [others] unfairly on a scale from 0 to 10, where 0 represents “completely unwilling to do so” and 10 represents “very willing to do so.” You may select any integer between 0 and 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.	Please indicate how well the following statement describes you on a scale from 0 to 10, where 0 represents “does not describe me at all” and 10 represents “describes me perfectly”: “If I am treated very unfairly, I will seek revenge at the first opportunity, even if it involves a personal cost.” You may select any integer between 0 and 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.	Please indicate how well the following statement describes you on a scale from 0 to 10, where 0 represents “does not describe me at all” and 10 represents “describes me perfectly”: “I am willing to reciprocate a favor when someone has done me one.” You may select any integer between 0 and 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 are lost in an unfamiliar area and ask a stranger for directions. The stranger offers to take you to your destination, incurring a cost of approximately [Amount A] [Local currency], but

		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].	refuses any payment. You have six presents with you, with values ranging from [Min amount] to [Max amount] [Local currency]. Please indicate your choice: Option (1): No present. Option (2): Present worth [Amount B] [Local currency]. Option (3): Present worth [Amount C] [Local currency]. Option (4): Present worth [Amount D] [Local currency]. Option (5): Present worth [Amount E] [Local currency]. Option (6): Present worth [Amount F] [Local currency]. Option (7): 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.	Please indicate your willingness to contribute to charitable causes without expecting anything in return on a scale from 0 to 10, where 0 represents “completely unwilling to do so” and 10 represents “very willing to do so.” You may select any integer between 0 and 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.	Please indicate the amount you would donate to a good cause if you unexpectedly received [Amount] [Local currency]. You may choose any amount between 0 and [Amount].
Trust	Self-assessment	How well does each of the following statement describe you? “I assume that people have only the best intentions.” Please	Please indicate how well the following statement describes you on a scale from 0 to 10, where 0 represents “does not

	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.	describe me at all” and 10 represents “describes me perfectly”: “I tend to assume that people have good intentions.” You may select any integer between 0 and 10.
--	--	--

Notes: The survey items are contained in the user prompts. The system prompt is the same as our GPT-5.4-mini (English) benchmark.

Table G.2. Comparison between GPT-5.4-mini (English) and paraphrasing experiment

Model	Risk-taking	Patience	Negative reciprocity	Positive reciprocity	Altruism	Trust	Average
Individual-level Spearman’s correlations							
GPT-5.4-mini English	0.231	0.071	0.034	0.089	0.120	-0.004	0.090
Paraphrasing	0.243	0.041	0.042	0.059	0.125	0.048	0.093
Country-level Spearman’s correlations							
GPT-5.4-mini English	0.284	0.053	-0.157	0.230	0.320	0.090	0.137
Paraphrasing	0.319	-0.047	-0.226	0.100	0.326	0.077	0.092
Individual-level MSEs							
GPT-5.4-mini English	1.566	1.885	1.924	1.828	1.782	2.014	1.833
Paraphrasing	1.529	1.917	1.908	1.886	1.783	1.917	1.823
Country-level MSEs							
GPT-5.4-mini English	1.513	1.817	2.186	1.581	1.377	2.208	1.780
Paraphrasing	1.383	1.903	2.343	1.786	1.254	1.769	1.740

Notes: The GPT-5.4-mini (English) benchmark refers to the English-only condition (matching the paraphrasing experiment, which is also conducted in English); see Table G.1 for prompt templates. Both the individual-level and country-level preferences are normalized to mean 0 and standard deviation 1, before computing the MSEs.

Table G.3. Response-constraint experiment design (conducted in local languages)

Group	Template
Original experiment	You are <i>[age]</i> -year-old <i>[gender]</i> , living in <i>[region]</i> , <i>[country]</i> . You are <i>[subjective cognitive skills]</i> . You are taking a survey. Please make your own choice according to the question. Please begin with I will.
Response-constraint experiment	You are <i>[age]</i> -year-old <i>[gender]</i> , living in <i>[region]</i> , <i>[country]</i> . You are <i>[subjective cognitive skills]</i> . You are taking a survey. Please make your own choice according to the question. Select one option and output its number without anything else.

Notes: The persona is contained in the system prompt. The user prompt is the same as our GPT-5.4-mini benchmark.

Table G.4. Comparison between GPT-5.4-mini and response-constraint experiment

Model	Risk-taking	Patience	Negative reciprocity	Positive reciprocity	Altruism	Trust	Average
Individual-level Spearman's correlations							
GPT-5.4-mini	0.227	0.102	0.028	0.091	0.056	0.075	0.097
Response-constraint	0.242	0.135	0.048	0.059	0.033	0.073	0.098
Country-level Spearman's correlations							
GPT-5.4-mini	0.262	0.026	-0.026	0.215	-0.131	0.319	0.111
Response-constraint	0.381	0.201	-0.075	0.099	-0.209	0.197	0.099
Individual-level MSEs							
GPT-5.4-mini	1.572	1.826	1.990	1.811	1.939	1.875	1.836
Response-constraint	1.569	1.805	1.925	1.847	1.962	1.883	1.832
Country-level MSEs							
GPT-5.4-mini	1.612	1.734	2.142	1.475	2.199	1.471	1.772
Response-constraint	1.378	1.621	2.112	1.670	2.364	1.613	1.793

Notes: The GPT-5.4-mini benchmark refers to the local-language condition (matching the response-constraint experiment, which is also conducted in local languages); see Table G.3 for prompt templates. Both the individual-level and country-level preferences are normalized to mean 0 and standard deviation 1, before computing the MSEs.

Table G.5. Item-by-item comparison: paraphrasing experiment vs GPT-5.4-mini (English) (left), and response-constraint experiment vs GPT-5.4-mini (local language) (right).

Survey item	Paraphrasing experiment (in English)			Response-constraint experiment (in local languages)		
	Correlations	P-value	MSE	Correlations	P-value	MSE
Risk-taking: Self-assessment	0.820	<0.001	0.360	0.741	<0.001	0.517
Risk-taking: Staircase	0.428	<0.001	1.188	0.210	<0.001	1.686
Patience: Self-assessment	0.549	<0.001	0.904	0.696	<0.001	0.638
Patience: Staircase	0.305	<0.001	1.429	0.172	<0.001	1.733
Negative reciprocity: Willingness to act (yourself)	0.132	<0.001	1.713	0.495	<0.001	0.913
Negative reciprocity: Willingness to act (others)	0.243	<0.001	1.513	0.452	<0.001	1.052
Negative reciprocity: Self-assessment	0.442	<0.001	1.113	0.355	<0.001	1.472
Positive reciprocity: Hypothetical situation	0.058	<0.001	1.867	0.163	<0.001	1.667
Positive reciprocity: Self-assessment	0.287	<0.001	1.425	0.360	<0.001	1.428
Altruism: Willingness to act	0.478	<0.001	1.043	0.459	<0.001	1.206
Altruism: Hypothetical situation	0.988	<0.001	0.072	0.699	<0.001	0.466
Trust	0.117	<0.001	1.765	0.406	<0.001	1.238

Notes: The results of item-by-item comparison between the paraphrasing experiment and GPT-5.4-mini (English) are reported in the first three columns; the results of item-by-item comparison between the response-constraint experiment and GPT-5.4-mini (local language) are reported in the last three columns. All variables are normalized to mean 0 and standard deviation 1 before computing the MSEs.

Section H. 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*⁹, *Quarterly Journal of Economics*¹, and *Review of Economic Studies*¹⁰. 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, regarding 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, which is the significance criterion used in the original studies. As a robustness check, we also report the replication results at the stricter 5% 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, consistent with how these results are conventionally 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 H.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-5.4 (English) performs best, successfully reproducing 16 of 35 (45.7%) country-level findings in the same sign as the original GPS estimate at the 10% significance level. The weakest performance is obtained for GPT-4o-mini, which reproduces only a single finding. Under the 5% criterion, the corresponding replication rates still range from 9 to 14 of 18 findings; the best- and worst-performing models are unchanged, while counts for some intermediate models decrease slightly (e.g., GPT-4o: 12 at the 10% level vs 11 at the 5% level; see [Table H.2](#)).

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 ($P < 0.10$). 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 H.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 and GPT-5.4 English perform worst, reproducing 9 of 18

(50.0%) findings. Under the 5% criterion, the corresponding replication rates range from 9 to 14 of 18 findings.

Table H.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	GPT-5.4-mini	GPT-5.4-mini Eng	GPT-5.4	GPT-5.4 Eng	DeepSeek-V3	Gemini 2.0 Flash	Gemini 2.5 Flash	10% level replications	5% level replications
<i>Findings from Patience and comparative development</i>																
Patience	Economic development: log GDP per capita	0.63***	-0.30**	-0.03	0.25**	0.29**	-0.07	-0.08	-0.02	0.43***	0.58***	0.04	0.06	-0.17	4	4
	Physical capital: Log [capital stock p/c]	0.56***	-0.16	-0.04	0.19	0.29**	-0.03	-0.13	-0.13	0.40***	0.54***	0.17	0.07	-0.11	3	3
	Human capital: Skilled fraction over 25-year-old	0.56***	-0.29**	0.06	0.29**	0.27**	-0.07	-0.10	-0.01	0.24**	0.46***	0.02	0.14	-0.14	4	4
	Productivity: TFP	0.23*	-0.34***	-0.10	-0.15	0.17	0.08	0.10	-0.37***	-0.20	-0.07	-0.20	-0.15	-0.15	0	/
<i>Findings from Relationship of gender differences in preferences to economic development and gender equality</i>																
Gender difference in preference	Gender equality	0.55***	0.22*	-0.23*	0.62***	0.06	-0.02	0.08	-0.52***	-0.05	0.07	0.20*	-0.19	-0.02	3	1
	log GDP per capita	0.64***	0.00	0.01	0.54***	0.15	0.22*	0.05	-0.41***	-0.13	0.08	0.23**	0.09	0.06	3	2
<i>Findings from Global evidence on economic preferences</i>																
Patience	Hofstede long-term orientation	0.33**	-0.21	-0.19	0.07	-0.07	-0.24*	-0.40***	-0.31**	0.28**	0.22*	0.21	0.06	-0.27**	2	1
	Geographic conditions	0.43***	-0.16	-0.28**	-0.05	0.31**	-0.22	-0.45***	-0.09	0.18	0.41***	0.42***	0.09	-0.36**	3	3
	Absolute latitude	0.48***	-0.08	-0.06	0.08	0.34***	-0.07	-0.32***	-0.04	0.23*	0.50***	0.38***	0.17	-0.16	4	3
	Biology conditions	0.38***	-0.19	-0.30**	-0.01	0.22	-0.26*	-0.47***	-0.03	0.12	0.38***	0.42***	0.07	-0.36***	2	2
	Weak future time reference	-0.33***	0.22*	0.23*	-0.01	0.11	0.10	0.08	0.10	-0.10	-0.13	-0.07	-0.06	0.22*	0	0
	Pronoun drop	0.57***	-0.09	0.13	0.07	0.28**	0.05	0.28**	-0.03	0.20*	0.29**	-0.02	-0.02	0.05	4	3
	Share of protestants	0.45***	-0.12	0.14	0.22*	0.15	0.18	0.32***	0.13	0.27**	0.23**	-0.09	0.25**	0.20*	6	4
	Individualism	0.67***	-0.28**	0.01	0.13	0.27**	0.05	0.11	-0.02	0.26**	0.55***	0.01	0.10	0.03	3	3
	Family ties	-0.54***	0.17	-0.01	-0.10	-0.19	-0.09	0.27*	0.14	-0.33**	-0.44***	-0.26*	-0.42***	0.09	4	3
Risk-taking	Hofstede uncertainty avoidance	-0.32**	-0.20	-0.08	-0.08	-0.01	-0.23*	-0.14	0.04	-0.27**	-0.15	-0.03	0.03	0.22*	2	1
	Biology conditions	-0.36**	-0.03	-0.09	-0.22	-0.35**	-0.13	-0.29**	-0.13	-0.31**	-0.17	-0.08	-0.48***	-0.20	4	4
	Geographic conditions	-0.30**	-0.05	-0.15	-0.23	-0.35**	-0.16	-0.35**	-0.12	-0.46***	-0.24*	-0.15	-0.50***	-0.10	5	4
	Family ties	0.34**	0.07	0.07	0.30**	0.45***	0.27*	0.44***	0.15	0.38***	0.32**	0.14	0.56***	0.16	7	6
Positive reciprocity	Share of protestants	-0.15	-0.01	0.16	0.13	0.00	-0.01	0.05	0.11	0.02	0.01	-0.26**	-0.22*	-0.18	0(2)	/
	Biology conditions	0.28**	-0.12	0.08	0.23	-0.03	-0.09	-0.03	0.36**	-0.01	-0.06	0.03	-0.11	-0.02	1	1
Negative reciprocity	Geographic conditions	0.34**	0.05	-0.15	-0.34**	-0.15	-0.09	-0.04	-0.32**	-0.32**	-0.58***	0.12	-0.04	-0.15	0	0

