

Simulated Responses from Large Language Models Are Weak Proxies for Human Economic Preferences across 74 Countries

Corresponding Author: Professor Fadong Chen

This file contains all editorial decision letters in order by version, followed by all author rebuttals in order by version.

Attachments originally included by the reviewers as part of their assessment can be found at the end of this file.

Version 0:

Decision Letter:

12th March 2026

Dear Professor Chen,

Thank you once again for your manuscript, entitled "Large Language Models as Synthetic Respondents for Economic Preferences: A Global Audit", and for your patience during the peer review process.

Your Article has now been evaluated by 3 referees. You will see from their comments copied below that, although they find your work of potential interest, they have raised quite substantial concerns. In light of these comments, we cannot accept the manuscript for publication, but would be interested in considering a revised version if you are willing and able to fully address reviewer and editorial concerns.

We hope you will find the referees' comments useful as you decide how to proceed. If you wish to submit a substantially revised manuscript, please bear in mind that we will be reluctant to approach the referees again in the absence of major revisions. We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

To guide the scope of the revisions, the editors discuss the referee reports in detail within the team, including with the chief editor, with a view to (1) identifying key priorities that should be addressed in revision and (2) overruling referee requests that are deemed beyond the scope of the current study. We hope that you will find the prioritised set of referee points to be useful when revising your study. Please do not hesitate to get in touch if you would like to discuss these issues further.

Specifically, please prioritize the following:

1. Reviewer #3 has raised substantial concerns regarding the reliance on correlation as the primary outcome measure and has recommended considering mean squared error (MSE) as an alternative. Please address these concerns comprehensively by conducting the additional analyses suggested by Reviewer #3.
2. As highlighted by both Reviewer #1 and Reviewer #2, the generalizability of your findings remains a key issue. In your revision, please respond to these concerns through clearer framing and by incorporating the robustness checks recommended by Reviewer #2.
3. Reviewers #1 and #2 also noted the absence of any exploration of underlying mechanisms and provided specific suggestions for additional analyses. We agree that such analyses would substantially strengthen the manuscript's contributions. Please incorporate these mechanistic analyses as part of your revision.
4. In addition, there are significant concerns about potential data contamination and temporal mismatches between data collection and analysis. Please conduct the analyses recommended by Reviewer #2 to adequately address these issues in your revised manuscript.

Finally, your revised manuscript must comply fully with our editorial policies and formatting requirements. Failure to do so will result in your manuscript being returned to you, which will delay its consideration. To assist you in this process, I have attached a checklist that lists all of our requirements. If you have any questions about any of our policies or formatting, please don't hesitate to contact me.

If you wish to submit a suitably revised manuscript, we would hope to receive it within 2 months. I would be grateful if you could contact us as soon as possible if you foresee difficulties with meeting this target resubmission date.

With your revision, please:

- Include a "Response to the editors and reviewers" document detailing, point-by-point, how you addressed each editor and referee comment. If no action was taken to address a point, you must provide a compelling argument. When formatting this document, please respond to each reviewer comment individually, including the full text of the reviewer comment verbatim followed by your response to the individual point. This response will be used by the editors to evaluate your revision and sent back to the reviewers along with the revised manuscript.
- Highlight all changes made to your manuscript or provide us with a version that tracks changes.

Please use the link below to submit your revised manuscript and related files:

Link Redacted

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

Thank you for the opportunity to review your work. Please do not hesitate to contact me if you have any questions or would like to discuss the required revisions further.

Sincerely,

Nature Human Behaviour

Reviewer expertise:

Reviewer #1: Behavioural & experimental economics, human-algorithm interaction

Reviewer #2: Behavioral Science, Machine Learning, AI for Science

Reviewer #3: Public Opinion, LLM - human comparison

REVIEWER COMMENTS:

Reviewer #1 (Remarks to the Author):

Key results: Please summarise what you consider to be the outstanding features of the work.

This paper audits whether LLMs can serve as synthetic respondents for economic preferences by replicating the Global Preference Survey across 74 countries and eight LLMs, generating approximately 11.2 million responses. Three features stand out. First, the scale is impressive — roughly 560,000 synthetic participants matched one-to-one to ~70,000 human respondents. Second, the paper identifies a "micro-to-macro gap": LLMs reproduce demographic gradients (~77% consistency) but fail at cultural-level patterns (~50%, essentially chance). Third, it provides a mechanistic account — demographic information drives individual-level alignment while country labels contribute little, and prediction errors covary with GDP and digital infrastructure rather than population size. The paper also documents gender bias amplification and reduced response diversity, and articulates important equity implications. Overall, it provides a clear, evidence-based argument for why human data remain irreplaceable for cross-cultural and macro-level inference.

Validity: Does the manuscript have flaws which should prohibit its publication?

The manuscript is methodologically careful and provides robustness checks for nearly every conclusion. The one-to-one prompting design with localized-language prompts is well-motivated and clearly documented. The inclusion of an English-only condition as a robustness check is a sensible design choice. The statistical analyses are appropriate for the claims made. While there are areas where the analysis could be extended (discussed below under Suggested Improvements), none of the issues I identify would prohibit publication.

Originality and significance: What are the major claims of the paper? Do you think that they represent a significant advance in the field?

The central claim is that LLM-based synthetic respondents are unreliable for macro-level inference, policy evaluation, or cross-cultural comparison, despite showing systematic individual-level alignment. I view this as a significant advance across multiple dimensions.

Most prior work evaluates LLM alignment at a single level and often within WEIRD contexts. This paper reframes the question — from "Can LLMs replicate human responses?" to "At what level?" — and introduces the micro-to-macro gap as a conceptual framework. Evidence-based advance: The dataset is among the largest in this literature, and the GPS benchmark is experimentally validated and widely used. Policy-related advance: The finding that errors covary with economic development raises important concerns about digital inequity and a policy dilemma where developers lack incentives to optimize for lower-income regions.

Contribution to LLM evaluation: The paper demonstrates how representative surveys can serve as behavioral benchmarks, complementing existing evaluation suites focused on coding and reasoning.

Given the high between-model correlations and shared Transformer architecture, the core findings likely generalize to near-future models. I believe this paper will be of broad interest across economics, psychology, computational social science, and the AI fairness community.

Data & methodology: Please comment on the validity of the approach, quality of the data and quality of presentation. The one-to-one persona-based prompting design with localized-language prompts, temperature set to zero, and GPS-consistent scoring ensures comparability with the human benchmark. The GPS is the most comprehensive cross-country dataset on experimentally validated economic preferences, and the cultural dimensions (WVS, Hofstede, Schwartz) are well-established.

Preregistration:

The authors do not report preregistration, which is understandable given that no new human data were collected. All analyses are conducted on existing survey data and LLM-generated outputs. This is not a concern.

Appropriate use of statistics and treatment of uncertainties:

The statistical tests are appropriate throughout.

Custom code:

The data is deposited on OSF, but I did not find the code.

Conclusions:

Do you find that the conclusions and data interpretation are robust, valid and reliable?

The conclusions follow logically from the analyses and are stated with appropriate caution. The authors are careful to characterize alignment as "systematic but modest" rather than making stronger claims. The discussion of potential applications (hypothesis generation, study-design diagnostics) is balanced, and the limitations section is candid about the possibility that successful predictions may partly reflect memorization, since the GPS was conducted in 2012 and its content may appear in LLM training data. My main remaining question concerns generalizability to future models, but the between-model consistency analyses provide reasonable, if indirect, evidence on this point.

Suggested improvements:

I offer the following suggestions, organized as minor revisions, that I believe would strengthen the paper:

Clarify economic preferences for a multidisciplinary audience. The six preference domains are fundamental in economics but may not be self-evident to readers from other disciplines. I recommend adding a paragraph in the introduction that explains why these specific preferences matter for individual behavior, market outcomes, and policy design, with concrete examples, to improve accessibility for the broad readership of *Nature Human Behaviour*.

Discuss the potential impact of Google Translate on results. The localized-language prompts are generated using Google Translate API (line 522). Machine translation quality varies substantially across languages, and translation errors could introduce systematic noise that differentially affects lower-resource languages, potentially confounding the GDP/IDI-related patterns in prediction errors. The authors should briefly discuss whether translation quality could be an alternative explanation for some of the observed cross-country variation in alignment.

Address the GPS vintage more directly. The GPS was conducted in 2012, over a decade before any of the LLMs were developed. Beyond training data contamination, it is possible that preferences in some countries have shifted since 2012, meaning that misalignment could partly reflect temporal change rather than model failure. A brief discussion of this would be valuable.

Elaborate on the temperature-zero design choice. Setting temperature to zero is appropriate for reproducibility, but it systematically reduces response variability. This may partly explain the reduced response diversity documented in Appendix G. A brief discussion of whether moderate temperature settings might yield closer distributional matches to human responses would be useful.

Consider variance decomposition. Reporting what fraction of variance in human preferences is explained by LLM-predicted preferences (e.g., R^2 values) would give readers a clearer sense of practical significance beyond correlation coefficients.

Engage with recent related work. Two recent economics working papers are directly relevant. Wu, Xi, and Xie (2025, NBER WP 34308) propose an LLM-based survey framework for U.S. inflation expectations using date-restrictive prompting and recover human-comparable treatment effects in a single-country setting. The contrast is instructive: LLMs may perform well where extensive high-quality training data exist but poorly in culturally specific, data-sparse contexts. Asirvatham, Moksiki, and Shleifer (2026, NBER WP 34834) examine LLMs as measurement tools in economics, providing complementary evidence. Discussing these papers would situate the study more fully within the rapidly evolving literature.

Strengthen the discussion of amplification. The finding that LLMs amplify demographic effects (Figure 4) is important but receives relatively brief treatment. The authors could discuss whether amplification varies systematically by model architecture, country, or cultural context.

Exploration of Error Mechanisms: While the paper identifies that errors track economic development, further investigation into the specific mechanisms underlying these geographical and cultural biases could be beneficial. This might involve qualitative analysis of LLM outputs from different regions or comparing the training data characteristics of LLMs across languages/cultures.

Fine-tuning: Explore whether targeted fine-tuning of LLMs on more diverse or culturally-specific datasets could improve their performance in capturing human preferences, especially at the cultural level.

Temporal Dynamics: The protocol is designed to track model updates over time. Presenting initial data or a discussion on how LLM performance has evolved with different model versions could add another layer of insight.

Ethical Implications: Given the findings, a more in-depth discussion on the ethical implications of using LLMs for policy-making, especially in diverse global contexts, could be valuable.

References: Does this manuscript reference previous literature appropriately?

The reference list is comprehensive and appropriate. I recommend adding:

Jing Cynthia Wu, Jin Xi, and Shihan Xie, "LLM Survey Framework: Coverage, Reasoning, Dynamics, Identification," NBER Working Paper 34308 (2025).

Hemanth Asirvatham, Elliott Moksiki, and Andrei Shleifer, "GPT as a Measurement Tool," NBER Working Paper 34834 (2026).

Wang, D., Huang, D., Shen, H. et al. A large-scale comparison of divergent creativity in humans and large language models. *Nat Hum Behav* (2025).

Clarity and context: Is the abstract clear, accessible? Are abstract, introduction and conclusions appropriate?

The abstract is clear and effectively summarizes the multi-level findings. The introduction is well-structured but would benefit from a

dedicated paragraph explaining what economic preferences are and why they matter, as noted above. The conclusions are appropriate and avoid overstating the findings.

Please indicate any particular part of the manuscript, data, or analyses that you feel is outside the scope of your expertise, or that you were unable to assess fully.

None.

Please address any other specific question asked by the editor via email.

None.

Summary recommendation:

This is a well-executed study that makes an important conceptual and empirical contribution. The micro-to-macro gap is a novel and valuable finding, and the scale of the audit is impressive.

Reviewer #1 (Remarks on code availability):

The authors provide the data for main results in the paper. However, I didn't find the detailed code to generate these results. Therefore, I am not able to assess the code.

Reviewer #2 (Remarks to the Author):

See attached PDF

Reviewer #3 (Remarks to the Author):

This is a technically impressive paper. The authors use several LLMs to construct virtual respondents that adopt the personas of roughly 70,000 individuals. They then compare LLM responses to the human GPS survey by reporting pooled correlations between human and LLM preferences. They find systematic correlations, but the correlations are weak. Moreover, they find variation that raises questions about the use of LLMs to replace humans.

I agree with the conclusions. And I applaud the work that was done. However, I struggled to understand why the authors were doing what they were doing. The analyses were not well-justified and were more complicated and tangential than needed.

For example, the authors begin by comparing the correlation between the LLM predicted and GPS surveyed respondents. But why is this the right metric? Correlation measures linear association, but that does not seem to be the quantity of interest in this research design. In this study each human respondent is paired with an LLM replicant, so the actual human response associated with each silicon response is known. If you are going to use a silicon respondent, isn't the relevant question: how far the predicted responses are from the true responses. If so, why not measure that directly using a measure such as mean squared error (MSE)?

As is, the correlation seems like it could be very misleading. A high correlation only indicates that two measures covary in a linear fashion across observations, not that the measures for a given (x,y) pair are close to one another in covariate space. But if the goal is to evaluate whether silicon respondents can stand in for human respondents, the key issue is whether the imputations/estimates of the individual responses are unbiased. Measures that are systematically biased can still produce high correlations if the bias affects all observations in a similar way.

On a related point, work by Bisbee et al. suggests that LLM predictions tend to be homogenizing. Even if there is a high correlation across observations, they may fail to reflect the heterogeneity in human responses. But this is again easy to evaluate in this setting by comparing the standard deviation of human responses with the standard deviation of the matched silicon respondents. If LLM responses compress variation, the standard deviations should be noticeably smaller. This matters because if we use them in second-stage analyses (e.g., regressions) then the variation in the silicon responses matters!

Moreover, if the authors want to examine how performance varies across indices or countries, they could simply predict MSE in point estimates or differences in standard deviations as a function of those characteristics.

I had the same set of reactions to Figure 3. Why do you want to, or need to, focus on correlations? If LLMs are meant to replace human respondents, the relevant benchmark is how closely they reproduce human responses. The natural statistic in Figure 3 would therefore be the average MSE in predicted responses across models within each sample (i.e., do the models give similar imputations?).

In the second half of the paper, when turning to demographic level consistency, the authors report Cohen's d across several preference domains. The results are presented in Figures 4 and 5. I confess that the wall of plots are very hard to interpret. It is not clear, for example, what to make of the considerable variation in Figure 5 across indices, models, and orientations.

I think Figure 6 comes closest to the type of analysis that seems appropriate, but it is still more complicated than necessary. If the silicon responses matched human responses shouldn't the regression coefficient simply be 1 - meaning that there is a high correlation? That said, this still has the issue with conflating linear association with similarity. A cleaner strategy would be to model the difference between the human and silicon responses directly. For example, one could predict the distance between the two responses as a function of question type, model type, country, demographics, or other relevant features. This would provide a direct assessment of where and why the predictions fail.

Stepping back, the data is impressive, but I struggle to comprehend whether the analysis is the optimal design for evaluating the claim. If the goal is to evaluate how closely LLMs approximate human responses, the analysis should be both simpler and different from what is currently presented.

What I would recommend doing:

Analysis 1: Calculate the average MSE between human and silicon respondents by country, index, or other relevant categories. This directly assesses bias. Correlation cannot do this because it measures linear association rather than accuracy.

Analysis 2: Examine whether LLMs reproduce variation in human responses. For each subset, compute the standard deviation of

human responses and the standard deviation of the corresponding silicon responses and compare them. This matters because the amount of variation affects precision in downstream analyses.

Analysis 3: Examine how these quantities vary across models and contexts. Using the same metric, such as MSE, allows a straightforward comparison across models.

Analysis 4: After characterizing the overall patterns, analyze whether the errors follow systematic patterns. For example, predict the absolute difference between human and silicon responses as a function of individual demographics, country characteristics, model type, and other relevant factors. This decomposes the sources of error. The analysis could be conducted separately for each question, which is probably preferable given the variation across items, or pooled with question fixed effects.

At the end of the day, I have very high confidence that the substantive conclusion is correct, but I am not convinced that the current paper provides the best demonstration of that claim. That said, the authors have assembled a creative framework that could be used to evaluate this question directly, but doing so would require a different set of analyses than those currently reported.

Version 1:

Decision Letter:

Our ref: NATHUMBEHAV-26020673A

16th June 2026

Dear Dr. Chen,

Thank you for submitting your revised manuscript "Large Language Models as Synthetic Respondents for Economic Preferences: A Global Audit" (NATHUMBEHAV-26020673A). It has now been seen by the original referees and their comments are below. As you can see, the reviewers find that the paper has improved in revision. We will therefore be happy in principle to publish it in Nature Human Behaviour, pending minor revisions to satisfy the referees' final requests and to comply with our editorial and formatting guidelines.

We are now performing detailed checks on your paper and will send you a checklist detailing our editorial and formatting requirements within a couple of weeks. Please do not upload the final materials and make any revisions until you receive this additional information from us.

Please do not hesitate to contact me if you have any questions.

Sincerely,

Nature Human Behaviour

Reviewer #1 (Remarks to the Author):

I thank the authors for a careful and substantial revision. The revised manuscript is much stronger than the previous version. In particular, the shift to MSE as the primary accuracy metric directly addresses the concern that correlations can be misleading for evaluating matched synthetic respondents. The new analyses also make the empirical message clearer: LLM-predicted preferences are not random, but the alignment is modest; models agree with one another more than they agree with the human benchmark; demographic gradients are partly recovered but often amplified; and country- and culture-level patterns remain poorly captured.

The added GPT-5.4 and GPT-5.4-mini experiments, prompt-component ablations, error decompositions, paraphrasing and response-format checks, and diversity analyses collectively provide a much more convincing audit of the use of LLMs as synthetic respondents. I especially appreciate the response to the earlier request for direct prediction-error analysis and the new evidence that response diversity is substantially compressed relative to human responses. These additions clarify the boundary conditions of the approach and make the manuscript more useful for the broad behavioral-science audience of Nature Human Behaviour.

My remaining comments are minor, but I think they should be addressed before publication.

1. Please temper the mechanistic and causal wording in several places. The ablation results are informative, but they are conducted primarily with GPT-5.4-mini and under one persona-prompting design. Similarly, the country-level error decompositions are observational. Phrases such as demographic information "drives" alignment, demographic information "accounts for most" of alignment, country labels "contribute little," and LLMs "lack sufficient cultural sensitivity" are directionally plausible but a little too strong. I recommend phrasing these as diagnostic evidence under the present prompting protocol: for example, "in these ablations, removing demographic information reduces alignment more than removing country information," or "the results are consistent with the interpretation that the models condition more effectively on demographic cues than on country labels." This would preserve the conclusion while avoiding overclaiming mechanism.

2. The contamination and memorization results should also be stated more cautiously. The paraphrasing and English-only checks are useful and make it unlikely that memorization is the primary source of the reported alignment. However, null differences under these perturbations do not show that memorization is not a driver at all. In the supplement, the statement that memorization "is not a driver of alignment" should be softened to "we find no evidence that memorization is the primary driver of alignment" or similar. This is especially important because the manuscript itself rightly notes that GPS material may have appeared in training data and that the resulting alignment estimates may be an upper bound.

3. Please include the GPT-4o-mini zero-response robustness information in the manuscript or supplement, not only in the response letter. The supplement notes that an initial programming error treated "0" responses as missing and re-collected them until a non-zero value was obtained. The response letter states that the affected local-language GPT-4o-mini condition was retained but that excluding it does not change the headline analyses. This is reassuring, but readers should have access to the sensitivity check in the published supplementary material. A short table or paragraph reporting the exclusion check would be sufficient.

4. Please clarify the scope of the prompt-ablation evidence. The ablations are valuable, but they are not a general test of all possible ways to encode country or culture in LLM prompts. A country label alone may be a weak operationalization of cultural context, and stronger country/culture prompts, retrieval-augmented prompts, or explicit cultural calibration could plausibly behave differently. I do not think the authors need to run these additional analyses, but the conclusion should be explicitly limited to the current persona-prompting design.

5. Some claims comparing newer and older models should avoid overinterpreting null tests. The addition of GPT-5.4 and GPT-5.4-mini is useful, and the absence of clear improvement is interesting. Still, "newer models do not appear to achieve stronger alignment" is safer than implying that newer models cannot improve. The manuscript should consistently frame these results as evidence within the sampled models, collection windows, prompting protocol, and preference task.

6. The code and data availability statement is appreciated. To ensure full reproducibility, please clarify that the deposited materials contain the exact prompts, model identifiers, data collection dates, temperature parameters, parsing and extraction workflows, and all scripts required to replicate every headline figure. More importantly, I was unable to locate the associated code and data via the current OSF link.

Reviewer #2 (Remarks to the Author):

I was Reviewer 2 in the previous round. I have read the authors' responses and revisions in detail. In my view, the authors have undertaken a serious major revision that makes the paper substantially stronger. The revision moves the contribution from a mostly descriptive audit (that is, that LLMs correlate weakly with human economic preferences) toward a more diagnostic claim: LLMs recover some demographic-level structure from prompting, but fail more seriously at country- and cultural-level inference. This pattern appears robust across models and is not obviously explained by prompt format or simple memorization.

The main remaining limitation is that the proposed "mechanism" is still not mechanistic in the stronger interpretability sense. It does not concern internal representations, training-data attribution, or latent computational mechanisms. Rather, it is a prompt-level result: which persona fields matter for observed alignment. Precisely because these deeper mechanisms are not identified, the paper offers relatively few actionable recommendations for LLM researchers who may be interested in how to train, fine-tune, or steer LLMs toward better individual-level alignment.

Relatedly, the prompt-level implications for designing AI agents or digital twins seem somewhat pessimistic. The results suggest that even emphasizing demographic-level structure does not lead to substantially better individual-level alignment. A reader interested in designing better digital twins may therefore leave with a largely negative conclusion about how to improve prompts, rather than with a clear set of constructive design principles.

Overall, I think the authors have responded seriously and constructively to the previous concerns, and the revised manuscript is substantially improved. My remaining concern is mainly about the scope and framing of the contribution: the paper is stronger as an audit and diagnostic study of current LLM synthetic respondents than as a mechanistic account or practical guide for improving them. I leave it to the editorial team to evaluate how important this type of contribution is for Nature Human Behaviour's interdisciplinary readership.

Reviewer #3 (Remarks to the Author):

Thank you for your responsive review. Using MSE rather than correlations makes the paper much more convincing and I am happy to support its' publication. Congratulations.

One final suggestion is that it might be worth citing a recent chapter from an upcoming volume on Public Opinion and AI by the American Political Science Association to help further contextualizing the contribution of the paper alongside a larger set of considerations about the use of AI in the study of public opinion and the implications of replacing human opinion with synthetic opinion.

Clinton, Joshua et al. 2026. "Public Opinion in the Age of AI." APSA Preprints. doi: 10.33774/apsa-2026-1rg21-v3.

Version 2:

Decision Letter:

Dear Professor Chen,

We are pleased to inform you that your Article "Simulated Responses from Large Language Models Are Weak Proxies for Human Economic Preferences across 74 Countries", has now been accepted for publication in Nature Human Behaviour.

Authors may need to take specific actions to achieve compliance with funder and institutional open access mandates. If your research is supported by a funder that requires immediate open access (e.g. according to [Plan S principles](https://www.springernature.com/gp/open-science/plan-s-compliance) or the [NIH public access policy](https://www.springernature.com/gp/open-science/us-federal-agency-compliance)) then you should select the gold OA route, and we will direct you to the compliant route where possible. Because authors warrant under our subscription licensing terms that they haven't committed to licensing any version of their article under a licence inconsistent with the terms of our agreement – including the applicable embargo period – publication under the subscription model isn't suitable for authors whose funders require no embargo.

Once your manuscript is typeset and you have completed the appropriate grant of rights, you will receive a link to your electronic proof via email with a request to make any corrections within 48 hours. If, when you receive your proof, you cannot meet this deadline, please inform us at rjsproduction@springernature.com immediately. Once your paper has been scheduled for online publication, the Nature press office will be in touch to confirm the details.

Acceptance of your manuscript is conditional on all authors' agreement with our publication policies (see <http://www.nature.com/nathumbehav/info/gta>). In particular your manuscript must not be published elsewhere and there must be no announcement of the work to any media outlet until the publication date (the day on which it is uploaded onto our web site).

If you have posted a preprint on any preprint server, please ensure that the preprint details are updated with a publication reference, including the DOI and a URL to the published version of the article on the journal website.

An online order form for reprints of your paper is available at <https://www.nature.com/reprints/author-reprints.html>. All co-authors, authors' institutions and authors' funding agencies can order reprints using the form appropriate to their geographical region.

We welcome the submission of potential cover material (including a short caption of around 40 words) related to your manuscript; suggestions should be sent to Nature Human Behaviour as electronic files (the image should be 300 dpi at 210 x 297 mm in either TIFF or JPEG format). Please note that such pictures should be selected more for their aesthetic appeal than for their scientific content, and that colour images work better than black and white or grayscale images. Please do not try to design a cover with the Nature Human Behaviour logo etc., and please do not submit composites of images related to your work. I am sure you will understand that we cannot make any promise as to whether any of your suggestions might be selected for the cover of the journal.

You can now use a single sign-on for all your accounts, view the status of all your manuscript submissions and reviews, access usage statistics for your published articles and download a record of your refereeing activity for the Nature journals.

To assist our authors in disseminating their research to the broader community, our SharedIt initiative provides you with a unique shareable link that will allow anyone (with or without a subscription) to read the published article. Recipients of the link with a subscription will also be able to download and print the PDF.

As soon as your article is published, you will receive an automated email with your shareable link.

In approximately 10 business days you will receive an email with a link to choose the appropriate publishing options for your paper and our Author Services team will be in touch regarding any additional information that may be required.

If you have any questions about our publishing options, costs, Open Access requirements, or our legal forms, please contact ASJournals@springernature.com

We look forward to publishing your paper.

With best regards,

Nature Human Behaviour

P.S. Click on the following link if you would like to recommend Nature Human Behaviour to your librarian
<https://www.nature.com/subscriptions/recommend.html#forms>

** Visit the Springer Nature Editorial and Publishing website at http://editorial-jobs.springernature.com?utm_source=ejp_NHumB_email&utm_medium=ejp_NHumB_email&utm_campaign=ejp_NHumB for more information about our career opportunities. If you have any questions please click http://editorial-jobs.springernature.com?utm_source=ejp_NHumB_email&utm_medium=ejp_NHumB_email&utm_campaign=ejp_NHumB

href="mailto:editorial.publishing.jobs@springernature.com">here.**

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Reviewer 1's Comments

1. Key results: Please summarise what you consider to be the outstanding features of the work.

This paper audits whether LLMs can serve as synthetic respondents for economic preferences by replicating the Global Preference Survey across 74 countries and eight LLMs, generating approximately 11.2 million responses. Three features stand out. First, the scale is impressive — roughly 560,000 synthetic participants matched one-to-one to ~70,000 human respondents. Second, the paper identifies a “micro-to-macro gap”: LLMs reproduce demographic gradients (~77% consistency) but fail at cultural-level patterns (~50%, essentially chance). Third, it provides a mechanistic account — demographic information drives individual-level alignment while country labels contribute little, and prediction errors covary with GDP and digital infrastructure rather than population size. The paper also documents gender bias amplification and reduced response diversity, and articulates important equity implications. Overall, it provides a clear, evidence-based argument for why human data remain irreplaceable for cross-cultural and macro-level inference.

Response: We sincerely thank you for this generous and accurate summary of our work. We are particularly grateful that you recognize the three features we hoped would stand out: the scale of the audit, the micro-to-macro gap as a unifying conceptual framing, and the mechanistic decomposition showing that individual-level alignment is primarily driven by demographic information rather than by country labels, while prediction errors covary with economic development and digital infrastructure.

Your careful reading and detailed, constructive feedback have guided our revisions throughout. In particular, your comments have helped us refine our analyses and sharpen the discussion, such as clarifying the role of economic preferences for a multidisciplinary readership, acknowledging the temporal mismatch between the GPS wave and current LLMs, and expanding the analysis of underlying mechanisms, with new ablation experiments and a quantitative error decomposition underpinning the revised Discussion.

We have addressed each of your comments thoroughly in the revised manuscript, and we believe these revisions have significantly strengthened the paper.

2. Validity: Does the manuscript have flaws which should prohibit its publication?

The manuscript is methodologically careful and provides robustness checks for nearly every conclusion. The one-to-one prompting design with localized-language prompts is well-motivated and clearly documented. The inclusion of an English-only condition as a robustness check is a sensible design choice. The statistical analyses are appropriate for the claims made. While there are areas where the analysis could be extended (discussed below under Suggested Improvements), none of the issues I identify would prohibit publication.

Response: We are very grateful for your assessment that the manuscript is methodologically careful and that none of the issues you identify would prohibit publication. We particularly appreciate your endorsement of our one-to-one persona-based prompting design with localized-language prompts, the English-only condition as a robustness check, and the appropriateness of our statistical analyses. Although you generously characterize the remaining issues as not blocking publication, we have nevertheless addressed each of your suggested improvements in detail, with the corresponding revisions to the main text (including Methods) and Appendix described in the point-by-point responses below.

3. Originality and significance: What are the major claims of the paper? Do you think that they represent a significant advance in the field?

The central claim is that LLM-based synthetic respondents are unreliable for macro-level inference, policy evaluation, or cross-cultural comparison, despite showing systematic individual-level alignment. I view this as a significant advance across multiple dimensions.

Most prior work evaluates LLM alignment at a single level and often within WEIRD contexts. This paper reframes the question — from “Can LLMs replicate human responses?” to “At what level?” — and introduces the micro-to-macro gap as a conceptual framework.

Evidence-based advance: The dataset is among the largest in this literature, and the GPS benchmark is experimentally validated and widely used.

Policy-related advance: The finding that errors covary with economic development raises important concerns about digital inequity and a policy dilemma where developers lack incentives to optimize for lower-income regions.

Contribution to LLM evaluation: The paper demonstrates how representative surveys can serve as behavioral benchmarks, complementing existing evaluation suites focused on coding and reasoning.

Given the high between-model correlations and shared Transformer architecture, the core findings likely generalize to near-future models. I believe this paper will be of broad interest across economics, psychology, computational social science, and the AI fairness community.

Response: We sincerely thank you for this thoughtful and encouraging assessment, and especially for situating our work along the four dimensions of advance you describe. We are grateful that you recognize our central contribution in reframing the question from whether LLMs can replicate human responses to the level at which such replication is reliable.

We also appreciate your recognition of the scale and benchmark quality of our audit, which combines the largest cross-country LLM evaluation in this literature with the experimentally validated GPS dataset. We are grateful for your highlighting of the policy-

relevance of our findings, namely, that prediction errors covary with economic development and raise concerns about digital inequity. The corresponding policy implications are discussed in paragraph 7 of the revised Discussion:

“Despite these constraints, our individual- and demographic-level alignment results highlight several promising scientific uses of LLMs as synthetic respondents. First, LLMs can be used to propose and probe latent motivational structures underlying human behavior, particularly in WEIRD populations, complementing formal decision theory (Bhatia, 2024; Bhatia et al., 2025; Binz et al., 2025; Wu et al., 2025; Xie et al., 2025; Zhu et al., 2025). Second, LLMs can accelerate early-stage hypothesis generation and study-design diagnostics, helping researchers identify promising interventions or measurement strategies before turning to human samples. They can help identify when cross-cultural extrapolation may fail (Henrich et al., 2010; Lowes & Nunn, 2025; Moscona et al., 2026), and flag measurement-context mismatches (Cohn et al., 2019; Huynh et al., 2020; Yang et al., 2023). They may also serve as a fast and low-cost environment for follow-up experiments and robustness checks (Guilbeault et al., 2025), and provide a safer testbed for exploring ethically sensitive counterfactuals, such as mechanisms that increase or reduce intergroup hostility, before designing human-subject experiments. Beyond synthetic respondents, a complementary application is to use LLMs as a measurement tool (Asirvatham et al., 2026; Fang et al., 2025), expanding social scientists’ toolkit.”

Finally, we share your view that representative surveys can serve as behavioral benchmarks complementing existing coding- and reasoning-focused LLM evaluations, and we are encouraged by your assessment that the work will be of interest across economics, psychology, computational social science, and the AI fairness community.

4. Data & methodology: Please comment on the validity of the approach, quality of the data and quality of presentation.

The one-to-one persona-based prompting design with localized-language prompts, temperature set to zero, and GPS-consistent scoring ensures comparability with the human benchmark. The GPS is the most comprehensive cross-country dataset on experimentally validated economic preferences, and the cultural dimensions (WVS, Hofstede, Schwartz) are well-established.

Response: We sincerely thank you for this positive assessment of our data and methodology. We are pleased that you recognize the validity of our one-to-one persona-based prompting design with localized-language prompts, our use of temperature = 0, and the GPS-consistent scoring that ensures comparability with the human benchmark. As you note in your Comment 11, the temperature-zero choice may bear on response diversity; we therefore additionally report a high-temperature robustness experiment in our response to that comment and in [Appendix J](#).

We are also grateful for your recognition of the GPS as the most comprehensive cross-country benchmark of experimentally validated economic preferences, and of the well-defined cultural-dimension datasets (WVS, Hofstede, Schwartz) we draw on. In the revised manuscript, we have expanded our robustness checks: in addition to the existing English-only condition, we now report a response-constraint experiment and a paraphrasing experiment designed to rule out alternative explanations of the observed alignment patterns (such as prompt-format sensitivity and potential data contamination). The detailed designs and results of these new experiments are reported in our response to your Comment 7 below and in the corresponding Methods and Appendix sections of the revised manuscript.

5. Preregistration: The authors do not report preregistration, which is understandable given that no new human data were collected. All analyses are conducted on existing survey data and LLM-generated outputs. This is not a concern.

Appropriate use of statistics and treatment of uncertainties: The statistical tests are appropriate throughout.

Response: We sincerely thank you for this assessment. We are grateful for your understanding regarding preregistration. Our study relies on existing survey data and LLM-generated outputs, and therefore does not lend itself to preregistration in the conventional sense. We also appreciate your positive evaluation of our statistical analyses and treatment of uncertainty.

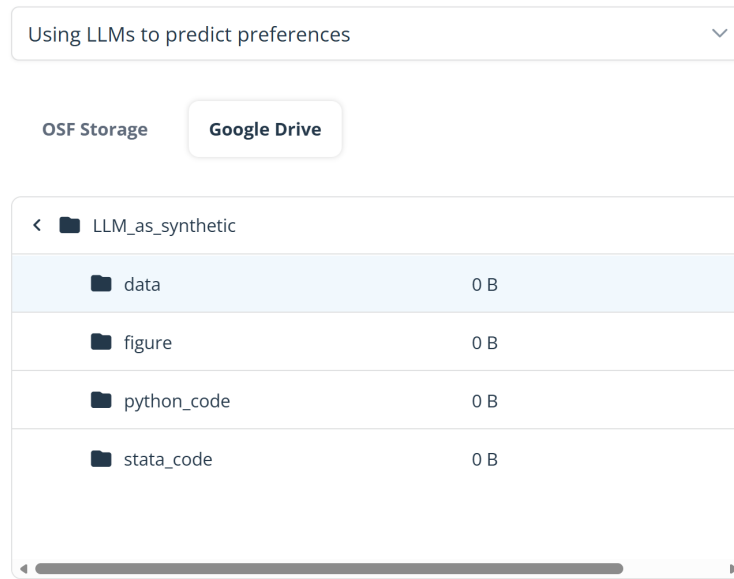
6. Custom code: The data is deposited on OSF, but I did not find the code.

Response: We sincerely thank you for catching this and apologize for the confusion. All custom code, data, and project documentation are in fact already deposited at the same OSF location, i.e.,

https://osf.io/pt2n4/overview?view_only=b705f153817542249d6e04bd511882f7

The reason the code may not have been visible is that the full set of LLM-generated responses across all conditions exceeds the storage capacity of OSF, so at submission we hosted the entire project on Google Drive and linked it from the OSF page rather than uploading the files to OSF directly.

To access the complete set of files, please click the Google Drive button on the OSF project page (see screenshot below). One known display issue with the OSF and Google Drive bridge is that all files are listed with a size of 0 B, which can make the folder appear empty; this is purely a display artifact and does not affect file accessibility. The code and data can all be accessed, downloaded, and executed normally.



7. Conclusions: Do you find that the conclusions and data interpretation are robust, valid and reliable?

The conclusions follow logically from the analyses and are stated with appropriate caution. The authors are careful to characterize alignment as “systematic but modest” rather than making stronger claims. The discussion of potential applications (hypothesis generation, study-design diagnostics) is balanced, and the limitations section is candid about the possibility that successful predictions may partly reflect memorization, since the GPS was conducted in 2012 and its content may appear in LLM training data. My main remaining question concerns generalizability to future models, but the between-model consistency analyses provide reasonable, if indirect, evidence on this point.

Response: Thank you for raising both the potential memorization (i.e., data-contamination) concern and the broader question of generalizability to future models. We address each concern in turn below.

Generalizability to newer models. To directly probe whether more recent models close the alignment gap we report in the main text, we conducted additional experiments using GPT-5.4 and GPT-5.4-mini, the most capable models available at the time of this revision

(released March 2026). At our initial submission, the GPT-5.4 series had not been released, hence the results from the GPT-5.4 series could be seen as an extrapolation.

As reported in [Figure 2](#) (mean squared errors, MSEs) in the revised main text and [Figure B.1](#) (Spearman’s correlations) in the revised Appendix (both figures reproduced below for your convenience), we find no statistically significant evidence that GPT-5.4 or GPT-5.4-mini outperform the earlier OpenAI models included in our main analysis (GPT-3.5-turbo, GPT-4o-mini, GPT-4.1-mini, and GPT-4o). Specifically, two-sided *t*-tests comparing the new and earlier OpenAI models yield no significant differences at either level or for either metric (individual-level MSE: $P = 0.400$; country-level MSE: $P = 0.474$; individual-level correlation: $P = 0.507$; country-level correlation: $P = 0.608$). While we acknowledge that null results from a small set of model comparisons cannot rule out future improvements, the absence of any detectable gain across four metrics for the most recent frontier models strengthens the generalizability of our main conclusions.