	Absolute latitude	0.23**	-0.07	-0.28**	-0.36***	-0.12	0.01	-0.06	-0.38***	-0.15	-0.51***	0.05	-0.12	-0.32***	0	0
	Biology conditions	0.31**	0.12	-0.22	-0.33**	-0.23	-0.07	-0.06	-0.34**	-0.34**	-0.57***	0.00	-0.06	-0.21	0	0
	Armed conflicts	0.32***	0.20*	-0.15	-0.21*	-0.06	-0.16	-0.13	-0.22*	-0.23**	-0.30***	-0.01	-0.04	-0.30***	1	0
Altruism	Agricultural suitability	-0.21*	0.01	0.16	-0.16	-0.20*	-0.17	0.01	-0.11	0.01	-0.16	-0.03	0.00	-0.25**	2	/
	Crop suitability	-0.21*	-0.09	0.11	-0.05	-0.32***	-0.23*	-0.07	-0.17	-0.09	-0.25**	-0.13	-0.07	-0.37***	4	/
	Family ties	0.27*	0.42***	0.07	0.17	0.57***	0.38***	0.28*	0.36**	0.40***	0.60***	0.39***	0.37**	0.49***	10	/
Trust	WVS Trust	0.51***	0.29**	0.13	0.16	0.00	0.01	0.27**	-0.03	0.08	-0.17	0.12	0.04	0.21	2	2
	Geographic conditions	0.41***	0.14	0.18	0.22	-0.42***	-0.01	0.19	-0.07	-0.10	-0.68***	-0.08	0.15	0.04	0	0
	Absolute latitude	0.28**	0.00	0.12	-0.02	-0.52***	-0.24**	0.14	-0.18	-0.23**	-0.76***	-0.03	0.10	-0.09	0	0
	Agricultural suitability	-0.47***	0.04	0.01	0.21*	-0.10	-0.16	-0.06	0.21*	-0.03	-0.08	0.13	-0.26**	-0.14	1	1
	Crop suitability	-0.36***	0.00	0.01	0.17	-0.26**	-0.26**	-0.02	0.12	-0.13	-0.32***	0.14	-0.25**	-0.16	4	4
	Biology conditions	0.46***	0.02	0.15	0.23	-0.50***	-0.09	0.15	-0.04	-0.14	-0.66***	-0.09	0.16	0.06	0	0
	Weak future time reference	-0.23*	-0.10	-0.20*	-0.27**	0.01	0.17	-0.04	-0.04	-0.17	-0.05	-0.09	-0.04	-0.23**	3	/
Successful replications at the 10% level			4	1	7	14	6	7	2	14	16	7(8)	8(9)	5	91(93)	
Successful replications at the 5% level			1	0	5	11	1	6	1	11	12	4	7	0		59

Notes: Entries are two-tailed Pearson correlation coefficients. * $P < 0.10$, ** $P < 0.05$, *** $P < 0.01$; Bold coefficients indicate successful replications at the 10% significance level under our sign- and significance-based criterion. The 10% significance threshold follows the criterion used in the original studies; the last two rows report the number of successful replications under the 10% and 5% thresholds, respectively. 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% (or 5%) level. When replicating the results in Falk and Hermle (2018), we conduct principal component analysis separately for each model's data. Numbers in parentheses report counts when the Protestant-share special case (see text) is included. Because the results are selected based on statistical significance at the 10% level, some are not significant at the 5% level ($0.05 < P < 0.10$), and these cases are indicated by “/.”

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

		GPS	GPT-3.5-turbo	GPT-4o-mini	GPT-4o-mini Eng	GPT-4o	GPT-4.1-mini	GPT-5.4-mini	GPT-5.4-mini Eng	GPT-5.4	GPT-5.4 Eng	DeepSeek-V3	Gemini 2.0 Flash	Gemini 2.5 Flash	10% level replications	5% level replications
Risk-taking	Female	-0.18***	-0.00	-0.16***	-0.10***	-0.35***	-0.30***	-0.38***	-0.45***	-0.33***	-0.28***	-0.31***	-0.34***	-0.23***	11	11
	Sub. Math skills	0.05***	0.22***	0.22***	0.20***	0.20***	0.24***	0.21***	0.22***	0.21***	0.22***	0.20***	0.19***	0.18***	12	12
	Age	-0.19	-0.19	0.79***	0.06	-2.82***	-0.19	-1.31***	-0.95***	0.23	0.61***	0.35**	-4.88***	-2.93***	4	4
	Age square	-1.11***	-0.70***	-2.83***	-2.32***	0.30	-1.90***	-0.79***	-1.30***	-3.00***	-3.41***	-2.89***	1.80***	0.38*		
Patience	Female	-0.07***	0.04**	-0.04***	0.05***	0.02*	-0.11***	-0.07***	-0.07***	0.06***	0.14***	-0.14***	0.07***	0.06***	5	5
	Sub. Math skills	0.03***	0.20***	0.20***	0.29***	0.27***	0.25***	0.17***	0.18***	0.27***	0.27***	0.23***	0.17***	0.23***	12	12
	Age	0.77***	-0.04	1.94***	2.65***	1.73***	1.67***	0.45**	1.97***	2.35***	2.14***	1.09***	1.77***	1.05***	11	11
	Age square	-1.51***	-0.48*	-2.22***	-2.58***	-2.59***	-2.45***	-0.92***	-2.18***	-2.76***	-2.20***	-1.74***	-2.31***	-1.30***		
Negative reciprocity	Female	-0.14***	0.13***	-0.10***	-0.26***	-0.13***	-0.31***	-0.06***	-0.19***	-0.18***	-0.17***	-0.08***	-0.30***	-0.17***	11	11
	Sub. Math skills	0.04***	0.12***	0.09***	0.15***	0.05***	0.16***	0.04***	0.05***	-0.06***	-0.06***	0.04***	0.05***	-0.04***	9	9
	Age	-0.43**	-0.35	1.71***	2.65***	4.07***	3.69***	1.89***	4.11***	1.93***	0.76***	2.66***	0.66***	3.31***	0	0
	Age square	-0.40**	-0.10	-3.64***	-6.11***	-6.88***	-4.57***	-2.73***	-5.64***	-2.52***	-1.71***	-3.89***	-2.46***	-5.48***		
Positive reciprocity	Female	0.05***	-0.01	0.02**	0.05***	0.03*	0.02	0.00	0.13***	-0.03*	-0.00	-0.04**	0.01	0.10***	5	4
	Sub. Math skills	0.04***	0.15***	0.15***	0.18***	0.10***	0.19***	0.11***	0.16***	0.09***	0.10***	0.07***	0.08***	0.01	11	11
	Age	1.14***	0.92***	1.77***	0.85***	2.83***	1.75***	0.89***	0.90***	1.22***	0.84***	1.01***	1.43***	1.40***	9	9
	Age square	-1.28***	-1.01***	-1.45***	-0.28	-2.24***	-1.66***	-0.59***	-0.20	-0.52**	0.72***	-0.49**	-1.05***	-0.75***		
Altruism	Female	0.10***	0.09***	0.28***	0.36***	0.42***	0.24***	0.27***	0.59***	0.40***	0.34***	0.24***	0.45***	0.38***	12	12
	Sub. Math skills	0.04***	0.13***	0.17***	0.25***	0.08***	0.23***	0.15***	0.18***	0.14***	0.17***	0.09***	0.17***	0.12***	12	12
	Age	0.03	-0.54**	0.54**	0.67***	1.84***	-0.62***	-0.80***	0.03	-0.19	-0.92***	-1.08***	0.17	0.74***	0	0
	Age square	-0.01	0.75***	0.40*	1.36***	-0.79**	0.71***	1.07***	0.64**	0.86**	2.08***	1.81***	0.52**	0.36		
Trust	Female	0.07***	0.03	0.26***	0.44***	0.35***	0.25***	0.16***	0.24***	0.24***	0.21***	0.32***	0.19***	0.24***	11	11
	Sub. Math skills	0.05***	0.12***	0.08***	-0.05***	-0.12***	0.04***	0.01	0.02***	-0.07***	-0.05***	-0.01	-0.08***	-0.12***	4	4
	Age	0.36	0.79***	-0.71***	-4.59***	-1.51***	-1.09***	-0.78***	1.94***	-2.30***	-5.11***	-0.86***	-1.28***	-0.07	1	1
	Age square	0.05	-0.72***	1.37***	5.96***	2.07***	1.26***	0.85***	-1.58***	2.45***	5.16***	1.26***	1.65***	0.25		
Successful replications at the 10% level			9	14	12	12	14	12	13	11	9	12	11	11	140	
Successful replications at the 5% level			9	14	12	11	14	12	13	11	9	12	11	11		139

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 ¹. 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 at the 10% significance level. The 10% significance threshold follows the criterion used in the original studies;

the last two rows report the number of successful replications under the 10% and 5% thresholds, respectively. 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.