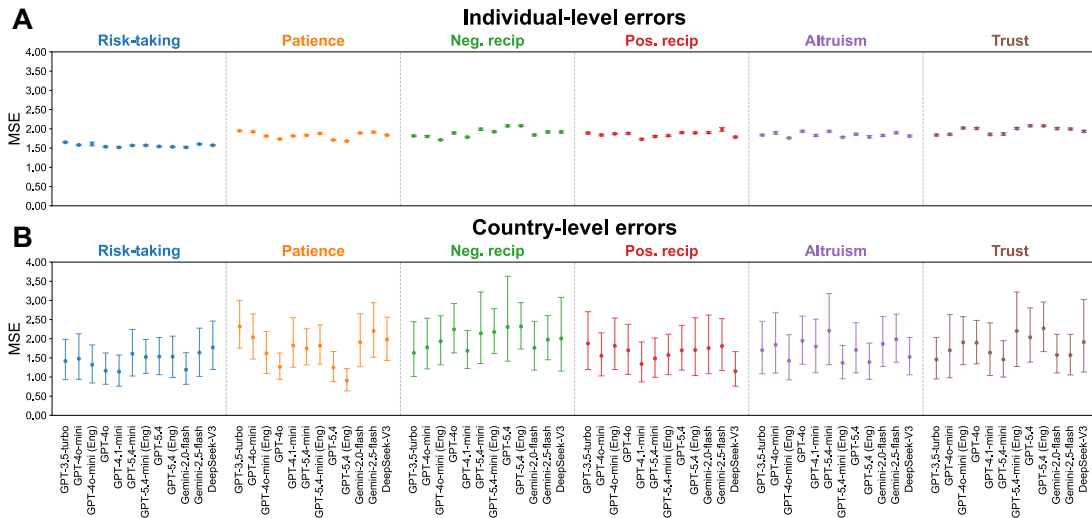


Figure 2. Mean squared errors (MSEs) between LLM-predicted and GPS-surveyed preferences. (A) Individual-level MSEs. Each point represents the MSE between LLM-predicted and GPS-surveyed preferences, pooling all individual observations across 74 countries ($N > 65,000$; the number of observations varies by preference domain due to missing values in the original GPS). **(B) Country-level MSEs.** Each point represents the MSE between country-level average LLM-predicted and GPS-surveyed preferences ($N = 74$). 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. Error bars indicate 95% bootstrapped CIs in both panels.

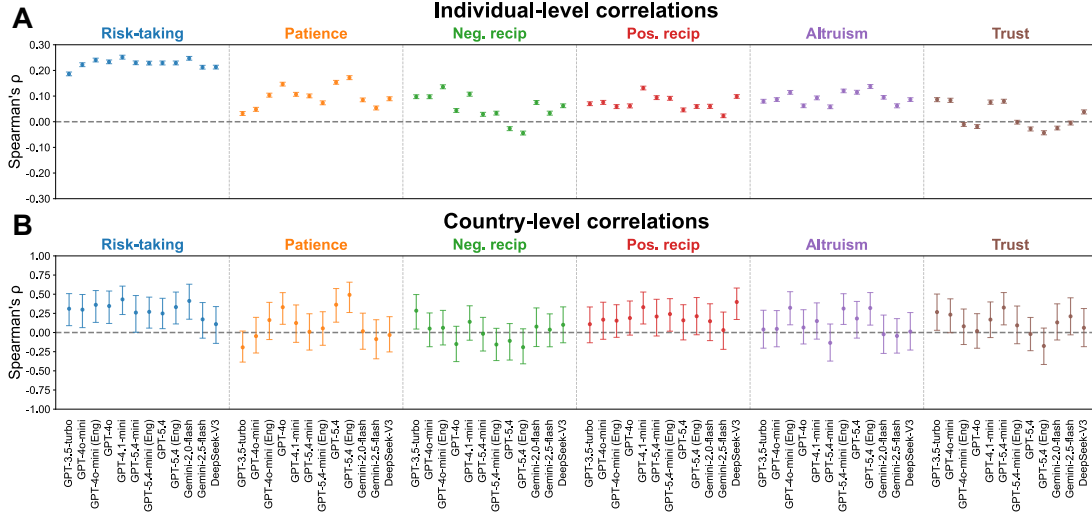


Figure B.1. Alignment between LLM-predicted and GPS-surveyed preferences. (A) Individual-level alignment. Each point represents the Spearman correlation coefficient between GPS-surveyed and LLM-predicted preferences, pooling all individual observations across 74 countries ($N > 65,000$; the number of 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.

Robustness to potential memorization. To address the potential memorization problem, we applied two strategies: (i) a paraphrasing experiment, and (ii) three English-only experiments using GPT-4o-mini, GPT-5.4-mini (newly added), and GPT-5.4 (newly added) instead of local languages. If contamination were the primary driver of alignment, we would expect substantial degradation in either the paraphrasing or the English condition; the absence of such degradation would be inconsistent with the hypothesis that contamination is the primary driver, although it could not rule out a partial role.

As shown in [Appendix Figures G.1](#) and [G.2](#) (both reproduced below), we do not detect statistically significant differences between local-language and English conditions on either set of metrics (two-sided paired t -tests pooling GPT-4o-mini, GPT-5.4-mini, and

GPT-5.4; individual-level correlations: $P = 0.903$; country-level correlations: $P = 0.307$; individual-level MSEs: $P = 0.488$; country-level MSEs: $P = 0.570$).

Compared with GPT-5.4-mini (English), individual-level correlations under paraphrasing are slightly higher for risk-taking, negative reciprocity, altruism, and trust, and slightly lower for patience and positive reciprocity. The MSE evidence is similarly mixed, with values varying across preferences. Pooling the individual- and country-level evidence from both metrics, we find no systematic effect of paraphrasing on alignment (two-sided paired t -test between GPT-5.4-mini and GPT-5.4-mini with paraphrasing; individual-level correlations: $P = 0.841$; country-level correlations: $P = 0.148$; individual-level MSEs: $P = 0.669$; country-level MSEs: $P = 0.694$). Detailed coefficients and MSEs are reported in [Appendix Table G.2](#) (reproduced below). Taken together, these two experiments give no evidence that memorization or data contamination is a primary driver of the observed alignment at either the individual or country level.

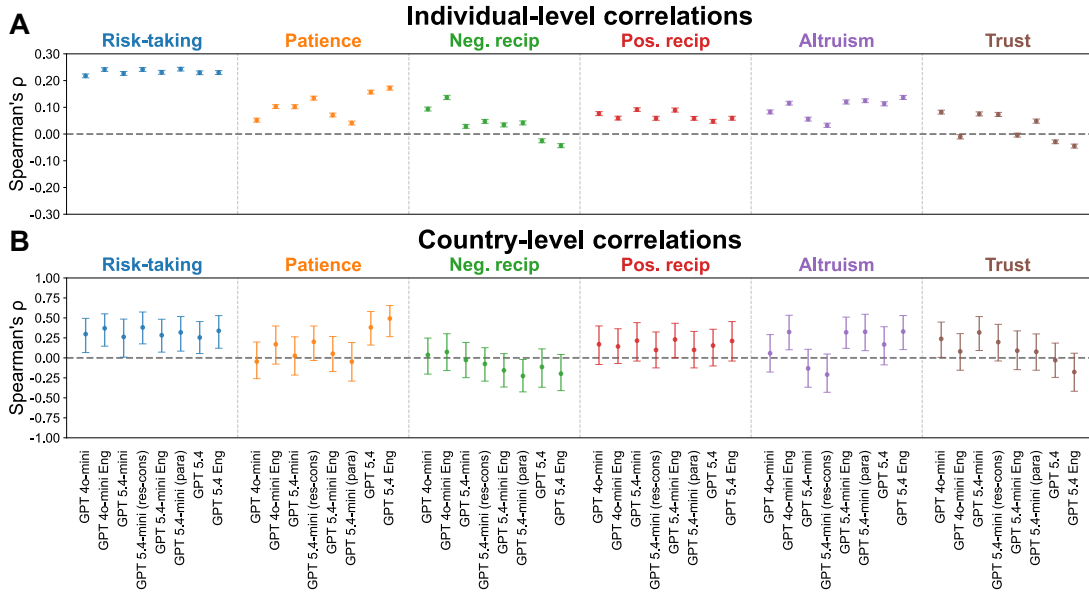


Figure G.1. Spearman correlations between LLM-predicted and GPS-surveyed preferences. **(A) Individual-level correlations.** Each point displays the Spearman's correlation coefficient between GPS-surveyed and LLM-predicted preferences, pooling all individual observations across 74 countries ($N > 65,000$, 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.

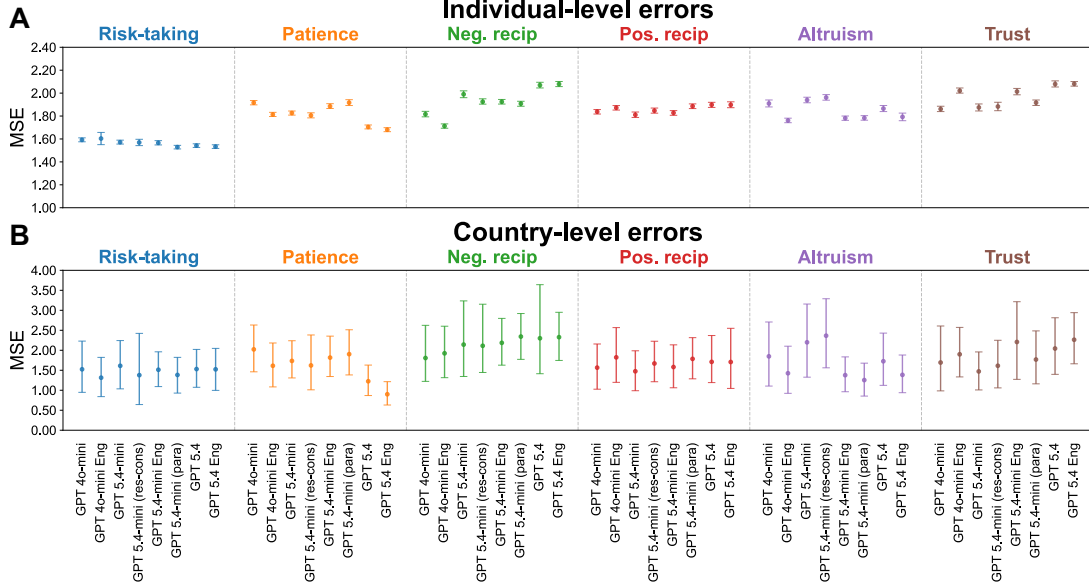


Figure G.2. Mean squared errors (MSEs) between LLM-predicted and GPS-surveyed preferences. (A) Individual-level MSEs. Each point displays the MSE between GPS-surveyed and LLM-predicted preferences, pooling all individual observations across 74 countries ($N > 65,000$, due to missing values in the original 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.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.

8. Clarify economic preferences for a multidisciplinary audience. The six preference domains are fundamental in economics but may not be self-evident to readers from other disciplines. I recommend adding a paragraph in the introduction that explains why these specific preferences matter for individual behavior, market outcomes, and policy design, with concrete examples, to improve accessibility for the broad readership of *Nature Human Behaviour*.

Response: Thank you for this helpful suggestion. We agree that the six preference domains, while standard in economics, deserve explicit motivation for the broad readership of *Nature Human Behaviour*. We have therefore expanded the introduction (paragraph 4) to introduce the six GPS preferences and their economic relevance:

“... The GPS also categorizes economic preferences into six distinct domains, namely risk-taking, patience, positive and negative reciprocity, altruism, and trust, that span the core dimensions of economic decision-making (Falk et al., 2018, 2023). Risk-taking captures attitudes toward uncertainty, while patience governs intertemporal trade-offs, predicting portfolio choice, entrepreneurial entry, educational attainment, and saving behavior (Dohmen et al., 2010; Sutter et al., 2013). Positive and negative reciprocity capture the inclination to reward kindness and punish unkindness, underpinning contract enforcement and workplace cooperation (Fehr & Gächter, 2000). Altruism drives charitable giving and pro-social behavior (Andreoni, 1990), and trust is a key correlate of financial development and economic growth (Algan & Cahuc, 2010). Together, these preferences shape aggregate market outcomes (e.g., savings rates, charitable giving, contract enforcement) and the effectiveness of policy interventions

ranging from retirement-savings nudges to public-good provision. Validating LLM-generated responses against this six-dimensional benchmark therefore provides a stringent test of whether synthetic respondents can support both behavioral science and applied policy research.”

9. Discuss the potential impact of Google Translate on results. The localized-language prompts are generated using Google Translate API (line 522). Machine translation quality varies substantially across languages, and translation errors could introduce systematic noise that differentially affects lower-resource languages, potentially confounding the GDP/IDI-related patterns in prediction errors. The authors should briefly discuss whether translation quality could be an alternative explanation for some of the observed cross-country variation in alignment.

Response: Thank you for raising the potential impact of Google Translate. Although using GPT-4o-mini as a translator would have been less costly, we chose Google Translate for its superior translation quality, particularly on low-resource languages (Zhu et al., 2024). Prior work in computer science has likewise relied on the Google Translate API to construct multilingual benchmarks (Upadhyay & Upadhya, 2023; Wang et al., 2025). To directly address this concern, we introduce two newly added experiments.

Newly added English-only experiments. We value your concern about residual translation noise and have therefore conducted a range of English-only experiments to control for systematic noise introduced by variation in machine-translation quality across languages. Unlike the original experiments, these new conditions deliver every prompt in English rather than in the respondent’s local language. The revised description of the experimental design now reads:

“Overall, our “virtual GPS” tests twelve LLM configurations, comprising nine local-language conditions (GPT-3.5-turbo, GPT-4o-mini, GPT-4o, GPT-4.1-mini, GPT-5.4-mini, GPT-5.4, Gemini-2.0 Flash, Gemini-2.5 Flash, DeepSeek-V3) and three English-only

conditions (GPT-4o-mini, GPT-5.4-mini and GPT-5.4; robustness check for prompt language), with each configuration completing the original GPS module for about 70,000 personas. We conduct eight follow-up experiments using GPT-5.4-mini: five prompt-component ablations, one paraphrasing variant, one response-constraint variant, and one high-temperature variant.”

Robustness check on translation quality. We use the consistency index (Spearman’s correlations for each country) and inconsistency index (MSE for each country) to measure the alignment and misalignment in the revised manuscript. We find that GDP/IDI-related patterns in prediction error are robust to both metrics and English-only sample.

(1) Robustness check of association between GDP/IDI and alignment. We report the GDP/IDI-related patterns in alignment in [Appendix E.2](#). In [Tables E.3](#) and [E.4](#) (both reproduced below for your convenience), we estimate a series of OLS regressions that examine which set of country-level proxies is more strongly associated with alignment, using two different metrics. If the pattern between prediction error and GDP/IDI is driven by machine-translation quality, that would disappear in our English-only experiment subsample.

a) Empirical evidence on the whole sample. The pattern is robust on the whole sample, consistent with our initial manuscript. 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 ($P < 0.001$). 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 not robustly different from zero across specifications. Because these four country-level variables are highly collinear, we do not include them jointly in a single specification. Instead, we conduct Principal Component Analysis (PCA) separately for the two sets of variables and extract the first principal component for economic development and for population. Column (5) of [Table 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 population is essentially zero and not statistically significant.

b) Empirical evidence from the English-only subsample. By design, the newly added English-only experiments are not affected by the potential systematic noise introduced by machine translation. Column (6) of [Table E.3](#) reports this specification on the English-only experiments. 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](#)).

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)	0.008*** (0.001)				
ICT Development Index			0.070*** (0.005)	0.077*** (0.006)		
log population	0.001 (0.001)		-0.001 (0.001)			
log number of internet users		-0.001 (0.001)		-0.001* (0.001)		
Score for economic development					0.009*** (0.001)	0.006*** (0.002)
Score for population					-0.001 (0.001)	-0.003** (0.001)
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 robust standard errors in parentheses. All regressions include preference fixed effects, model fixed effects, and their interaction terms. * $P < 0.10$, ** $P < 0.05$, *** $P < 0.01$.

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***	0.009***				

	(0.001)	(0.001)				
ICT Development Index			0.077***	0.083***		
			(0.005)	(0.006)		
log population	0.001		-0.001*			
	(0.001)		(0.001)			
log number of internet users		-0.001*		-0.001**		
		(0.001)		(0.001)		
Score for economic development					0.010***	0.007***
					(0.001)	(0.001)
Score for population					-0.001**	-0.004***
					(0.001)	(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 robust standard errors in parentheses. All regressions include preference fixed effects, model fixed effects, and their interaction terms. * $P < 0.10$, ** $P < 0.05$, *** $P < 0.01$.

2) Robustness check of association between GDP/IDI and prediction error. We report the GDP/IDI-related patterns in prediction error in [Appendix E.3](#). We next repeat the analysis using the inconsistency index as the dependent variable ([Table E.5](#), reproduced below).

a) Empirical evidence on the whole sample. As reported in [Table E.5](#), in all specifications, higher economic development is associated with lower inconsistency. 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.

b) Empirical evidence from the English-only subsample. 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 distinguishable from zero ($b = 0.015$, $P = 0.251$). As a robustness check, we redefine the inconsistency index as the MSE between sampling-weighted country-level average preferences in the LLM and GPS data. [Table E.6](#) (reproduced below) reports regression results using this alternative definition. All findings still stand.

Taken together, these results indicate that translation quality is unlikely to be the primary driver of the cross-country alignment patterns we document. LLMs achieve higher levels of alignment (i.e., smaller absolute prediction errors) for countries with higher levels of economic development, and larger deviations for countries with lower levels of development; population size and the number of internet users play at most a secondary role. Within the English-only subsample, the association between population-related variables and alignment varies across specifications, but the economic-related variables show a consistent pattern. The consistency of these patterns across the local-language and English-only experiments therefore rules out machine-translation quality as a leading explanation for the documented cross-country variation.

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*** (0.017)	-0.205*** (0.019)				
ICT Development Index			-1.343*** (0.131)	-1.494*** (0.150)		
log population	-0.020** (0.009)		-0.006 (0.009)			
log number of internet users		-0.011 (0.010)		-0.006 (0.010)		
Score for economic development					-0.201*** (0.019)	-0.058*** (0.017)
Score for population					-0.007 (0.010)	0.015 (0.013)
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 robust standard errors in parentheses. All regressions include preference fixed effects, model fixed effects, and their interaction terms. * $P < 0.10$, ** $P < 0.05$, *** $P < 0.01$.

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*** (0.036)	-0.260*** (0.039)				

ICT Development Index			-1.885*** (0.270)	-2.141*** (0.296)		
log population	0.018 (0.025)		0.048* (0.027)			
log number of internet users		0.042 (0.027)		0.051* (0.027)		
Score for economic development					-0.264*** (0.038)	-0.172*** (0.057)
Score for population					0.048* (0.027)	0.104* (0.053)
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 robust standard errors in parentheses. All regressions include preference fixed effects, model fixed effects, and their interaction terms. * $P < 0.10$, ** $P < 0.05$, *** $P < 0.01$.

10. Address the GPS vintage more directly. The GPS was conducted in 2012, over a decade before any of the LLMs were developed. Beyond training data contamination, it is possible that preferences in some countries have shifted since 2012, meaning that misalignment could partly reflect temporal change rather than model failure. A brief discussion of this would be valuable.

Response: We appreciate your raising the temporal-mismatch concern that people’s economic preferences may drift over time. Prior work suggests that economic preferences are largely stable within an individual’s lifespan (Meier & Sprenger, 2015; Schildberg-Hörisch, 2018) and over long country-level historical horizons (Galor & Özak, 2016). Nevertheless, cohort-level secular shifts since 2012 remain a plausible source of misalignment that is conceptually distinct from data contamination. We have added this point to the Discussion, which now reads:

“Several limitations qualify our conclusions. Because GPS was conducted in 2012, portions of its content may appear in LLM training data. Successful predictions could therefore partially reflect memorization, making our alignment estimates an upper

bound. To alleviate this concern, we ran a paraphrasing experiment and a range of English-only experiments (see Appendix G). If memorization drove alignment, paraphrasing items and translating them from the original survey languages into English should both weaken match with training data; neither manipulation produced systematic changes in individual- or country-level alignment. Additionally, secular shifts in economic preferences over the past decade could lead our analysis to underestimate alignment for current cohorts. To address both concerns simultaneously, future large-scale surveys and intervention experiments could be designed as synchronous digital-twin panels in domains where training data contamination is less likely, helping to disentangle memorization from genuine reasoning.”

Importantly, temporal mismatch also cannot explain our central findings that LLMs systematically recover individual preferences and demographic gradients. If cohort-level secular shifts since 2012 were the main driver of misalignment, both demographic and cultural gradients should degrade in parallel. Instead, we observe a sharp asymmetry: alignment is high at the demographic level but substantially lower at the cultural level (sign-based match ratio: 78.7% vs 56.6%; CI-based consistency ratio: 77.8% vs 48.1%). This is hard to reconcile with a uniform temporal-shift mechanism.

11. Elaborate on the temperature-zero design choice. Setting temperature to zero is appropriate for reproducibility, but it systematically reduces response variability. This may partly explain the reduced response diversity documented in Appendix G. A brief discussion of whether moderate temperature settings might yield closer distributional matches to human responses would be useful.

Response: Thank you for raising this point. We agree that setting temperature to zero, while appropriate for reproducibility, mechanically narrows the conditional output distribution and therefore likely contributes to the reduced response diversity documented in [Appendix J](#). We have addressed this in two places.

(1) Acknowledge the trade-off. First, we now flag the trade-off explicitly in the main-text paragraph (paragraph 6 of the Discussion) on practical constraints and direct readers to the appendix:

“... The other is limited response diversity. Compared with human responses in GPS, LLM-generated responses are more concentrated (Bisbee et al., 2024; D. Wang et al., 2026), and this concentration is itself patterned by country-level development (see Appendix J). Though part of this homogenization likely reflects our decision to set temperature to zero for reproducibility, increasing the temperature to 1 only partially restores response diversity, which still remains substantially below human levels (see Appendix J).”

(2) A robustness check experiment. We conducted a high-temperature experiment ($T = 1$) using GPT-5.4-mini in localized languages as a robustness check. 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: $P = 0.946$; country-level correlations: $P = 0.910$; individual-level MSEs: $P = 0.995$; country-level MSEs: $P = 0.880$). Detailed correlation coefficients and MSEs are reported in [Table J.3](#) (reproduced below for your convenience).

Turning to diversity, [Figure J.3](#) (reproduced below for your convenience) 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, $P < 0.001$; pooling all countries and preferences). However, the diversity ratio at $T = 1$ remains significantly below 1 (for each preference, $P < 0.001$). 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 (Chen et al., 2023; A. Wang et al., 2025; D. Wang et al., 2026; S. Wang et al., 2025).

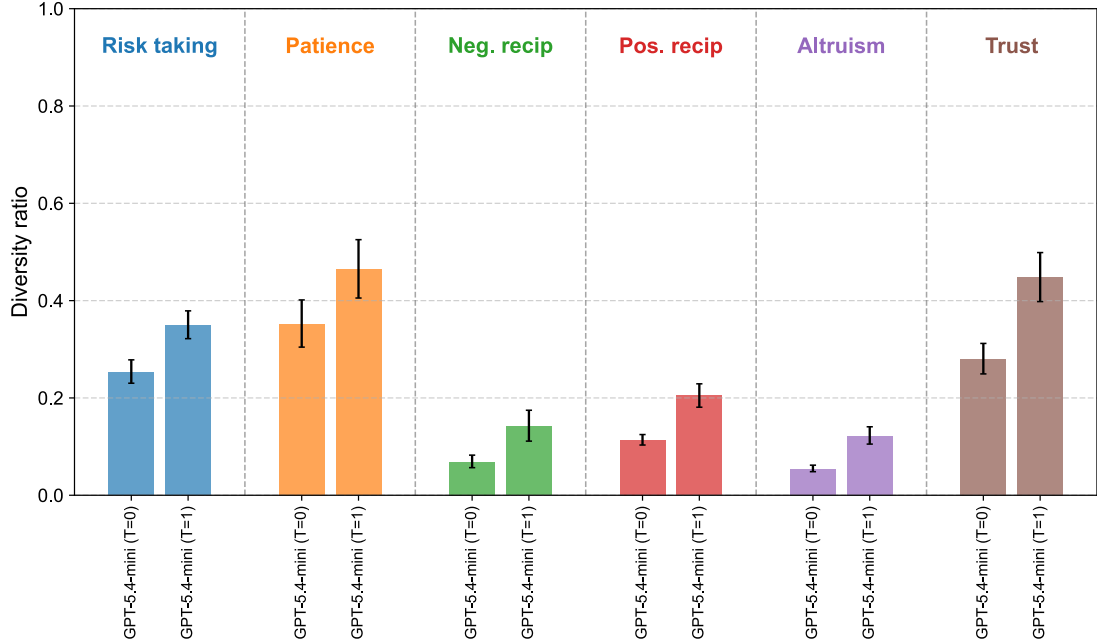


Figure J.3. Diversity ratios for GPT-5.4-mini at temperature 0 and temperature 1, across preference domains. Bars show the mean diversity ratio across countries for each combination of preference domain and temperature setting ($T = 0$ vs $T = 1$), generated by GPT-5.4-mini. The diversity ratio is defined 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.3. 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	0.090
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	1.612	1.734	2.142	1.475	2.199	1.471	1.772

($T = 0$)							
GPT-5.4-mini	1.452	1.738	2.159	1.475	2.158	1.473	1.743
($T = 1$)							

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

12. Consider variance decomposition. Reporting what fraction of variance in human preferences is explained by LLM-predicted preferences (e.g., R^2 values) would give readers a clearer sense of practical significance beyond correlation coefficients.

Response: We appreciate your advice on variance decomposition. Reporting the explanatory power of LLM-predicted preferences would help readers better understand the practical significance. Beyond the correlation coefficients and MSEs, we reported the R^2 values in the main text:

“Finally, to interpret the above results, we calculate the proportion of variance in GPS-surveyed preferences explained by LLM-predicted preferences. We estimate this by running OLS regressions of GPS-surveyed preferences on LLM-predicted preferences at both the individual and country levels, and calculating the resulting R^2 . 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 (see Appendix Tables B.1–B.2). Again, the higher R^2 for country-level predictions likely reflects the fact that aggregating responses within countries removes a lot of noise compared to the individual-level data. The best-performing combinations remain modest: at the individual level, GPT-4.1-mini reaches 5.8% on risk-taking; at the country level, GPT-5.4 (English) reaches 29.2% on patience. Nonetheless, both sets of R^2 s are very low, indicating that LLM-predicted preferences exhibit only weak positive alignment with the human benchmark.”

Specifically, we estimate R^2 from individual-level and country-level OLS regressions of GPS-surveyed preferences on LLM-predicted preferences. 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. Across preference domains, the explanatory power is higher for risk-taking than for trust. These findings are consistent with evidence based on other measures. For your convenience, we reproduce the full [Tables B.1](#) and [B.2](#) below.

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.

13. Engage with recent related work. Two recent economics working papers are directly relevant. Wu, Xi, and Xie (2025, NBER WP 34308) propose an LLM-based survey framework for U.S. inflation expectations using date-restrictive prompting and recover human-comparable treatment effects in a single-country setting. The contrast is instructive: LLMs may perform well where extensive high-quality training data exist but poorly in culturally specific, data-sparse contexts. Asirvatham, Mokski, and Shleifer (2026, NBER WP 34834) examine LLMs as measurement tools in economics, providing complementary evidence. Discussing these papers would situate the study more fully within the rapidly evolving literature.

Response: We thank you for pointing us to these two important and timely contributions. We have integrated both into the Discussion. Wu, Xi, and Xie (2025) now appears in our list of prior work showing that LLMs can be used to probe latent motivational structures in WEIRD populations, and Asirvatham, Mokski, and Shleifer (2026) is cited alongside Fang, Li, and Lu (2025) as a complementary line of work that applies LLMs as a measurement tool rather than as a synthetic respondent. We also share your reading that the contrast between Wu, Xi, and Xie (2025) and our findings is instructive: LLMs may perform well in high-resource, single-country settings where extensive training data exist,

but their reliability degrades in culturally specific, data-sparse contexts. The relevant Discussion paragraphs now read:

*“Despite these constraints, our individual- and demographic-level alignment results highlight several promising scientific uses of LLMs as synthetic respondents. First, LLMs can be used to propose and probe latent motivational structures underlying human behavior, particularly in WEIRD populations, complementing formal decision theory (Bhatia, 2024; Bhatia et al., 2025; Binz et al., 2025; **Wu et al., 2025**; Xie et al., 2025; Zhu et al., 2025). Second, LLMs can accelerate early-stage hypothesis generation and study-design diagnostics, helping researchers identify promising interventions or measurement strategies before turning to human samples. They can help identify when cross-cultural extrapolation may fail (Henrich et al., 2010; Lowes & Nunn, 2025; Moscona et al., 2026), and flag measurement-context mismatches (Cohn et al., 2019; Huynh et al., 2020; Yang et al., 2023). They may also serve as a fast and low-cost environment for follow-up experiments and robustness checks (Guilbeault et al., 2025), and provide a safer testbed for exploring ethically sensitive counterfactuals, such as mechanisms that increase or reduce intergroup hostility, before designing human-subject experiments. Beyond synthetic respondents, a complementary application is to use LLMs as a measurement tool (Asirvatham et al., 2026; Fang et al., 2025), expanding social scientists’ toolkit.”*

14. Strengthen the discussion of amplification. The finding that LLMs amplify demographic effects (Figure 4) is important but receives relatively brief treatment. The authors could discuss whether amplification varies systematically by model architecture, country, or cultural context.

Response: We thank you for highlighting the amplified demographic differences. To address this point, we re-designed Figure 4 (reproduced below for your convenience) so the amplification reads off directly: any point above the 45° line indicates that the LLM-

predicted demographic effect exceeds the GPS-surveyed effect. We also state the amplification explicitly in the main text, which now reads:

“Although LLMs reproduce the direction of most demographic differences, they do so with systematically inflated magnitudes. Specifically, the effect sizes estimated from LLM-predicted preferences are larger than their survey-based counterparts (Bisbee et al., 2024). This is especially the case for risk-taking and patience (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 (Exley et al., 2025; A. Wang et al., 2025).”

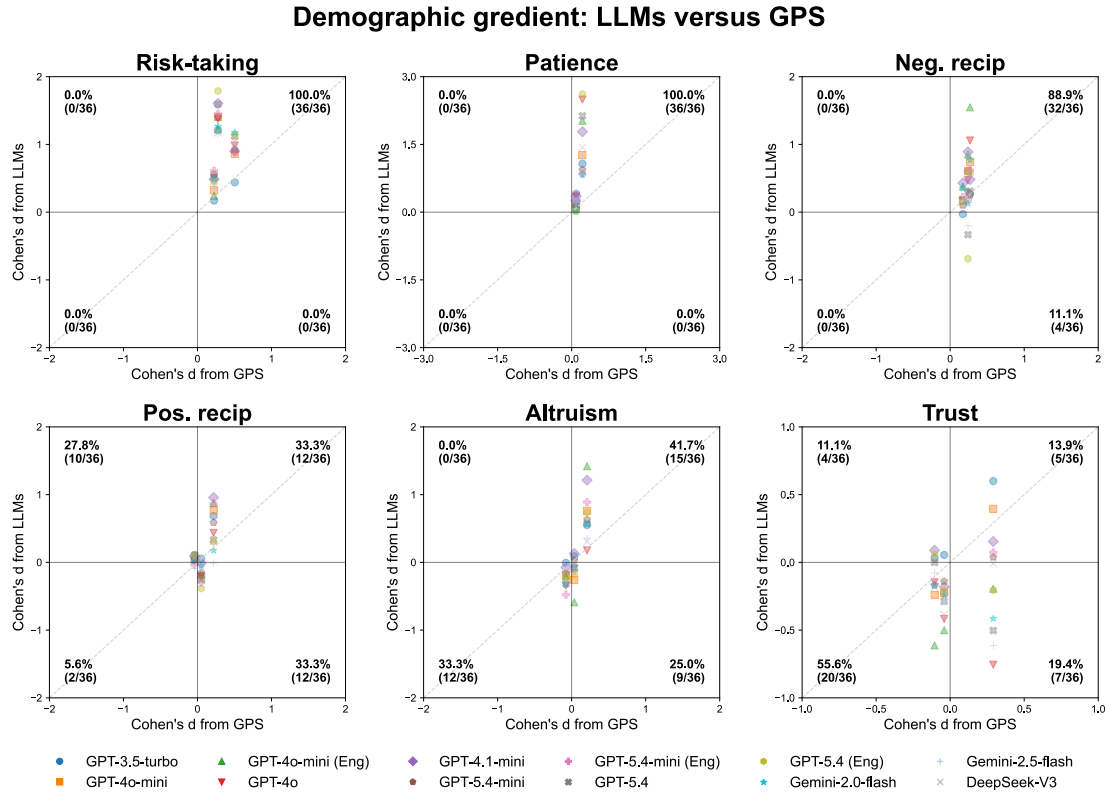


Figure 4. Demographic gradients in GPS-surveyed and LLM-predicted preferences. Each of the six panels corresponds to one preference domain. Within each panel, the x-axis shows effect sizes (Cohen's d) estimated from GPS-surveyed preferences and the y-axis shows the corresponding effect sizes estimated from LLM-predicted preferences. Demographic variables are encoded as binary indicators for gender (male vs female), age (< 60 vs ≥ 60), and cognitive ability (≥ 5 vs < 5). Each point represents one (LLM $\times$ demographic dimension) pair; with 12 LLM configurations and three demographic dimensions (gender, age, and cognitive ability), each panel

contains 36 points. Positive effect sizes indicate higher preference levels for males (vs females), younger (vs older), and higher cognitive ability (vs lower). The dashed 45° line denotes exact consistency between LLM- and GPS-derived effect sizes in both sign and magnitude. The percentage in each quadrant reports the share of the 36 (LLM × demographic dimension) pairs falling in that quadrant, and Quadrants I (upper-right) and III (lower-left) jointly indicate sign-consistent alignment with GPS. Marker shape and color denote the LLM configuration (see legend). Axis ranges differ across panels for visualization clarity and should not be compared across panels.

Beyond the overall result, we trace the origin of the amplification through one diagnostic claim and one speculative claim. The diagnostic claim is that the amplification arises from demographic information in the system prompts, independent of country information. We ran a range of diagnostic ablation experiments; the corresponding system prompts are reported in [Table F.1](#) (reproduced below). The speculative claim is that the amplification co-varies with the cultural background. We construct a gender-stereotype index from multiple resources and find that the gender bias (i.e., the extent of amplification in the gender domain) is associated with this index, though we treat the latter as marginal and speculative evidence.

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.

Amplification arises from demographic information. If demographic information in system prompts produces the demographic gradient, rather than country information or random noise, then the effect sizes of partial-persona conditions should be attenuated, and those of country-persona conditions could be attenuated more. We therefore expect: effect sizes (full-persona) $\approx$ effect sizes (demographic-persona) $>$ effect sizes (partial-persona) $\approx$ effect sizes (country-persona).

Here, effect size (partial-persona) 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.

As shown in Figure 7B (reproduced below), effect sizes from the demographic-persona condition are close to those from the full-persona condition ($M_{\text{full-persona}} = 0.303$ vs $M_{\text{demographic-persona}} = 0.303$). Effect sizes are substantially attenuated in the partial-persona and country-persona conditions ($M_{\text{partial-persona}} = 0.046$ vs $M_{\text{country-persona}} = 0.016$, corresponding to 84.8% and 94.7% attenuation, respectively).

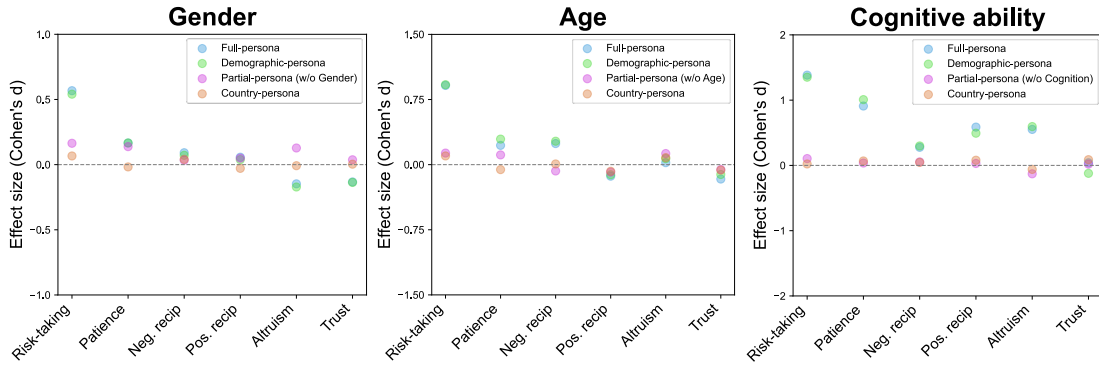


Figure 7(B). Demographic effect sizes across ablation conditions. Each subpanel corresponds to one demographic dimension (gender, age, cognitive ability). Each point plots a Cohen's d effect size of a demographic variable (female vs male, older vs younger, lower vs higher cognitive ability) estimated from LLM-predicted preferences under each ablation condition. For each subpanel, four conditions are shown: full-persona, demographic-persona, country-persona, and the matched partial-persona (e.g., 'w/o Gender' in the Gender subpanel; see legend in each panel). y-axis ranges differ across subpanels to accommodate effect-size magnitudes.