Section I. 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 ¹¹⁻¹⁴. Building on this work, we investigate whether gender differences in economic preferences are amplified by LLMs. Following the estimation approach of Falk and Hermle (2018), we standardize 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 I.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 than men (12 out of 12 models, Benjamini-Hochberg FDR corrected $P < 0.05$, two-tailed paired *t*-tests) and display higher trust (11 out of 12 models, Benjamini-Hochberg FDR corrected $P < 0.05$, 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 ⁴, in this case, beliefs about women’s prosocial orientations ⁵.

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 [Section K](#) 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. Finally, we extract 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 are uncorrelated with the

intensity of gender stereotypes across countries. Panel A of [Figure I.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 I.1](#) shows the Spearman correlations between these gender-bias measures and the composite gender-stereotype index. All correlations for risk-taking (12/12), patience (12/12), and negative reciprocity (12/12) are negative, while most correlations for positive reciprocity (11/12), altruism (12/12), and trust (11/12) 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, 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 I.1. Differences between LLM-predicted and GPS-surveyed gender gaps in preferences (paired *t*-tests across countries)

	Risk-taking			Patience			Negative reciprocity			Positive reciprocity			Altruism			Trust		
	Diff	95% CI	adj. P	Diff	95% CI	adj. P	Diff	95% CI	adj. P	Diff	95% CI	adj. P	Diff	95% CI	adj. P	Diff	95% CI	adj. P
GPT-3.5-turbo	0.19	[0.15, 0.24]	<0.001	0.12	[0.07, 0.17]	<0.001	0.41	[0.33, 0.49]	<0.001	-0.03	[-0.08, 0.02]	0.280	0.06	[0.00, 0.12]	0.050	-0.04	[-0.09, 0.01]	0.132
GPT-4o-mini	0.01	[-0.04, 0.05]	0.775	0.03	[-0.02, 0.07]	0.318	0.03	[-0.01, 0.08]	0.160	0.00	[-0.04, 0.03]	0.870	0.28	[0.23, 0.32]	<0.001	0.29	[0.24, 0.33]	<0.001
GPT-4o-mini English	0.08	[0.04, 0.12]	<0.001	0.12	[0.08, 0.15]	<0.001	-0.13	[-0.17, -0.09]	<0.001	0.02	[-0.02, 0.05]	0.409	0.28	[0.24, 0.32]	<0.001	0.37	[0.33, 0.40]	<0.001
GPT-4o	-0.23	[-0.27, -0.18]	<0.001	0.09	[0.05, 0.12]	<0.001	-0.01	[-0.07, 0.06]	0.838	0.01	[-0.04, 0.06]	0.705	0.43	[0.36, 0.50]	<0.001	0.35	[0.31, 0.39]	<0.001
GPT-4.1-mini	-0.14	[-0.18, -0.10]	<0.001	-0.05	[-0.09, 0.00]	0.092	-0.27	[-0.34, -0.21]	<0.001	-0.04	[-0.08, 0.01]	0.134	0.18	[0.13, 0.24]	<0.001	0.29	[0.24, 0.35]	<0.001
GPT-5.4-mini	-0.24	[-0.28, -0.20]	<0.001	-0.01	[-0.05, 0.03]	0.719	0.05	[-0.01, 0.10]	0.132	-0.05	[-0.10, -0.01]	0.040	0.27	[0.21, 0.33]	<0.001	0.20	[0.14, 0.26]	<0.001
GPT-5.4-mini English	-0.30	[-0.33, -0.26]	<0.001	0.00	[-0.04, 0.03]	0.843	-0.08	[-0.13, -0.03]	0.006	0.10	[0.07, 0.14]	<0.001	0.53	[0.47, 0.58]	<0.001	0.17	[0.12, 0.21]	<0.001
GPT-5.4	-0.18	[-0.22, -0.13]	<0.001	0.14	[0.10, 0.18]	<0.001	-0.12	[-0.19, -0.04]	0.005	-0.10	[-0.15, -0.05]	<0.001	0.31	[0.25, 0.37]	<0.001	0.28	[0.21, 0.34]	<0.001
GPT-5.4 English	-0.13	[-0.17, -0.08]	<0.001	0.22	[0.18, 0.25]	<0.001	-0.05	[-0.10, 0.01]	0.134	-0.04	[-0.08, -0.00]	0.066	0.22	[0.18, 0.26]	<0.001	0.17	[0.12, 0.21]	<0.001
Gemini 2.0 Flash	-0.19	[-0.23, -0.15]	<0.001	0.15	[0.10, 0.20]	<0.001	-0.22	[-0.28, -0.17]	<0.001	-0.03	[-0.07, 0.02]	0.244	0.45	[0.39, 0.52]	<0.001	0.18	[0.14, 0.22]	<0.001
Gemini 2.5 Flash	-0.08	[-0.13, -0.02]	0.012	0.13	[0.10, 0.17]	<0.001	-0.05	[-0.12, 0.01]	0.141	0.13	[0.09, 0.17]	<0.001	0.39	[0.34, 0.45]	<0.001	0.23	[0.19, 0.28]	<0.001
DeepSeek-V3	-0.17	[-0.21, -0.12]	<0.001	-0.07	[-0.12, -0.02]	0.008	0.02	[-0.06, 0.10]	0.686	-0.09	[-0.15, -0.04]	0.003	0.24	[0.17, 0.32]	<0.001	0.44	[0.37, 0.51]	<0.001

Notes: The number of countries is 74, except for altruism in DeepSeek-V3, where $n = 73$. Diff 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). The *P*-values are adjusted by Benjamini-Hochberg FDR method. Positive values indicate that LLMs predict women to have higher preference levels than those observed in the GPS survey data.

Table I.2. Associations between gender gaps and gender-stereotype index.

	Risk-taking			Patience			Negative reciprocity			Positive reciprocity			Altruism			Trust		
	<i>Rho</i>	adj. P	95% CI	<i>Rho</i>	adj. P	95% CI	<i>Rho</i>	adj. P	95% CI	<i>Rho</i>	adj. P	95% CI	<i>Rho</i>	adj. P	95% CI	<i>Rho</i>	adj. P	95% CI
GPT-3.5-turbo	-0.06	0.781	[-0.33, 0.21]	-0.27	0.176	[-0.50, -0.00]	-0.06	0.781	[-0.33, 0.22]	0.03	0.886	[-0.22, 0.29]	-0.25	0.222	[-0.49, 0.03]	-0.12	0.578	[-0.39, 0.15]
GPT-4o-mini	-0.19	0.354	[-0.45, 0.08]	-0.6	<0.001	[-0.77, -0.38]	-0.13	0.546	[-0.39, 0.14]	0.16	0.417	[-0.09, 0.41]	0.04	0.881	[-0.23, 0.29]	-0.21	0.294	[-0.44, 0.07]
GPT-4o-mini English	-0.52	<0.001	[-0.71, -0.27]	-0.43	0.008	[-0.64, -0.20]	-0.59	<0.001	[-0.76, -0.34]	0.42	0.011	[0.17, 0.62]	0.36	0.035	[0.10, 0.59]	-0.01	0.956	[-0.30, 0.27]
GPT-4o	-0.13	0.535	[-0.42, 0.14]	-0.11	0.608	[-0.36, 0.16]	-0.18	0.362	[-0.44, 0.10]	-0.06	0.781	[-0.33, 0.20]	-0.18	0.362	[-0.45, 0.10]	-0.51	<0.001	[-0.71, -0.25]
GPT-4.1-mini	0.23	0.242	[-0.04, 0.50]	-0.21	0.294	[-0.46, 0.06]	-0.02	0.94	[-0.28, 0.28]	0.01	0.951	[-0.28, 0.29]	0.24	0.222	[-0.03, 0.50]	-0.07	0.777	[-0.34, 0.23]
GPT-5.4-mini	0.09	0.702	[-0.19, 0.38]	-0.03	0.916	[-0.27, 0.21]	-0.23	0.242	[-0.47, 0.03]	0.00	0.971	[-0.25, 0.26]	-0.18	0.365	[-0.44, 0.11]	0.12	0.549	[-0.13, 0.35]
GPT-5.4-mini English	0.18	0.362	[-0.10, 0.47]	0.28	0.176	[0.01, 0.52]	-0.39	0.017	[-0.61, -0.14]	-0.22	0.28	[-0.48, 0.08]	0.40	0.016	[0.14, 0.63]	0.28	0.17	[0.00, 0.52]
GPT-5.4	0.04	0.849	[-0.23, 0.32]	0.14	0.51	[-0.13, 0.39]	-0.11	0.578	[-0.38, 0.16]	0.20	0.335	[-0.03, 0.41]	-0.26	0.213	[-0.48, -0.00]	-0.13	0.541	[-0.36, 0.12]
GPT-5.4 English	-0.08	0.764	[-0.38, 0.24]	0.36	0.035	[0.12, 0.56]	-0.66	<0.001	[-0.80, -0.46]	0.16	0.407	[-0.14, 0.44]	0.23	0.242	[-0.05, 0.47]	0.43	0.008	[0.18, 0.63]
Gemini 2.0 Flash	-0.07	0.774	[-0.33, 0.23]	-0.19	0.354	[-0.45, 0.07]	-0.44	0.008	[-0.65, -0.21]	-0.14	0.501	[-0.42, 0.16]	-0.24	0.222	[-0.49, 0.03]	-0.07	0.774	[-0.34, 0.19]
Gemini 2.5 Flash	-0.11	0.578	[-0.39, 0.18]	0.28	0.17	[-0.01, 0.52]	-0.24	0.24	[-0.47, 0.02]	-0.01	0.951	[-0.26, 0.25]	-0.17	0.378	[-0.42, 0.13]	-0.26	0.213	[-0.53, 0.02]
DeepSeek-V3	-0.05	0.792	[-0.32, 0.21]	-0.07	0.774	[-0.35, 0.20]	0.17	0.378	[-0.09, 0.43]	0.06	0.781	[-0.23, 0.33]	-0.24	0.222	[-0.51, 0.03]	-0.18	0.362	[-0.42, 0.09]

Notes: All correlations are computed across 56 countries with non-missing data on both the gender gap 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. The *Rho* column shows the Spearman correlation between these country-level gender gaps and the composite gender-stereotype index. The *P*-values are adjusted by Benjamini-Hochberg FDR method.

Table I.3. Associations between gender biases and gender-stereotype index.