A linear mixed-effects regression across model $\times$ preference $\times$ demographic-dimension combinations confirms this pattern. As reported in [Table F.3](#) (reproduced below), compared with the country-persona condition, the coefficient of partial-persona is not significant at the 5% level, while both coefficients of full-persona and demographic-persona are significant ($P < 0.001$) in the linear mixed-effects model. [Table F.4](#) (reproduced below) reports the detailed effect sizes estimated from each condition. The ablation experiments thus provide solid causal evidence that demographic information drives the demographic gradient.

Table F.3. Linear mixed-effects model of effect sizes (Cohen's d)

	DV: Cohen's d
Condition	
Full-persona	0.287*** (0.079)
Demographic-persona	0.287*** (0.079)
Partial-persona	0.030 (0.079)
Preference	
Risk-taking	0.434*** (0.097)
Patience	0.166* (0.097)
Negative reciprocity	0.025 (0.097)
Positive reciprocity	-0.014 (0.097)
Trust	-0.138 (0.097)
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. Standard errors are in parentheses. * $P < 0.10$, ** $P < 0.05$, *** $P < 0.01$.

Table F.4. 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.

Amplification covaries with the cultural background. For the speculative claim, we relate gender gaps in LLM-predicted preferences to a country-level index of gender stereotypes. We first draw on gender-stereotype items from the World Values Survey (WVS) and use principal component analysis (PCA) to extract the first principal component as a WVS-based gender-stereotype index (see [Appendix K](#) for variable definitions). We then combine this WVS index with three gender-equality indicators, the Global Gender Gap Index (GGGI), the UNDP Gender Inequality Index (GII), and the UNDP Gender Development Index (GDI), and 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](#) (reproduced below) 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.

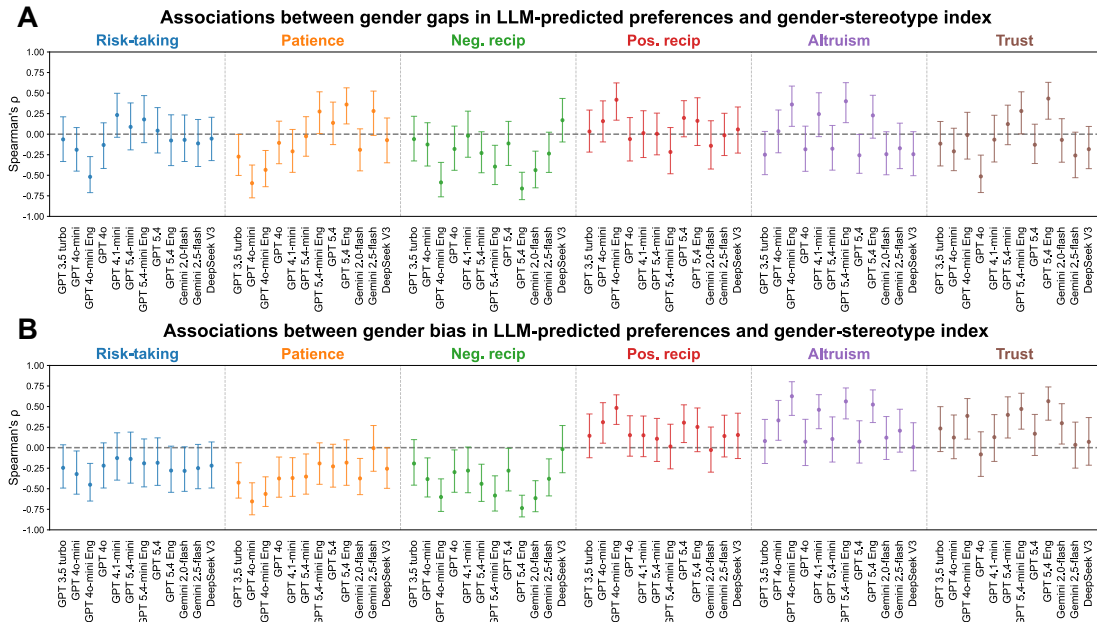


Figure I.1. Associations between gender gaps, gender biases, and gender-stereotype index. (A) Gender gaps in LLM-predicted preferences and the gender-stereotype index. For each model and preference domain, we estimate country-specific gender gaps in preferences using the regression specification of Falk 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. Error bars indicate 95% bootstrap CIs in both panels.

To directly address the three dimensions of variation you raise:

(1) Model architecture. Across the 12 LLM configurations, we find some signs of agreement among different model families (GPT vs Gemini vs DeepSeek) as reported in [Table I.1 \(reproduced below\)](#). Most models overpredict women’s altruism and trust. From GPT-4o, the GPT models start underpredicting women’s risk-taking, like Gemini and DeepSeek.

(2) Country. Our country-level analysis is reported in [Appendix I](#) and summarized above. We focus on gender because gender norms have the most established cross-country quantitative indicators; analogous indices for age and cognitive ability across 74 countries are not, to our knowledge, available with comparable coverage.

(3) Cultural context. The composite gender-stereotype index combines four cultural and institutional indicators (WVS gender-stereotype items, GGGI, GII, and GDI). The systematic association between this index and gender bias for risk-taking, patience, and negative reciprocity (all 12/12 negative; see Panel B of [Figure I.1](#)) indicates that amplification covaries with cultural context, at least along the gender dimension.

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

	Difference in gender gap (LLM – GPS)					
	Risk-taking	Patience	Negative reciprocity	Positive reciprocity	Altruism	Trust
GPT-3.5-turbo	0.19***	0.12***	0.41***	-0.03	0.06**	-0.04
GPT-4o-mini	0.01	0.03	0.03	-0.00	0.28***	0.29***
GPT-4o-mini English	0.08***	0.12***	-0.13***	0.02	0.28***	0.37***
GPT-4o	-0.23***	0.09***	-0.01	0.01	0.43***	0.35***
GPT-4.1-mini	-0.14***	-0.05*	-0.27***	-0.04	0.18***	0.29***
GPT-5.4-mini	-0.24***	-0.01	0.05	-0.05**	0.27***	0.20***

GPT-5.4-mini English	-0.30***	-0.00	-0.08***	0.10***	0.53***	0.17***
GPT-5.4	-0.18***	0.14***	-0.12***	-0.10***	0.31***	0.28***
GPT-5.4 English	-0.13***	0.22***	-0.05	-0.04**	0.22***	0.17***
Gemini 2.0 Flash	-0.19***	0.15***	-0.22***	-0.03	0.45***	0.18***
Gemini 2.5 Flash	-0.08***	0.13***	-0.05	0.13***	0.39***	0.23***
DeepSeek-V3	-0.17***	-0.07***	0.02	-0.09***	0.24***	0.44***

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

15. **Exploration of Error Mechanisms:** While the paper identifies that errors track economic development, further investigation into the specific mechanisms underlying these geographical and cultural biases could be beneficial. This might involve qualitative analysis of LLM outputs from different regions or comparing the training data characteristics of LLMs across languages/cultures.

Response: We appreciate your suggestions on probing the mechanisms behind the geographic and cultural error patterns. In the revised manuscript, we approach this question from two complementary angles: (1) a quantitative decomposition of prediction errors into demographic, country-characteristic, and country-fixed-effect components; and (2) indirect evidence from prior work and disclosure from OpenAI and Anthropic.

Quantitative decomposition of prediction errors. We first decompose the prediction errors at the individual level by regressing the average individual-level MSE across the six preferences on three sets of predictors: demographic attributes, country-level economic and infrastructural characteristics (GDP, IDI, population, internet users), and country fixed effects (see [Methods](#)). [Figure 6C](#) (reproduced below for your convenience) reveals two patterns. First, in every model the country fixed effects explain the largest share of the prediction error, ranging from 3% to 37% across models. This finding is consistent with our earlier observation that demographic information in the persona drives most of the alignment, leaving country context as the dominant residual source of misalignment. Second, among observable country characteristics, economic-related variables (GDP, IDI)

explain a substantially larger share of the error than population-related variables (internet users, population size), although the absolute explanatory power of either set remains modest (no more than about 7% in any model–preference cell). Panel B of Figure 6 separately reports Spearman correlations between the country-level inconsistency index and four country characteristics; GDP per capita and IDI are negatively correlated with MSE in the majority of model–preference combinations, whereas population size and the number of internet users show no systematic association.

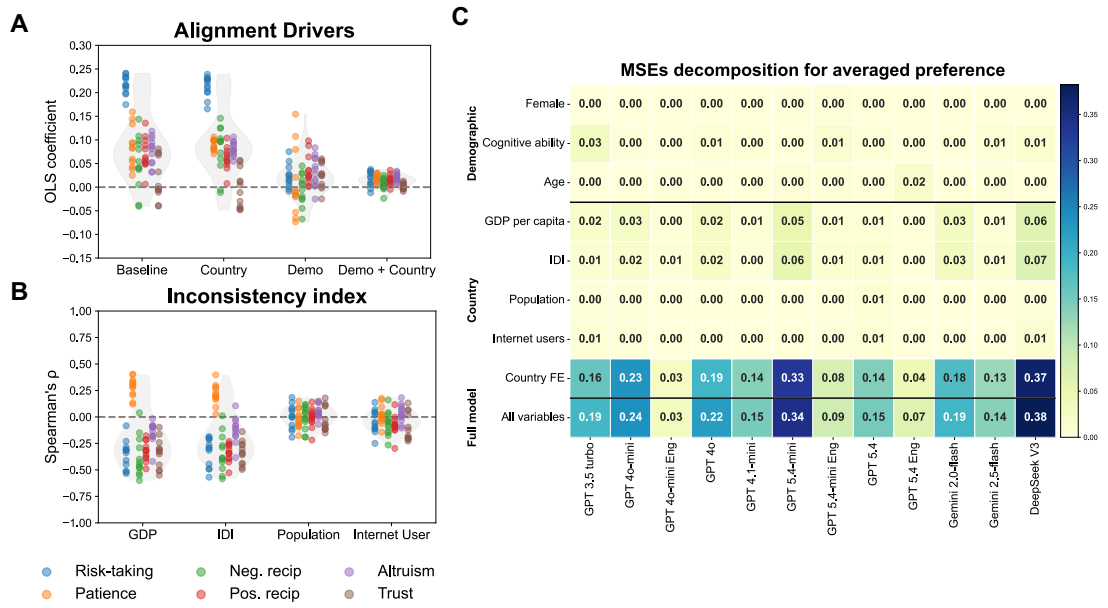


Figure 6. Decomposing the sources of LLM-GPS misalignment. (A) Regression coefficients from models regressing GPS-surveyed preferences on LLM-predicted preferences. Each point represents an OLS coefficient estimated on pooled individual-level data for a given model $\times$ preference combination (12 LLMs $\times$ 6 preferences). The four categories on the x-axis correspond to four alternative regression specifications: baseline (no controls), with country indicators, with demographic characteristics, and with both sets jointly. **(B)** Spearman correlations between the inconsistency index (the GPS sampling-weighted within-country mean of individual-level MSEs) and four country characteristics: log GDP per capita, the Information and Communications Technology Development Index (IDI), log population, and log number of internet users. Each point represents a Spearman correlation coefficient computed across 74 countries. Each characteristic shows 72 points (12 LLMs $\times$ 6 preferences). **(C)** Error decomposition of LLM-predicted preferences. For each LLM (column), separate OLS regressions are estimated with the average individual-level MSE across six preferences as dependent variable and each variable category (row) as independent variable. Cell values report the R^2 of each single-predictor regression.

Figures E.1–E.6 report the decomposition results separately for each model and each preference. The pattern is consistent: demographic variables and country characteristics explain at most about 6% of the prediction error in any cell (the highest value occurs for trust under DeepSeek-V3), whereas country fixed effects explain 3% to 40%. Across all models and preferences, the average explanatory power of economic-related variables is roughly an order of magnitude larger than that of population-related variables (mean R^2 of 0.020 vs 0.002; two-sided t -test, $P < 0.001$).

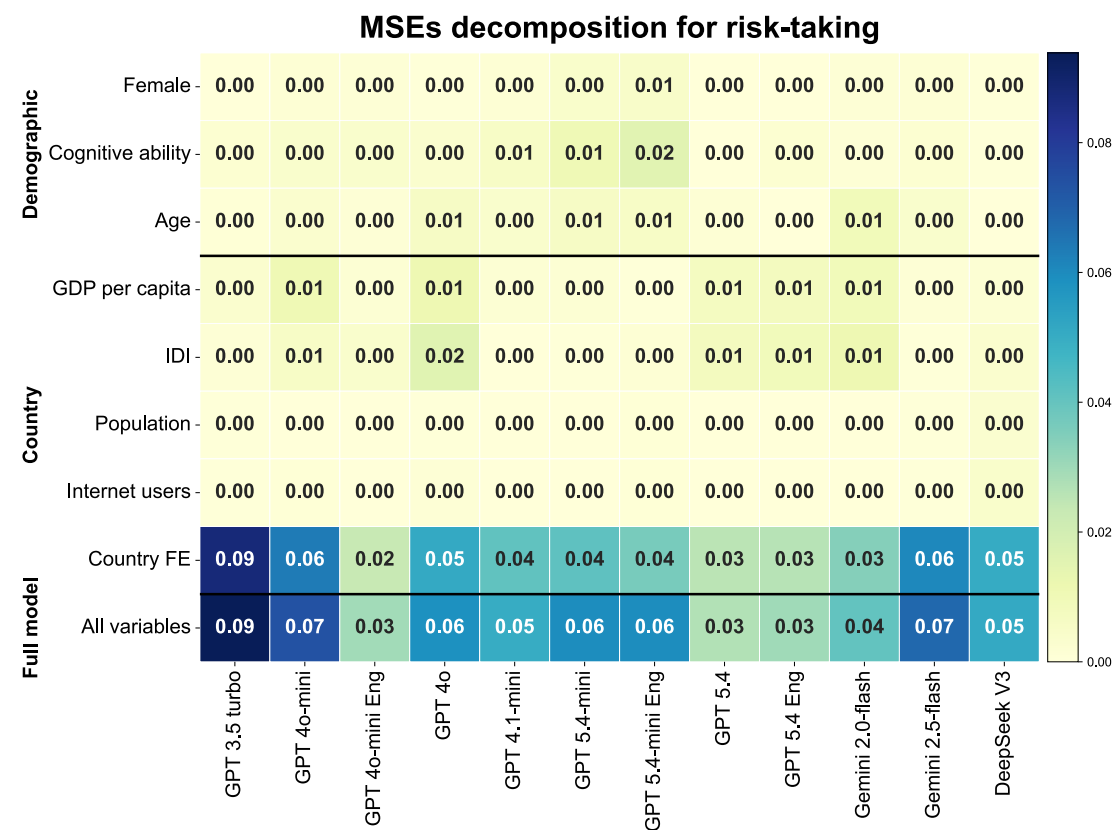


Figure E.1. Error decomposition of LLM-predicted preferences (risk-taking). The number in each cell represents the R^2 estimated from a regression where the individual-level MSE for risk-taking is the dependent variable and corresponding variables indicated by the row label are the independent variables.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

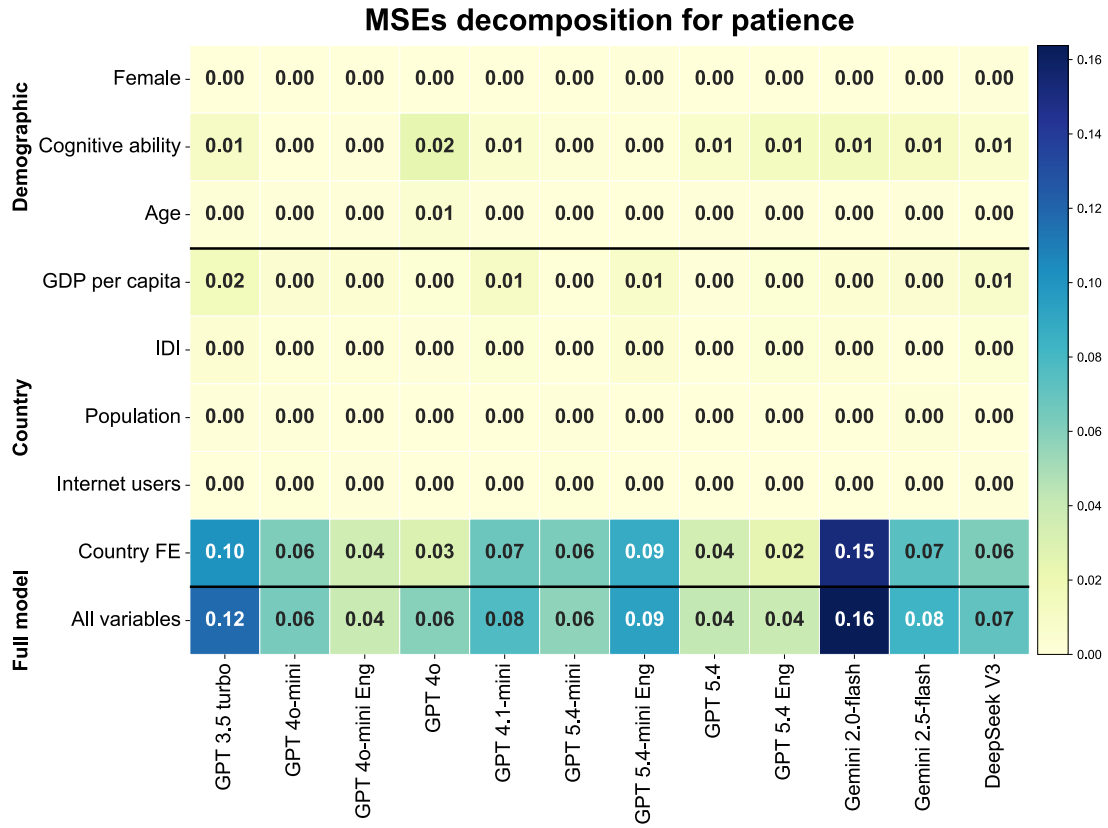


Figure E.2. Error decomposition of LLM-predicted preferences (patience). The number in each cell represents the R^2 estimated from a regression where the individual-level MSE for patience is the dependent variable and corresponding variables indicated by the row label are the independent variables. Columns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

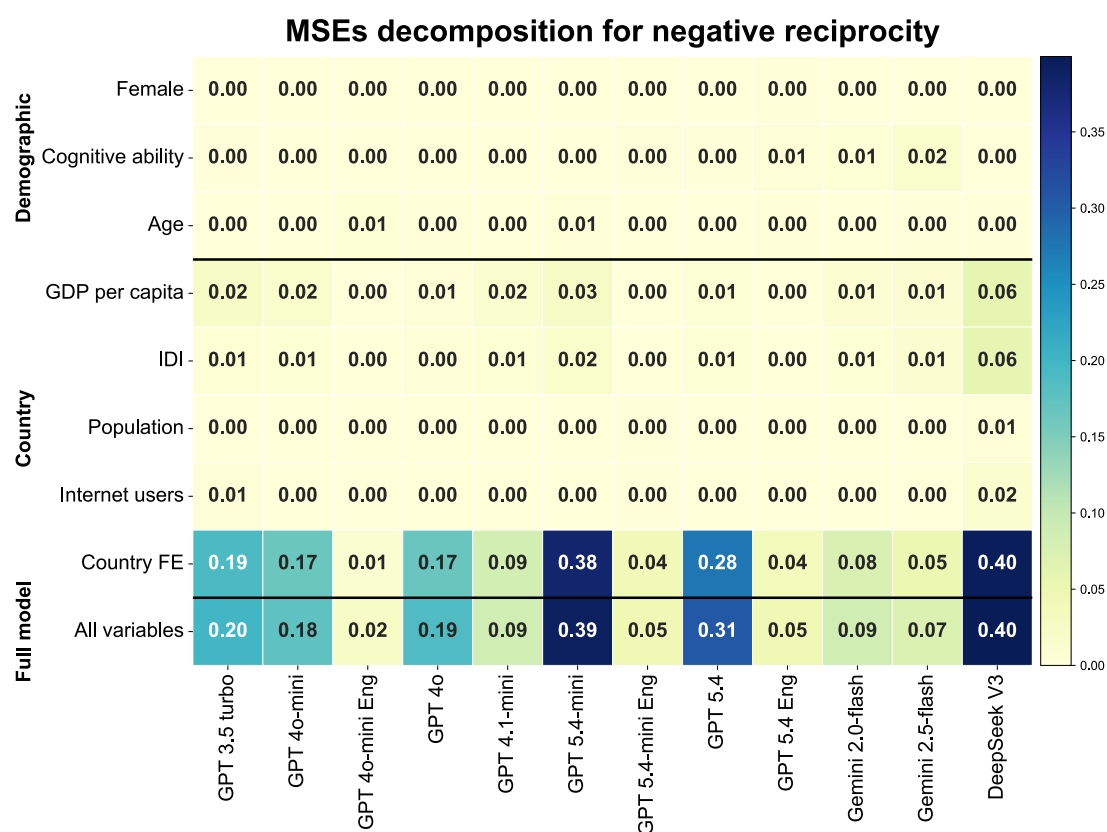


Figure E.3. Error decomposition of LLM-predicted preferences (negative reciprocity).
The number in each cell represents the R^2 estimated from a regression where the individual-level MSE for negative reciprocity is the dependent variable and corresponding variables indicated by the row label are the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

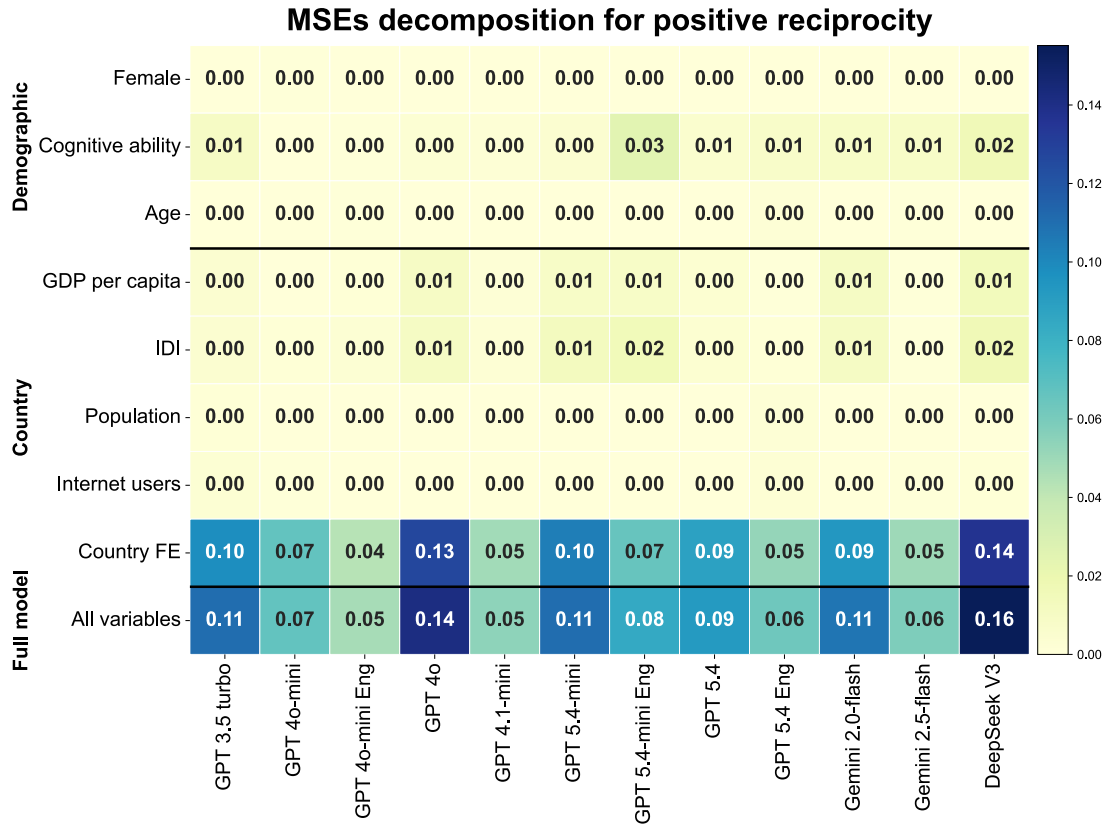


Figure E.4. Error decomposition of LLM-predicted preferences (positive reciprocity). The number in each cell represents the R^2 estimated from a regression where the individual-level MSE for positive reciprocity is the dependent variable and corresponding variables indicated by the row label are the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

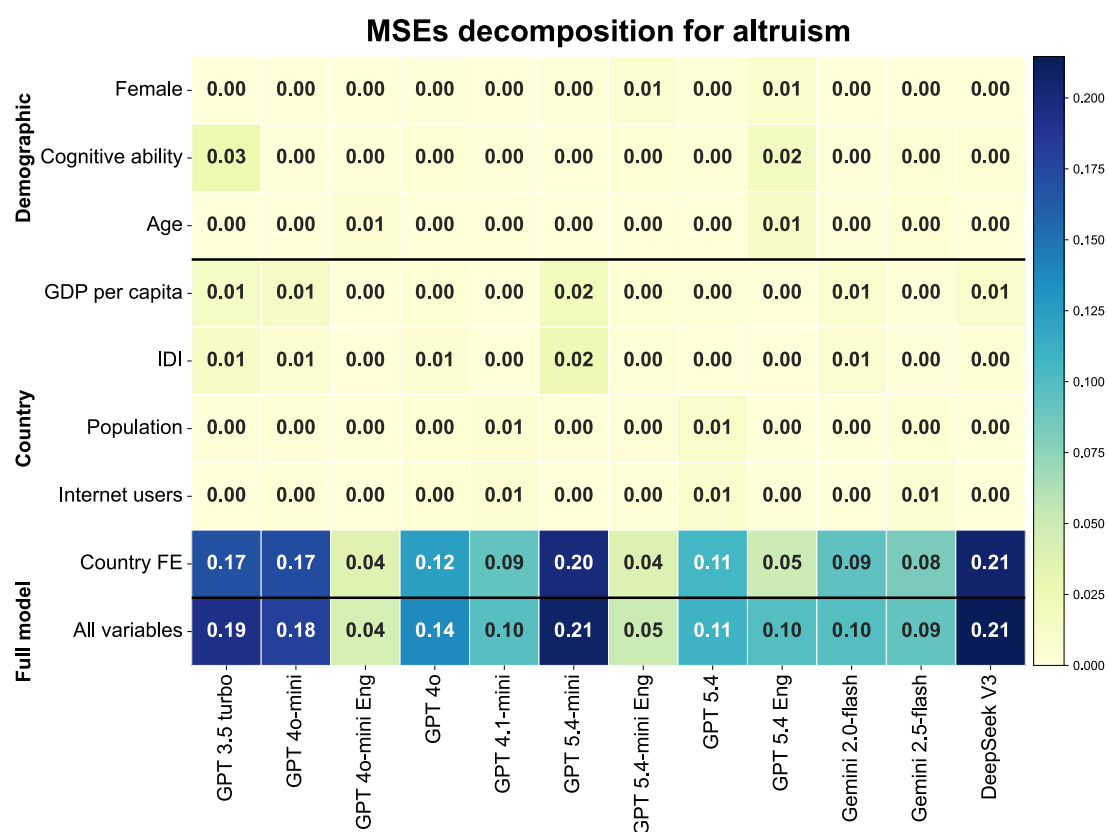


Figure E.5. Error decomposition of LLM-predicted preferences (altruism). The number in each cell represents the R^2 estimated from a regression where the individual-level MSE for altruism is the dependent variable and corresponding variables indicated by the row label are the independent variables. Columns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

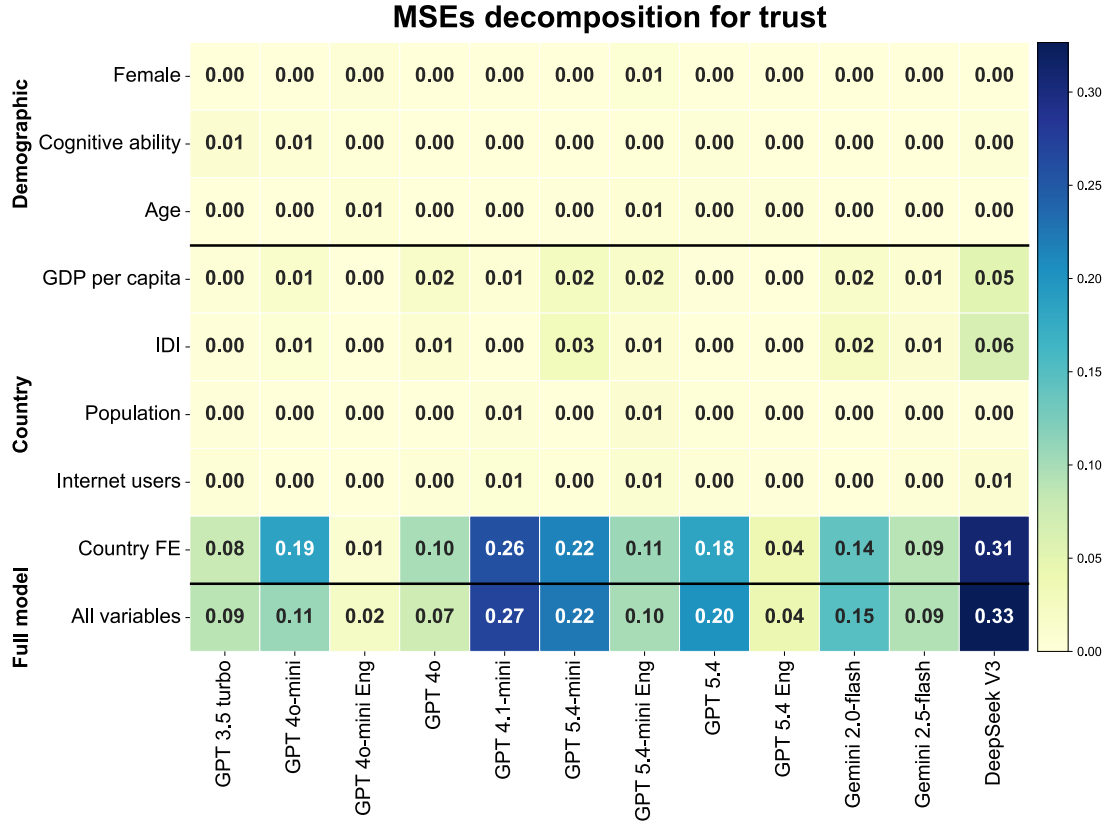


Figure E.6. Error decomposition of LLM-predicted preferences (trust). The number in each cell represents the R^2 estimated from a regression where the individual-level MSE for trust is the dependent variable and corresponding variables indicated by the row label are the independent variables. Columns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

Indirect evidence on error mechanisms. Although a direct inspection of training corpora would be more informative, it is not feasible: the pretraining data of major LLMs remain proprietary and inaccessible. We can still infer the underlying mechanism linking GDP to corpus characteristics and model optimization. Two strands of indirect evidence are nonetheless helpful to support our findings.

First, public disclosures from OpenAI and Anthropic show that the positive relationship between GDP and LLM usage is robust (Appel et al., 2025; Chatterji et al., 2025; Ruth et al., 2026). Hence, developers may have more training data from high-GDP countries, since these markets generate more user-interaction data used for model improvement (King et al., 2025). Consistent with this asymmetry, prior NLP research documents systematic underrepresentation of low-resource languages in pretraining resources (Joshi et al. 2020).

Second, developers have weaker commercial incentives to optimize and train models for users in low-GDP countries. The tokenizer is one example: it is a relatively separate component from pretraining and post-training and is crucial to model performance (Ali et al., 2024); cross-language disparities emerge at the tokenization stage, before a model is even invoked (Petrov et al., 2023); yet tokenizer design has remained largely unchanged across model generations and has received little optimization attention (Alqahtani et al., 2026; Dagan et al., 2024).

16. Fine-tuning: Explore whether targeted fine-tuning of LLMs on more diverse or culturally-specific datasets could improve their performance in capturing human preferences, especially at the cultural level.

Response: We thank you for this constructive suggestion. While targeted fine-tuning is beyond the scope of an audit study, we have responded along two complementary lines. First, we expanded the Discussion to point readers toward concrete intervention strategies documented in prior work. The revised paragraph now reads:

“Moreover, although we find systematic individual-level alignment, the mechanisms by which models reproduce these patterns remain opaque. At least two directions for future research are worth pursuing. The first is combining audits with targeted interventions, such as data augmentation in under-represented contexts, cross-lingual grounding, and domain-specific alignment, to test whether the micro-to-macro gap can be narrowed. Prior work shows that cultural- and country-level alignment can be improved by prompt injection (Pataranutaporn et al., 2025; Tao et al., 2024) and fine-tuning (Li et al., 2024; Suh et al., 2025), suggesting one promising route to addressing misalignment associated with country fixed effects. The second is to probe LLMs’ internal representations and apply mechanistic interventions such as steering, to improve alignment (Khanuja et al., 2026).”

Second, we conducted a prompt-component ablation (Appendix F). Specifically, we removed the demographic information in the system prompt, and only kept the country information. As a consequence, the country is the only information that LLMs could use to infer preferences. Although this does not modify model parameters as fine-tuning would, it increases the salience of country information.

However, we find that the individual-level correlations decrease from 0.096 (full-persona average) to 0.030 (country-persona average), a decrease of approximately 69% ($0.066/0.096$), and MSEs increase from 1.838 to 1.961, as reported in Figure F.1 (reproduced below). Turning to the country-level alignment, correlation coefficients in country-persona increase from 0.109 to 0.138. However, since most of the 95% confidence intervals overlap, we cannot conclude that they differ statistically. The result stands when using MSEs as the alternative metric, as reported in Figure F.2 (reproduced below).

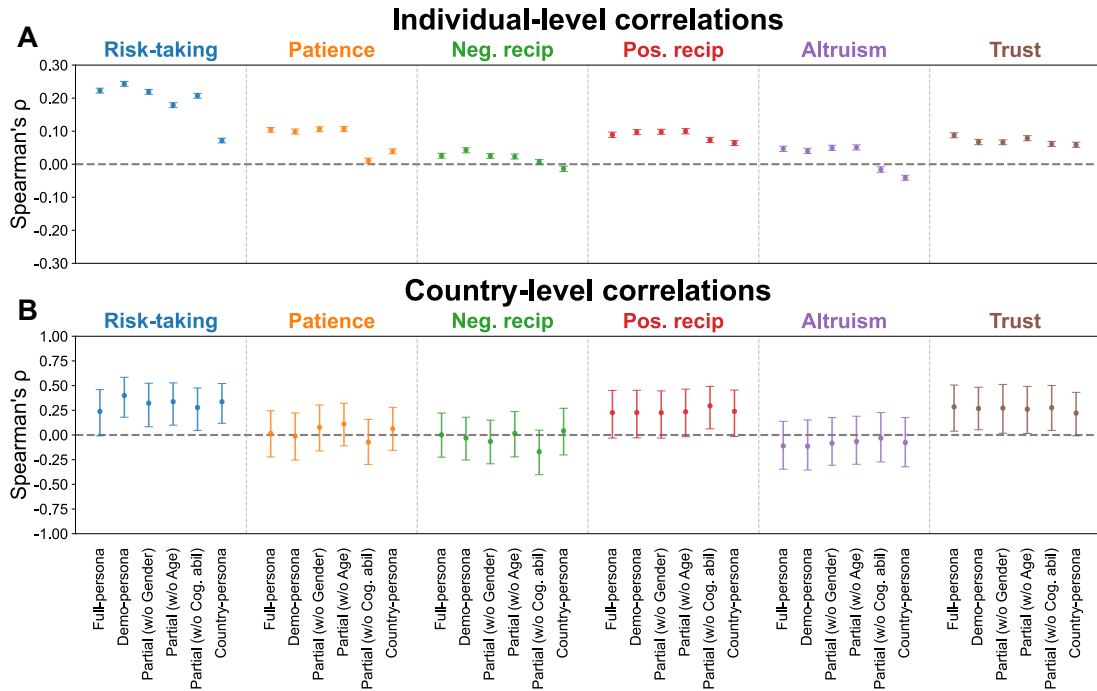


Figure F.1. Alignment between LLM-predicted and GPS-surveyed preferences. (A) Individual-level alignment. Each point displays the Spearman correlation coefficient between GPS-surveyed and LLM-predicted preferences, pooling all individual observations across 74 countries ($N > 65,000$, due to missing 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.

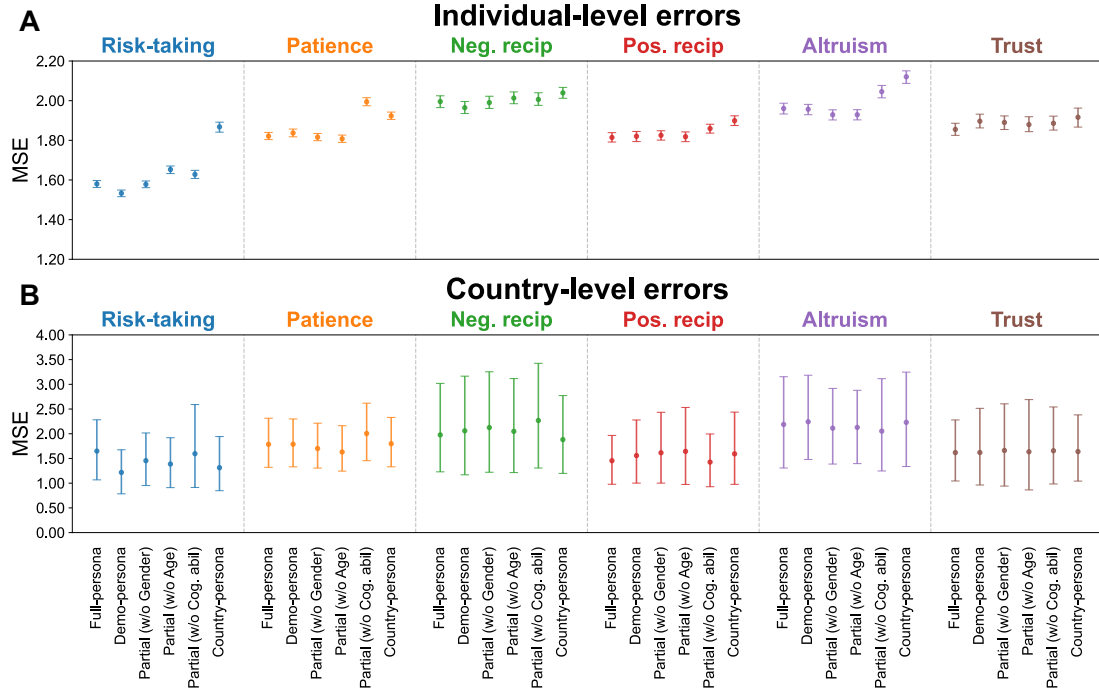


Figure F.2. Mean squared errors (MSEs) between LLM-predicted and GPS-surveyed preferences. (A) Individual-level MSEs. Each point displays the MSE between GPS-surveyed and LLM-predicted preferences, pooling all individual observations across 74 countries ($N > 65,000$, due to missing values in the original 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.

17. Temporal Dynamics: The protocol is designed to track model updates over time. Presenting initial data or a discussion on how LLM performance has evolved with different model versions could add another layer of insight.

Response: Thank you for raising this point. The revised manuscript already presents systematic comparisons across model generations and within model families. As stated in the second paragraph of the Results section:

“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 unpaired t-test, $P = 0.400$) or that Gemini-2.5 Flash outperforms Gemini-2.0 Flash (two-sided unpaired t-test, $P = 0.554$). Thus, newer models do not appear to achieve stronger individual-level alignment on economic preferences.”

Visual inspection of [Figure 2](#) (reproduced below for your convenience) also indicates no clear within-family scale-related improvement, with more-parameter variants (GPT-4o, GPT-5.4) showing no consistent advantage over their counterparts (GPT-4o-mini, GPT-5.4-mini). Together, these comparisons suggest that more recent or larger OpenAI and Gemini models do not deliver stronger alignment with human economic preferences within the scope of our audit. These cross-generation comparisons speak directly to the temporal evolution of LLM performance that you raise.

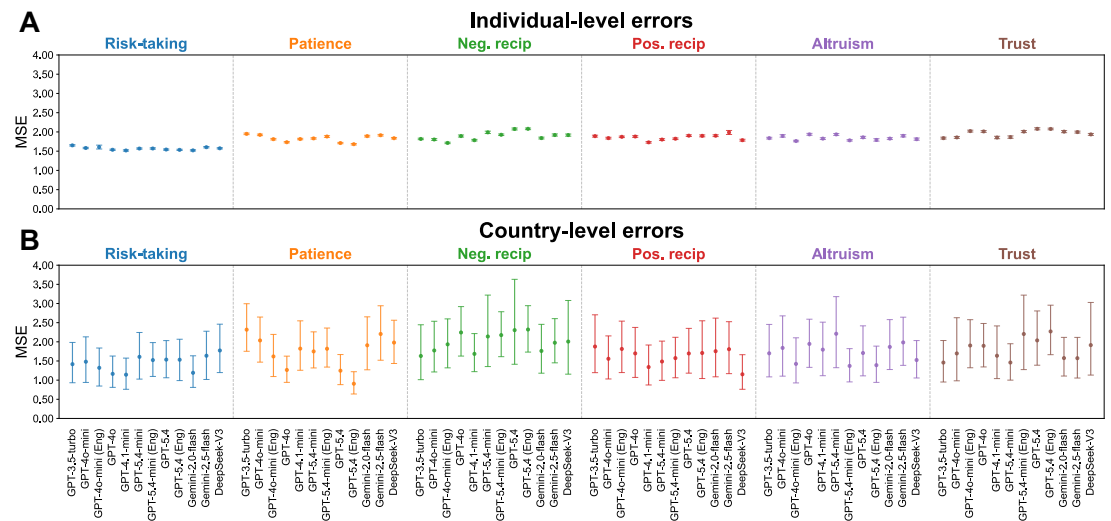


Figure 2. Mean squared errors (MSEs) between LLM-predicted and GPS-surveyed preferences. (A) Individual-level MSEs. Each point represents the MSE between LLM-predicted and GPS-surveyed preferences, pooling all individual observations across 74 countries ($N > 65,000$; the number of observations varies by preference domain due to missing values in the original GPS). **(B) Country-level MSEs.** Each point represents the MSE between country-level average LLM-

predicted and GPS-surveyed preferences ($N = 74$). 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. Error bars indicate 95% bootstrapped CIs in both panels.

18. Ethical Implications: Given the findings, a more in-depth discussion on the ethical implications of using LLMs for policy-making, especially in diverse global contexts, could be valuable.

Response: We agree this is an important dimension to develop further, and your comment prompted us to expand the ethical discussion of the manuscript. The revised Discussion now treats the ethical implications of LLM-based policy work along three lines. First, equity risks arise because LLM performance tracks economic development, raising distributional concerns when synthetic respondents are deployed globally. Second, LLM simulations have two complementary scientific uses: probing latent motivational structures in well-studied (WEIRD) populations, and flagging where cross-cultural extrapolation and standard measurement instruments are likely to fail. Third, LLM simulations expand the methodological toolkit for policy-relevant work, by enabling fast follow-up experiments, providing a safer testbed for ethically sensitive counterfactuals, and supporting LLM-based measurement.

The first line is anchored in our empirical finding that LLM alignment systematically tracks economic development. Paragraph 5 of the Discussion (existing in the original submission, retained and now connected to our cross-country results) makes this risk explicit:

“As stated in the Universal Declaration on Cultural Diversity, “Cultural diversity is as necessary for humankind..., is the common heritage of humanity.” Our results, however, expose a practical dilemma in LLM development. Although models perform relatively poorly in lower-income countries, developers have limited commercial incentives to optimize them for these contexts. The equity risks are nontrivial. Downstream,

developed countries may be able to leverage AI more effectively in research, productivity, and education, while users in developing countries receive answers that are less adapted to local contexts. Over time, if LLM performance systematically favors developed countries, cultural knowledge from developing countries may remain under-represented, encouraging homogenization and centralization of the global cultural ecosystem, eroding cross-cultural diversity (Abdurahman et al., 2024; Atari et al., 2023), and widening digital inequity (Capraro et al., 2024; Pataranutaporn et al., 2025). These risks underscore the need to align LLMs not only with universal ethical principles but also with diverse cultural, linguistic, and social contexts.”

This paragraph anchors the equity argument in our cross-country results and motivates the audit framework as a tool for surfacing such disparities before deployment.

The second line is reflected in Paragraph 7 of the revised Discussion. Paragraph 7 identifies two complementary scientific roles for LLM simulations, one is used to identify latent decision motivation in WEIRD populations, and the other one is used to design diagnostics.

“Despite these constraints, our individual- and demographic-level alignment results highlight several promising scientific uses of LLMs as synthetic respondents. First, LLMs can be used to propose and probe latent motivational structures underlying human behavior, particularly in WEIRD populations, complementing formal decision theory (Bhatia, 2024; Bhatia et al., 2025; Binz et al., 2025; Wu et al., 2025; Xie et al., 2025; Zhu et al., 2025). Second, LLMs can accelerate early-stage hypothesis generation and study-design diagnostics, helping researchers identify promising interventions or measurement strategies before turning to human samples. They can help identify when cross-cultural extrapolation may fail (Henrich et al., 2010; Lowes & Nunn, 2025; Moscona et al., 2026), and flag measurement-context mismatches (Cohn et al., 2019; Huynh et al., 2020; Yang et al., 2023).”

In both cases the simulations complement rather than substitute for human evidence in diverse global contexts, directing researchers toward where human data are most needed.

To make the logic of the examples concrete, we illustrate why LLMs could be used in design diagnostics in detail here. The measurement-context illustration draws on Cohn et al. (2019), who used email-based wallet-return rates as a measure of civic honesty across countries. In contexts where email is rarely used relative to chat (Huynh et al., 2020), this measure may systematically underestimate honesty in countries such as China (Yang et al., 2023), and LLM simulations can flag such mismatches before fielding human studies. The cross-cultural extrapolation illustration builds on the WEIRD critique. Much research in economics and psychology assumes shared fundamental cognitive processes and extrapolates from a single population. A growing literature challenges this assumption (Henrich et al., 2010; Lowes & Nunn, 2025; Moscona et al., 2026), and LLMs may help researchers diagnose where such extrapolation is most likely to fail.

The third line identifies a class of policy-relevant questions where LLM-based simulation offers an ethical advantage over human-subject research. The closing portion of Paragraph 7 develops this point along three concrete uses. It treats LLM simulations as a fast and low-cost environment for follow-up experiments and robustness checks, as a safer testbed for ethically sensitive counterfactuals such as manipulations of intergroup hostility, and as a measurement tool that expands the social-science toolkit beyond what human respondents can directly provide.

“They may also serve as a fast and low-cost environment for follow-up experiments and robustness checks (Guilbeault et al., 2025), and provide a safer testbed for exploring ethically sensitive counterfactuals, such as mechanisms that increase or reduce intergroup hostility, before designing human-subject experiments. Beyond synthetic respondents, a complementary application is to use LLMs as a measurement tool (Asirvatham et al., 2026; Fang et al., 2025), expanding social scientists’ toolkit.”

The revised Discussion (especially Paragraphs 5 and 7) frames the ethics of LLM use in policy-relevant settings as a balance. Synthetic respondents pose real distributional risks when treated as substitutes for diverse human populations. They also create useful methodological options when used as exploratory tools whose outputs are validated against, rather than replacing, human evidence. We hope the revised treatment

addressees the discussion you requested.

Reviewer 2's Comments

1. This manuscript presents a large-scale “virtual replication” of the Global Preference Survey (GPS), which measured economic preferences (risk-taking, patience, reciprocity, altruism, and trust) in nationally representative samples across 74 countries and 42 languages, alongside demographic covariates (e.g., age, gender, region/country, and subjective cognitive skills). The authors construct LLM “personas” by conditioning prompts on these demographic profiles (via system prompts) and administer the GPS items to multiple LLMs using deterministic decoding (temperature = 0). They evaluate correspondence between LLM-generated and human responses at multiple levels of aggregation.

The main empirical conclusion is that LLM responses show statistically detectable but generally weak alignment with human preferences at the individual level. Performance varies across domains (risk-taking is typically better reproduced than trust), across model families, and across sources of heterogeneity. Synthetic responses also appear more responsive to demographic cues (gender, age, cognitive ability) than to country or cultural context, suggesting that model “preference” variation may be driven disproportionately by demographic regularities rather than country-specific structure. The overall message is appropriately cautious: while LLMs can generate plausible survey responses conditioned on demographics, the variability and limited individual-level fidelity raise concerns about using such models as substitutes for human respondents, especially for cross-cultural inference.

I find the behavioral evaluation comprehensive and the dataset-scale benchmarking valuable. The paper offers a useful snapshot of how current LLMs behave under a straightforward persona-conditioning protocol.

Response: We sincerely appreciate your careful and thorough evaluation of our work, and we thank you for your constructive, detailed, and actionable feedback. Your comments have been especially instrumental in helping us illustrate the underlying mechanism, strengthen our robustness checks, and reinforce our generalization claim.

We have carefully addressed each of your comments in this revision, and we believe that the manuscript has been substantially improved as a result.

2. That said, I see two limitations that reduce the manuscript’s scientific contribution in its current form. First, the analysis is almost entirely behavioral. While this is an essential starting point, the paper does not leverage what makes LLMs unique as “synthetic respondents”: the possibility of more diagnostic, mechanism-oriented analyses. In particular, the study stops short of clarifying why some domains (e.g., trust) fail, why demographic gradients are amplified, or why country cues appear weak. For open-weight models, the authors could go beyond outcome comparisons and add diagnostic analyses (e.g., prompt sensitivity, representational similarity, or other probes tied to internal activations). Without at least one such analysis, the work reads more as an evaluation benchmark than as an explanation of the sources of alignment and misalignment.

Response: Thank you for your insightful comments. We agree that the initial version was largely behavioral, and we have substantially strengthened the diagnostic, mechanism-oriented analyses you recommended.

Summary of newly added experiments. In this revision, we are following your recommendation to design and conduct a series of robustness experiments, including prompt-component ablation experiments, a response-format constraint experiment, a paraphrasing experiment, and a high-temperature experiment using GPT-5.4 and GPT-5.4-mini, which are the state-of-the-art models available at the time of revision. We summarize the coverage and sample sizes of each condition in [Table A.1](#) (reproduced below for your convenience). These analyses move the manuscript beyond a purely behavioral evaluation toward an explanation of which prompt components and persona attributes drive alignment, addressing your questions on why country cues appear weak and why demographic gradients are amplified. We acknowledge that the relative weakness of the trust domain is harder to attribute, and we view this plausibly driven, at

least in part, by a structural feature of the GPS instrument. Trust is measured by a single self-assessment item, whereas the other preferences combine multiple items, so the trust signal is noisier ex ante. We flag a more targeted trust-domain mechanism analysis as future work.

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	86.73	74
GPT-5.4-mini (English)	69,671	86.72	74
GPT-5.4	69,681	86.73	74
GPT-5.4 (English)	69,681	86.73	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.

While [Table A.1](#) reports the sample sizes and country coverage of each condition, [Table R.1](#) summarizes the design rationale of the newly added experiments and the specific concerns each is intended to address.

Table R.1. Summary of newly added experiments.

New experiment	Model	Experiment objective
Replication experiment (in local languages and in English)	GPT-5.4 and GPT-5.4-mini	1) To address concerns about model- and time-dependence; 2) To address the concern on memorization.
Prompt ablation experiment (in local languages)	GPT-5.4-mini	To provide diagnostic evidence on the source of alignment.
Response-format constraints experiment (in local languages)	GPT-5.4-mini	To strengthen claims about generalization.
Paraphrasing experiment (in English)	GPT-5.4-mini	To address the concern about memorization.
High-temperature experiment (in local languages)	GPT-5.4-mini	To address the concern about diversity.

Prompt-components ablation experiments. To identify which prompt components drive alignment, following your recommendation, we design four types of ablation conditions using GPT-5.4-mini (the same model adopted in the main text): full-persona (all information), partial-persona (full-persona with one demographic attribute removed; three variants: gender, age, or cognitive ability), demographic-persona (only demographic information, no country), and country-persona (only country, no demographics). This yields six conditions in total. All ablation experiments are conducted in local languages, consistent with our original experiment. Detailed system prompt templates appear in [Table F.1](#) (reproduced below).

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	You are <i>[age]</i> -year-old <i>[gender]</i> , living in <i>[region]</i> , <i>[country]</i> .

(without cognitive ability)	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.

The paraphrasing experiment and English-only experiments. Following your advice, we address the data-contamination (i.e., memorization vs inference) issue directly. We address this concern through two complementary designs. First, we paraphrase the survey items, as shown in [Table G.1](#) (reproduced below), and run a paraphrasing experiment in English using GPT-5.4-mini. Second, going beyond your recommendation, we also run the GPS in English on three models (GPT-4o-mini, GPT-5.4-mini, and GPT-5.4) for which both English and local-language results are reported in our main analysis (Table A.1, Panel A). The logic is that if memorization were the dominant driver of alignment, performance should differ markedly between English (where the training corpus is densest) and the corresponding local-language conditions. Observing comparable alignment across the two settings therefore supports inference over memorization.

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	Please indicate your general willingness to take risks on 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.	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.
	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.
Patience	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.
	Willingness to act	How willing are you to punish someone who treats [you] unfairly, even if there may be costs for you? Please indicate your answer on a scale from 0 to 10. A 0 means “completely unwilling to do so”, and a 10 means “very willing to do so”. You can also use any number between 0 and 10 to indicate where you fall on the scale, using 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10.	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 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 [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	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.

		on the scale, using 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10.	
	Self-assessment	How well does each of the following statement describe you? "If I am treated very unjustly, I will take revenge at the first occasion, even if there is a cost to do so". Please indicate your answer on a scale from 0 to 10. A 0 means "does not describe me at all", and a 10 means "describes me perfectly". You can use any number between 0 and 10 to indicate where you fall on the scale, using 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10.	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.
	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.
Positive reciprocity	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	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 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

		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].	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 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 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.

Response-constraint experiment. We fully agree with your suggestion that a stricter response-format constraint could strengthen the generalization. Accordingly, we

conducted 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](#) (reproduced below).

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.

A high-temperature experiment. We fully agree with your suggestion that varying the decoding temperature could strengthen the generalization claim. Although deterministic decoding ($T = 0$) is a standard choice for reproducibility in scientific research, varying the temperature provides an additional generalization test. We therefore conducted a high temperature experiment with GPT-5.4-mini in localized languages at $T = 1$.

3. Second, the behavioral results are likely to be model- and time-dependent. Given rapid model updates and deployment changes, especially for closed-source systems, some findings may only hold for specific model snapshots and collection windows. API-mediated behavior can also vary across providers and deployments (e.g., hidden updates or serving optimizations), complicating cross-model comparisons and the durability of reported effect sizes. The paper would benefit from clearer framing and/or robustness checks separating conclusions likely to generalize (e.g., “demographic gradients > country effects”) from those that are likely contingent on the specific models and time periods studied.

Response: Thank you for highlighting this point. We agree that a clearer framing of which conclusions are likely to generalize, supported by robustness checks, would strengthen our contribution.

Among our findings, we identify two conclusions that are more likely to generalize. The first is the micro-to-macro gap (i.e., the misleading country-level synthetic data and the weak recoverability of cultural-level gradients). The second is the systematic relationship between prediction error and GDP per capita and the IDI, which indicates that LLMs exhibit larger prediction errors in lower-income countries, echoing the uneven geographic adoption of AI.

We further note that our response to your Comment 2 above describes a battery of robustness experiments newly added to this revision, including GPT-5.4 and GPT-5.4-mini replication, prompt-component ablation, paraphrasing, and response-constraint conditions (see Panel B of [Table A.1](#) and [Table R.1](#)), that collectively address the model- and time-dependence concerns raised here. The replication on the most recent GPT-5.4 series, in particular, shows that the two generalizable conclusions above continue to hold for the latest available models. These two conclusions are stated explicitly in Paragraphs 3–4 of the revised Discussion.

4. Overall, I view this as a valuable benchmark documenting current limits of demographic-conditioned LLM personas for simulating economic preferences at global scale. However, for an interdisciplinary venue such as *Nature Human Behaviour*, the present version may be limited in conceptual depth. A stronger contribution would either (i) add at least one mechanistic/diagnostic analysis to explain the observed failures and successes, or (ii) strengthen claims about generalization of the behavioral results via synchronized sampling, design variations (e.g., stricter response-format constraints), or targeted robustness tests.

Response: Thank you for this overarching comment and for outlining the two

complementary paths along which the manuscript could be strengthened. The revised version addresses both.

(1) Mechanistic/diagnostic analysis. We add a range of prompt-component ablation experiments (full-persona, demographic-persona, country-persona, and three partial-persona conditions removing one demographic field at a time) using GPT-5.4-mini. The ablation isolates which prompt elements drive the observed alignment and the demographic gradients, providing diagnostic evidence beyond correlational attribution. Detailed results, figures, and tables are reported in our response to your Comment 5 below and in [Appendix F](#).

(2) Generalization and robustness. We add four robustness experiments: (a) a paraphrasing experiment that rewords every survey item in English to test memorization versus inference; (b) a parallel English-language experiment across GPT-4o-mini, GPT-5.4-mini, and GPT-5.4 to test memorization versus inference by comparing against the local-language main analysis; and (c) a response-format constraint experiment in which the model is instructed to output only a number, addressing the stricter response-format constraint you suggest. (d) A high-temperature experiment ($T = 1$) testing whether deterministic decoding in the main analyses understates response variability. All four experiments leave the headline alignment patterns essentially unchanged. Detailed results are reported in our response to your Comment 6 below and in [Appendix G](#).

These additions target both the mechanistic depth and the generalization concerns you raised, and the revised manuscript is stronger as a result.

5. Detailed suggestions:

Add diagnostic analyses beyond regressions. At present, even the “sources of misalignment” section relies mainly on regressions over LLM responses. If neural-level analysis is feasible for open-weight models (e.g., DeepSeek), the authors could probe internal representations to test how persona attributes and country labels are encoded

and which components causally drive responses. If this is not feasible, a simpler but still informative alternative is prompt-component ablations (system persona vs country label vs individual demographic fields), combined with sensitivity analyses to quantify which prompt elements drive preference shifts beyond correlational attribution.

Response: We agree that adding diagnostic analyses beyond regressions could further strengthen our theoretical and practical contributions, and we thank you for the detailed and constructive comments. As we noted in our response to Comment 4 above, this addition addresses (1) mechanistic/diagnostic analysis.

While we recognize that neural-level analysis would provide the most comprehensive evidence, conducting such analysis is beyond our current computational resources, given the large parameter scale (e.g., DeepSeek-V3 with 671B parameters) and the required volume of API calls (approximately 1.4 million). Following your alternative recommendation, we therefore implement prompt-component ablation experiments.

(1) Ablation experiment design. The ablation experiments comprise six conditions organized into four types: (i) full-persona (the original prompt with country and all three demographic fields); (ii) three partial-persona conditions, each removing exactly one demographic field (gender, age, or cognitive ability) while keeping the country and the other two fields; (iii) a demographic-persona condition that retains all demographic fields but removes the country label; and (iv) a country-persona condition that retains only the country label and removes all demographic fields.

(2) Identify the alignment driver. We first examine which prompt component drives the alignment. If demographic information is the primary driver, alignment should decrease in the country-persona condition. [Figure F.1](#) (reproduced below for your convenience) reports the individual- and country-level Spearman’s correlations. Individual-level Spearman’s correlations decrease from 0.096 (full-persona) to 0.030 (country-persona), a drop of approximately 69% ($0.066/0.096$). In contrast, removing the country label has essentially no effect: the correlation moves slightly upward from 0.096 (full-persona) to 0.098 (demographic-persona). This pattern is in line with the regression evidence in [Appendix E.1](#),

where demographic variables absorb approximately 77% of the explanatory power of alignment. Using MSE as the alternative metric (Figure F.2, reproduced below), individual-level MSEs increase from 1.838 (full-persona) to 1.961 (country-persona), about 7% ($0.123/1.838$). For detailed changes of correlations and MSEs for each preference, see Table F.2 (reproduced below).

Although individual-level correlations decline in some partial-persona conditions, the overall evidence is mixed. For the cognitive ability dimension, individual-level correlations decrease and MSEs increase relative to full-persona, indicating that cognitive ability plays a substantive role in shaping LLM-predicted preferences. For the gender and age dimensions, however, the directions of changes in correlations and MSEs are not consistent across preference domains.

At the country level, correlations in the country-persona condition rise from 0.109 (full-persona) to 0.138. However, most of the 95% confidence intervals overlap, so we cannot conclude that the conditions differ statistically. The same conclusion holds when using MSEs as the alternative metric ($M_{\text{full-persona}} = 1.779$ vs $M_{\text{demographic-persona}} = 1.747$ vs $M_{\text{country-persona}} = 1.743$).

One possible explanation is that, even in the country-persona condition, LLMs remain insensitive to the country information, compared to the demographic information. Applying the same error-decomposition technique used in the main text (Figure 6C), we decompose MSEs separately for each ablation condition. As shown in Figure F.3, country fixed effects continue to explain most of the prediction errors across the six conditions (see Figures F.4–F.9 for per-preference error decomposition, reproduced below).

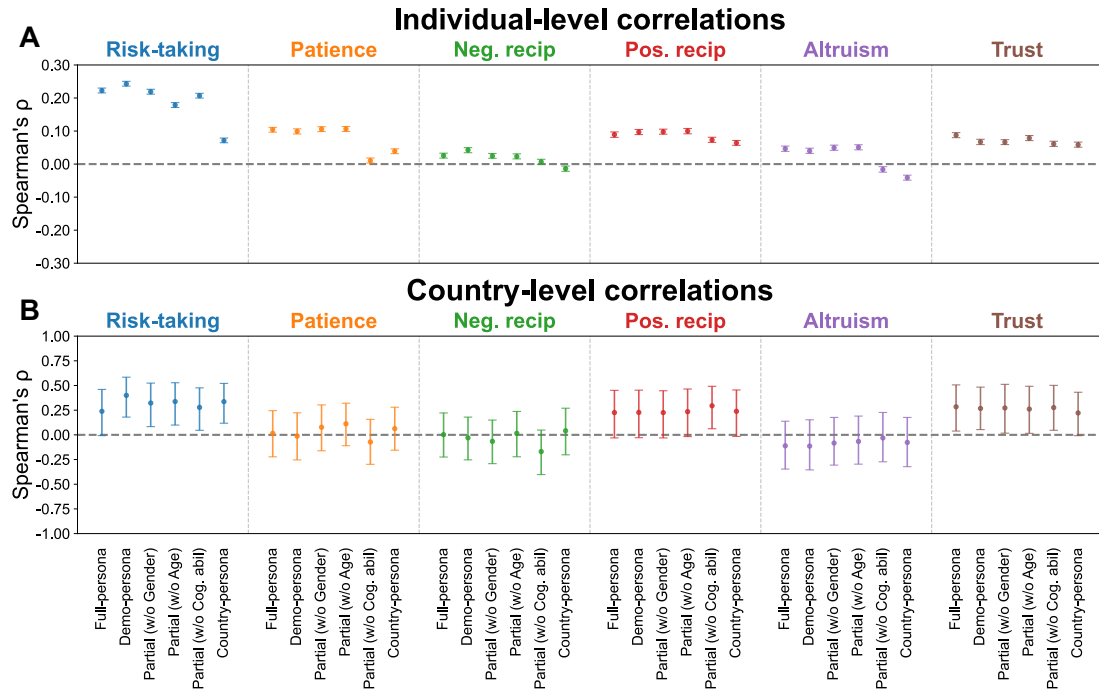


Figure F.1. Alignment between LLM-predicted and GPS-surveyed preferences. (A) Individual-level alignment. Each point displays the Spearman correlation coefficient between GPS-surveyed and LLM-predicted preferences, pooling all individual observations across 74 countries ($N > 65,000$, due to missing 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.

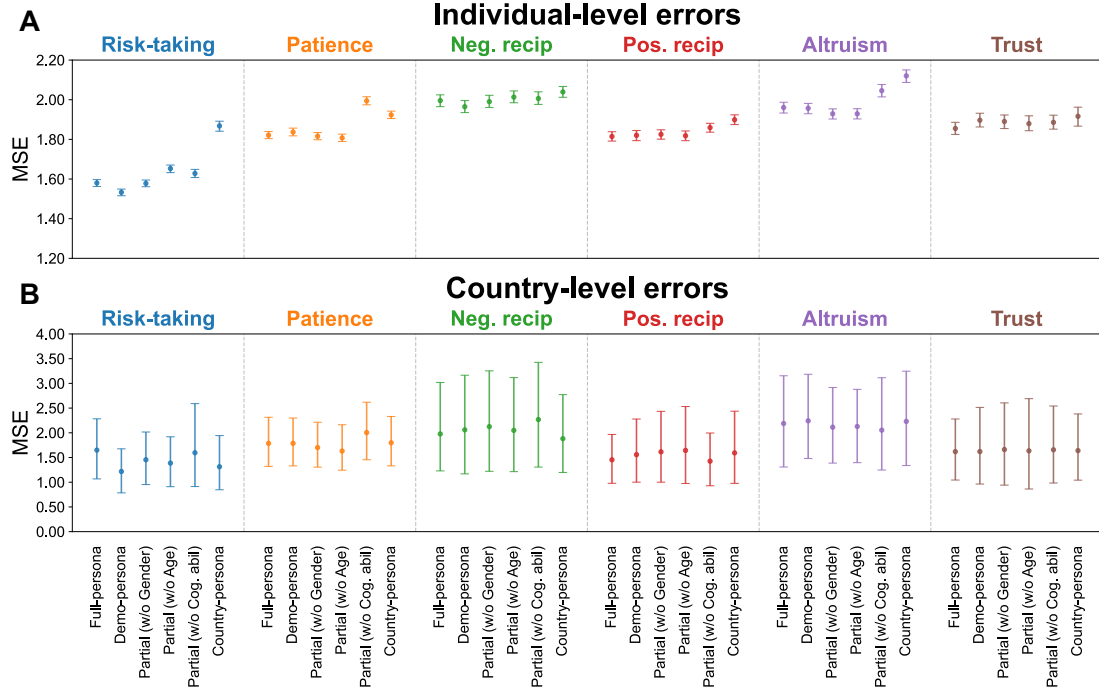


Figure F.2. Mean squared errors (MSEs) between LLM-predicted and GPS-surveyed preferences. (A) **Individual-level MSEs.** Each point displays the MSE between GPS-surveyed and LLM-predicted preferences, pooling all individual observations across 74 countries ($N > 65,000$, due to missing values in the 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.2. 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

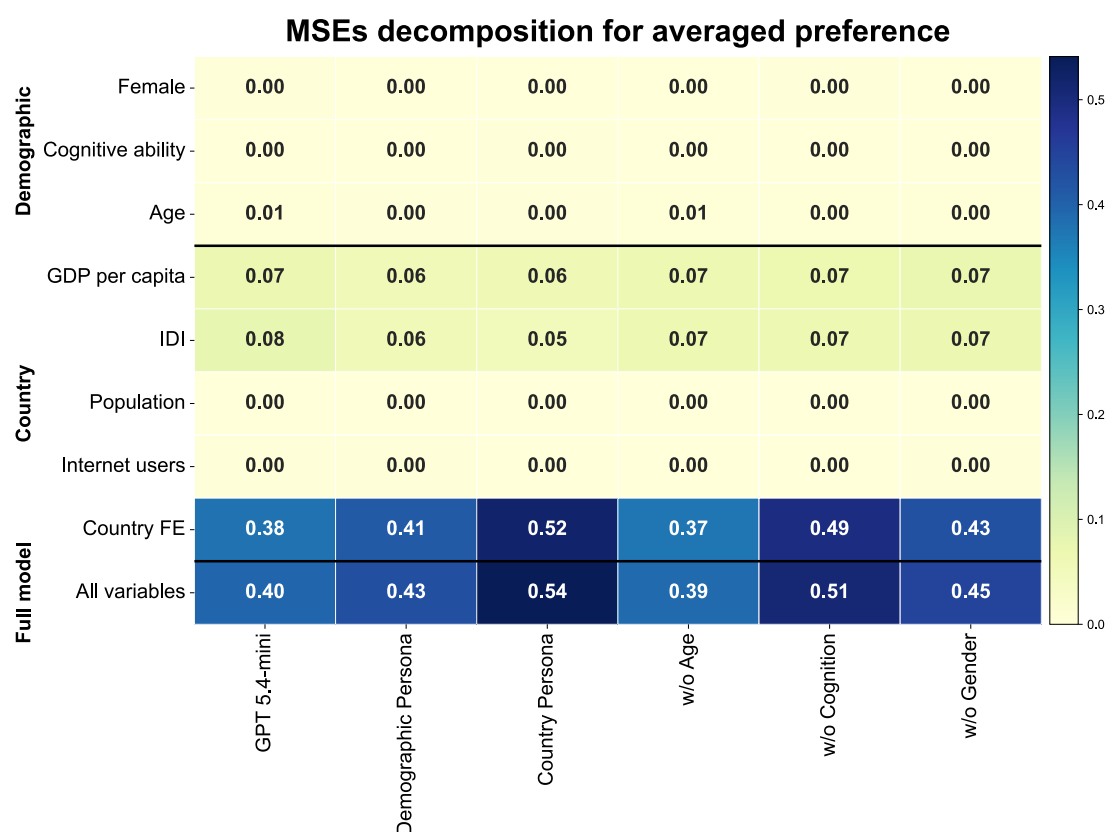


Figure F.3. Error decomposition pooled across LLM-predicted preferences. The number in each cell represents the R^2 estimated from a regression where the individual-level MSE pooled across six preferences is the dependent variable and corresponding variables indicated by the row label are the independent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

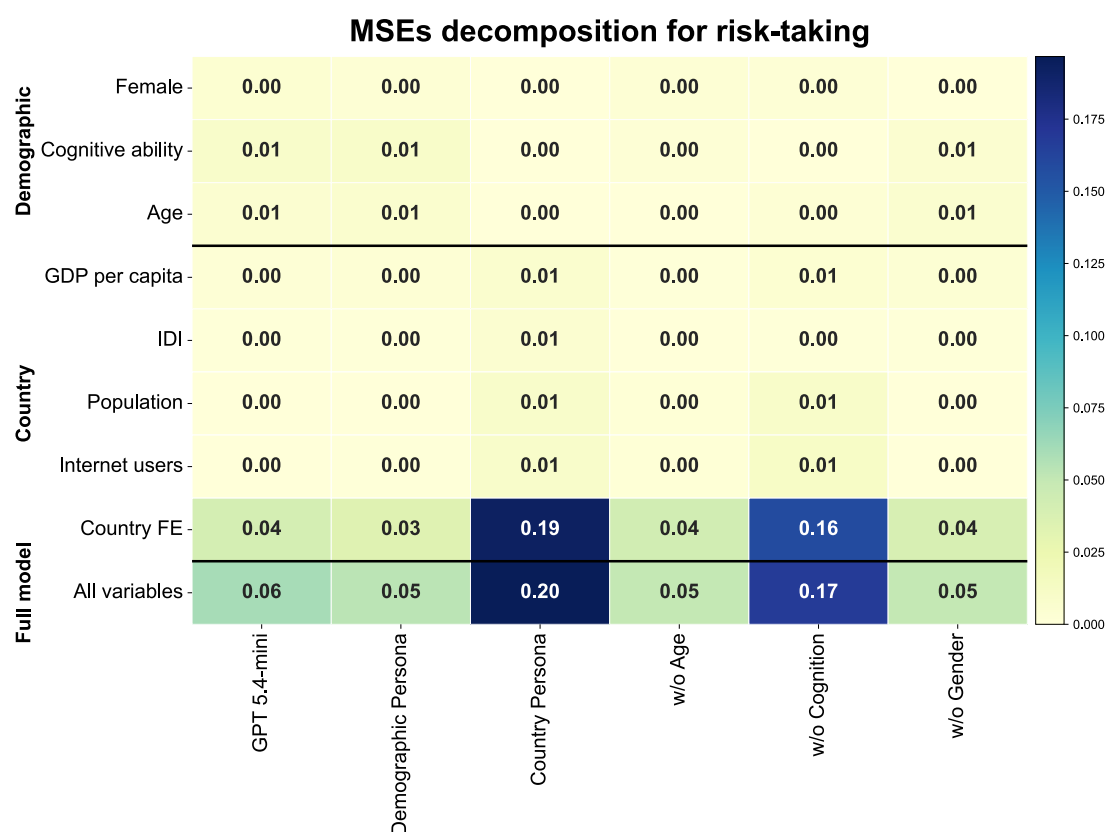


Figure F.4. Error decomposition of LLM-predicted preferences (risk-taking). The number in each cell represents the R^2 estimated from a regression where the individual-level MSE for risk-taking is the dependent variable and corresponding variables indicated by the row label are the independent variables.

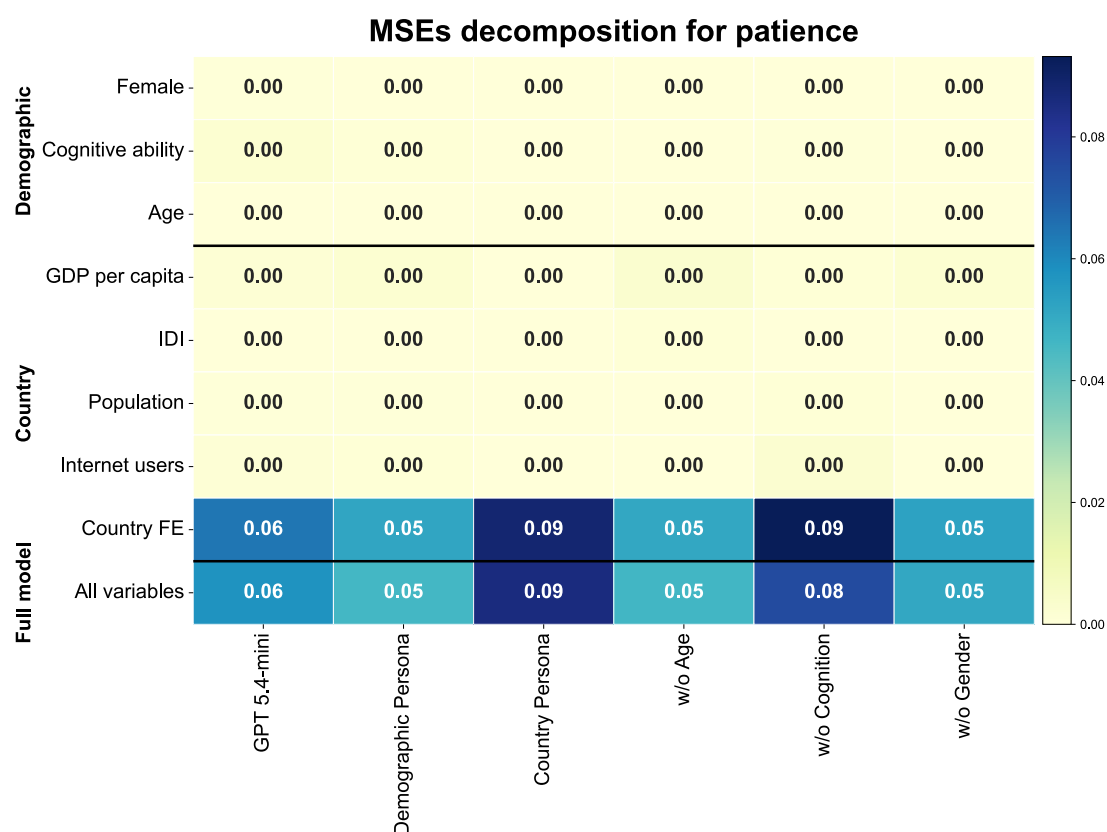


Figure F.5. Error decomposition of LLM-predicted preferences (patience). The number in each cell represents the R^2 estimated from a regression where the individual-level MSE for patience is the dependent variable and corresponding variables indicated by the row label are the independent variables.