	Risk-taking			Patience			Negative reciprocity			Positive reciprocity			Altruism			Trust		
	<i>Rho</i>	adj. <i>P</i>	95% CI	<i>Rho</i>	adj. <i>P</i>	95% CI	<i>Rho</i>	adj. <i>P</i>	95% CI	<i>Rho</i>	adj. <i>P</i>	95% CI	<i>Rho</i>	adj. <i>P</i>	95% CI	<i>Rho</i>	adj. <i>P</i>	95% CI
GPT-3.5-turbo	-0.25	0.129	[-0.49, 0.03]	-0.43	0.005	[-0.61, -0.18]	-0.19	0.250	[-0.46, 0.10]	0.14	0.392	[-0.12, 0.41]	0.08	0.635	[-0.19, 0.34]	0.23	0.154	[-0.05, 0.50]
GPT-4o-mini	-0.32	0.044	[-0.57, -0.04]	-0.65	<0.001	[-0.82, -0.43]	-0.38	0.014	[-0.60, -0.12]	0.31	0.054	[0.05, 0.55]	0.33	0.036	[0.09, 0.57]	0.12	0.454	[-0.13, 0.40]
GPT-4o-mini English	-0.45	0.003	[-0.65, -0.19]	-0.56	<0.001	[-0.72, -0.35]	-0.60	<0.001	[-0.78, -0.38]	0.48	0.001	[0.28, 0.64]	0.63	<0.001	[0.39, 0.80]	0.39	0.013	[0.10, 0.60]
GPT-4o	-0.22	0.180	[-0.49, 0.06]	-0.37	0.015	[-0.60, -0.11]	-0.30	0.062	[-0.54, -0.03]	0.15	0.373	[-0.10, 0.39]	0.07	0.659	[-0.22, 0.34]	-0.08	0.635	[-0.35, 0.19]
GPT-4.1-mini	-0.13	0.444	[-0.40, 0.18]	-0.37	0.016	[-0.59, -0.12]	-0.28	0.080	[-0.55, 0.01]	0.15	0.373	[-0.11, 0.38]	0.46	0.002	[0.23, 0.64]	0.13	0.444	[-0.17, 0.40]
GPT-5.4-mini	-0.14	0.411	[-0.43, 0.19]	-0.35	0.023	[-0.57, -0.08]	-0.44	0.003	[-0.65, -0.20]	0.11	0.511	[-0.17, 0.36]	0.11	0.519	[-0.18, 0.37]	0.40	0.010	[0.12, 0.62]
GPT-5.4-mini English	-0.19	0.250	[-0.48, 0.11]	-0.19	0.250	[-0.45, 0.06]	-0.58	<0.001	[-0.77, -0.34]	0.01	0.943	[-0.26, 0.29]	0.56	<0.001	[0.35, 0.73]	0.47	0.002	[0.23, 0.66]
GPT-5.4	-0.18	0.267	[-0.46, 0.12]	-0.23	0.164	[-0.48, 0.04]	-0.28	0.080	[-0.53, -0.01]	0.3	0.058	[0.06, 0.52]	0.07	0.659	[-0.19, 0.33]	0.17	0.312	[-0.10, 0.40]
GPT-5.4 English	-0.28	0.080	[-0.54, 0.02]	-0.18	0.269	[-0.46, 0.10]	-0.74	<0.001	[-0.84, -0.58]	0.25	0.122	[-0.05, 0.48]	0.52	<0.001	[0.31, 0.70]	0.57	<0.001	[0.33, 0.74]
Gemini 2.0 Flash	-0.28	0.080	[-0.53, 0.01]	-0.37	0.015	[-0.57, -0.13]	-0.61	<0.001	[-0.78, -0.40]	-0.03	0.883	[-0.30, 0.25]	0.12	0.454	[-0.14, 0.38]	0.3	0.062	[0.04, 0.53]
Gemini 2.5 Flash	-0.25	0.125	[-0.50, 0.04]	-0.01	0.966	[-0.29, 0.27]	-0.38	0.014	[-0.59, -0.14]	0.14	0.398	[-0.11, 0.39]	0.21	0.212	[-0.05, 0.47]	0.03	0.865	[-0.25, 0.31]
DeepSeek-V3	-0.22	0.180	[-0.49, 0.07]	-0.26	0.117	[-0.49, -0.00]	-0.02	0.935	[-0.31, 0.27]	0.15	0.368	[-0.13, 0.42]	0.01	0.966	[-0.28, 0.30]	0.07	0.659	[-0.21, 0.37]

Notes: All correlations are computed across 56 countries with non-missing data on both the gender bias 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). 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 – GPS gap). The *Rho* column shows the Spearman correlation between these country-level gender biases and the composite gender-stereotype index. The *P*-values are adjusted by Benjamini-Hochberg FDR method.

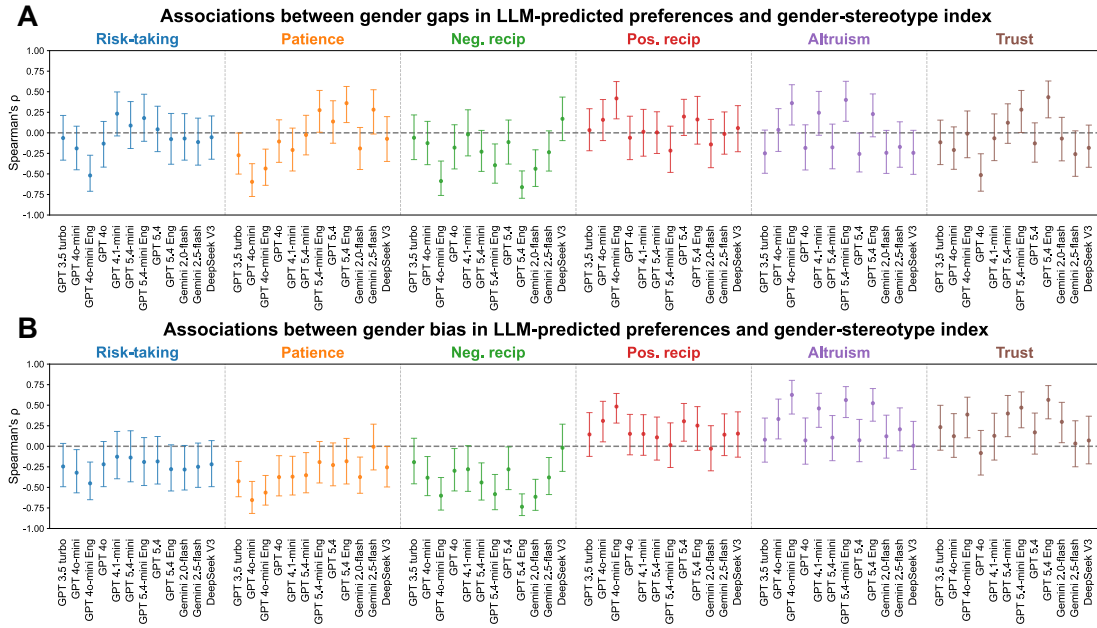


Figure I.1. Associations between gender gaps, gender biases, and gender-stereotype index. **(A) Gender gaps in LLM-predicted preferences and the gender-stereotype index.** For each model and preference domain, we estimate country-specific gender gaps in preferences using the regression specification 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 the 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 – GPS gap). Each point in Panel B reports the Spearman correlation between these gender-bias measures and the composite gender-stereotype index. All correlations are computed across 56 countries with non-missing data on both the gender gap, gender bias and the gender-stereotype index. Error bars indicate 95% bootstrap CIs in both panels.

Section J. 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”¹⁵. 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 data⁴. 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 0.224 (risk-taking: 0.218; patience: 0.376; negative reciprocity: 0.112; positive reciprocity: 0.147; altruism: 0.083; trust: 0.411). [Figure J.1](#) reports diversity ratios for each of the six preference domains and 12 models. All 72 model-preference combinations (6 domains $\times$ 12 LLMs) have mean ratios significantly below 1 (two-sided t -tests: $t(71) = -43.611$, $P < 0.001$, Cohen’s $d = -5.140$, 95% CI = [0.19, 0.26]; see [Table J.1](#)), indicating that LLMs never reach human levels of response diversity. Across all six domains, most ratios are below 0.5 (68 of 72 combinations), and for altruism all 12 ratios are below 0.2. Thus, LLM-generated responses are substantially less diverse than human responses. The detailed statistics are reported in [Table J.1](#).

Among the six preference domains, trust provides the cleanest comparison for two reasons. First, the GPS imputes missing values whenever a preference index is constructed from multiple items and does not flag which observations are imputed, so the count of distinct values for the multi-item preferences (risk-taking, patience, negative reciprocity,

positive reciprocity, altruism) confounds genuine response diversity with imputation artifacts. Trust, by contrast, is based on a single survey question and is therefore not subject to GPS imputation. Second, because trust is not imputed, we can recover the underlying raw responses by converting both LLM-predicted and GPS-surveyed normalized scores back to the original 0–10 scale, yielding 11 directly comparable response categories.

Even in this cleanest domain, the largest diversity ratio (for GPT-3.5-turbo) is 0.647, implying that the model covers, on average, about 7 of the 11 available response categories. Combined with earlier work showing that LLMs approximate “average-human” behavior while generating markedly less diverse responses ^{4,16}, our one-to-one country-model setup strengthens this conclusion: even where the diversity comparison is cleanest and the LLM performs best, response coverage is still substantially below the human benchmark.

Following Bisbee et al.(2024), we compute standard deviations for human and silicon respondents. As shown in [Figure J.2](#), the results are consistent with the homogenization pattern reported in the literature: LLM-predicted trust responses show substantially lower standard deviations (ranging from 0.295 to 1.690 across models) compared with the GPS benchmark (SD = 2.740). This confirms that LLMs compress response variation, even for the trust domain, where their rank-order accuracy is weakest.

We next examine how response diversity varies with country-level variables. [Tables J.2](#) and [J.3](#) report standardized regression coefficients from country-level regressions whose dependent variable is the diversity index (standardized to mean 0, SD 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 for the log of the GPS sample size and include model fixed effects. In [Table J.2](#), we additionally include preference-domain fixed effects and model-by-preference interaction effects.

When all preference domains are pooled (Table J.2), 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 0.161-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 0.147-standard-deviation decline ($P < 0.001$; column 3) when controlling for 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 0.043-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 0.039-standard-deviation decrease ($P < 0.001$; column 2), when controlling for 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 economic development indicators and on the two population indicators, and extract the first principal component from each set. In Column (5) of Table J.2, the first component of the economic development indicators is strongly negatively associated with the diversity index ($b = -0.169$, $P < 0.001$), whereas the first component of the population indicators has a much smaller negative coefficient ($b = -0.039$, $P < 0.001$). 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. Table J.3 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 -0.120 to -0.128 ($P < 0.001$; columns 1 and 2) and from -0.073 to -0.072 ($P < 0.05$; 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 -0.042 and -0.039 ($P < 0.10$) 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 (-0.105 , $P < 0.001$) than the population component (-0.039 , $P = 0.081$).

These results indicate that LLMs tend to produce less diverse responses when simulating respondents from more economically developed countries, while generating somewhat more varied responses for less developed countries. 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.

As described in the [Methods](#) ($T = 0$), a natural question is whether moderate temperature settings (e.g., $T = 0.7$ or $T = 1.0$) would yield diversity closer to the human benchmark. To answer this question, we conduct an experiment at a higher temperature ($T = 1.0$) with GPT-5.4-mini in localized language. We first verify that the higher temperature does not materially affect alignment between LLM-predicted preferences and GPS-surveyed preferences (two-sided paired t -test between GPT-5.4-mini at $T = 0$ and $T = 1$; individual-level correlations: $t(5) = -1.131$, $P = 0.309$, Cohen's $d = -0.040$, 95% CI = $[-0.01, 0.00]$; country-level correlations: $t(5) = 0.938$, $P = 0.391$, Cohen's $d = 0.067$, 95% CI = $[-0.02, 0.05]$; individual-level MSEs: $t(5) = -0.072$, $P = 0.945$, Cohen's $d = -0.004$, 95% CI = $[-0.02, 0.02]$; country-level MSEs: $t(5) = -1.090$, $P = 0.326$, Cohen's $d = -0.089$, 95% CI = $[-0.10, 0.04]$). Detailed correlation coefficients and MSEs are reported in [Table J.4](#).

Turning to diversity, [Figure J.3](#) shows that the diversity ratio increases from 0.187 (GPT-5.4-mini, $T = 0$) to 0.288 (GPT-5.4-mini, $T = 1$), a significant increase (two-sided paired t -test: $t(443) = 22.932$, $P < 0.001$, Cohen's $d = 0.529$, 95% CI = $[0.09, 0.11]$; pooling all countries and preferences). However, the diversity ratio at the high-temperature remains significantly below 1 (for each preference, Benjamini-Hochberg FDR corrected $P < 0.001$, see [Table J.5](#)). Together, these results indicate that although raising temperature does increase output diversity, it does not bring LLM-generated responses to diversity levels comparable to those of human respondents.

This pattern is consistent with recent systematic evaluations reporting that temperature variation has limited effects on response-distribution statistics and behavioral-task outcomes once persona conditioning and decoding strategy are held fixed

4,17–19.

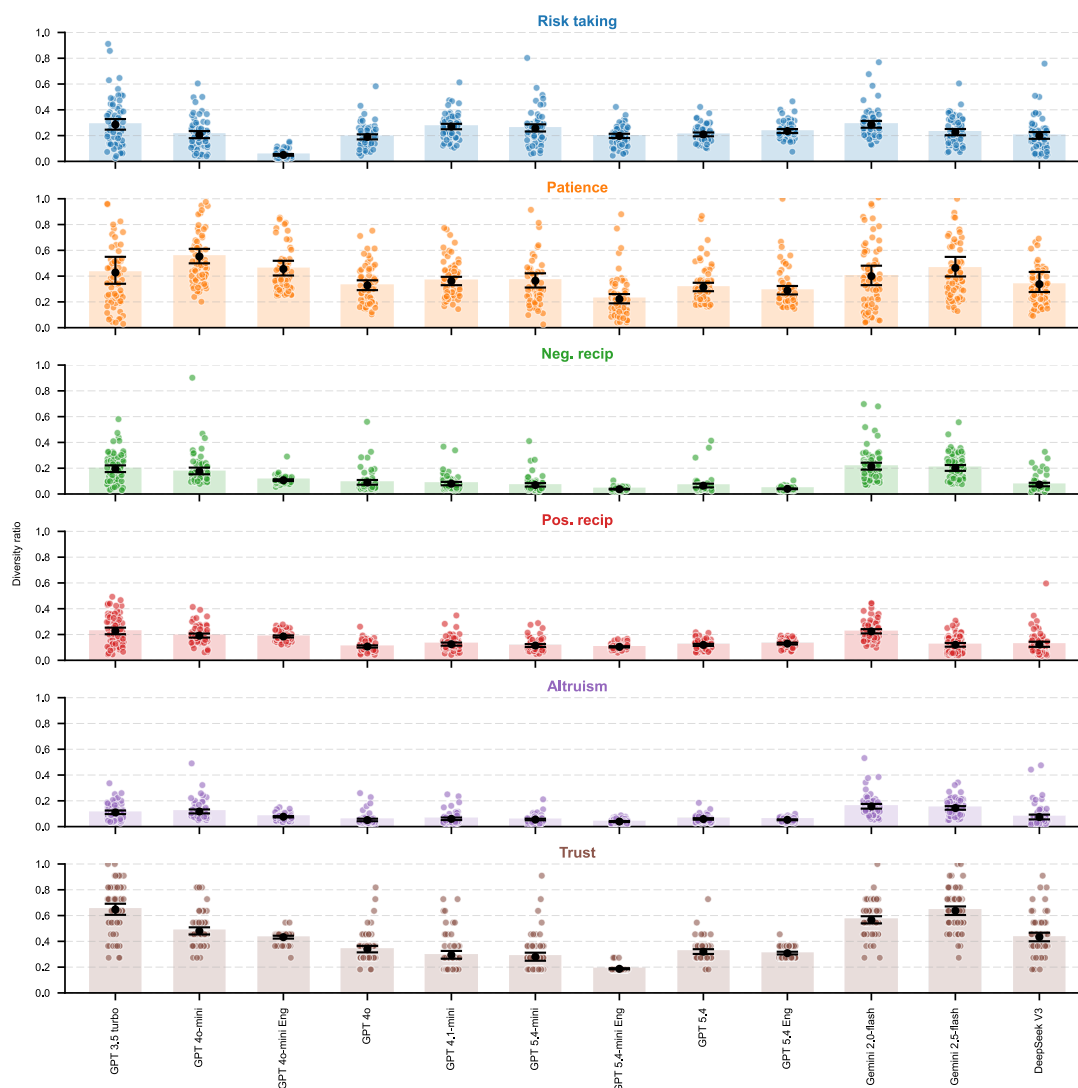


Figure J.1. Diversity ratios across preference domains and models. Bars show the mean diversity ratio ($n = 74$) across countries for each combination of preference domain and LLM. The diversity ratio is defined as the number of distinct values in the LLM-generated preference index divided by the number 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 95% bootstrapped CIs.

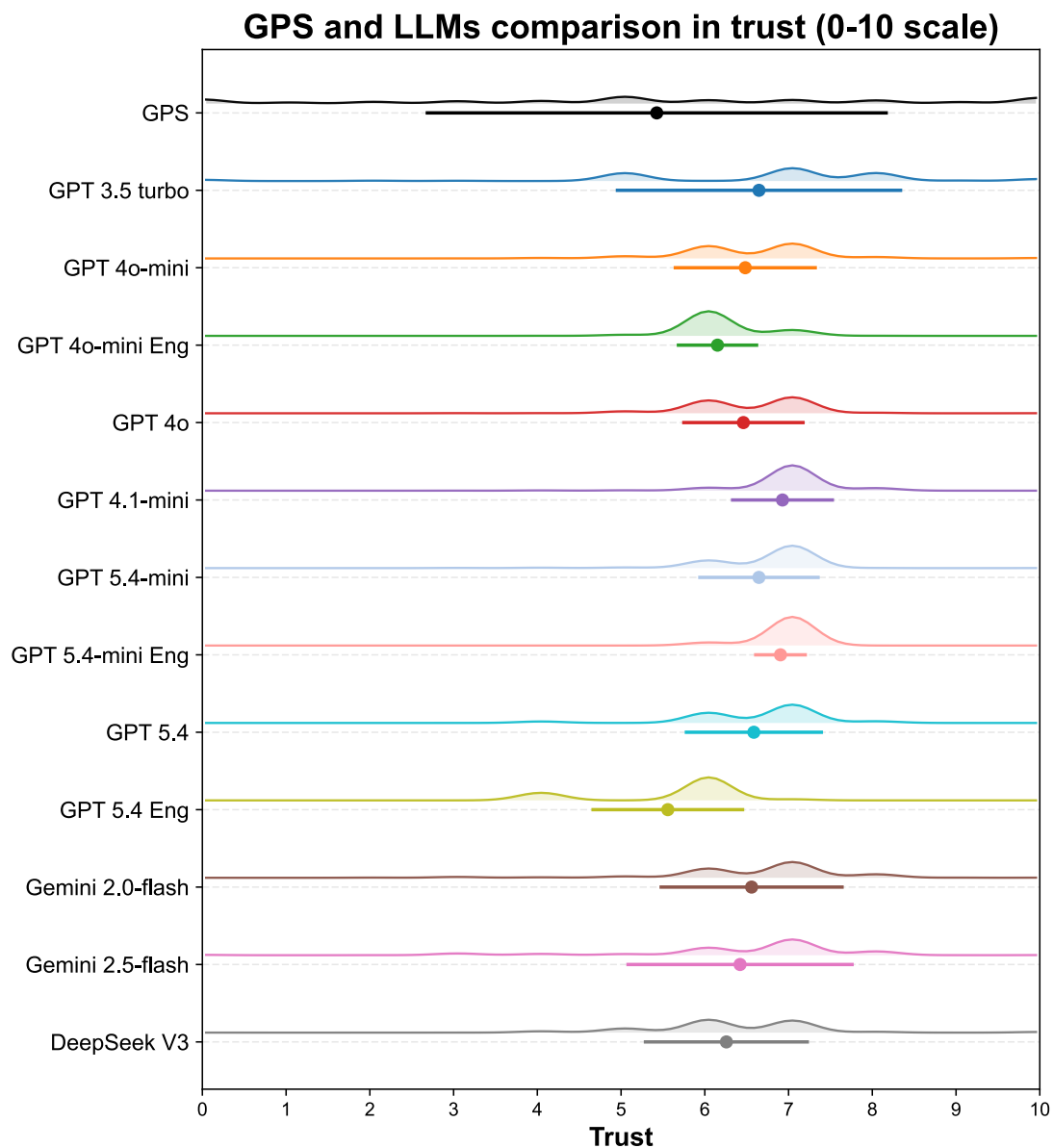


Figure J.2. Standard deviations of trust responses across LLMs and GPS. Bars show the individual-level standard deviation of trust responses for each model and the GPS benchmark, computed by pooling all individual observations across the 74 countries. Sample sizes are matched across models and GPS ($n = 65,852$).