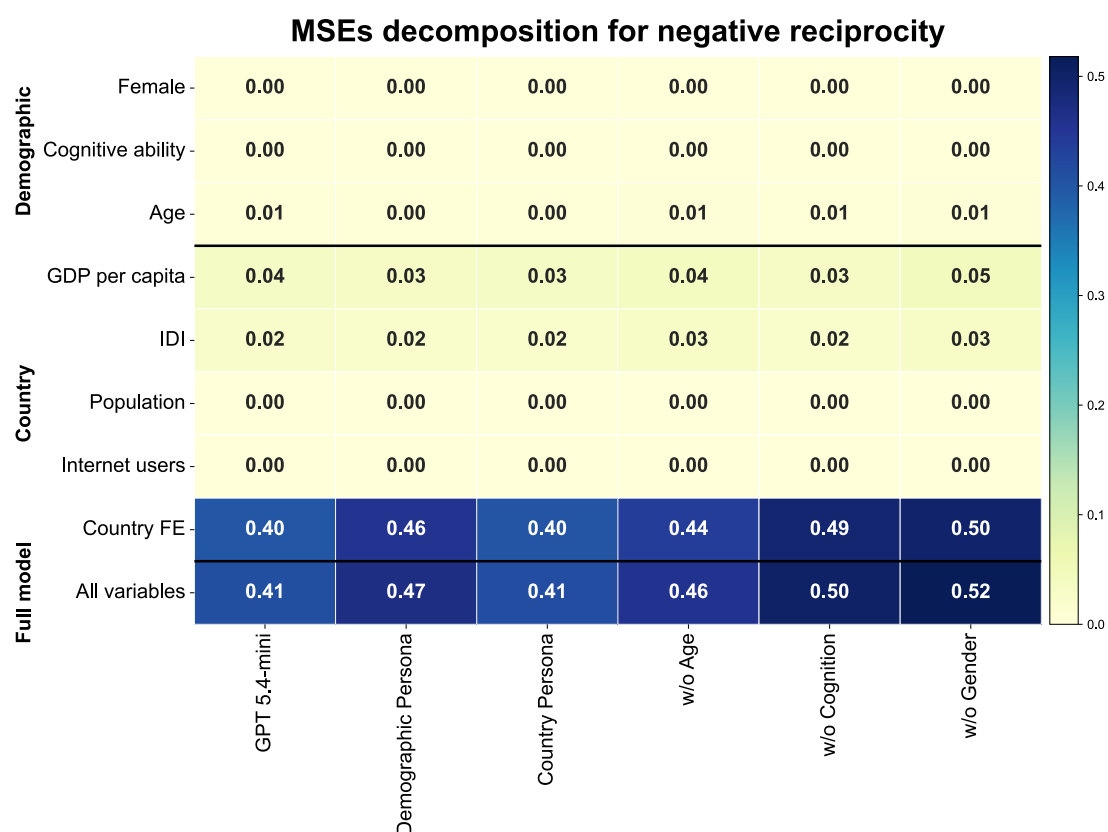


Figure F.6. Error decomposition of LLM-predicted preferences (negative reciprocity). The number in each cell represents the R^2 estimated from a regression where the individual-level MSE for negative reciprocity is the dependent variable and corresponding variables indicated by the row label are the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

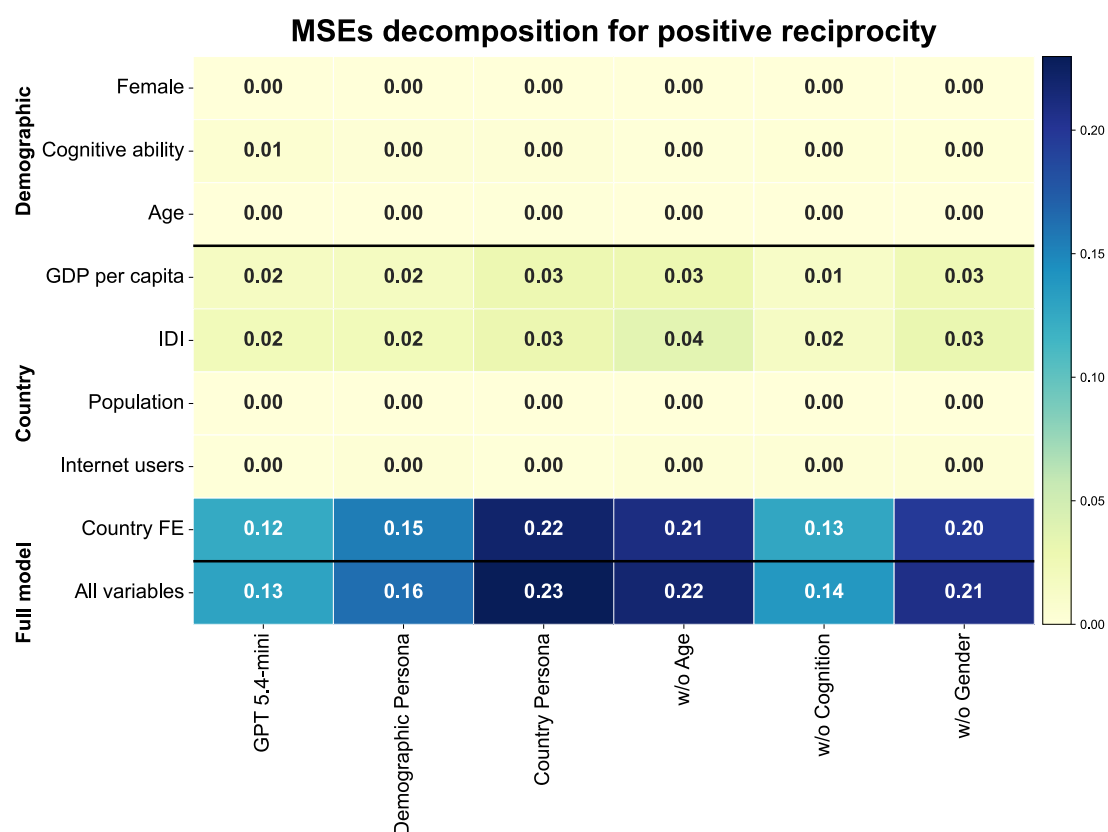


Figure F.7. Error decomposition of LLM-predicted preferences (positive reciprocity). The number in each cell represents the R^2 estimated from a regression where the individual-level MSE for positive reciprocity is the dependent variable and corresponding variables indicated by the row label are the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

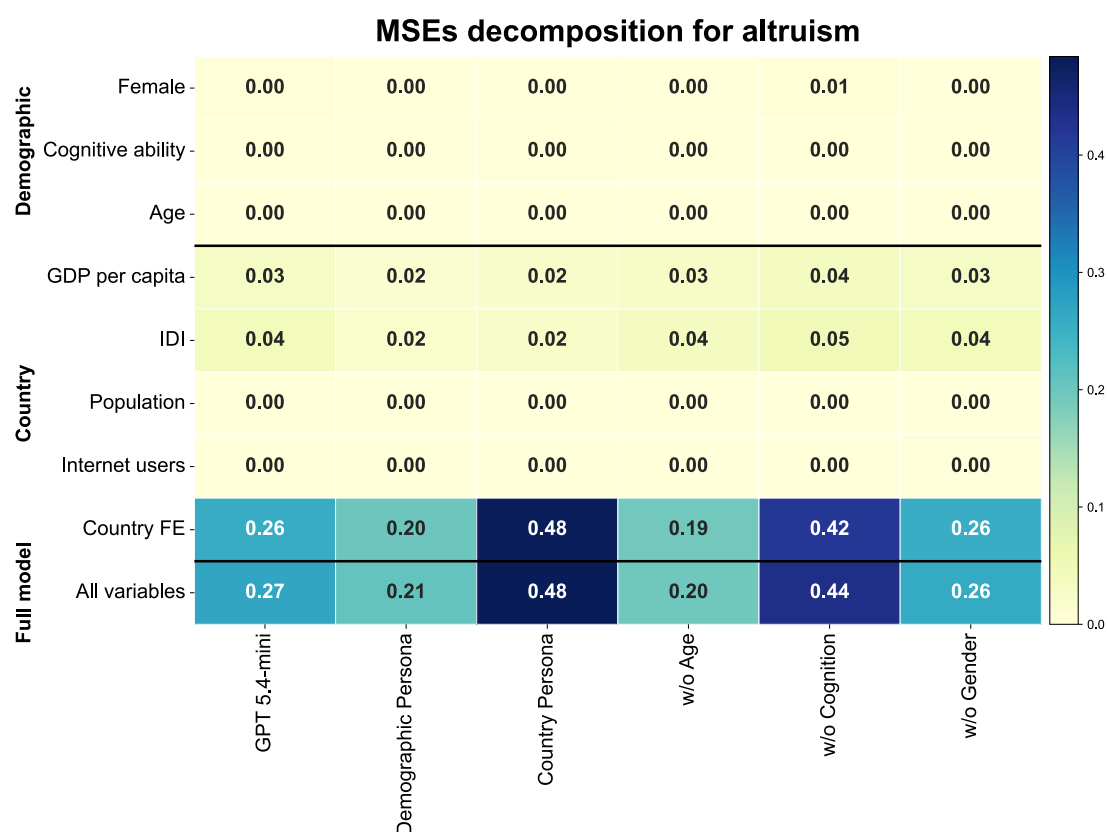


Figure F.8. Error decomposition of LLM-predicted preferences (altruism). The number in each cell represents the R^2 estimated from a regression where the individual-level MSE for altruism is the dependent variable and corresponding variables indicated by the row label are the independent variables.

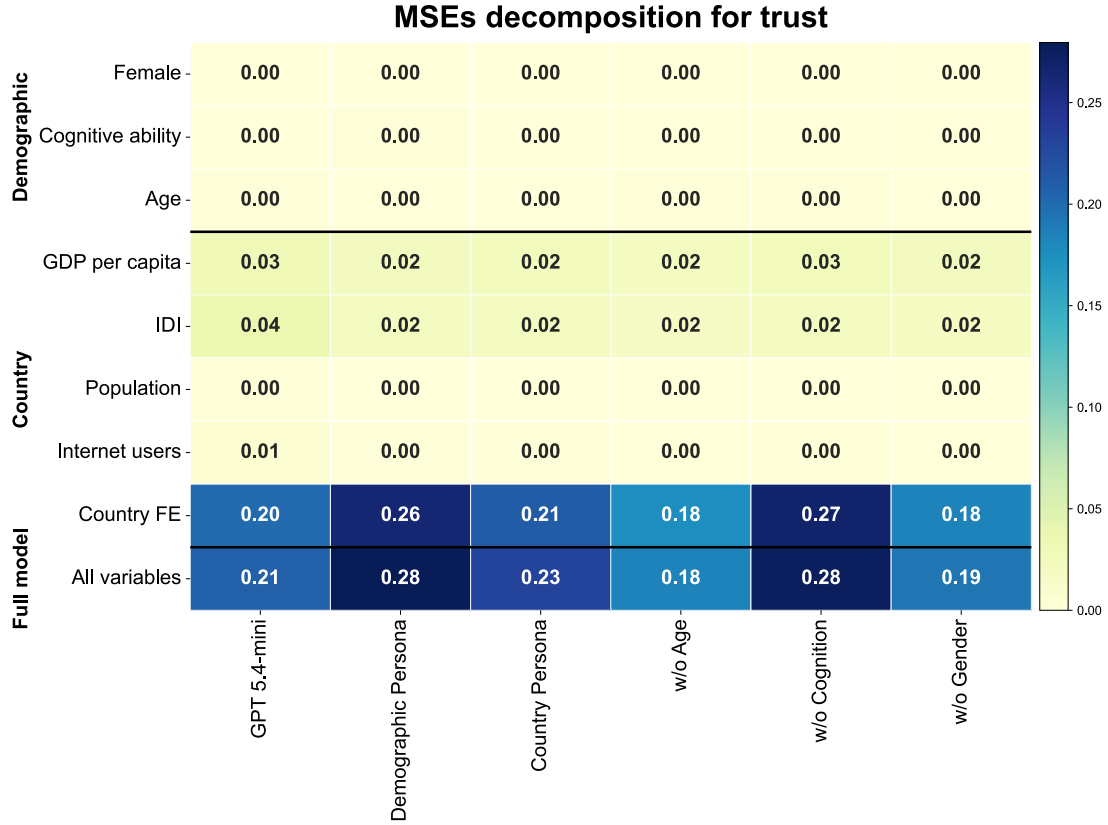


Figure F.9. Error decomposition of LLM-predicted preferences (trust). The number in each cell represents the R^2 estimated from a regression where the individual-level MSE for trust is the dependent variable and corresponding variables indicated by the row label are the independent variables.

(3) Attenuated demographic gradient. Next, we check whether the demographic gradients themselves are attenuated when demographic information is partially or fully removed from the prompt. We hypothesize, based on the alignment and error-decomposition results above, that the demographic information in system prompts (rather than country information or random noise) generates the demographic gradient. If so, the effect sizes of partial-persona should be attenuated, and those of country-persona should be attenuated further. Hence, we expect the following pattern: effect sizes (full-persona) $\approx$ effect sizes (demographic-persona) $>$ effect sizes (partial-persona) $\approx$ effect sizes (country-persona).

As shown in [Figure 7B](#) (reproduced below), the effect sizes estimated from the demographic-persona condition are essentially identical to those from full-persona (M_{full}).

persona = 0.303 vs $M_{\text{demographic-persona}} = 0.303$). The effect sizes from partial-persona and country-persona are similarly attenuated ($M_{\text{partial-persona}} = 0.046$ vs $M_{\text{country-persona}} = 0.016$). We then fit a linear mixed-effects model pooling the full-persona, demographic-persona, country-persona, and partial-persona conditions. As shown in Table F.3 (reproduced below), the linear mixed-effects model shows that, compared to the country-persona condition, the partial-persona coefficient is not significantly larger than the country-persona reference, while both the full-persona and demographic-persona coefficients are significantly larger than the country-persona reference ($P < 0.001$). Detailed effect sizes estimated from each condition are reported in Table F.4 (reproduced below).

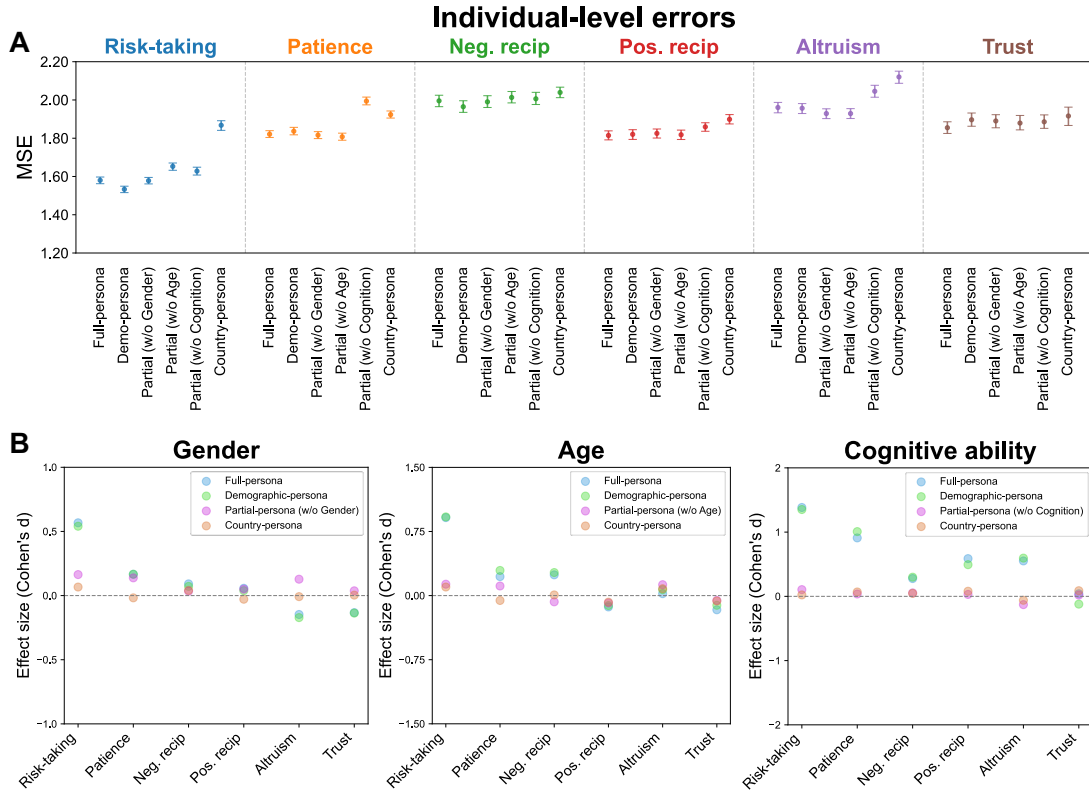


Figure 7. Ablation experiments isolating the contributions of country and demographic information in persona prompts. (A) Individual-level errors (MSEs) across ablation conditions. Each point represents the MSE between GPS-surveyed and LLM-predicted preferences, pooling all individual observations across 74 countries ($N > 65,000$) for a given model $\times$ preference combination, shown for six ablation conditions: full-persona (system prompt contains gender, age, cognitive ability, and country); demographic-persona (only demographic characteristics, country removed); three partial-persona conditions (full-persona with gender, age, or cognitive ability removed, respectively); and country-persona (only country, all demographic characteristics

removed). Error bars denote 95% bootstrapped CIs. **(B) Demographic effect sizes across ablation conditions.** Each subpanel corresponds to one demographic dimension (gender, age, cognitive ability). Each point plots a Cohen's d effect size of a demographic variable (female vs male, older vs younger, lower vs higher cognitive ability) estimated from LLM-predicted preferences under each ablation condition. For each subpanel, four conditions are shown: full-persona, demographic-persona, country-persona, and the matched partial-persona (e.g., 'w/o Gender' in the Gender subpanel; see legend in each panel). y-axis ranges differ across subpanels to accommodate effect-size magnitudes.

Table F.3. Linear mixed-effects model of effect sizes (Cohen's d)

DV: Cohen's d	
Condition	
Full-persona	0.287*** (0.079)
Demographic-persona	0.287*** (0.079)
Partial-persona	0.030 (0.079)
Preference	
Risk-taking	0.434*** (0.097)
Patience	0.166* (0.097)
Negative reciprocity	0.025 (0.097)
Positive reciprocity	-0.014 (0.097)
Trust	-0.138 (0.097)
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. Standard errors are in parentheses. * $P < 0.10$, ** $P < 0.05$, *** $P < 0.01$.

Table F.4. 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.

Taken together, the ablation experiments provide convergent mechanistic evidence that the demographic information in personas, rather than the country label, is the dominant driver of both the individual-level alignment we report and the demographic gradients. Removing the country label leaves alignment essentially unchanged. In contrast, removing demographic information collapses both alignment and demographic gradients. Due to LLMs' limited sensitivity to country information, it is difficult to achieve macro-level alignment and clear cultural gradients across countries. This finding directly supports the micro-to-macro gap interpretation we develop in the Discussion (see also our response to your Comment 3 above).

6. Engage the data-contamination issue more directly. Because GPS dates to 2012, it is plausible that some items, weights, or summaries entered LLM training corpora. Data contamination would typically bias toward higher apparent alignment, yet the paper reports limited alignment. This tension is worth addressing with at least one additional analysis. For example, the authors could test memorization vs inference by paraphrasing or perturbing item wording and checking whether inferred preferences are stable or brittle. Relatedly, it would be valuable to discuss (even conceptually, if new experiments are infeasible) what interventions might improve alignment (e.g., targeted finetuning or calibration) since practitioners often care less about whether off-the-shelf models align and more about whether alignment can be improved in a controlled way.

Response: Thank you for pointing out the data-contamination issue and, equally importantly, for the constructive question of whether and how alignment could be improved. We address these two parts in turn below, and additionally report a response-constraint experiment following your recommendation to engage in the data-contamination issue directly.

(1) Robustness to potential memorization. To probe whether the alignment we report could be an artifact of memorization rather than inference, we conducted two complementary experiments. First, we paraphrased every survey item ([Appendix Table G.1](#), reproduced below for your convenience) and ran a paraphrasing experiment in English using GPT-5.4-mini. Second, we ran three controlled experiments that differ from our main experiments only in being conducted in English rather than the local languages used in the GPS programme, using GPT-4o-mini, GPT-5.4-mini (newly added), and GPT-5.4 (newly added). If the alignment we report were primarily driven by memorization, we would expect correlations to drop sharply and MSEs to rise sharply under either variant.

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	Which option would you prefer: Option (1): A draw	Please choose between the following options: Option

	quantitative questions	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?	(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.
	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.
Patience	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 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 [you] 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.
	Willingness to act	How willing are you to	Please indicate your

		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.	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.
	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.
Positive Reciprocity	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	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 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.

		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].	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 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	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 tend to assume that people have good intentions.” You may select any integer between 0 and 10.

indicate where you fall on the scale, using 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, or 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.

(a) Robustness to the different language setting. As shown in [Figures G.1 and G.2](#) (reproduced below), we observe no statistically significant differences between models in local languages and in English on either set of metrics (two-sided paired t-test by pooling GPT-4o-mini, GPT-5.4-mini, and GPT-5.4; individual-level correlations: $P = 0.903$; country-level correlations: $P = 0.307$; individual-level MSEs: $P = 0.488$; country-level MSEs: $P = 0.570$). The detailed correlation coefficients and MSEs are reported in [Table G.2](#) (reproduced below).

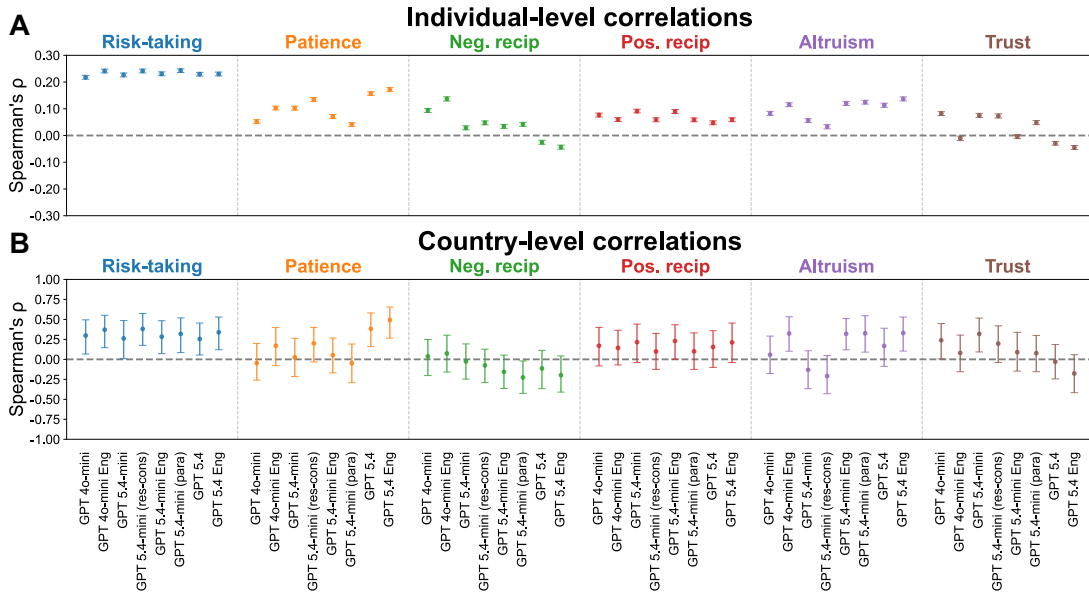


Figure G.1. Spearman correlations between LLM-predicted and GPS-surveyed preferences. **(A) Individual-level correlations.** Each point displays the Spearman's correlation coefficient between GPS-surveyed and LLM-predicted preferences, pooling all individual observations across 74 countries ($N > 65,000$, 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.

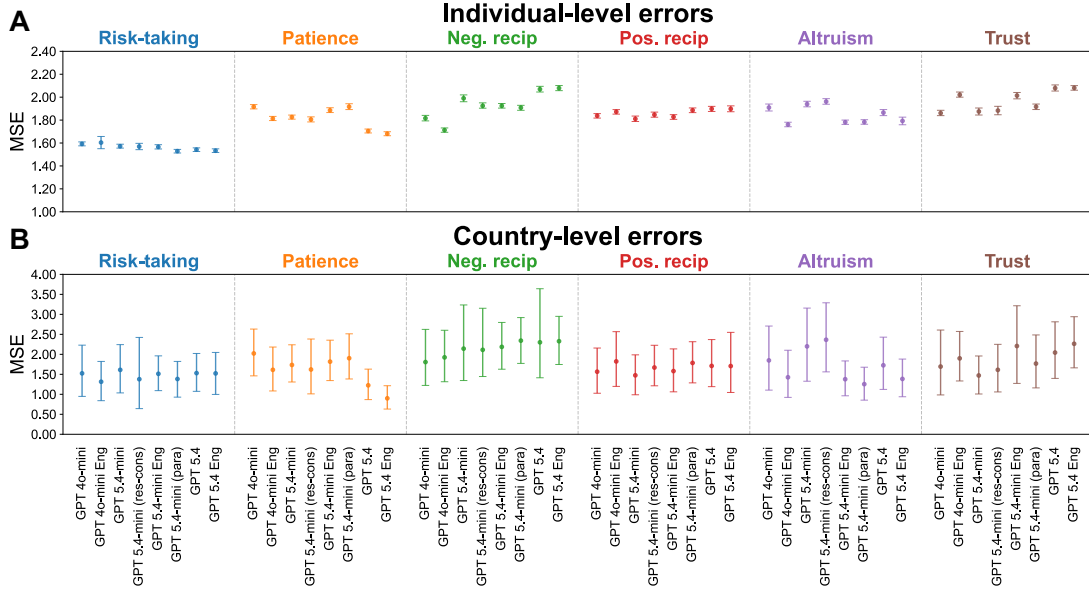


Figure G.2. Mean squared errors (MSEs) between LLM-predicted and GPS-surveyed preferences. (A) Individual-level MSEs. Each point displays the MSE between GPS-surveyed and LLM-predicted preferences, pooling all individual observations across 74 countries ($N > 65,000$, due to missing values in the original 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.

(b) Robustness to the paraphrased items. Compared to GPT-5.4-mini (English), individual-level correlations under paraphrasing are slightly higher for risk-taking, negative reciprocity, altruism, and trust, and slightly lower for patience and positive reciprocity. The MSE evidence is similarly mixed across preferences. Pooling the individual- and country-level evidence from both metrics, we find no systematic effect of paraphrasing on alignment (two-sided paired t-test between GPT-5.4-mini and GPT-5.4-mini with paraphrasing; individual-level correlations: $P = 0.841$; country-level correlations: $P = 0.148$; individual-level MSEs: $P = 0.669$; country-level MSEs: $P = 0.694$). Detailed coefficients and MSEs are reported in [Appendix Table G.2](#).

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.

(2) Robustness to prompt-format sensitivity. A separate but related concern is whether our results are sensitive to the specific phrasing of our system prompt. To probe this possibility, we conducted an additional response-constraint experiment using GPT-5.4-mini, in which the system prompt was modified to require only the option number as output, eliminating the free-text generation that the original prompt allowed (full prompt template in [Appendix Table G.3](#), reproduced below). If prompt format were the primary driver of alignment, we would expect substantial changes in correlations and MSEs under this constrained format. We do not observe any such systematic change in either Spearman's correlations or MSEs (two-sided paired t-test; individual-level correlations: $P = 0.892$; country-level correlations: $P = 0.828$; individual-level MSEs: $P = 0.812$; country-level MSEs: $P = 0.781$). The detailed correlation coefficients and MSEs are reported in [Table G.4](#). We interpret this as evidence that prompt-format sensitivity is unlikely to be the primary driver of the alignment patterns we report.

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-	You are <i>[age]</i> -year-old <i>[gender]</i> , living in <i>[region]</i> , <i>[country]</i> . You are

constraint experiment	[subjective cognitive skills]. 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.

(3) Stability of inferred preferences. Beyond these aggregate tests, we also examined the LLMs' output item-by-item ([Appendix Table G.5](#), reproduced below). Following the GPS programme, twelve raw responses are collected per respondent. Although the survey contains 20 items, five risk-taking items and five patience items follow a staircase procedure. This procedure resembles a binary search algorithm, in which only the final outcome is used as the variable of interest.

Because the paraphrasing experiment is conducted in English, we use GPT-5.4-mini (English) as its benchmark; because the response-constraint experiment is conducted in the local languages used in our main experiments, we use GPT-5.4-mini (local language) as its benchmark. The average Spearman correlation across the twelve items is 0.403 for the paraphrasing experiment and 0.434 for the response-constraint experiment. All

twelve item-level correlations are significantly larger than zero ($P < 0.001$ in every cell). Both averages exceed the overall between-model correlations (i.e., individual-level: 0.346) reported in [Appendix C](#) and the GPS-LLM correlations (i.e., individual-level: 0.094) reported in [Figure 2](#), indicating that the inferred preferences are stable across paraphrasing and response-constraint variants.

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.

(4) Robustness to temperature sensitivity. In our main experiments, temperature is set to zero for reproducibility, and we find that LLM responses are highly concentrated and less diverse than human response. A natural concern then arises: whether the finding is sensitive to the temperature setting.

We conducted a high-temperature experiment ($T = 1$) using GPT-5.4-mini in localized languages as a robustness check. 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: $P = 0.946$; country-level correlations: $P = 0.910$; individual-level MSEs: $P = 0.995$; country-level MSEs: $P = 0.880$). Detailed correlation coefficients and MSEs are reported in [Table J.3](#) (reproduced below).

Turning to diversity, [Figure J.3](#) (reproduced below) 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, $P < 0.001$; pooling all countries and preferences). However, the diversity ratio at the high-temperature remains significantly below 1 (for each preference, $P < 0.001$). 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 (Chen et al., 2023; A. Wang et al., 2025; D. Wang et al., 2026; S. Wang et al., 2025).

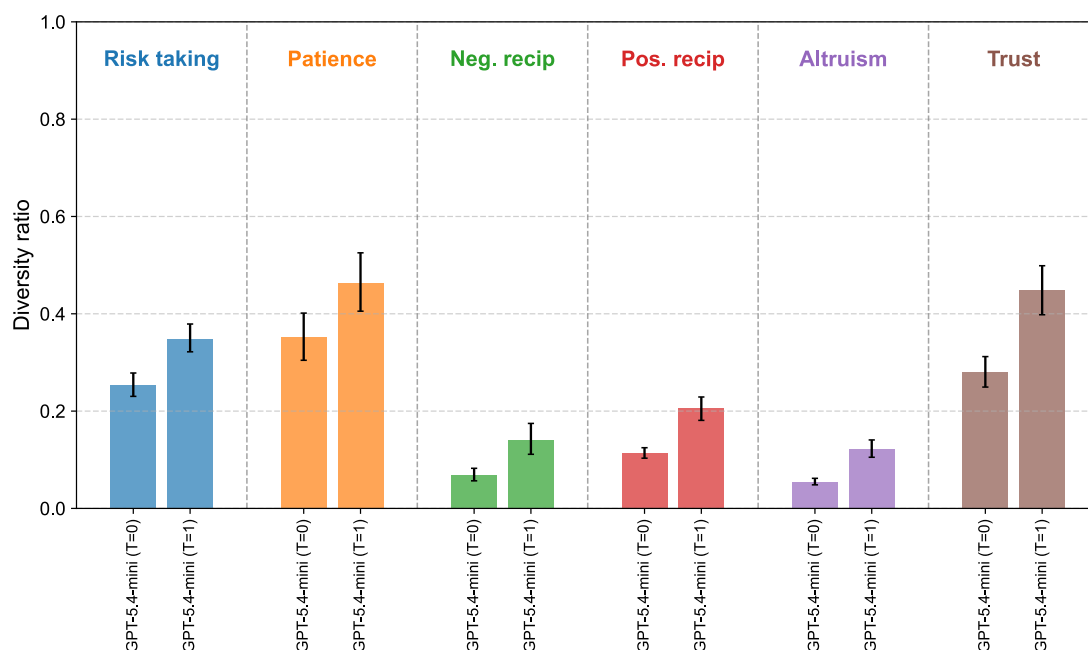


Figure J.3. Diversity ratios for GPT-5.4-mini at temperature 0 and temperature 1, across preference domains. Bars show the mean diversity ratio across countries for each combination

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.3. 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	0.090
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.

(5) Summary for robustness checks. Returning to the data-contamination concern, the paraphrasing and English-language experiments give no evidence that data contamination is a primary driver of the alignment we report at either the individual or country level. We have also expanded the Discussion to acknowledge contamination as a residual upper-bound concern (Paragraph 9 of the revised Discussion):

“Several limitations qualify our conclusions. Because GPS was conducted in 2012, portions of its content may appear in LLM training data. Successful predictions could therefore partially reflect memorization, making our alignment estimates an upper bound. To alleviate this concern, we ran a paraphrasing experiment and a range of English-only experiments (see Appendix G). If memorization drove alignment,

paraphrasing items and translating them from the original survey languages into English should both weaken match with training data; neither manipulation produced systematic changes in individual- or country-level alignment. Additionally, secular shifts in economic preferences over the past decade could lead our analysis to underestimate alignment for current cohorts. To address both concerns simultaneously, future large-scale surveys and intervention experiments could be designed as synchronous digital-twin panels in domains where training data contamination is less likely, helping to disentangle memorization from genuine reasoning.”

(6) Interventions to improve alignment. Beyond establishing that contamination is unlikely to be the primary driver of the patterns we report, you also asked the constructive question of which interventions could improve alignment going forward. We agree that this is the question practitioners ultimately care about, and we have expanded the Discussion accordingly. While a full empirical evaluation of intervention strategies lies beyond the scope of this paper, we now outline two complementary directions motivated by our findings. First, alignment can be targeted from the outside of the model, by combining audits with data augmentation in under-represented contexts, cross-lingual grounding, and domain-specific alignment to narrow the micro-to-macro gap; prior work suggests that cultural- and country-level alignment can be improved through prompt injection (Pataranutaporn et al. 2025; Tao et al. 2024) and fine-tuning (Li et al. 2024; Suh et al. 2025), which is a promising route to addressing the country fixed-effect misalignment we document. Second, alignment can be targeted from the inside of the model, by probing the internal representations associated with preference-relevant outputs and applying neural-level manipulation, such as steering (Khanuja et al. 2026). The full added paragraph appears as the penultimate paragraph of the revised Discussion.

“Moreover, although we find systematic individual-level alignment, the mechanisms by which models reproduce these patterns remain opaque. At least two directions for future research are worth pursuing. The first is combining audits with targeted interventions, such as data augmentation in under-represented contexts, cross-lingual grounding, and domain-specific alignment, to test whether the micro-to-macro gap can

be narrowed. Prior work shows that cultural- and country-level alignment can be improved by prompt injection (Pataranutaporn et al., 2025; Tao et al., 2024) and fine-tuning (Li et al., 2024; Suh et al., 2025), suggesting one promising route to addressing misalignment associated with country fixed effects. The second is to probe LLMs' internal representations and apply mechanistic interventions such as steering, to improve alignment (Khanuja et al., 2026)."

7. Minor comments:

Appendix Line 54-56 (GPT-4o-mini programming error): The manuscript notes a bug where "0" responses were treated as missing and resampled until non-zero, and that this was corrected in later collection rounds. Please clarify whether the buggy data remain in the analyses. If so, a simple robustness check excluding the affected GPT-4o-mini subset (or excluding the affected items) would increase confidence that headline conclusions are not driven by mixed "buggy + corrected" data. If recollection is infeasible, explicitly reporting sensitivity would help.

Response: Thank you for the careful reading and for catching this issue. We have taken several steps to address your concern.

In the initial GPT-4o-mini collection, when the model returned "0" on numeric items, the value was treated as missing and the item was resampled until a non-zero response was obtained. This affected only the local-language GPT-4o-mini condition; the GPT-4o-mini (English) condition collected later was unaffected. The bug was corrected in all subsequent collection rounds for every other model.

OpenAI announced that several sub-models in the GPT-4o-mini family would be deprecated as of July 23, 2026. Rather than re-running a model scheduled for deprecation, we collected GPT-5.4 and GPT-5.4-mini as supplementary conditions in the revised manuscript, which provides stronger external validity by anchoring our conclusions to

the most recent models. The original GPT-4o-mini condition is retained so that readers can still see the full set of twelve LLM-prompting conditions.

In our initial analysis, we retained the buggy data, since the ‘0’ responses account for only a very small proportion of the dataset. [Table R.2](#) 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 proportion of zero responses exceeding 1%. For every other model, the proportion of zero responses remains below 1%. Notably, GPT-4o-mini (English), which shares the same base model as the affected GPT-4o-mini condition, returns zero in 0% of cases. These patterns indicate that the corrected sampling procedure is working as intended and the GPT-4o-mini model rarely outputs zero on these items. This provides an upper bound on how much the resampling could influence our data.