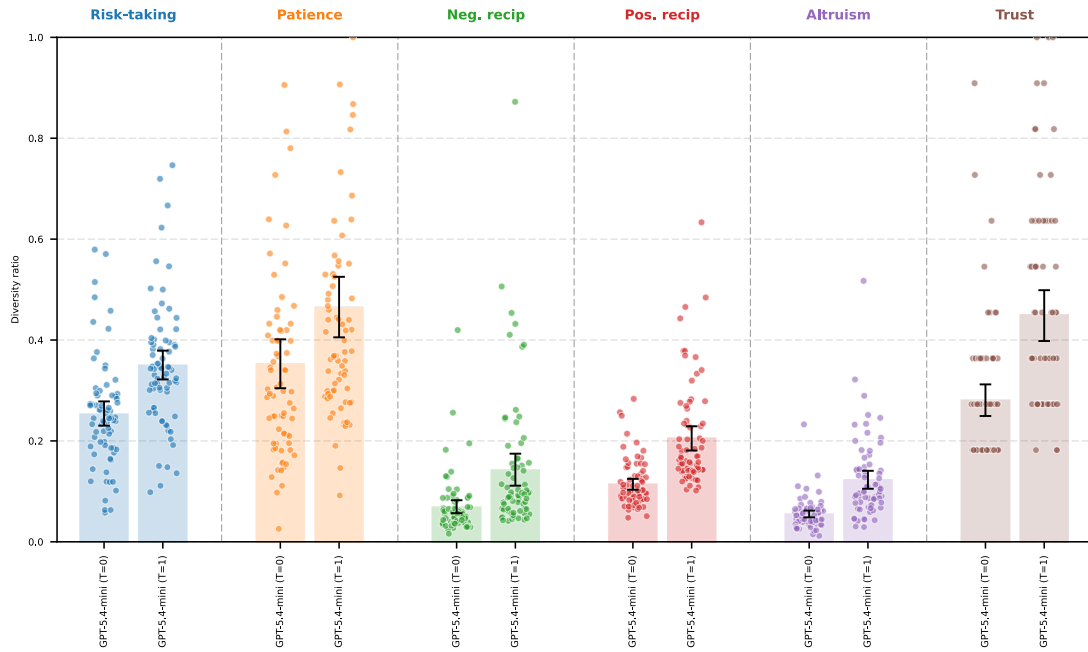


Figure J.3. Diversity ratios for GPT-5.4-mini at temperature 0 and temperature 1, across preference domains. Bars show the mean diversity ratio ($n = 74$) across countries for each combination of preference domain and temperature setting ($T = 0$ vs $T = 1$), generated by GPT-5.4-mini. The diversity 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 the model does not reach human levels of response diversity. Error bars indicate 95% bootstrapped CIs.

Table J.1. The diversity ratio across each model and each preference.

	Risk-taking			Patience			Negative reciprocity			Positive reciprocity			Altruism			Trust		
	Mean	adj. P	95% CI	Mean	adj. P	95% CI	Mean	adj. P	95% CI	Mean	adj. P	95% CI	Mean	adj. P	95% CI	Mean	adj. P	95% CI
GPT-3.5-turbo	0.29	<0.001	[0.25, 0.33]	0.43	<0.001	[0.34, 0.55]	0.20	<0.001	[0.17, 0.22]	0.23	<0.001	[0.20, 0.25]	0.11	<0.001	[0.10, 0.12]	0.65	<0.001	[0.61, 0.69]
GPT-4o-mini	0.21	<0.001	[0.18, 0.23]	0.55	<0.001	[0.50, 0.61]	0.18	<0.001	[0.15, 0.20]	0.19	<0.001	[0.18, 0.21]	0.12	<0.001	[0.10, 0.13]	0.48	<0.001	[0.45, 0.51]
GPT-4o-mini English	0.05	<0.001	[0.04, 0.06]	0.46	<0.001	[0.40, 0.52]	0.11	<0.001	[0.10, 0.11]	0.18	<0.001	[0.18, 0.19]	0.08	<0.001	[0.07, 0.08]	0.43	<0.001	[0.42, 0.44]
GPT-4o	0.19	<0.001	[0.17, 0.21]	0.33	<0.001	[0.29, 0.37]	0.09	<0.001	[0.07, 0.11]	0.11	<0.001	[0.10, 0.12]	0.05	<0.001	[0.04, 0.06]	0.34	<0.001	[0.31, 0.36]
GPT-4.1-mini	0.27	<0.001	[0.25, 0.29]	0.36	<0.001	[0.33, 0.39]	0.08	<0.001	[0.07, 0.09]	0.13	<0.001	[0.11, 0.14]	0.06	<0.001	[0.05, 0.07]	0.29	<0.001	[0.27, 0.33]
GPT-5.4-mini	0.26	<0.001	[0.23, 0.29]	0.36	<0.001	[0.31, 0.42]	0.07	<0.001	[0.06, 0.08]	0.11	<0.001	[0.10, 0.13]	0.05	<0.001	[0.05, 0.06]	0.28	<0.001	[0.25, 0.31]
GPT-5.4-mini English	0.20	<0.001	[0.18, 0.21]	0.22	<0.001	[0.19, 0.26]	0.04	<0.001	[0.03, 0.04]	0.10	<0.001	[0.10, 0.11]	0.04	<0.001	[0.04, 0.04]	0.19	<0.001	[0.18, 0.19]
GPT-5.4	0.21	<0.001	[0.19, 0.22]	0.31	<0.001	[0.28, 0.35]	0.06	<0.001	[0.05, 0.08]	0.12	<0.001	[0.11, 0.13]	0.06	<0.001	[0.05, 0.07]	0.32	<0.001	[0.30, 0.34]
GPT-5.4 English	0.24	<0.001	[0.22, 0.25]	0.29	<0.001	[0.26, 0.32]	0.04	<0.001	[0.04, 0.04]	0.13	<0.001	[0.12, 0.14]	0.05	<0.001	[0.05, 0.06]	0.31	<0.001	[0.30, 0.32]
Gemini 2.0 Flash	0.29	<0.001	[0.26, 0.31]	0.40	<0.001	[0.33, 0.48]	0.21	<0.001	[0.19, 0.24]	0.22	<0.001	[0.21, 0.24]	0.16	<0.001	[0.14, 0.18]	0.57	<0.001	[0.54, 0.59]
Gemini 2.5 Flash	0.23	<0.001	[0.20, 0.25]	0.46	<0.001	[0.40, 0.55]	0.20	<0.001	[0.18, 0.22]	0.12	<0.001	[0.11, 0.13]	0.14	<0.001	[0.13, 0.16]	0.64	<0.001	[0.60, 0.67]
DeepSeek-V3	0.20	<0.001	[0.17, 0.23]	0.34	<0.001	[0.28, 0.43]	0.07	<0.001	[0.06, 0.09]	0.12	<0.001	[0.10, 0.14]	0.07	<0.001	[0.06, 0.09]	0.43	<0.001	[0.40, 0.47]

Notes: The number of countries is 74. The mean column shows 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. We use two-sided individual *t*-tests to compare the difference between the mean diversity ratio and 1. The *P*-values are adjusted by Benjamini-Hochberg FDR method, and the 95% CIs are computed from bootstrapped method.

Table J.2. 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]	-0.161*** (<0.001) [-0.181, -0.142]	-0.167*** (<0.001) [-0.189, -0.145]			
std. IDI			-0.147*** (<0.001) [-0.170, -0.124]	-0.149*** (<0.001) [-0.175, -0.123]	
std. log population	-0.043*** (<0.001) [-0.060, -0.026]		-0.037*** (<0.001) [-0.055, -0.020]		
std. log internet users		-0.039*** (<0.001) [-0.056, -0.022]		-0.036*** (<0.001) [-0.054, -0.019]	
std. score for economic development					-0.169*** (<0.001) [-0.194, -0.145]
std. score for population					-0.039*** (<0.001) [-0.057, -0.022]
Observations	5256	4896	5184	4896	4896
R^2	0.590	0.589	0.588	0.584	0.588

Notes: The dependent variable is the standardized diversity index (mean 0, standard deviation 1), pooled across all six preference domains and all models. 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 P -values in parentheses and 95% CIs in brackets. Regressions include controls for the log of the GPS sample size in each country, preference-domain fixed effects, model fixed effects, and model $\times$ preference interaction effects. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$

Table J.3. 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]	-0.128*** (<0.001) [-0.176, -0.079]	-0.120*** (<0.001) [-0.169, -0.071]			
std. IDI			-0.072* (0.011) [-0.127, -0.016]	-0.073** (0.010) [-0.128, -0.017]	
std. log population	-0.042 (0.051) [-0.083, 0.000]		-0.029 (0.192) [-0.072, 0.014]		
std. log internet users		-0.039 (0.085) [-0.082, 0.005]		-0.034 (0.130) [-0.078, 0.010]	
std. score for economic development					-0.105*** (<0.001) [-0.159, -0.052]
std. score for population					-0.039 (0.081) [-0.082, 0.005]
Observations	876	816	864	816	816
R^2	0.602	0.615	0.597	0.607	0.611

Notes: The dependent variable is the standardized diversity index (mean 0, standard deviation 1), pooled across all models. The sample is restricted to the trust domain. 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 P -values in parentheses and 95% CIs in brackets. Regressions include controls for the log of the GPS sample size in each country and model fixed effects. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.

Table J.4. Per-preference and average alignment metrics for GPT-5.4-mini at $T = 0$ vs $T = 1$

Model	Risk-taking	Patience	Negative reciprocity	Positive reciprocity	Altruism	Trust	Average
Individual-level Spearman's correlations							
GPT-5.4-mini ($T = 0$)	0.227	0.102	0.028	0.091	0.056	0.075	0.097
GPT-5.4-mini ($T = 1$)	0.229	0.098	0.031	0.081	0.059	0.066	0.094
Country-level Spearman's correlations							
GPT-5.4-mini ($T = 0$)	0.262	0.026	-0.026	0.215	-0.131	0.319	0.111
GPT-5.4-mini ($T = 1$)	0.340	0.025	-0.037	0.218	-0.121	0.314	0.123
Individual-level MSEs							
GPT-5.4-mini ($T = 0$)	1.572	1.826	1.990	1.811	1.939	1.875	1.836
GPT-5.4-mini ($T = 1$)	1.558	1.834	1.979	1.835	1.918	1.886	1.835
Country-level MSEs							
GPT-5.4-mini ($T = 0$)	1.612	1.734	2.142	1.475	2.199	1.471	1.772
GPT-5.4-mini ($T = 1$)	1.452	1.738	2.159	1.475	2.158	1.473	1.743

Notes: Both the individual-level and country-level preferences are normalized to mean 0 and standard deviation 1 before computing the MSEs.

Table J.5. Per-preference diversity ratio for GPT-5.4-mini at $T = 0$ vs $T = 1$

Preference	Temperature	Diversity ratio	Adjusted P-value	95% CI
Risk-taking	$T = 0$	0.25	<0.001	[0.23, 0.28]
	$T = 1$	0.35	<0.001	[0.32, 0.38]
Patience	$T = 0$	0.35	<0.001	[0.30, 0.40]
	$T = 1$	0.46	<0.001	[0.41, 0.53]
Negative reciprocity	$T = 0$	0.07	<0.001	[0.06, 0.08]
	$T = 1$	0.14	<0.001	[0.11, 0.17]
Positive reciprocity	$T = 0$	0.11	<0.001	[0.10, 0.12]
	$T = 1$	0.21	<0.001	[0.18, 0.23]
Altruism	$T = 0$	0.05	<0.001	[0.05, 0.06]
	$T = 1$	0.12	<0.001	[0.11, 0.14]
Trust	$T = 0$	0.28	<0.001	[0.25, 0.31]
	$T = 1$	0.45	<0.001	[0.40, 0.50]

Notes: The number of countries is 74. 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. We use two-sided individual t -tests to compare the difference between the mean diversity ratio and 1. The P -values are adjusted by Benjamini-Hochberg FDR method, and the 95% CIs are computed from bootstrapped method.

Section K. Variable Definitions

Log [GDP per capita]. Natural logarithm of GDP per capita in 2022, taken from the World Bank Development Indicators. In [Section 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 Values Survey website (<https://www.worldvaluessurvey.org/WVSEVSjoint2017.jsp>). We use the Joint EVS/WVS 2017-2022 dataset (v5.0; June 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 component 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 Inequality 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, take 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 [Section I](#).

Hofstede cultural dimensions. Six cultural dimensions (version 2015-12-08) in Hofstede's Cultural Theory taken from the 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.researchgate.net/publication/304715744_The_7_Schwartz_cultural_value_orientation_scores_for_80_countries).

Section L. Parametric and Non-parametric Tests

As described in [Methods](#) (Statistical analysis), we use Welch's t -tests, which do not assume equal variances, for comparisons between independent samples. Because normality cannot be guaranteed for small samples, we repeat all t -tests using nonparametric alternatives (Wilcoxon rank-sum tests for unpaired comparisons and Wilcoxon signed-rank tests for paired comparisons). Binomial and chi-square tests do not rely on normality assumptions. For binomial tests, the null hypothesis is that the success probability equals 0.5. Unless otherwise specified, all tests were two-sided, with no adjustment for multiple comparisons.

[Table L.1](#) and [Table L.2](#) report the inferential statistics for the analyses presented in the main text and the Supplementary Information, respectively. The first column presents results from the parametric analyses, and the second column presents the corresponding non-parametric results. The third column assesses consistency between the two approaches: a check mark (✓) indicates that the estimated effects have the same direction and the same significance classification (that is, both $P < 0.05$ or both $P \geq 0.05$).

Table L1. Comparison of parametric and non-parametric estimation result in the Main text

Inference in the main text		
Parametric test result	Non-parametric test result	Result consistency
Turning to model comparisons, we find no evidence that newer models like GPT-5.4 and GPT-5.4-mini outperform earlier OpenAI models (i.e., GPT-3.5-turbo, GPT-4o-mini, GPT-4.1-mini, and GPT-4o; two-sided Welch's t -test, $t(41.32) = 0.851$, Benjamini-Hochberg FDR corrected $P = 0.726$, Cohen's $d = 0.241$, 95% CI = [-0.05, 0.12]) or that Gemini-2.5 Flash outperforms Gemini-2.0 Flash (two-sided Welch's t -test, $t(9.83) = 0.612$, Benjamini-Hochberg FDR corrected $P = 0.726$, Cohen's $d = 0.353$, 95% CI = [-0.15, 0.26]).	Turning to model comparisons, we find no evidence that newer models like GPT-5.4 and GPT-5.4-mini outperform earlier OpenAI models (i.e., GPT-3.5-turbo, GPT-4o-mini, GPT-4.1-mini, and GPT-4o; two-sided Wilcoxon rank-sum test, Benjamini-Hochberg FDR corrected $P = 0.515$) or that Gemini-2.5 Flash outperforms Gemini-2.0 Flash (two-sided Wilcoxon rank-sum test, Benjamini-Hochberg FDR corrected $P = 0.515$).	✓
The mean country-level MSE across preferences is 1.722, slightly lower than its individual-level counterpart ($M_{\text{country}} = 1.722$ vs $M_{\text{individual}} = 1.827$, two-sided paired t -test, $t(71) = -3.744$, $P < 0.001$, Cohen's $d = -0.425$, 95% CI = [-0.16, -0.05]).	The mean country-level MSE across preferences is 1.722, slightly lower than its individual-level counterpart ($M_{\text{country}} = 1.722$ vs $M_{\text{individual}} = 1.827$, two-sided Wilcoxon signed-rank test, $P = 0.001$).	✓
This analysis shows that LLM-predicted preferences explain only 1.3% of the individual-level variance and 4.2% of the country-level variance in human preferences (two-sided paired t -test, $t(71) = 5.196$, $P < 0.001$, Cohen's $d = 0.722$, 95% CI = [0.02, 0.04]).	This analysis shows that LLM-predicted preferences explain only 1.3% of the individual-level variance and 4.2% of the country-level variance in human preferences (two-sided Wilcoxon signed-rank test, $P < 0.001$).	✓
The mean individual-level between-model MSE across preferences is 1.377 (risk-taking: 0.770; patience: 1.049; negative reciprocity: 1.679; positive reciprocity: 1.579; altruism: 1.329; trust: 1.856), lower than the corresponding LLM-GPS MSE of 1.827 (two-sided Welch's t -test: $t(313.31) = -16.569$, $P < 0.001$, Cohen's $d = -1.158$, 95% CI = [-0.50, -0.40]).	The mean individual-level between-model MSE across preferences is 1.377 (risk-taking: 0.770; patience: 1.049; negative reciprocity: 1.679; positive reciprocity: 1.579; altruism: 1.329; trust: 1.856), lower than the corresponding LLM-GPS MSE of 1.827 (two-sided Wilcoxon rank-sum test, $P < 0.001$).	✓
The mean country-level between-model MSE is 1.491 (risk-taking: 1.423; patience: 1.500; negative reciprocity: 1.628; positive reciprocity: 1.485; altruism: 1.199; trust: 1.709), lower than the country-level LLM-GPS MSE of 1.722 (two-sided Welch's t -test, $t(115.62) = -5.468$, $P < 0.001$, Cohen's $d = -0.600$, 95% CI = [-0.32, -0.15]).	The mean country-level between-model MSE is 1.491 (risk-taking: 1.423; patience: 1.500; negative reciprocity: 1.628; positive reciprocity: 1.485; altruism: 1.199; trust: 1.709), lower than the country-level LLM-GPS MSE of 1.722 (two-sided Wilcoxon rank-sum test, $P < 0.001$).	✓

Notes: The first column presents results from parametric analyses, and the second column presents the corresponding non-parametric results. The third column assesses consistency between the two approaches: a consistency mark (✓) indicates that the estimated effects have the same direction and the same significance classification (that is, both $P < 0.05$ or both $P \geq 0.05$).

Table L2. Comparison of parametric and non-parametric estimation results in the Supplementary Information

Inference in the Supplementary Information		
Parametric test result	Non-parametric test result	Result consistency
Turning to model comparisons, we find no evidence that GPT-5.4 and GPT-5.4-mini outperform earlier OpenAI models (i.e., GPT-3.5-turbo, GPT-4o-mini, GPT-4.1-mini, and GPT-4o; two-sided unpaired t -test on the individual-level correlations across the six preference domains: $t(42.56) = -0.669$, Benjamini-Hochberg FDR corrected $P = 0.658$, Cohen's $d = -0.189$, 95% CI = [-0.06, 0.03]) or that Gemini 2.5 Flash outperforms Gemini 2.0 Flash (two-sided unpaired t -test: $t(9.81) = -0.553$; Benjamini-Hochberg FDR corrected $P = 0.658$, Cohen's $d = -0.319$, 95% CI = [-0.13, 0.08]).	Turning to model comparisons, we find no evidence that GPT-5.4 and GPT-5.4-mini outperform earlier OpenAI models (i.e., GPT-3.5-turbo, GPT-4o-mini, GPT-4.1-mini, and GPT-4o; two-sided Wilcoxon rank-sum test on the individual-level correlations across the six preference domains: $P = 0.646$) or that Gemini 2.5 Flash outperforms Gemini 2.0 Flash (two-sided Wilcoxon rank-sum test, $P = 0.646$).	✓
We also do not find that models with more parameters (i.e., GPT-4o, GPT-5.4 and GPT-5.4 Eng) outperform their counterparts with fewer parameters (i.e., GPT-4o-mini, GPT-5.4-mini and GPT-5.4-mini English; two-sided unpaired t -test: $t(30.21) = -0.447$, Benjamini-Hochberg FDR corrected $P = 0.658$, Cohen's $d = -0.149$, 95% CI = [-0.07, 0.04]).	We also do not find that models with more parameters (i.e., GPT-4o, GPT-5.4 and GPT-5.4 Eng) outperform their counterparts with fewer parameters (i.e., GPT-4o-mini, GPT-5.4-mini and GPT-5.4-mini English; two-sided Wilcoxon rank-sum test, Benjamini-Hochberg FDR corrected $P = 0.646$).	✓
Compared with their individual-level counterparts, the country-level correlations are on average slightly larger ($M_{\text{country}} = 0.137$ vs $M_{\text{individual}} = 0.094$, two-sided paired t -test: $t(71) = -2.956$, $P = 0.004$, Cohen's $d = -0.341$, 95% CI = [-0.07, -0.01]), but they remain modest in absolute magnitude, and many model-preference pairs do not reach conventional significance thresholds.	Compared with their individual-level counterparts, the country-level correlations are on average slightly larger ($M_{\text{country}} = 0.137$ vs $M_{\text{individual}} = 0.094$, two-sided Wilcoxon signed-rank test, $P = 0.007$), but they remain modest in absolute magnitude, and many model-preference pairs do not reach conventional significance thresholds.	✓
The split-half internal consistency of LLM-predicted preferences is slightly but significantly higher than that of the GPS-surveyed data ($\text{mean}_{\text{GPS}} = 0.984$ vs $\text{mean}_{\text{LLMs}} = 0.992$, two-sided Welch's t -test: $t(8.00) = 4.353$, $P = 0.002$, Cohen's $d = 1.142$, 95% CI = [0.00, 0.01]). This supports the feasibility of aggregating individual-level LLM-predicted preferences into country-level average preference indices.	The split-half internal consistency of LLM-predicted preferences is slightly but significantly higher than that of the GPS-surveyed data ($\text{mean}_{\text{GPS}} = 0.984$ vs $\text{mean}_{\text{LLMs}} = 0.992$, two-sided Wilcoxon rank-sum test, $P < 0.001$). This supports the feasibility of aggregating individual-level LLM-predicted preferences into country-level average preference indices.	✓
The between-model correlations are higher than GPS-LLM correlations (two-sided Welch's t -test; individual-level: 0.346 vs 0.094, $t(326.38) = 17.726$, $P < 0.001$, Cohen's $d = 1.224$, 95% CI = [0.22, 0.28]; country-level: 0.262 vs 0.137, $t(105.74) = 5.822$, $P < 0.001$, Cohen's $d = 0.692$, 95% CI = [0.08, 0.17]).	The between-model correlations are higher than GPS-LLM correlations (individual-level: 0.346 vs 0.094, two-sided Wilcoxon rank-sum test, $P < 0.001$; country-level: 0.262 vs 0.137, two-sided Wilcoxon rank-sum test, $P < 0.001$).	✓
The between-model MSEs are lower than GPS-LLM MSEs (two-sided Welch's t -test; individual-level: 1.377 vs 1.827, $t(313.31) = -16.569$, $P < 0.001$, Cohen's $d = -1.158$, 95% CI = [-0.50, -0.40]; country-level: 1.491 vs 1.722, $t(115.62) = -5.468$, $P < 0.001$, Cohen's $d = -0.600$, 95% CI = [-0.32, -0.15]).	The between-model MSEs are lower than GPS-LLM MSEs (individual-level: 1.377 vs 1.827, two-sided Wilcoxon rank-sum test, $P < 0.001$; country-level: 1.491 vs 1.722, two-sided Wilcoxon rank-sum test, $P < 0.001$).	✓
We also find that the explanatory power of GDP and IDI (i.e., economic-related	We also find that the explanatory power of GDP and IDI (i.e.,	✓