We re-ran each of our four headline analyses without the “buggy data” (GPT-4o-mini, in localized languages), as a robustness check:

- (1) **Individual- and country-level alignment.** Because Spearman’s correlations and MSEs are computed within each model, removing one condition cannot affect the metrics for the others. Average individual-level correlation moves from 0.094 to 0.093 and country-level correlation from 0.137 to 0.138; average individual-level MSE moves from 1.827 to 1.828 and country-level MSE from 1.722 to 1.722. Country-level correlations remain higher than individual-level (two-sided paired t -test, $P = 0.005$), and country-level MSEs remain lower than individual-level ($P < 0.001$). Our first claim is therefore not driven by GPT-4o-mini.
- (2) **Between-model consistency.** Because pairwise correlations and MSEs are computed within each pair of models, dropping one model leaves the remaining pairs unchanged. Overall between-model Spearman’s correlations (individual-level: from 0.347 to 0.346; country-level: from 0.262 to 0.258) and MSEs (individual-level: from 1.377 to 1.377; country-level: from 1.491 to 1.491) move

only marginally.

- (3) **Demographic versus cultural gradient.** Because effect sizes (Cohen’s d) are computed within each model, the gradient comparison is not sensitive to GPT-4o-mini. Both the consistency ratio and the match ratio show the same pattern when GPT-4o-mini is excluded (demographic-level match ratio: from 78.7% to 77.2%; cultural-level match ratio: from 56.6% to 56.9%; demographic-level consistency ratio: from 77.8% to 77.3%; cultural-level consistency ratio: from 48.1% to 48.2%). Both cultural-level ratios remain significantly lower than corresponding demographic-gradient ratios ($P < 0.001$).
- (4) **Association between prediction error and GDP/IDI.** The prediction error decomposition introduced in the revised manuscript ([Figure 6C](#)) is computed within each model, so the result is not sensitive to GPT-4o-mini.

In the revised manuscript, we also introduce the ablation experiment based on GPT-5.4-mini. The result we find in the ablation experiments is independent from the GPT-4o-mini data. Taken together, these checks indicate that our conclusions are not driven by mixed “buggy + corrected” data.

Table R.2. Percentage 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).

8. Clarify use-cases: substitution vs hypothesis generation. The paper is appropriately skeptical about using LLMs as substitutes for survey inference, while also suggesting they may be useful for hypothesis generation or scenario testing. It would help to sharpen this distinction and avoid over-generalizing from a single operationalization. Even one concrete example illustrating when/why the hypothesis-generation use-case remains viable (despite weak macro/cultural validity) would strengthen the narrative.

Response: Thank you for this important suggestion. We agree that distinguishing the use of LLMs as substitutes for human respondents from their use as tools for hypothesis generation is critical to avoid over-generalizing from a single operationalization. We address this in paragraph 7 of the revised Discussion, which sharpens the use-case taxonomy into three distinct constructive uses (probing latent motivational structures, hypothesis generation and study-design diagnostics, and a safer testbed for ethically sensitive counterfactuals) and retains two concrete examples through targeted citations. The first example concerns the diagnosis of cross-cultural extrapolation failure (Henrich et al., 2010; Lowes & Nunn, 2025; Moscona et al., 2026). The second concerns the flagging of measurement-context mismatches (Cohn et al., 2019; Huynh et al., 2020; Yang et al., 2023). The relevant passage reads:

“Despite these constraints, our individual- and demographic-level alignment results highlight several promising scientific uses of LLMs as synthetic respondents. First, LLMs can be used to propose and probe latent motivational structures underlying human behavior, particularly in WEIRD populations, complementing formal decision theory (Bhatia, 2024; Bhatia et al., 2025; Binz et al., 2025; Wu et al., 2025; Xie et al., 2025; Zhu et al., 2025). Second, LLMs can accelerate early-stage hypothesis generation and study-design diagnostics, helping researchers identify promising interventions or measurement strategies before turning to human samples. They can help identify when cross-cultural extrapolation may fail (Henrich et al., 2010; Lowes & Nunn, 2025; Moscona et al., 2026), and flag measurement-context mismatches (Cohn et al., 2019; Huynh et al., 2020; Yang et al., 2023). They may also serve as a fast and low-cost environment for follow-up experiments and robustness checks (Guilbeault et al., 2025),

and provide a safer testbed for exploring ethically sensitive counterfactuals, such as mechanisms that increase or reduce intergroup hostility, before designing human-subject experiments. Beyond synthetic respondents, a complementary application is to use LLMs as a measurement tool (Asirvatham et al., 2026; Fang et al., 2025), expanding social scientists' toolkit."

To make the logic of these examples concrete, we describe both illustrations in detail here. The measurement-context illustration draws on Cohn et al. (2019), who used email-based wallet-return rates as a measure of civic honesty across countries. In contexts where email is rarely used relative to chat (Huynh et al., 2020), this measure may systematically underestimate honesty in countries such as China (Yang et al., 2023), and LLM simulations can flag such mismatches before fielding human studies. The cross-cultural extrapolation illustration builds on the WEIRD critique. Much research in economics and psychology assumes shared fundamental cognitive processes and extrapolates from a single population. A growing literature challenges this assumption (Henrich et al., 2010; Lowes & Nunn, 2025; Moscona et al., 2026), and LLMs may help researchers diagnose where such extrapolation is most likely to fail.

Both illustrations highlight the same underlying logic. Even though our results show that LLMs do not reliably reproduce cross-cultural variation in preferences, the ease of running synthetic conditions makes them a low-cost prior for surfacing hypotheses and for flagging measurement-context mismatches before researchers commit to a costly human study.

To keep the Discussion compact, we have retained these illustrations in the response rather than in the manuscript itself. If the reviewer prefers, we are happy to add the worked examples back into paragraph 7 of the Discussion.

Reviewer 3's Comments

1. This is a technically impressive paper. The authors use several LLMs to construct virtual respondents that adopt the personas of roughly 70,000 individuals. They then compare LLM responses to the human GPS survey by reporting pooled correlations between human and LLM preferences. They find systematic correlations, but the correlations are weak. Moreover, they find variation that raises questions about the use of LLMs to replace humans. I agree with the conclusions. And I applaud the work that was done.

Response: Thank you for this generous assessment and for agreeing with our central conclusion that current LLMs do not provide a reliable substitute for human respondents in cross-cultural research.

Your comments have been especially instrumental for the revision. They prompted us to (i) replace correlation with mean squared error as the primary accuracy metric, (ii) add a direct test of whether silicon respondents compress response variation relative to human respondents, and (iii) redesign [Figures 4](#) and [5](#) so that the recovery of demographic and cultural gradients can be read off the plots at a glance.

We address each of your comments in turn below, and we believe these revisions have sharpened the empirical strategy of the paper.

2. However, I struggled to understand why the authors were doing what they were doing. The analyses were not well-justified and were more complicated and tangential than needed.

For example, the authors begin by comparing the correlation between the LLM predicted and GPS surveyed respondents. But why is this the right metric? Correlation measures linear association, but that does not seem to be the quantity of interest in this research design. In this study each human respondent is paired with an LLM replicant, so the actual human response associated with each silicon response is known. If you are going to use a

silicon respondent, isn't the relevant question: how far the predicted responses are from the true responses. If so, why not measure that directly using a measure such as mean squared error (MSE)?

As is, the correlation seems like it could be very misleading. A high correlation only indicates that two measures covary in a linear fashion across observations, not that the measures for a given (x,y) pair are close to one another in covariate space. But if the goal is to evaluate whether silicon respondents can stand in for human respondents, the key issue is whether the imputations/estimates of the individual responses are unbiased. Measures that are systematically biased can still produce high correlations if the bias affects all observations in a similar way.

Response: We appreciate this important insight. You are correct that correlation and accuracy are conceptually distinct, and that a direct measure of prediction error is more appropriate for evaluating whether silicon respondents can stand in for human respondents. We fully agree that mean squared error (MSE) should serve as a primary metric.

Before turning to MSE, we first clarify a possible source of confusion. Throughout the paper, our primary correlation measure is the Spearman rank correlation rather than the Pearson correlation. Unlike Pearson correlation, which assumes a linear relationship, Spearman correlation captures monotonic, rank-order association and is therefore more appropriate for ordinal preference data.

We retain Spearman correlation as a complementary measure because correlation and MSE capture distinct aspects of alignment. As you note, correlation tells us whether silicon respondents who are predicted to score higher on a given preference also tend to score higher in the GPS data, while MSE tells us how far the predicted values are from the true values. A model could achieve low MSE while having low correlation, or vice versa. Because part of our analysis examines whether LLMs recover the ranking of preferences across demographic groups and countries, for example whether a country that ranks high on patience in the GPS also ranks high in the LLM data, rank-based measures remain

informative for that purpose. In the revised manuscript we therefore report both metrics, but **we foreground MSE as the primary accuracy metric** and use Spearman correlation as a complementary measure of ordinal consistency.

Specifically, we compute the MSE between each GPS-surveyed and LLM-predicted preference pair, pooling across all matched individual observations within each model and preference domain combination. We also compute country-level MSE by comparing country-level weighted means, as shown in [Figure 2](#) (reproduced below for your convenience) in the revised manuscript.

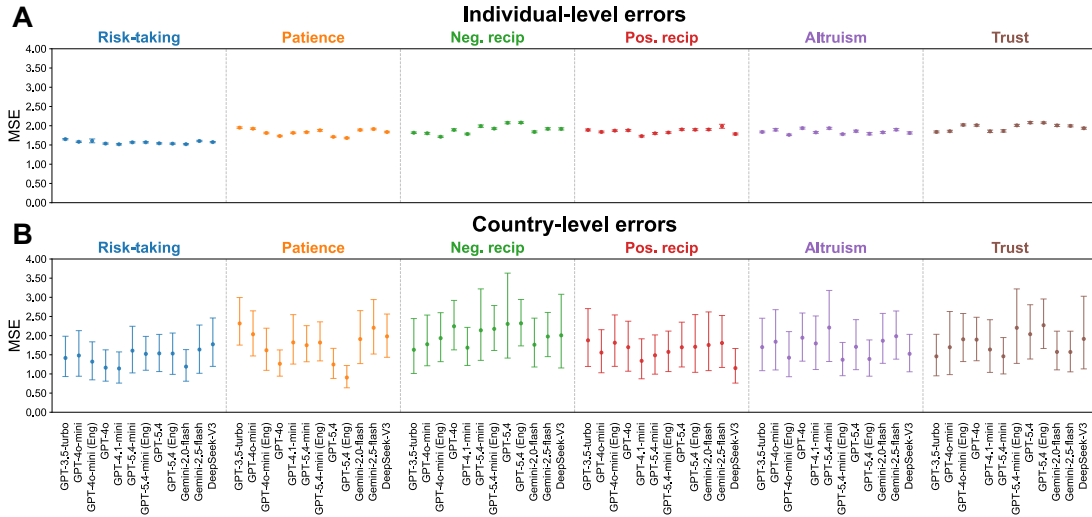


Figure 2. Mean squared errors (MSEs) between LLM-predicted and GPS-surveyed preferences. (A) Individual-level MSEs. Each point represents the MSE between LLM-predicted and GPS-surveyed preferences, pooling all individual observations across 74 countries ($N > 65,000$; the number of observations varies by preference domain due to missing values in the original GPS). **(B) Country-level MSEs.** Each point represents the MSE between country-level average LLM-predicted and GPS-surveyed preferences ($N = 74$). 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. Error bars indicate 95% bootstrapped CIs in both panels.

The MSE results ([Figure 2](#)) confirm and sharpen our original conclusions. The mean of 72 individual-level MSEs (6 preference domains $\times$ 12 LLMs) is 1.827 (95% CI = [1.793, 1.858]). A natural benchmark is the MSE between two independent variables following a standard normal distribution (mean 0 and standard deviation 1), which equals 2. Because

the 95% confidence interval of the observed MSEs excludes this benchmark, we conclude that individual-level LLM-predicted preferences are aligned with GPS-surveyed preferences, though only weakly: compared with the standard deviation of 1, the MSEs are nearly double, indicating systematic but modest individual-level alignment. This pattern echoes our result based on correlations, that 63 out of 72 (87.5%) correlations (6 preference domains $\times$ 12 LLM conditions) are significantly positive. Across preference domains, MSEs are generally lowest for risk-taking (average MSEs across all models: risk-taking: 1.567; patience: 1.832; negative reciprocity: 1.897; positive reciprocity: 1.859; altruism: 1.847; trust: 1.963). The lowest individual-level MSE is for GPT-4.1-mini on risk-taking (MSE = 1.518).

The mean country-level MSEs across preferences is 1.722, lower than their individual-level counterparts ($M_{\text{country}} = 1.722$ vs $M_{\text{individual}} = 1.827$, two-sided paired t -test, $P < 0.001$), although we caution that the two are not directly comparable because the country-level data are constructed as sampling-weighted averages. Yet the MSEs remain large, compared with their standard deviations. 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.

Although we now foreground MSE as the primary metric in place of correlation, the substantive conclusions remain unchanged. This provides evidence for the robustness of our findings.

3. On a related point, work by Bisbee et al. suggests that LLM predictions tend to be homogenizing. Even if there is a high correlation across observations, they may fail to reflect the heterogeneity in human responses. But this is again easy to evaluate in this setting by comparing the standard deviation of human responses with the standard deviation of the matched silicon respondents. If LLM responses compress variation, the standard deviations should be noticeably smaller. This matters because if we use them in

second-stage analyses (e.g., regressions) then the variation in the silicon responses matters!

Response: We agree that assessing whether LLMs compress response variation is critical, particularly because, as you emphasize, the variation in synthetic responses directly affects the validity of any downstream regression analysis.

We have now added a systematic comparison of response dispersion between human and silicon respondents. A practical constraint is that for five of the six GPS preference domains (risk-taking, patience, negative reciprocity, positive reciprocity, and altruism), the published preference indices involve imputation of missing values and global-level standardization (mean = 0, variance = 1), which removes the original response scale and makes standard deviation comparisons uninformative for these domains.

To access the raw human response, we first checked the variable list of Gallup World Poll microdata through the University of Pennsylvania's library, but we did not find raw response in the list. We then contacted the original authors of the GPS (Prof. Armin Falk, Prof. Uwe Sunde, and Prof. Thomas Dohmen) to request access to the individual-level raw survey responses. They confirmed that the GPS data are only released in the currently published format, and that raw item-level responses are not available for external research.

Hence, given that raw responses are unavailable, we focus on the trust domain since trust is measured by a single, non-imputed survey item on an 11-point scale (0–10). We convert GPS-surveyed responses back to the original 0–10 scale and, following Bisbee et al. (2024), compute standard deviations for human and silicon respondents. As shown in [Figure J.2](#) (reproduced below for your convenience), 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 in the trust domain, consistent with the homogenization pattern reported in prior work.

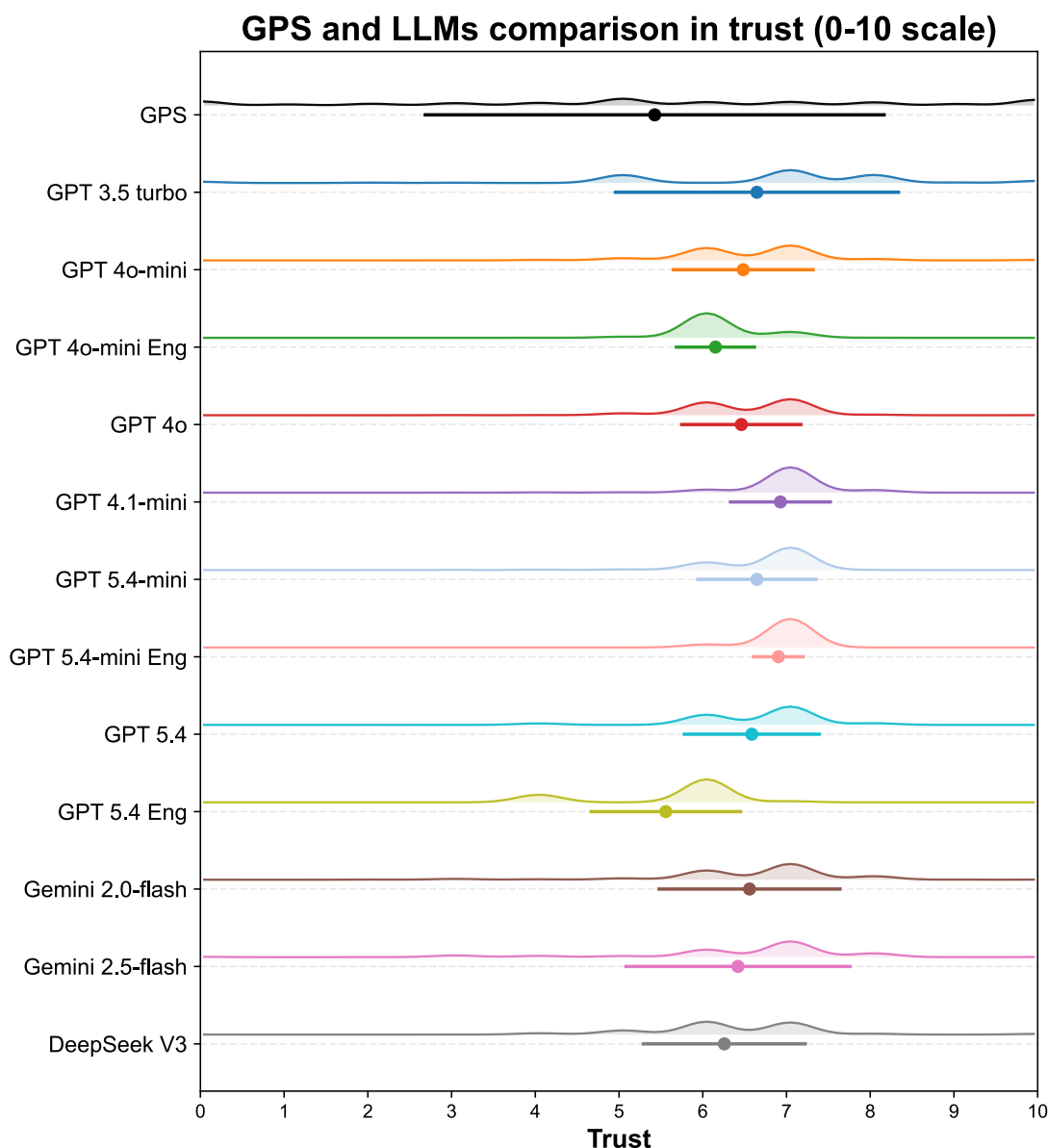


Figure J.2. Standard deviations of trust responses across LLMs and GPS. Bars show the individual-level standard deviation of trust responses for each model and the GPS benchmark, computed by pooling all individual observations across the 74 countries. Sample sizes are matched across models and GPS.

As supplementary evidence, in [Appendix J](#) we used the number of unique response values within each country as a proxy for response diversity across all six preference domains. Averaged across countries and models, the diversity ratio (number of distinct LLM values divided by number of distinct GPS values) is 0.224, and all 72 model–preference combinations have ratios significantly below 1 (two-sided t -tests, $P < 0.001$). This analysis

also reveals that the reduction in diversity is more pronounced in high-income countries, where LLM outputs are more deterministic.

We fully agree that if LLM responses compress variance, this could introduce systematic bias into second-stage analyses by understating standard errors and overestimating statistical power (Bisbee et al. 2024). Hence, we changed our measurement of the demographic and cultural gradients from the consistency ratio (a 95% CI based metric) to the match ratio (a sign-direction based metric). We have also modified our introduction to explicitly point out this potential problem. The new paragraph reads:

“Despite this promise, concerns remain about using LLMs as surrogates for human respondents, given inherent algorithm biases such as identity flattening (A. Wang et al., 2025), systematic skew toward Western, Educated, Industrialized, Rich, and Democratic (WEIRD) populations (Abdurahman et al., 2024; Atari et al., 2023; Crockett & Messeri, 2025; Lin, 2025; Tao et al., 2024), overestimated statistical power (Bisbee et al., 2024; Cui et al., 2025), and other potential risks (Y. Chen et al., 2023; Gao et al., 2025). Prior work has focused largely on political opinions (Argyle et al., 2023; Bisbee et al., 2024; Ceron et al., 2024; Jiang et al., 2024; Röttger et al., 2024; Rozado, 2024; Santurkar et al., 2023), and social and moral attitudes (Abdurahman et al., 2024; Atari et al., 2023; Durmus et al., 2024; Li et al., 2024; Sinacola et al., 2025; Tao et al., 2024), while paying much less attention to economic preferences, even though these preferences play a central role in linking individual behavior to market outcomes and social welfare, and constitute one of the most compelling use cases for LLMs in efforts to improve individual and societal well-being.”

Given the importance of the homogenization problem in LLMs, we have also modified the Discussion to explicitly point out this concern. The new paragraph reads:

“Beyond alignment, we identify two additional practical constraints [...] The other is limited response diversity. Compared with human responses in GPS, LLM-generated responses are more concentrated (Bisbee et al., 2024; D. Wang et al., 2026), and this concentration is itself patterned by country-level development (see Appendix J). Though

part of this homogenization likely reflects our decision to set temperature to zero for reproducibility, increasing the temperature to 1 only partially restores response diversity, which still remains substantially below human levels (see Appendix J)."

We thank you again for raising this point. In the revised manuscript, we explicitly note this homogenization as a practical limitation of LLMs as synthetic respondents (Discussion, paragraph on "limited response diversity"), and direct readers to [Appendix J](#) for the full diagnostic.

4. Moreover, if the authors want to examine how performance varies across indices or countries, they could simply predict MSE in point estimates or differences in standard deviations as a function of those characteristics.

Response: We thank you for this constructive suggestion. We agree that directly modeling prediction error and response-diversity as functions of country/index characteristics provides a sharper diagnostic of where and why LLMs fail. The revised manuscript now implements both forms of this analysis.

(1) Modeling MSEs as functions of country/index characteristics. Following your recommendation, we predict MSE as a function of country characteristics and provide these results in three ways.

a. Correlation-based results. [Figure 6B](#) (reproduced below for your convenience) reports Spearman correlations between the country-level inconsistency index (the weighted country means of individual-level MSEs) and four country-level features (log GDP per capita, the ICT Development Index (IDI), log population, and log number of internet users). The inconsistency index is negatively correlated with GDP (59 of 72 model-preference combinations, 81.9%) and IDI (58 of 72, 80.6%), but uncorrelated with population size or number of internet users (population: 34/72 negative, 47.2%; internet

users: 43/72 negative, 59.7%). Economic development and digital infrastructure are therefore the dominant country-level drivers of LLM-human misalignment.

b. OLS coefficient results. Besides the correlation results, we report regressions of country-level inconsistency index on four country characteristics in [Appendix E.3](#). Columns (1)–(4) of [Table E.5](#) (reproduced below) show that 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 smaller in magnitude: between -0.020 and -0.006 for log population and between -0.011 and -0.006 for log number of internet users.

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.5](#) reports regressions including both principal components. In the PCA specification, 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](#) 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 economic-development score remains significantly negative ($b = -0.058$, $P < 0.001$), whereas the population score is not statistically distinguishable from zero ($b = 0.015$, $P = 0.251$).

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](#) (reproduced below for your convenience) reports regression results using this alternative definition. Under this alternative definition, economic-development variables remain strongly and negatively associated with the inconsistency index.

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*** (0.017)	-0.205*** (0.019)				
ICT Development Index			-1.343*** (0.131)	-1.494*** (0.150)		
log population	-0.020** (0.009)		-0.006 (0.009)			
log number of internet users		-0.011 (0.010)		-0.006 (0.010)		
Score for economic development					-0.201*** (0.019)	-0.058*** (0.017)
Score for population					-0.007 (0.010)	0.015 (0.013)
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 robust standard errors in parentheses. All regressions include preference fixed effects, model fixed effects, and their interaction terms. * $P < 0.10$, ** $P < 0.05$, *** $P < 0.01$.

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*** (0.036)	-0.260*** (0.039)				
ICT Development Index			-1.885*** (0.270)	-2.141*** (0.296)		
log population	0.018 (0.025)		0.048* (0.027)			
log number of internet users		0.042 (0.027)		0.051* (0.027)		
Score for economic development					-0.264*** (0.038)	-0.172*** (0.057)
Score for population					0.048* (0.027)	0.104* (0.053)
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 robust standard errors in parentheses. All regressions include preference fixed effects, model fixed effects, and their interaction terms. * $P < 0.10$, ** $P < 0.05$, *** $P < 0.01$.

c. OLS R^2 results. A third approach is to measure how much MSE can be predicted by single dimensions. [Figure 6C](#) (reproduced below) decomposes individual-level MSE further by regressing the average individual-level MSE across the six preference domains on demographic variables, country-level continuous variables, and country fixed effects, and reporting the R^2 of each regression.

Country fixed effects explain the largest share of error variance, ranging from 3% to 37% across models, substantially more than demographic information; among continuous country variables, economic indicators (GDP, IDI) explain more than population-based variables (up to nearly 7%, from DeepSeek-V3 for IDI). This identifies why predictions fail: errors are concentrated at the country level rather than the demographic level, indicating that LLMs are systematically insensitive to country labels.

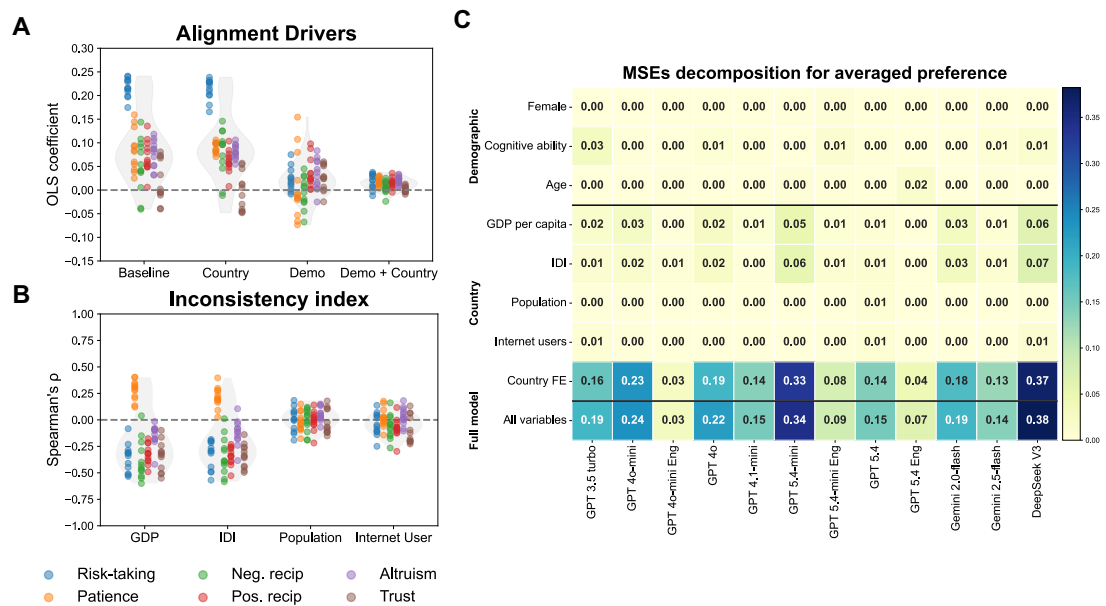


Figure 6. Decomposing the sources of LLM-GPS misalignment. **(A)** Regression coefficients from models regressing GPS-surveyed preferences on LLM-predicted preferences. Each point represents an OLS coefficient estimated on pooled individual-level data for a given model $\times$ preference combination (12 LLMs $\times$ 6 preferences). The four categories on the x-axis correspond to four alternative regression specifications: baseline (no controls), with country indicators, with demographic characteristics, and with both sets jointly. **(B)** Spearman correlations between the inconsistency index (the GPS sampling-weighted within-country mean of individual-level MSEs) and four country characteristics: log GDP per capita, the Information and Communications

Technology Development Index (IDI), log population, and log number of internet users. Each point represents a Spearman correlation coefficient computed across 74 countries. Each characteristic shows 72 points (12 LLMs $\times$ 6 preferences). **(C)** Error decomposition of LLM-predicted preferences. For each LLM (column), separate OLS regressions are estimated with the average individual-level MSE across six preferences as dependent variable and each variable category (row) as independent variable. Cell values report the R^2 of each single-predictor regression.

(2) Modeling response dispersion as functions of country/index characteristics.

The second analysis you suggest, [Appendix J](#) (Tables J.1 and J.2, reproduced below) regresses the country-level diversity index on standardized economic-development and population indicators, controlling for the log of sample sizes of each country, preference-domain fixed effects, model fixed effects, and model $\times$ preference interaction effects throughout. The diversity index is a homogenization metric defined in [Appendix J](#) as the ratio of distinct LLM response values to distinct GPS response values within each country. For trust, where the original 0–10 response scale is preserved, this dispersion metric is directly comparable to the SD-difference framing in your Comment 3; for the other five preferences, where global standardization removes the original scale, the diversity-index proxy provides the closest available equivalent. The first principal component of the economic-development indicators is strongly negatively associated with response diversity ($b = -0.169$, $P < 0.001$ for all preferences pooled; $b = -0.105$, $P < 0.001$ for the trust-only specification), whereas the population-component coefficient is much smaller in both specifications ($b = -0.039$). This shows that the SD compression documented in our response to Comment 3 above is itself heterogeneous across countries, with more economically developed contexts producing more deterministic LLM outputs.

Together, these analyses move beyond aggregate accuracy and provide the direct, characteristic-level diagnostic of where LLM performance deteriorates that you recommend.

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

	(1)	(2)	(3)	(4)	(5)
std. log [GDP p/c]	-0.161*** (0.010)	-0.167*** (0.011)			
std. IDI			-0.147*** (0.012)	-0.149*** (0.013)	
std. log population	-0.043*** (0.009)		-0.037*** (0.009)		
std. log internet users		-0.039*** (0.009)		-0.036*** (0.009)	
std. score for economic development					-0.169*** (0.013)
std. score for population					-0.039*** (0.009)
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 robust standard errors in parentheses. Regressions include controls for the log of the GPS sample size in each country, preference-domain fixed effects, model fixed effects, and model $\times$ preference interaction effects. * $P < 0.10$, ** $P < 0.05$, *** $P < 0.01$.

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

	(1)	(2)	(3)	(4)	(5)
std. log [GDP p/c]	-0.128*** (0.025)	-0.120*** (0.025)			
std. IDI			-0.072** (0.028)	-0.073*** (0.028)	
std. log population	-0.042* (0.021)		-0.029 (0.022)		
std. log internet users		-0.039* (0.022)		-0.034 (0.022)	
std. score for economic development					-0.105*** (0.027)
std. score for population					-0.039* (0.022)
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 six preference domains and 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 robust standard errors in parentheses. Regressions include controls for the log of the GPS sample size in each country and model fixed effects. * $P < 0.10$, ** $P < 0.05$, *** $P < 0.01$.

5. I had the same set of reactions to Figure 3. Why do you want to, or need to, focus on correlations? If LLMs are meant to replace human respondents, the relevant benchmark is how closely they reproduce human responses. The natural statistic in Figure 3 would therefore be the average MSE in predicted responses across models within each sample (i.e., do the models give similar imputations?).

Response: We agree that for Figure 3 the MSE is a more appropriate metric for assessing how similar models' imputations are to one another. We have therefore replaced the correlation-based Figure 3 with an MSE-based version. The new Figure 3 (reproduced below for your convenience) reports the between-model MSE (i.e., how far apart different LLMs' predictions are from each other for the same individual) and we juxtapose these values against the LLM-GPS MSEs already reported in Figure 2 (and tabulated by preference in Appendix Table C.2). If models converge on similar but inaccurate predictions, between-model MSE would be low while model-GPS MSE remains high, which is precisely what we observe.

The new Figure 3 shows that 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 individual-level LLM-GPS MSEs (1.377 vs 1.827, two-sided t -test, $P < 0.001$; see Appendix Table C.2 for comparison of each preference, reproduced below).

At the country level (Appendix Figure C.3, reproduced below), the mean 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), again lower than the country-level LLM-GPS MSEs (1.491 vs 1.722, two-sided t -test, $P < 0.001$; see Appendix Table C.2 for comparison of each preference).

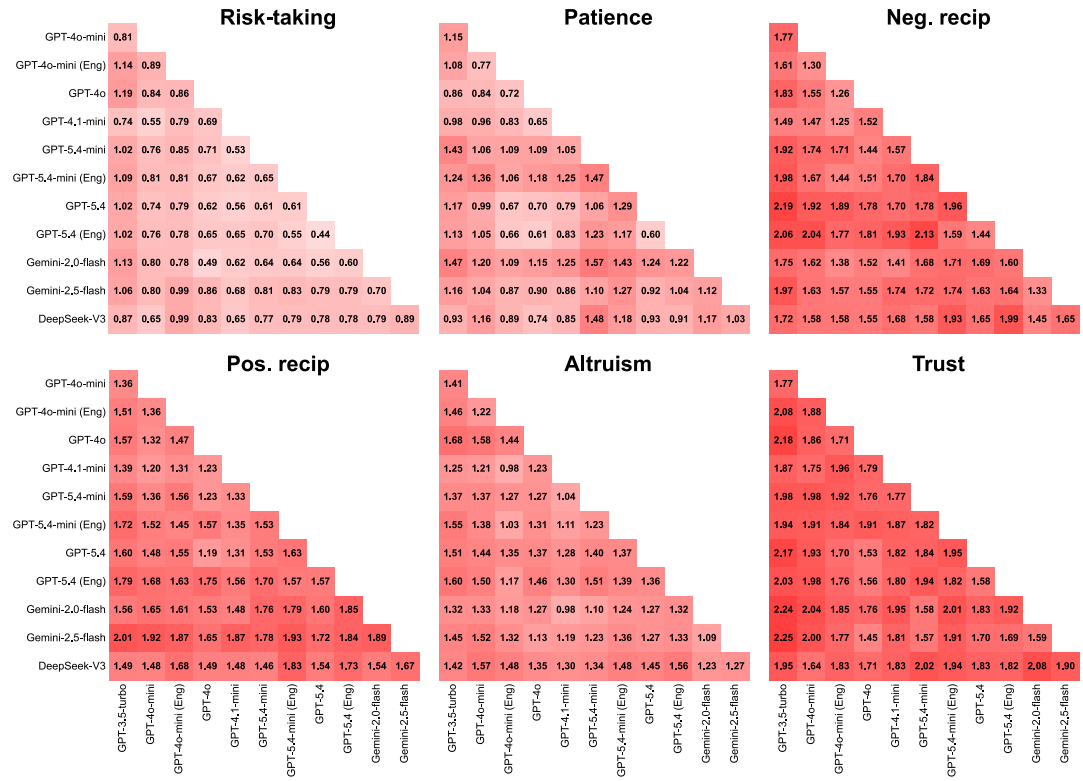


Figure 3. Individual-level between-model MSEs of LLM-predicted preferences. Each cell of the 12×12 matrix reports the MSE between two LLM configurations, computed by pooling individuals across all 74 countries. Individual-level preference indices are standardized to mean 0 and standard deviation 1 prior to MSE computation. Color intensity encodes MSE magnitude, with lower values indicating greater consistency between models.

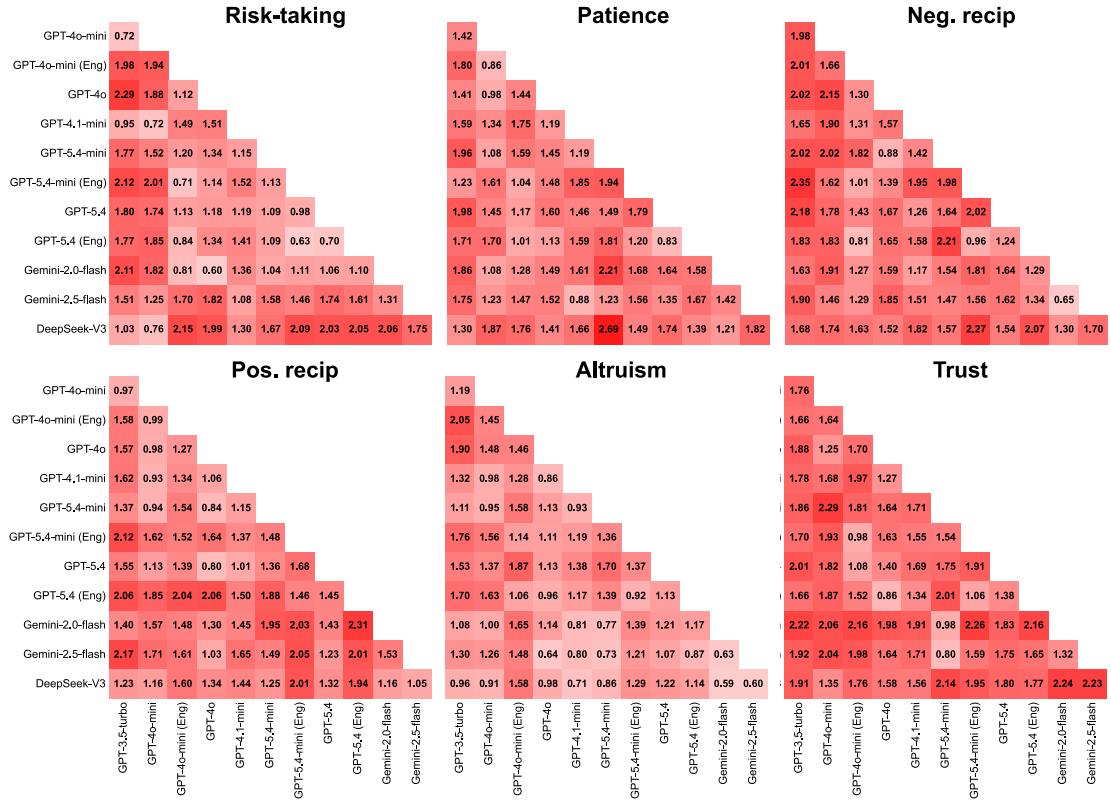


Figure C.3. Country-level between-model MSEs. Each cell reports the MSE between two LLM configurations, computed using country-level averages of LLM-predicted preferences (weighted by GPS ex-post sampling weights). Country-level preference indices are standardized to mean 0 and standard deviation 1 prior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

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

Preference	GPS-LLM MSEs	Between-model MSEs	P-value
Individual-level			
Risk-taking	1.567	0.770	<0.001
Patience	1.832	1.049	<0.001
Negative reciprocity	1.897	1.679	<0.001
Positive reciprocity	1.859	1.579	<0.001
Altruism	1.847	1.329	<0.001
Trust	1.963	1.856	0.003
Country-level			
Risk-taking	1.444	1.423	0.794
Patience	1.740	1.500	0.081
Negative reciprocity	1.996	1.628	<0.001

Positive reciprocity	1.622	1.485	0.082
Altruism	1.730	1.199	<0.001
Trust	1.802	1.709	0.319

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.