variables) is substantially larger than that of population-related variables (0.020 vs 0.002; two-sided t -test: $t(23.92) = 4.351, P < 0.001$, Cohen's $d = 1.256$, 95% CI = [0.01, 0.03]).	economic-related variables) is substantially larger than that of population-related variables (0.020 vs 0.002; two-sided Wilcoxon rank-sum test, $P < 0.001$).	
We do not observe differences between models in local languages and in English in our individual- and country-level metrics (two-sided paired t -test by pooling GPT-4o-mini, GPT-5.4-mini, and GPT-5.4; individual-level correlations: $t(17) = 0.124, P = 0.903$, Cohen's $d = 0.014$, 95% CI = [-0.02, 0.02]; country-level correlations: $t(17) = 1.052, P = 0.307$, Cohen's $d = 0.232$, 95% CI = [-0.04, 0.12]; individual-level MSEs: $t(17) = -0.709, P = 0.488$, Cohen's $d = -0.090$, 95% CI = [-0.06, 0.03]; country-level MSEs: $t(17) = -0.580, P = 0.570$, Cohen's $d = -0.138$, 95% CI = [-0.22, 0.12]).	We do not observe differences between models in local languages and in English in our individual- and country-level metrics (two-sided Wilcoxon signed-rank test by pooling GPT-4o-mini, GPT-5.4-mini, and GPT-5.4; individual-level correlations: $P = 0.609$; country-level correlations: $P = 0.393$; individual-level MSEs: $P = 0.580$; country-level MSEs: $P = 0.702$.	✓
Combining the individual- and country-level evidence from two metrics, we do not find evidence that paraphrasing systematically changes the alignment (two-sided paired t -test between GPT-5.4-mini and GPT-5.4-mini with paraphrasing; individual-level correlations: $t(5) = -0.211, P = 0.841$, Cohen's $d = -0.033$, 95% CI = [-0.04, 0.03]; country-level correlations: $t(5) = 1.707, P = 0.148$, Cohen's $d = 0.230$, 95% CI = [-0.02, 0.11]; individual-level MSEs: $t(5) = 0.454, P = 0.669$, Cohen's $d = 0.066$, 95% CI = [-0.05, 0.07]; country-level MSEs: $t(5) = 0.417, P = 0.694$, Cohen's $d = 0.110$, 95% CI = [-0.21, 0.29]).	Combining the individual- and country-level evidence from two metrics, we do not find evidence that paraphrasing systematically changes the alignment (two-sided Wilcoxon signed-rank test between GPT-5.4-mini and GPT-5.4-mini with paraphrasing; individual-level correlations: $P = 0.844$; country-level correlations: $P = 0.219$; individual-level MSEs: $P = 0.844$; country-level MSEs: $P = 1.000$).	✓
We do not find evidence that the response-constraint instruction systematically changes the alignment (two-sided paired t -test between GPT-5.4-mini and GPT-5.4-mini with response-constraint; individual-level correlations: $t(5) = -0.143, P = 0.892$, Cohen's $d = -0.020$, 95% CI = [-0.03, 0.02]; country-level correlations: $t(5) = 0.228, P = 0.828$, Cohen's $d = 0.060$, 95% CI = [-0.12, 0.15]; individual-level MSEs: $t(5) = 0.250, P = 0.812$, Cohen's $d = 0.026$, 95% CI = [-0.03, 0.04]; country-level MSEs: $t(5) = -0.293, P = 0.781$, Cohen's $d = -0.060$, 95% CI = [-0.20, 0.16]).	We do not find evidence that the response-constraint instruction systematically changes the alignment (two-sided Wilcoxon signed-rank test between GPT-5.4-mini and GPT-5.4-mini with response-constraint; individual-level correlations: $P = 1.000$; country-level correlations: $P = 1.000$; individual-level MSEs: $P = 1.000$; country-level MSEs: $P = 0.844$).	✓
Figure J.1 reports diversity ratios for each of the six preference domains and 12 models. The mean diversity ratio across the 72 model-preference combinations (6 domains $\times$ 12 LLMs) is significantly below 1 (two-sided one-sample t -test: $t(71) = -43.611, P < 0.001$, Cohen's $d = -5.140$, 95% CI = [0.19, 0.26]), indicating that LLMs do not reach human levels of response diversity.	Figure J.1 reports diversity ratios for each of the six preference domains and 12 models. The mean diversity ratio across the 72 model-preference combinations (6 domains $\times$ 12 LLMs) is significantly below 1 (two-sided Wilcoxon signed-rank test, $P < 0.001$), indicating that LLMs do not reach human levels of response diversity.	✓
We first verify that the higher temperature does not materially affect alignment between LLM-predicted preferences and GPS-surveyed preferences (two-sided paired t -test between GPT-5.4-mini at $T = 0$ and $T = 1$; individual-level correlations: $t(5) = -1.131, P = 0.309$, Cohen's $d = -0.040$, 95% CI = [-0.01, 0.00]; country-level correlations: $t(5) = 0.938, P = 0.391$, Cohen's $d = 0.067$, 95% CI = [-0.02, 0.05]; individual-level MSEs: $t(5) = -0.072, P = 0.945$, Cohen's $d = -0.004$, 95% CI = [-0.02, 0.02]; country-level MSEs: $t(5) = -1.090, P = 0.326$, Cohen's $d = -0.089$, 95% CI = [-0.10, 0.04]).	We first verify that the higher temperature does not materially affect alignment between LLM-predicted preferences and GPS-surveyed preferences (two-sided Wilcoxon signed-rank test between GPT-5.4-mini at $T = 0$ and $T = 1$; individual-level correlations: $P = 0.438$; country-level correlations: $P = 0.688$; individual-level MSEs: $P = 0.844$; country-level MSEs: $P = 1.000$).	✓

Turning to diversity, Figure J.3 shows that the diversity ratio increases from 0.187 (GPT-5.4-mini, $T = 0$) to 0.288 (GPT-5.4-mini, $T = 1$), a significant increase (two-sided paired t -test: $t(443) = 22.932$, $P < 0.001$, Cohen's $d = 0.529$, 95% CI = [0.09, 0.11]; pooling all countries and preferences).	Turning to diversity, Figure J.3 shows that the diversity ratio increases from 0.187 (GPT-5.4-mini, $T = 0$) to 0.288 (GPT-5.4-mini, $T = 1$), a significant increase (two-sided Wilcoxon signed-rank test: $P < 0.001$; pooling all countries and preferences).	✓
---	---	---

Notes: The first column presents results from parametric analyses, and the second column presents the corresponding non-parametric results. The third column assesses consistency between the two approaches: a consistency mark (✓) indicates that the estimated effects have the same direction and the same significance classification (that is, both $P < 0.05$ or both $P \geq 0.05$).

References

1. Falk, A. *et al.* Global evidence on economic preferences. *Q. J. Econ.* **133**, 1645–1692 (2018).
2. Aftiss, A., Lamsiyah, S., Schommer, C. & El Alaoui, S. O. Empirical evaluation of pre-trained language models for summarizing moroccan darija news articles. in *Proceedings of the 4th Workshop on Arabic Corpus Linguistics (WACL-4)* 77–85 (Association for Computational Linguistics, Abu Dhabi, UAE, 2025).
3. Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B. & Larson, J. M. Synthetic replacements for human survey data? The perils of large language models. *Polit. Anal.* **32**, 401–416 (2024).
4. Wang, A., Morgenstern, J. & Dickerson, J. P. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nat. Mach. Intell.* **7**, 400–411 (2025).
5. Exley, C. L., Hauser, O. P., Moore, M. & Pezzuto, J.-H. Believed gender differences in social preferences. *Q. J. Econ.* **140**, 403–458 (2025).
6. Kaplan, J. *et al.* Scaling laws for neural language models. Preprint at <https://doi.org/10.48550/arXiv.2001.08361> (2020).
7. Villalobos, P. *et al.* Will we run out of data? Limits of LLM scaling based on human-generated data. Preprint at <https://doi.org/10.48550/arXiv.2211.04325> (2022).
8. Guilbeault, D., Delecourt, S. & Desikan, B. S. Age and gender distortion in online media and large language models. *Nature* **646**, 1129–1137 (2025).
9. Falk, A. & Hermle, J. Relationship of gender differences in preferences to economic development and gender equality. *Science* **362**, eaas9899 (2018).
10. Sunde, U. *et al.* Patience and comparative development. *Rev. Econom. Stud.* **89**, 2806–2840 (2022).
11. Kotek, H., Dockum, R. & Sun, D. Gender bias and stereotypes in large language models. in *Proceedings of the ACM Collective Intelligence Conference* 12–24 (ACM, Delft Netherlands, 2023).
12. Fulgu, R. A. & Capraro, V. Surprising gender biases in GPT. *Comput. Hum. Behav. Rep.* **16**, 100533 (2024).
13. Armstrong, L., Liu, A., MacNeil, S. & Metaxa, D. The silicon ceiling: auditing GPT’s race and gender biases in hiring. in *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* 1–18 (ACM, San Luis Potosi Mexico, 2024).
14. Zack, T. *et al.* Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. *Lancet Digit. Health* **6**, e12–e22 (2024).
15. Grossmann, I. *et al.* AI and the transformation of social science research. *Science* **380**, 1108–1109 (2023).
16. Mei, Q., Xie, Y., Yuan, W. & Jackson, M. O. A turing test of whether AI chatbots are behaviorally similar to humans. *Proc. Natl. Acad. Sci.* **121**, e2313925121 (2024).
17. Chen, Y., Liu, T. X., Shan, Y. & Zhong, S. The emergence of economic rationality of GPT. *Proc. Natl. Acad. Sci.* **120**, e2316205120 (2023).
18. Wang, D., Huang, D., Shen, H. & Uzzi, B. A large-scale comparison of divergent creativity in humans and large language models. *Nat. Hum. Behav.* **10**, 531–540 (2026).
19. Wang, S. *et al.* When experimental economics meets large language models: evidence-

based tactics. Preprint at <https://doi.org/10.48550/arXiv.2505.21371> (2025).