These results suggest that LLMs make broadly similar predictions to one another at both levels, but these shared predictions align only weakly with human economic preferences, indicating a common but imperfect underlying mechanism for how current LLMs represent and extrapolate such preferences. These MSE-based comparisons strengthen the original finding from the correlation matrices: models are internally consistent but externally misaligned, a pattern that is now documented using both associational (correlation) and accuracy-based (MSE) metrics.

6. In the second half of the paper, when turning to demographic level consistency, the authors report Cohen's *d* across several preference domains. The results are presented in Figures 4 and 5. I confess that the wall of plots are very hard to interpret. It is not clear, for example, what to make of the considerable variation in Figure 5 across indices, models, and orientations.

Response: We appreciate this feedback. The original figures were indeed dense and difficult to parse. We have comprehensively revised the visualization strategy for both Figures 4 and 5 in the revised manuscript.

In the revised figures (Figure 4 and Figure 5, reproduced below for your convenience), we plot the Cohen's *d* effect sizes estimated from GPS-surveyed data on the *x*-axis and the corresponding effect sizes from LLM-predicted data on the *y*-axis. A 45-degree reference line indicates perfect recovery of effect sizes. This format makes it immediately visible (1) whether LLMs recover the *direction* of demographic and cultural effects (points in the

correct quadrants), (2) whether LLMs recover the *magnitude* of these effects (proximity to the 45-degree line), and (3) whether LLMs systematically *amplify* effects (points lying beyond the 45-degree line). For each panel, we report the percentage of points in each of the four quadrants, providing both a summary of directional consistency and a transparent view of where mismatches occur.

This visualization makes the two central patterns in our data immediately apparent. For demographic gradients (Figure 4), most points cluster in the correct quadrants, with an overall match ratio of 78.7% (95% CI = [72.6%, 84.0%]), but many points lie above the 45° line in Quadrant I (and below it in Quadrant III), indicating systematic amplification of demographic effects by LLMs. This pattern is most pronounced for risk-taking (34 out of 36 LLM-demographic combinations show amplification) and patience (30 out of 36).

For cultural dimensions (Figure 5), points are dispersed much more widely, and many fall in the wrong quadrants, visually confirming the significantly lower cultural-level match ratio of 56.6% (95% CI = [53.6%, 59.6%]; $\chi^2(1) = 35.89$, $P < 0.001$). The considerable variation you noted across indices, models, and orientations reflects a substantive finding: LLMs are markedly better at recovering demographic gradients than cross-cultural heterogeneity in economic preferences.

Demographic gradient: LLMs versus GPS

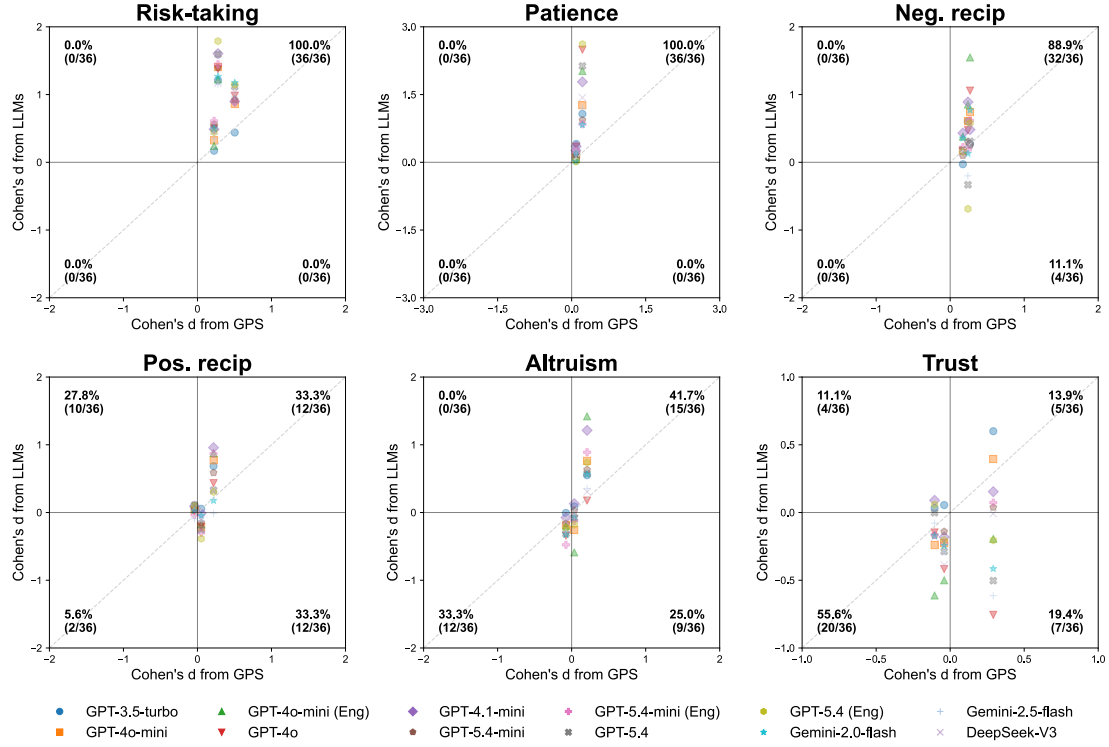


Figure 4. Demographic gradients in GPS-surveyed and LLM-predicted preferences. Each panel corresponds to one of the six preference domains. Within each panel, the x-axis shows effect sizes (Cohen's d) estimated from GPS-surveyed preferences and the y-axis shows the corresponding effect sizes estimated from LLM-predicted preferences. Demographic variables are encoded as binary indicators for gender (male vs female), age (< 60 vs ≥ 60), and cognitive ability (≥ 5 vs < 5). Each point represents one (LLM $\times$ demographic dimension) pair; with 12 LLM configurations and three demographic dimensions (gender, age, and cognitive ability), each panel contains 36 points. Positive effect sizes indicate higher preference levels for males (vs females), younger (vs older), and higher cognitive ability (vs lower). The dashed 45° line denotes exact consistency between LLM- and GPS-derived effect sizes in both sign and magnitude. The percentage in each quadrant reports the share of the 36 (LLM $\times$ demographic dimension) pairs falling in that quadrant, and Quadrants I (upper-right) and III (lower-left) jointly indicate sign-consistent alignment with GPS. Marker shape and color denote the LLM configuration (see legend). Axis ranges differ across panels for visualization clarity and should not be compared across panels.

Cultural gradient: LLMs versus GPS

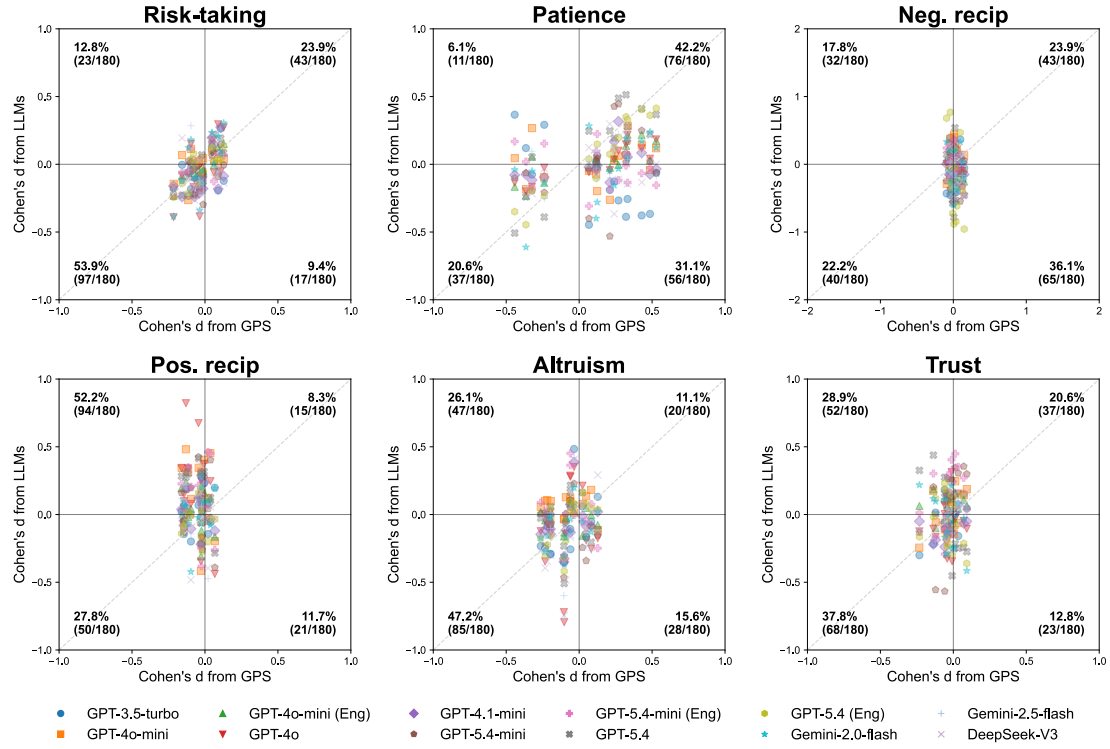


Figure 5. Cultural-level consistency and gradients in GPS-surveyed and LLM-predicted preferences. Each panel corresponds to one of the six preference domains. Within each panel, the x -axis shows effect sizes (Cohen's d) estimated from GPS-surveyed preferences and the y -axis shows the corresponding effect sizes estimated from LLM-predicted preferences. For each cultural dimension, countries are split at the median into low and high groups, and individual respondents inherit the classification of their country. Each point represents one (LLM $\times$ cultural dimension) pair; with 12 LLM configurations and 15 cultural dimensions, each panel contains 180 points. Positive effect sizes indicate higher preference levels in high-group countries relative to low-group countries. The dashed 45° line denotes exact consistency between LLM- and GPS-based effect sizes in both sign and magnitude. The percentage in each quadrant reports the share of the 180 (LLM $\times$ cultural dimension) pairs falling in that quadrant, and Quadrants I (upper-right) and III (lower-left) jointly indicate sign-consistent alignment with GPS. Marker shape and color denote the LLM configuration (see legend). Axis ranges differ across panels for visualization clarity and should not be compared across panels.

7. I think Figure 6 comes closest to the type of analysis that seems appropriate, but it is still more complicated than necessary. If the silicon responses matched human responses shouldn't the regression coefficient simply be 1 - meaning that there is a high correlation? That said, this still has the issue with conflating linear association with similarity. A cleaner strategy would be to model the difference between the human and silicon responses directly. For example, one could predict the distance between the two responses as a function of question type, model type, country, demographics, or other relevant features. This would provide a direct assessment of where and why the predictions fail.

Response: Thank you for this constructive suggestion. We agree on both points: (i) under perfect alignment the coefficient on LLM-predicted preferences should equal 1, and (ii) regression-on-LLM-predictions still conflates linear association with similarity. We have therefore restructured the analysis around direct modeling of the prediction error, exactly as you recommend.

(1) Reframe the role of Figure 6A. To clarify the role of Figure 6A in the revised manuscript (reproduced below for your convenience): this panel is no longer intended as a stand-alone test of whether LLM predictions equal human responses (which is now assessed directly via MSE in Figure 2; see our responses to Comments 2 and 9). Instead, Figure 6A serves a diagnostic purpose. By progressively adding country indicators and demographic controls, it shows which set of variables accounts for the observed alignment. The finding that country indicators leave the coefficient essentially unchanged ($b = 0.000, P = 0.961$) while demographic controls substantially attenuate it ($b = -0.066, P < 0.001$) reveals that the alignment is driven primarily by demographic information rather than country context. We have rewritten the surrounding text to make this diagnostic interpretation explicit and added a range of prompt-component ablation experiments to provide more directly diagnostic evidence.

(2) Modeling the distance between two preferences. Following your recommendation, we now directly model the distance between LLM-predicted and GPS-surveyed responses as a function of country, demographic, and model features. This is implemented in two

complementary ways:

a. Country-level error modeling (Figure 6B). We construct an inconsistency index, the country-level MSE between LLM-predicted and GPS-surveyed preferences, and relate it to four country characteristics (log GDP per capita, IDI, log population, log internet users). The inconsistency index is negatively correlated with GDP (59/72, 81.9%) and IDI (58/72, 80.6%), but uncorrelated with population size or number of internet users. This directly identifies where predictions fail: in lower-income, less digitally developed countries.

b. Individual-level error decomposition (Figure 6C). We regress the individual-level MSE (averaged across the six preference domains) on demographic characteristics, country characteristics, and country fixed effects, and report the R^2 for each predictor block. Country fixed effects explain the largest share of error variance (3%–37% across models), substantially more than demographic information; among country-level continuous predictors, economic variables (GDP, IDI) explain more than population-based variables (up to nearly 7%, from DeepSeek-V3). This directly identifies why predictions fail: errors are explained mainly by the country-level characteristics rather than the demographic-level variables, indicating that LLMs are systematically insensitive to country context.

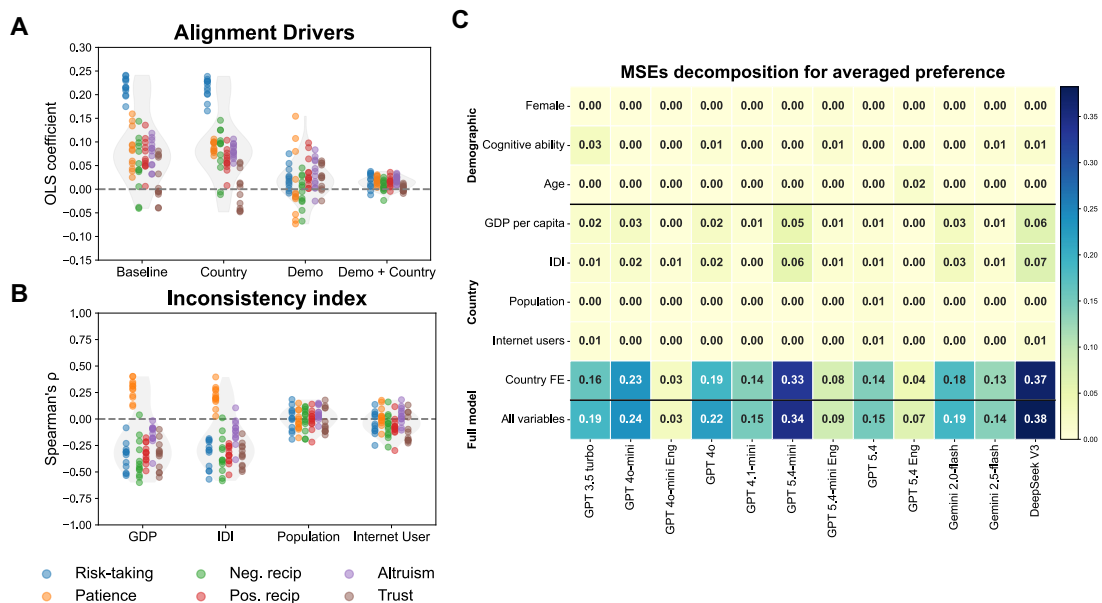


Figure 6. Decomposing the sources of LLM-GPS misalignment. (A) Regression coefficients from models regressing GPS-surveyed preferences on LLM-predicted preferences. Each point

represents an OLS coefficient estimated on pooled individual-level data for a given model $\times$ preference combination (12 LLMs $\times$ 6 preferences). The four categories on the x -axis correspond to four alternative regression specifications: baseline (no controls), with country indicators, with demographic characteristics, and with both sets jointly. **(B)** Spearman correlations between the inconsistency index (the GPS sampling-weighted within-country mean of individual-level MSEs) and four country characteristics: log GDP per capita, the Information and Communications Technology Development Index (IDI), log population, and log number of internet users. Each point represents a Spearman correlation coefficient computed across 74 countries. Each characteristic shows 72 points (12 LLMs $\times$ 6 preferences). **(C)** Error decomposition of LLM-predicted preferences. For each LLM (column), separate OLS regressions are estimated with the average individual-level MSE across six preferences as dependent variable and each variable category (row) as independent variable. Cell values report the R^2 of each single-predictor regression.

(3) Diagnostic ablation experiments. To move beyond correlational evidence, we additionally conduct a series of ablation experiments (a new section, “Diagnostic evidence from ablation experiments”, in the revised manuscript) that causally identify the role of country versus demographic information. Removing country information from the prompt leaves individual-level MSE essentially unchanged (1.835 vs 1.838 baseline, $P = 0.820$), whereas removing demographic information significantly increases MSE (1.961 vs 1.838, $P = 0.020$). These results confirm that the demographic component of the prompt is what drives the limited alignment we observe, while country labels in the prompt are largely ignored.

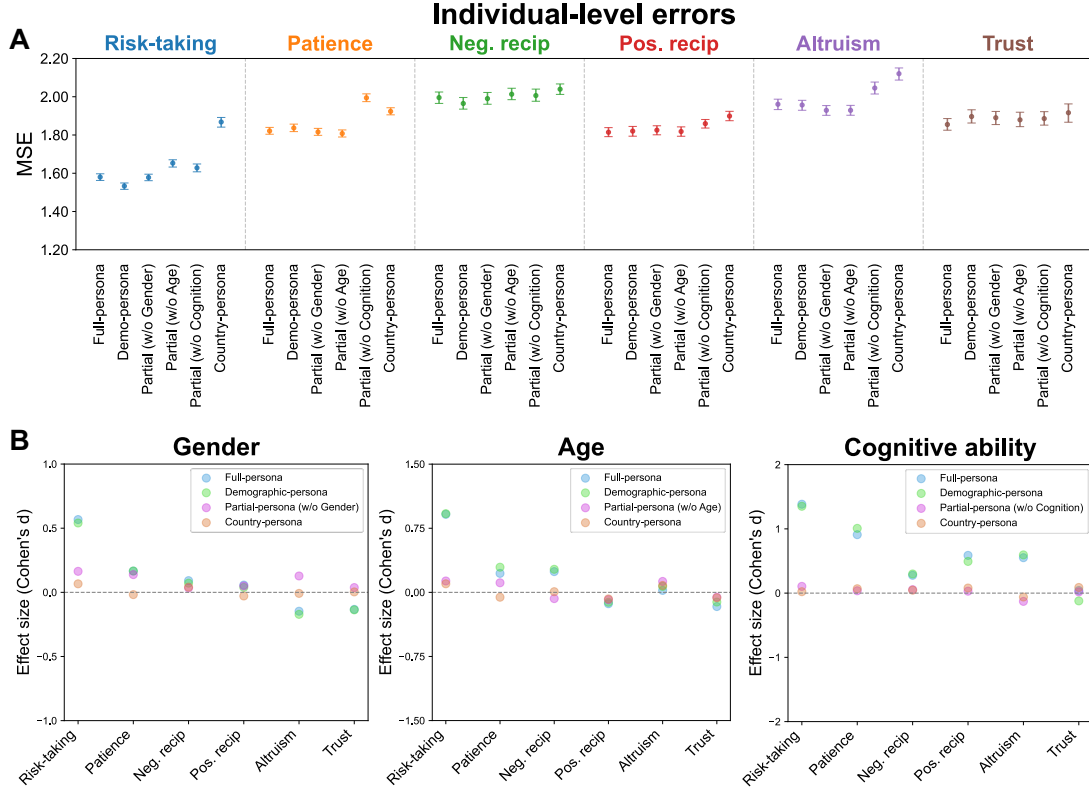


Figure 7. Ablation experiments isolating the contributions of country and demographic information in persona prompts. (A) Individual-level errors (MSEs) across ablation conditions. Each point represents the MSE between GPS-surveyed and LLM-predicted preferences, pooling all individual observations across 74 countries ($N > 65,000$) for a given model $\times$ preference combination, shown for six ablation conditions: full-persona (system prompt contains gender, age, cognitive ability, and country); demographic-persona (only demographic characteristics, country removed); three partial-persona conditions (full-persona with gender, age, or cognitive ability removed, respectively); and country-persona (only country, all demographic characteristics removed). Error bars denote 95% bootstrapped CIs. **(B) Demographic effect sizes across ablation conditions.** Each subpanel corresponds to one demographic dimension (gender, age, cognitive ability). Each point plots a Cohen's d effect size of a demographic variable (female vs male, older vs younger, lower vs higher cognitive ability) estimated from LLM-predicted preferences under each ablation condition. For each subpanel, four conditions are shown: full-persona, demographic-persona, country-persona, and the matched partial-persona (e.g., 'w/o Gender' in the Gender subpanel; see legend in each panel). y-axis ranges differ across subpanels to accommodate effect-size magnitudes.

Together, these revisions implement exactly the simple but direct strategy you proposed. Figure 6A diagnoses what carries the apparent alignment, showing that the coefficient is

sustained by demographic information in the prompt rather than by country context; Figures 6B and 6C then model the prediction error directly, identifying where the predictions fail (low-GDP/IDI countries) and why they fail (LLMs are systematically insensitive to country context); and the ablation experiments in Figure 7 provide mechanistic evidence on the prompt components responsible for these patterns. We are grateful for this suggestion, which substantially strengthened the analytical rigor of the paper.

8. Stepping back, the data is impressive, but I struggle to comprehend whether the analysis is the optimal design for evaluating the claim. If the goal is to evaluate how closely LLMs approximate human responses, the analysis should be both simpler and different from what is currently presented.

Response: Thank you for this candid and constructive framing. We agree that the original analytical design conflated two distinct objectives (measuring linear association and measuring accuracy) and was unnecessarily complicated as a result. Taking your call for a *simpler and different* analysis seriously, we have restructured the entire analytical strategy around a single, transparent question: *how closely do LLM-predicted responses approximate human responses?*

Concretely, the revised manuscript adopts mean squared error (MSE) between LLM-predicted and GPS-surveyed responses as the primary accuracy metric throughout, computed at both the individual level (pooled across all matched observations within each model $\times$ preference combination) and the country level (sampling-weighted country means). Correlation-based statistics are now retained only as complementary measures of rank-order consistency rather than as primary accuracy evidence. This shift directly addresses your central concern that linear association cannot be conflated with similarity.

Building on this MSE foundation, we have implemented all four analyses you recommend:

- **Analysis 1** (MSE-based accuracy): now the central evidence in the new [Figure 2](#) (see our response to Comments 2 and 9).
- **Analysis 2** (dispersion comparison): added as a systematic homogenization analysis on the original 0–10 trust scale, complemented by a response-diversity proxy across all six domains ([Appendix J](#); see our response to Comments 3 and 10).
- **Analysis 3** (cross-model and cross-context comparison using the same metric): implemented as between-model MSE alongside model–GPS MSE in the new [Figure 3](#) (see our response to Comments 5 and 11).
- **Analysis 4** (direct error modeling): now implemented in the redesigned [Figure 6](#), where we (i) regress the country-level inconsistency index on country characteristics (GDP, IDI, population, internet users) and (ii) decompose individual-level MSE into demographics, country covariates, and country fixed effects (see our response to Comments 4, 7, and 12). We further provide mechanistic evidence for the source of (mis)alignment through a new ablation-experiment section in the main text ([Figure 7](#)).

Together, these revisions yield a simpler analytical design (one primary metric, one direct question), a more direct assessment of accuracy (MSE rather than correlation), and a more diagnostic treatment of where and why predictions fail (error modeling and ablations). We respond in detail to each of your specific recommendations in Comments 9–12 below, and provide a closing synthesis in our response to Comment 13.

9. What I would recommend doing:

[Analysis 1: Calculate the average MSE between human and silicon respondents by country, index, or other relevant categories. This directly assesses bias. Correlation cannot do this because it measures linear association rather than accuracy.](#)

Response: We agree with this recommendation. MSE directly assesses bias in a way that correlation cannot. As detailed in our response to your Comment 2 above, we have

adopted MSE as the primary accuracy metric throughout the revised manuscript. Specifically, we compute MSE between each GPS-surveyed and LLM-predicted preference pair at the individual level (pooled across all matched observations within each model $\times$ preference combination) and at the country level (comparing sampling-weighted country means). The mean individual-level MSE across all 72 model $\times$ preference combinations is 1.827 (95% CI = [1.793, 1.858]), and the mean country-level MSE is 1.722. We further break down MSE by preference domain (risk-taking: 1.567; patience: 1.832; negative reciprocity: 1.897; positive reciprocity: 1.859; altruism: 1.847; trust: 1.963) and by model. These results are presented in the new [Figure 2](#) (reproduced below for your convenience) and discussed in detail in the revised Results section. Correlation (Spearman) is retained only as a complementary measure of rank-order consistency.

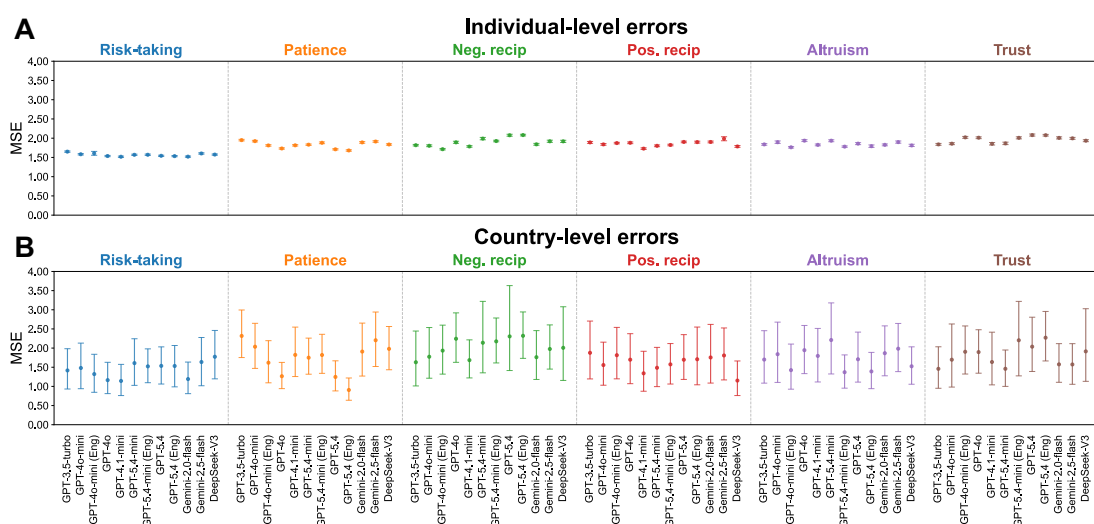


Figure 2. Mean squared errors (MSEs) between LLM-predicted and GPS-surveyed preferences. (A) Individual-level MSEs. Each point represents the MSE between LLM-predicted and GPS-surveyed preferences, pooling all individual observations across 74 countries ($N > 65,000$; the number of observations varies by preference domain due to missing values in the original GPS). **(B) Country-level MSEs.** Each point represents the MSE between country-level average LLM-predicted and GPS-surveyed preferences ($N = 74$). 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. Error bars indicate 95% bootstrapped CIs in both panels.

10. Analysis 2: Examine whether LLMs reproduce variation in human responses. For each subset, compute the standard deviation of human responses and the standard deviation of the corresponding silicon responses and compare them. This matters because the amount of variation affects precision in downstream analyses.

Response: We agree with this recommendation. Compressed variation would bias regression coefficients in any downstream analysis using synthetic data, making dispersion comparison essential. As detailed in our response to your Comment 3 above, we have now conducted a systematic dispersion comparison.

Following your suggestion, we apply the strategy developed by Bisbee et al. (2024) to the trust domain, the one GPS index measured on a clean, non-imputed 0–10 scale. 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), confirming the homogenization pattern (see [Appendix Figure J.2](#), reproduced below for your convenience). For the remaining five preference domains, the published GPS indices involve imputation and global standardization that remove the original response scale, precluding a direct SD comparison.

As supplementary evidence, we use the ratio between number of unique response values from LLMs and GPS within each country as a diversity proxy across all six domains ([Appendix J](#)). The mean diversity ratio is 0.224, and all 72 model × preference combinations have ratios significantly below 1 (two-sided *t*-tests, $P < 0.001$). The revised Discussion now addresses the implications of this homogenization for downstream uses of LLM-generated data.

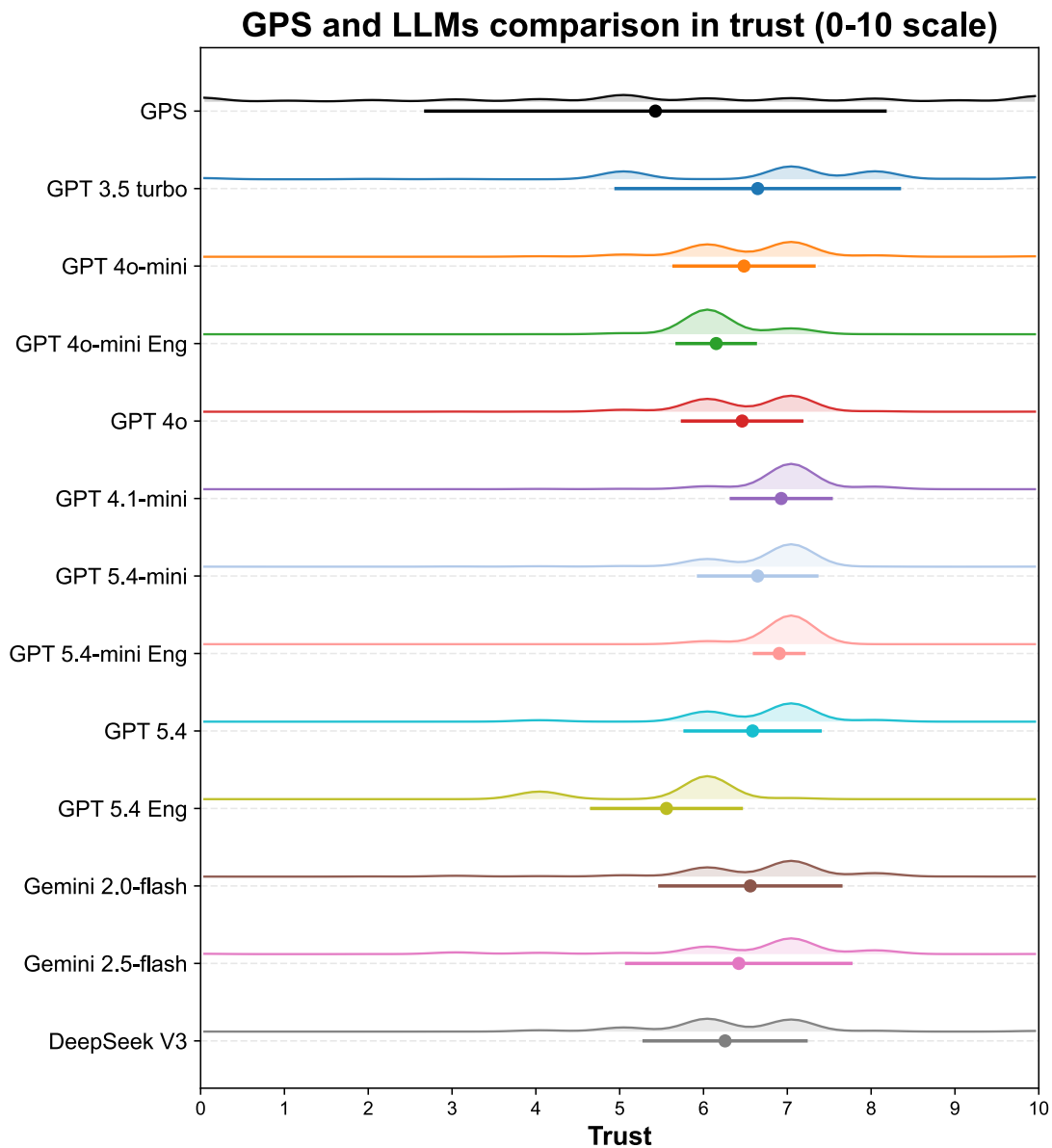


Figure J.2. Standard deviations of trust responses across LLMs and GPS. Bars show the individual-level standard deviation of trust responses for each model and the GPS benchmark, computed by pooling all individual observations across the 74 countries. Sample sizes are matched across models and GPS.

11. Analysis 3: Examine how these quantities vary across models and contexts. Using the same metric, such as MSE, allows a straightforward comparison across models.

Response: We agree with this recommendation. A common metric enables transparent cross-model and cross-context comparisons. As detailed in our response to your Comment 5 above, we now compute between-model MSEs alongside LLM–GPS MSEs to

examine how prediction accuracy and inter-model agreement vary across models and preference domains.

The mean individual-level between-model MSE is 1.377, significantly lower than the mean LLM-GPS MSE of 1.827 (two-sided t -test, $P < 0.001$), indicating that models converge on similar predictions that nonetheless diverge from human responses. This pattern holds at the country level as well (between-model MSE: 1.491 vs LLM-GPS MSE: 1.722, $P < 0.001$). We further break down these comparisons by preference domain (e.g., risk-taking shows the lowest average between-model MSE at 0.770, while trust shows the highest at 1.856). These MSE-based comparisons are presented in the new [Figure 3](#) (reproduced below for your convenience) and discussed in the revised Results section.

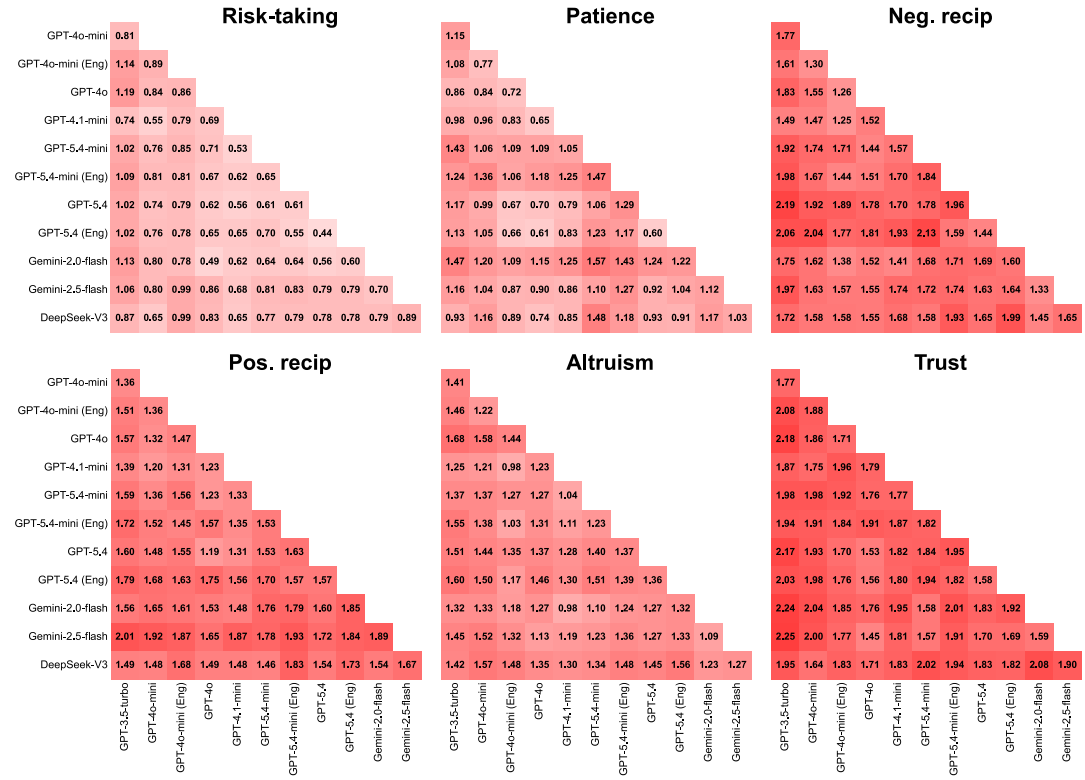
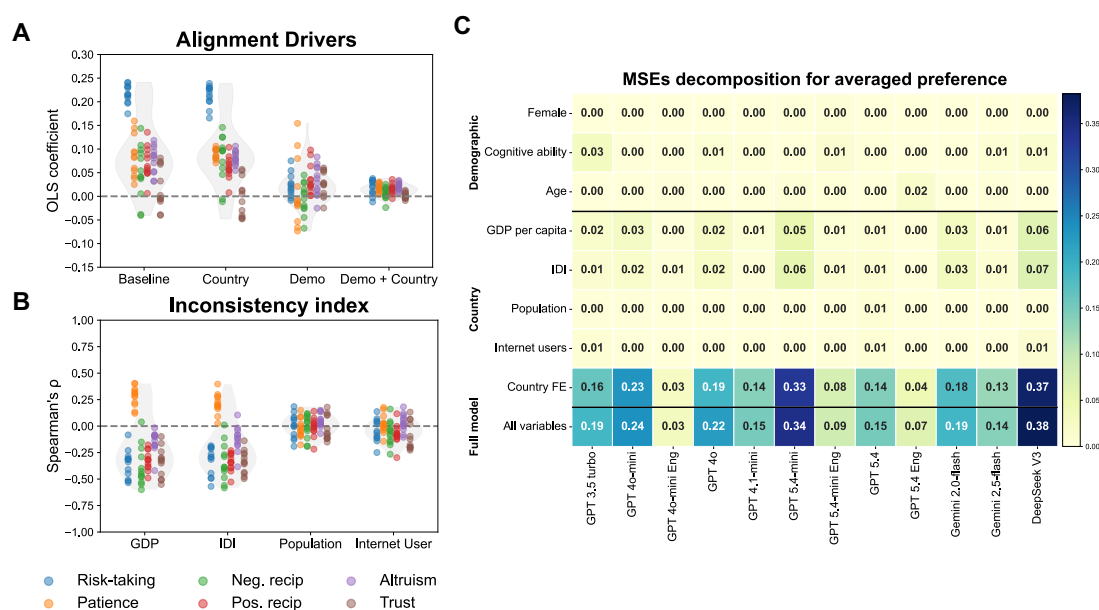


Figure 3. Individual-level between-model MSEs of LLM-predicted preferences. Each cell of the 12×12 matrix reports the MSE between two LLM configurations, computed by pooling individuals across all 74 countries. Individual-level preference indices are standardized to mean 0 and standard deviation 1 prior to MSE computation. Color intensity encodes MSE magnitude, with lower values indicating greater consistency between models. The corresponding country-level between-model MSE matrix is reported in [Appendix Figure C.3](#).

12. Analysis 4: After characterizing the overall patterns, analyze whether the errors follow systematic patterns. For example, predict the absolute difference between human and silicon responses as a function of individual demographics, country characteristics, model type, and other relevant factors. This decomposes the sources of error. The analysis could be conducted separately for each question, which is probably preferable given the variation across items, or pooled with question fixed effects.



inconsistency index (the GPS sampling-weighted within-country mean of individual-level MSEs) and four country characteristics: log GDP per capita, the Information and Communications Technology Development Index (IDI), log population, and log number of internet users. Each point represents a Spearman correlation coefficient computed across 74 countries. Each characteristic shows 72 points (12 LLMs $\times$ 6 preferences). **(C)** Error decomposition of LLM-predicted preferences. For each LLM (column), separate OLS regressions are estimated with the average individual-level MSE across six preferences as dependent variable and each variable category (row) as independent variable. Cell values report the R^2 of each single-predictor regression.

Specifically, we provide three complementary decompositions in the main text:

- **Country-level error modeling** (Figure 6B). We examine the relationship between the country-level inconsistency index (GPS sampling-weighted mean of individual-level mean squared errors, computed per country) and four country characteristics (log GDP per capita, IDI, log population, log number of internet users). The index is negatively correlated with GDP (59/72 model–preference combinations, 81.9%) and IDI (58/72, 80.6%), but essentially uncorrelated with population size or internet users, identifying *where* predictions fail: in lower-income and less digitally developed countries. The regression evidence is provided in Appendix E.
- **Individual-level error decomposition** (Figure 6C). We regress the individual-level MSE (averaged across the six preference domains) on three predictor blocks, demographic characteristics, country-level continuous covariates, and country fixed effects, and report the R^2 attributable to each variable (indicated by the row’s name), separately for each model. Country fixed effects explain by far the largest share of error variance (3%–37% across models), substantially more than demographic information or country-level continuous covariates (up to nearly 7%, from DeepSeek-V3). This identifies why predictions fail: errors concentrate at the country level rather than the demographic level, indicating systematic insensitivity to country context.
- **Per-domain breakdown.** Following your advice for analyses with respect to preference-level variation, the above individual-level error decomposition results

reported are based on averaged analyses across six preferences. Per-domain decompositions of country-characteristic effects on prediction errors are reported in [Figure 6B](#) and [Appendix Figures E.1–E.6](#).

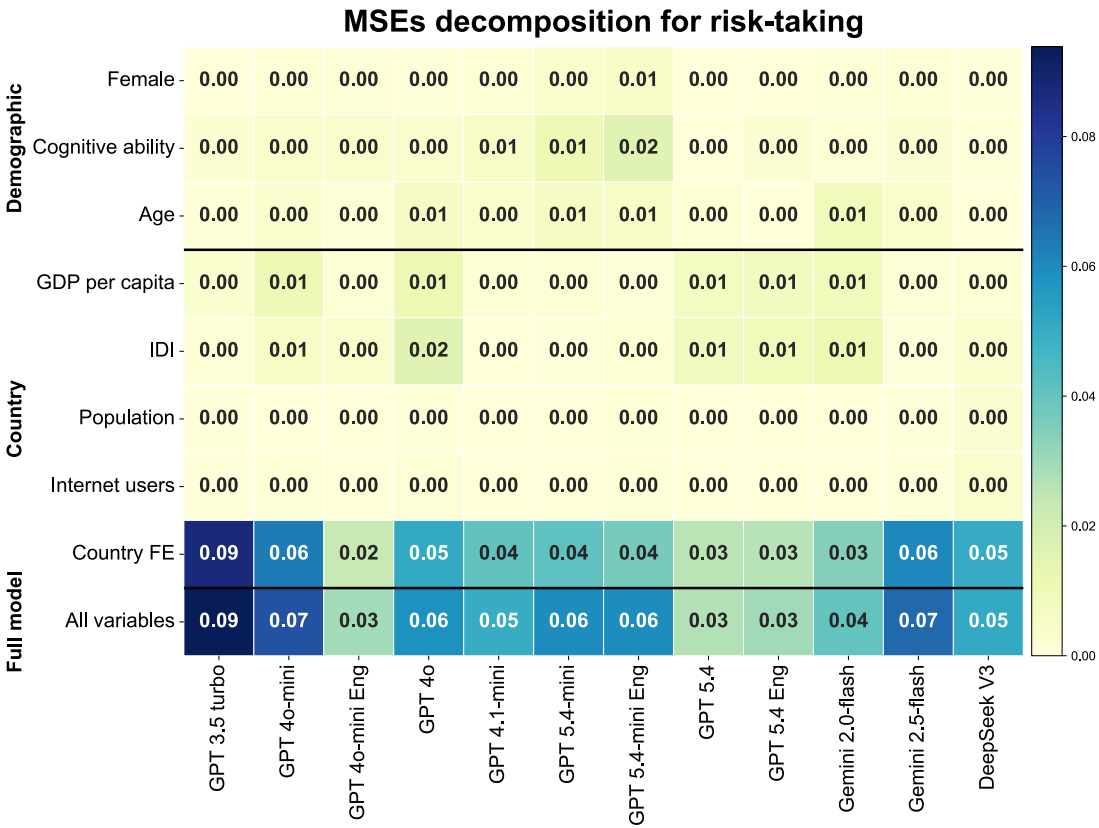


Figure E.1. Error decomposition of LLM-predicted preferences (risk-taking). The number in each cell represents the R^2 estimated from a regression where the individual-level MSE for risk-taking is the dependent variable and corresponding variables indicated by the row label are the independent variables.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

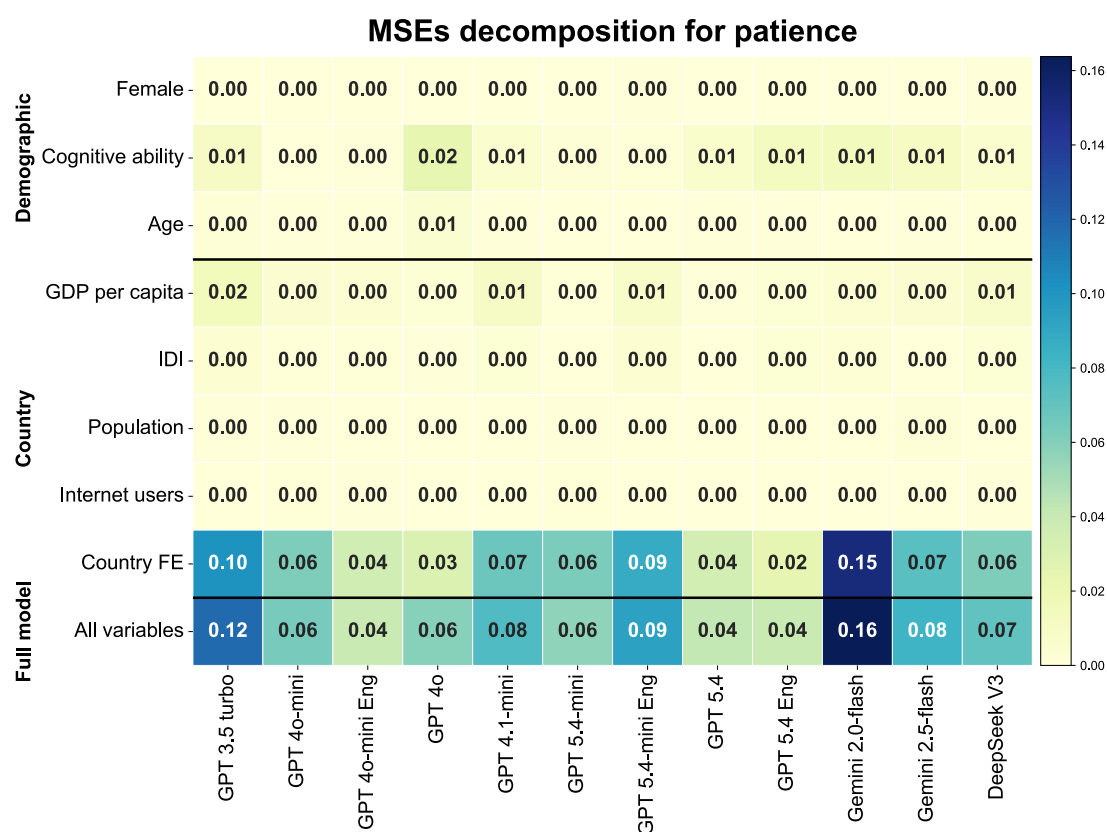


Figure E.2. Error decomposition of LLM-predicted preferences (patience). The number in each cell represents the R^2 estimated from a regression where the individual-level MSE for patience is the dependent variable and corresponding variables indicated by the row label are the independent variables. Columns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

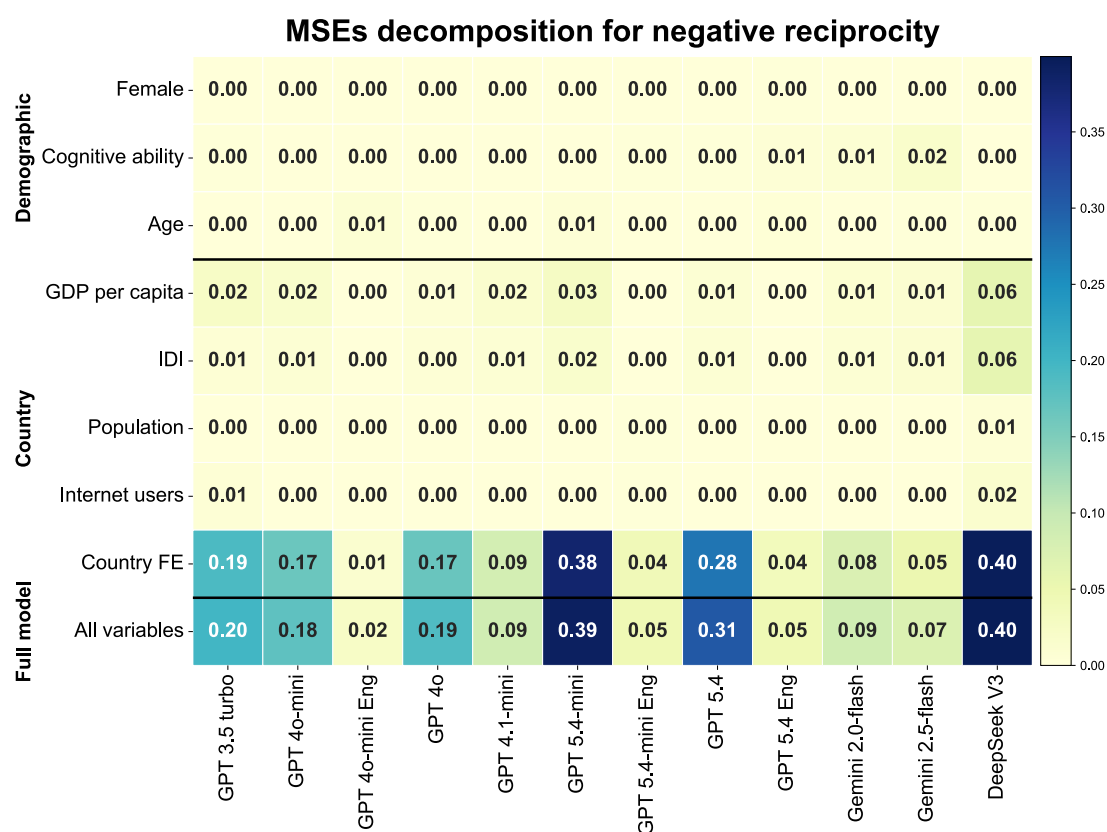


Figure E.3. Error decomposition of LLM-predicted preferences (negative reciprocity). The number in each cell represents the R^2 estimated from a regression where the individual-level MSE for negative reciprocity is the dependent variable and corresponding variables indicated by the row label are the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

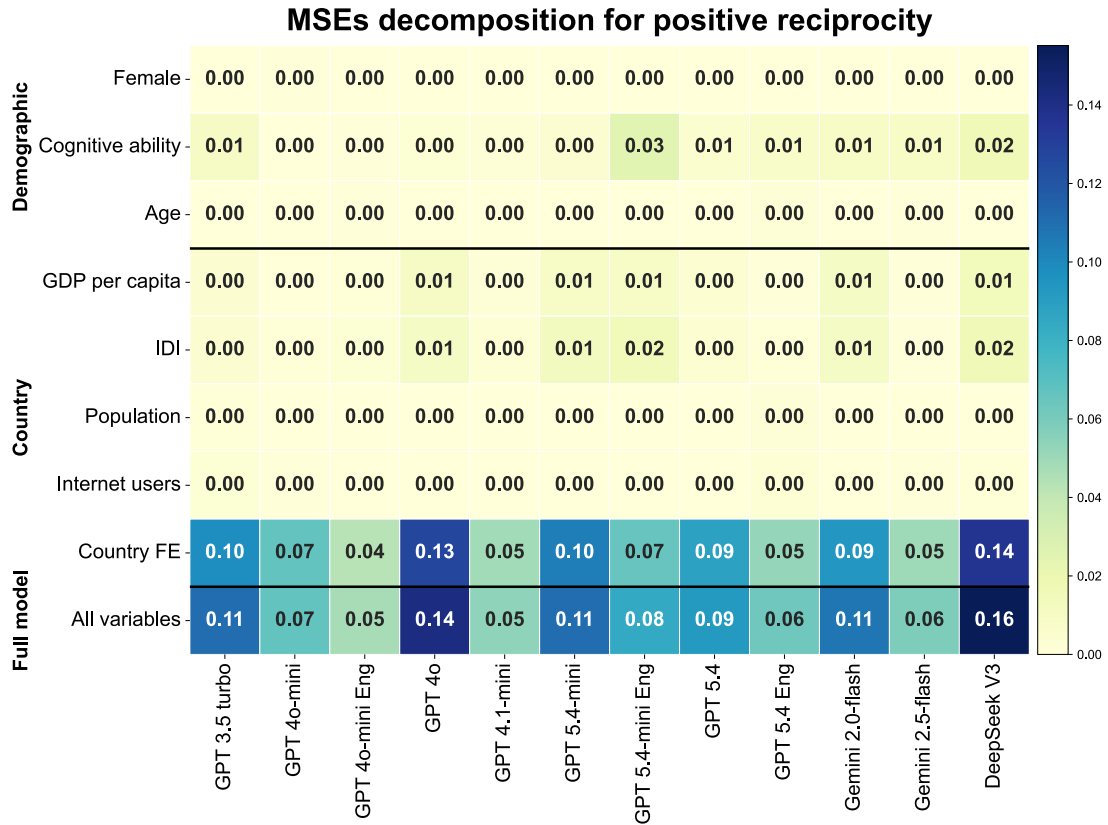


Figure E.4. Error decomposition of LLM-predicted preferences (positive reciprocity). The number in each cell represents the R^2 estimated from a regression where the individual-level MSE for positive reciprocity is the dependent variable and corresponding variables indicated by the row label are the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

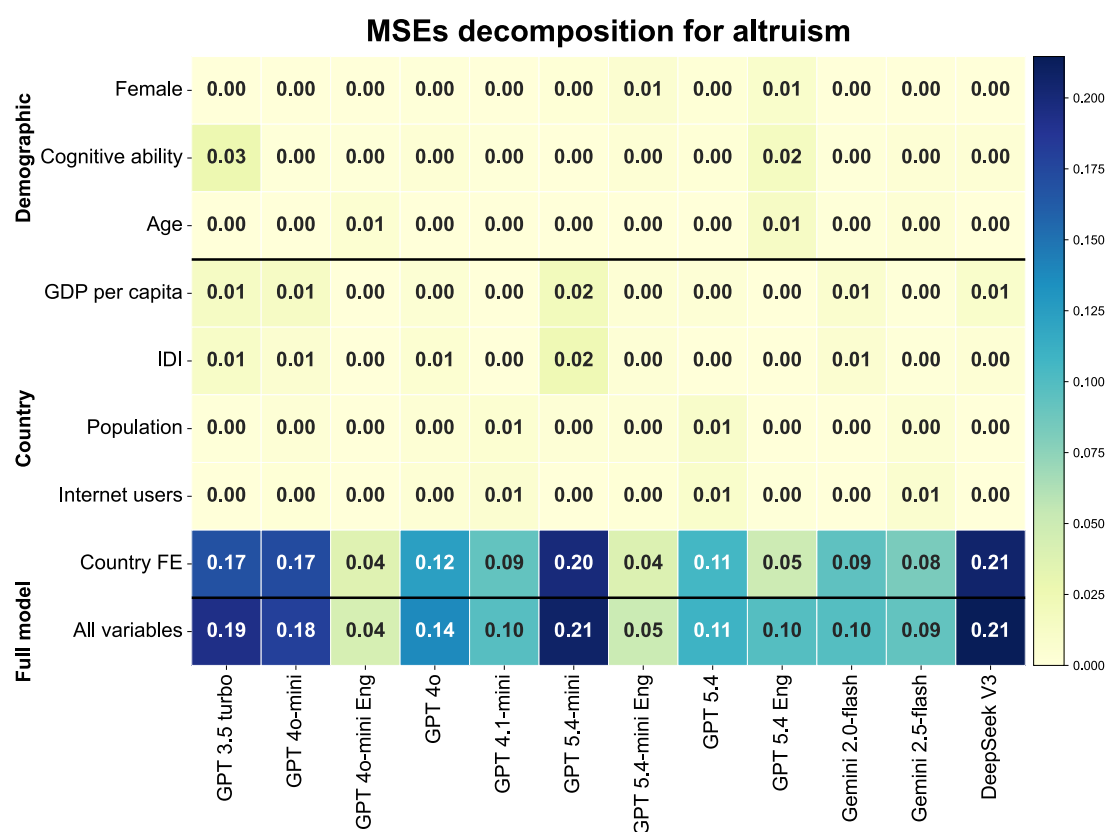


Figure E.5. Error decomposition of LLM-predicted preferences (altruism). The number in each cell represents the R^2 estimated from a regression where the individual-level MSE for altruism is the dependent variable and corresponding variables indicated by the row label are the independent variables. Columns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

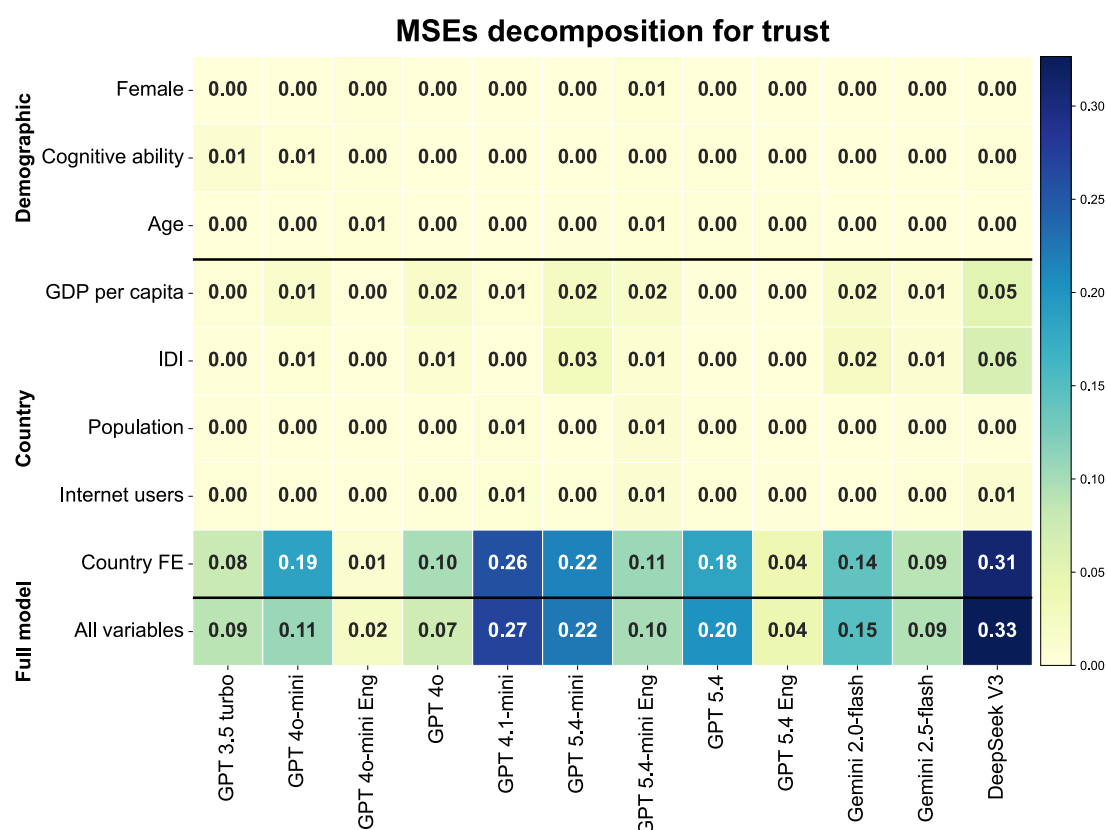


Figure E.6. Error decomposition of LLM-predicted preferences (trust). The number in each cell represents the R^2 estimated from a regression where the individual-level MSE for trust is the dependent variable and corresponding variables indicated by the row label are the independent variables. Columns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

13. At the end of the day, I have very high confidence that the substantive conclusion is correct, but I am not convinced that the current paper provides the best demonstration of that claim. That said, the authors have assembled a creative framework that could be used to evaluate this question directly, but doing so would require a different set of analyses than those currently reported.

Response: We are encouraged by your confidence in our substantive conclusion and your recognition of the framework’s potential. Following your recommendations, we have substantially restructured the analytical strategy in the revised manuscript.

As described in our responses above, we have (1) adopted MSE as the primary accuracy metric (Analysis 1; see our response to your Comments 2 and 9), (2) added systematic

dispersion comparisons to assess response homogenization (Analysis 2; see our response to your Comments 3 and 10), (3) implemented MSE-based cross-model comparisons to examine how prediction accuracy varies across models and contexts (Analysis 3; see our response to your Comments 5 and 11), and (4) conducted error decomposition analyses to identify systematic patterns in prediction errors as a function of demographics, country characteristics, and model type (Analysis 4; see our response to your Comments 4, 7, and 12).

We believe these revisions now provide the rigorous, metric-appropriate demonstration that you envisioned, while preserving the creative framework that you recognized as valuable.

Reviewer 1's Comments

1. I thank the authors for a careful and substantial revision. The revised manuscript is much stronger than the previous version. In particular, the shift to MSE as the primary accuracy metric directly addresses the concern that correlations can be misleading for evaluating matched synthetic respondents. The new analyses also make the empirical message clearer: LLM-predicted preferences are not random, but the alignment is modest; models agree with one another more than they agree with the human benchmark; demographic gradients are partly recovered but often amplified; and country- and culture-level patterns remain poorly captured.

The added GPT-5.4 and GPT-5.4-mini experiments, prompt-component ablations, error decompositions, paraphrasing and response-format checks, and diversity analyses collectively provide a much more convincing audit of the use of LLMs as synthetic respondents. I especially appreciate the response to the earlier request for direct prediction-error analysis and the new evidence that response diversity is substantially compressed relative to human responses. These additions clarify the boundary conditions of the approach and make the manuscript more useful for the broad behavioral-science audience of *Nature Human Behaviour*.

My remaining comments are minor, but I think they should be addressed before publication.

Response: We sincerely thank the reviewer for the careful evaluation of our manuscript and for this encouraging assessment of the revision. We are glad that the reviewer found the shift to MSE, the additional model experiments, the prompt-component ablations, the error decompositions, and the diversity and robustness analyses to have strengthened the audit and clarified the boundary conditions of the approach. The reviewer's earlier comments were very helpful in sharpening the empirical contribution of our study.

We have addressed each of the remaining comments in detail below.

2. Please temper the mechanistic and causal wording in several places. The ablation results are informative, but they are conducted primarily with GPT-5.4-mini and under one persona-prompting design. Similarly, the country-level error decompositions are observational. Phrases such as demographic information "drives" alignment, demographic information "accounts for most" of alignment, country labels "contribute little," and LLMs "lack sufficient cultural sensitivity" are directionally plausible but a little too strong. I recommend phrasing these as diagnostic evidence under the present prompting protocol: for example, "in these ablations, removing demographic information reduces alignment more than removing country information," or "the results are consistent with the interpretation that the models condition more effectively on demographic cues than on country labels." This would preserve the conclusion while avoiding overclaiming mechanism.

Response: We agree that the ablation analyses should be presented as diagnostic evidence under the present persona-prompting protocol, and that the country-level error decompositions are observational rather than causal. We have therefore tempered the relevant wording throughout the manuscript, including the Abstract, Results, Discussion, and Supplementary Information.

- For the observational analysis, the revised manuscript now reads:

"These results indicate that demographic information in the prompts accounts for more of the individual-level alignment than country labels do." and "These results are consistent with the interpretation that models condition more effectively on demographic cues than on country labels". We further summarize that, "Together, these analyses indicate that, compared with country label, demographic information in prompt accounts for more of the individual-level alignment, whereas the prediction errors that remain are concentrated along country lines and track indicators of economic development rather than corpus-size proxies."

- For the ablation analysis, we now report the results as follows:

“These results are consistent with the interpretation that models condition more effectively on demographic cues than on country labels (see Supplementary Figures F.3–F.9).”

- In the Discussion, we similarly describe the ablation experiments as providing *“diagnostic evidence that demographic information in prompts contributes more to predictive accuracy than country information.”* We have also removed the broader statement that LLMs “lack sufficient cultural sensitivity” from the main text and supplementary information.
- In the Abstract, the corresponding sentence now reads: *“In ablation and error decompositions, removing demographic information reduces alignment more than removing country labels, and performance is weakest in economically undeveloped and less digitally connected countries.”* In the Supplementary Information, we likewise revised Supplementary Information E.4 (*“demographic information accounts for more of the individual-level alignment than country information does”*) and Supplementary Information F.2 (*“If alignment relies primarily on demographic information, ...”*).

These revisions temper the mechanistic and causal interpretation while retaining the empirical pattern observed under the present design.

3. The contamination and memorization results should also be stated more cautiously. The paraphrasing and English-only checks are useful and make it unlikely that memorization is the primary source of the reported alignment. However, null differences under these perturbations do not show that memorization is not a driver at all. In the supplement, the statement that memorization “is not a driver of alignment” should be softened to “we find no evidence that memorization is the primary driver of alignment” or similar. This is especially important because the manuscript itself rightly notes that GPS material may have appeared in training data and that the resulting alignment estimates may be an upper bound.

Response: We agree that the paraphrasing and English-only experiments cannot rule out memorization entirely, and that null differences under these perturbations do not show that memorization plays no role. We have therefore softened the statement in the Supplementary Information, which now reads:

“Taking these two experiments’ results into account, we find no evidence that memorization (or data contamination) is the primary driver of alignment, either at the individual or at the country level.”

The main text likewise presents these checks in evidential terms:

“If memorization primarily drove alignment, both manipulations should weaken the match with training data. We found no evidence that either manipulation systematically changed individual- or country-level alignment (see Supplementary Tables G.1–G.5).”

In addition, the Discussion retains the qualification that portions of the GPS content may appear in LLM training data, that successful predictions could therefore partially reflect memorization, and that our alignment estimates should be interpreted as an upper bound.

4. Please include the GPT-4o-mini zero-response robustness information in the manuscript or supplement, not only in the response letter. The supplement notes that an initial programming error treated "0" responses as missing and re-collected them until a non-zero value was obtained. The response letter states that the affected local-language GPT-4o-mini condition was retained but that excluding it does not change the headline analyses. This is reassuring, but readers should have access to the sensitivity check in the published supplementary material. A short table or paragraph reporting the exclusion check would be sufficient.

Response: We agree that readers should have access to this robustness information in the published supplementary material. We have therefore added to Supplementary Information A, immediately following the description of the initial programming error, a

paragraph together with a new Supplementary Table A.4 that quantifies the potential influence of this error on our data, as we reported in the previous response letter. We report the proportion of zero responses for each model and each affected survey item in the corrected data. The newly added paragraph reads:

“In our analysis, we retained the data influenced by the programming error, since the ‘0’ responses account for only a very small proportion of the dataset. [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 proportion of zero responses exceeding 1%. For every other model, the proportion of zero responses remains below 1%. Notably, GPT-4o-mini (English), which shares the same base model as the affected GPT-4o-mini condition, returns zero in 0% of cases. These patterns indicate that the GPT-4o-mini model rarely outputs zero on these items. This provides an upper bound on how much the resampling could influence our data.”

Table A.4. Percentage 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).

5. Please clarify the scope of the prompt-ablation evidence. The ablations are valuable, but they are not a general test of all possible ways to encode country or culture in LLM prompts. A country label alone may be a weak operationalization of cultural context, and stronger country/culture prompts, retrieval-augmented prompts, or explicit cultural calibration could plausibly behave differently. I do not think the authors need to run these additional analyses, but the conclusion should be explicitly limited to the current persona-prompting design.

Response: We fully agree that the ablation experiments do not provide a general test of all possible ways of encoding country or cultural context, and that stronger cultural prompts, retrieval-augmented prompting, or explicit cultural calibration could plausibly behave differently. We have revised the Discussion to explicitly limit the interpretation of the ablation evidence to the present persona-prompting design. The revised Discussion now reads:

"...We note that our ablation experiments are conducted primarily with GPT-5.4-mini under a single persona-prompting design and therefore do not constitute a general test of all possible ways of encoding country or cultural context, such as richer cultural descriptions, retrieval-augmented prompts, or explicit cultural calibration. A country label alone may be a weak operationalization of cultural context, and stronger operationalization might steer LLMs to behave differently. Prior work shows that cultural- and country-level alignment can be improved by prompt injection (Pataranutaporn et al., 2025; Tao et al., 2024) and fine-tuning (Li et al., 2024; Suh et al., 2025), suggesting one promising route to addressing misalignment associated with country fixed effects."

6. Some claims comparing newer and older models should avoid overinterpreting null tests. The addition of GPT-5.4 and GPT-5.4-mini is useful, and the absence of clear improvement is interesting. Still, "newer models do not appear to achieve stronger alignment" is safer than implying that newer models cannot improve. The manuscript

should consistently frame these results as evidence within the sampled models, collection windows, prompting protocol, and preference task.

Response: We have revised the relevant statement in the main text:

“Turning to model comparisons, we find no evidence that newer models like GPT-5.4 and GPT-5.4-mini outperform earlier OpenAI models ... or that Gemini-2.5 Flash outperforms Gemini-2.0 Flash.”

The Supplementary Information uses the same evidential formulation: *“Thus, we find no evidence that advanced models achieve stronger individual-level alignment on economic preferences.”* These statements describe the absence of detected improvement among the models examined in this study and in the present economic preference context, and they do not imply that newer or future models cannot achieve stronger alignment.

7. The code and data availability statement is appreciated. To ensure full reproducibility, please clarify that the deposited materials contain the exact prompts, model identifiers, data collection dates, temperature parameters, parsing and extraction workflows, and all scripts required to replicate every headline figure. More importantly, I was unable to locate the associated code and data via the current OSF link.

Response: We carefully rechecked the OSF deposit and confirmed that all files have been uploaded correctly and are accessible via the link provided in the Data and Code Availability statements. We apologize for the difficulty the reviewer experienced; we have verified from multiple devices and networks that the link now resolves correctly to the project page, where all materials can be viewed and downloaded. We also confirm that the deposited materials contain everything required for full reproducibility.

Reviewer 2's Comments

1. I was Reviewer 2 in the previous round. I have read the authors' responses and revisions in detail. In my view, the authors have undertaken a serious major revision that makes the paper substantially stronger. The revision moves the contribution from a mostly descriptive audit (that is, that LLMs correlate weakly with human economic preferences) toward a more diagnostic claim: LLMs recover some demographic-level structure from prompting, but fail more seriously at country- and cultural-level inference. This pattern appears robust across models and is not obviously explained by prompt format or simple memorization.

Response: We sincerely thank the reviewer for the careful reading of our responses and revisions, and for this positive assessment. We are glad that the reviewer finds the revision has moved the contribution from a primarily descriptive audit toward a more diagnostic claim, that LLMs recover some demographic-level structure from prompting but fail more seriously at country- and cultural-level inference, and that this pattern appears robust across models and is not obviously explained by prompt format or simple memorization. The reviewer's comments in the previous round played a key role in shaping these analyses, and we address the remaining comments point by point below.

2. The main remaining limitation is that the proposed "mechanism" is still not mechanistic in the stronger interpretability sense. It does not concern internal representations, training-data attribution, or latent computational mechanisms. Rather, it is a prompt-level result: which persona fields matter for observed alignment. Precisely because these deeper mechanisms are not identified, the paper offers relatively few actionable recommendations for LLM researchers who may be interested in how to train, fine-tune, or steer LLMs toward better individual-level alignment.

Relatedly, the prompt-level implications for designing AI agents or digital twins seem somewhat pessimistic. The results suggest that even emphasizing demographic-level structure does not lead to substantially better individual-level alignment. A reader

interested in designing better digital twins may therefore leave with a largely negative conclusion about how to improve prompts, rather than with a clear set of constructive design principles.

Response: We agree that our analyses do not identify mechanisms in the stronger interpretability sense: we provide no direct evidence on internal representations, training-data attribution, or latent computational processes. What the prompt-component ablations offer is a prompt-level diagnosis, evidence on which persona fields matter for observed alignment under the present design, and we have been careful to frame it as such throughout the revised manuscript (e.g., the Results section "Diagnostic evidence from ablation experiments").

Regarding actionable recommendations for LLM researchers, we agree that we do not directly test how models should be trained, fine-tuned, or steered. However, our analyses identify a concrete target for such efforts: the error decomposition shows that the remaining prediction errors are systematic and country-structured, concentrated in country fixed effects that track economic development rather than corpus-size proxies. The Discussion now points to two corresponding routes: *"Prior work shows that cultural- and country-level alignment can be improved by prompt injection (Pataranutaporn et al., 2025; Tao et al., 2024) and fine-tuning (Li et al., 2024; Suh et al., 2025), suggesting one promising route to addressing misalignment associated with country fixed effects. The second is to probe LLMs' internal representations and apply mechanistic interventions such as steering, to improve alignment (Khanuja et al., 2026)."*

We also appreciate the concern that readers interested in designing digital twins may draw a largely negative conclusion. We have clarified in the Discussion that the implications are not uniformly negative: *"Despite these constraints, our individual- and demographic-level alignment results suggest several possible scientific uses of LLM-simulated responses, although these possibilities require further empirical validation. First, LLMs may be used to propose and probe latent motivational structures underlying human behavior, particularly in WEIRD populations, complementing formal decision theory^{5,84–88}. Second, LLMs may accelerate early-stage hypothesis generation and study-design*

diagnostics...” When the research objective is to examine demographic-level patterns rather than to achieve country-level fidelity, the persona-prompting approach evaluated here may still serve as a useful experimental tool, consistent with recent work showing that LLM-simulated data can recover the results of social science experiments (Ashokkumar et al., 2026).

3. Overall, I think the authors have responded seriously and constructively to the previous concerns, and the revised manuscript is substantially improved. My remaining concern is mainly about the scope and framing of the contribution: the paper is stronger as an audit and diagnostic study of current LLM synthetic respondents than as a mechanistic account or practical guide for improving them. I leave it to the editorial team to evaluate how important this type of contribution is for Nature Human Behaviour’s interdisciplinary readership.

Response: We agree that the manuscript’s strongest contribution is as a global audit and prompt-level diagnostic study of current LLMs as synthetic respondents, rather than as a mechanistic account or a comprehensive practical guide to improving them.

We have revised the framing throughout the manuscript to make this scope explicit. The Introduction now describes the study as “*a global, multi-level audit of LLM-generated synthetic data across individuals, demographic groups, countries, and cultures,*” and states that our contribution is to “*clarify the conditions under which LLM-generated responses*” align with human data.

The Discussion reframes the contribution in diagnostic terms: we “*compare LLM-generated responses with human data from the Global Preference Survey (GPS) across multiple levels...and thereby clarify the boundary conditions of their scientific utility,*” and the ablation results are presented under the heading “*Diagnostic evidence from ablation experiments*” rather than as a mechanistic account.

In addition, for readers seeking practical guidance on designing and using LLMs as synthetic respondents, we now cite a recently published reference in Nature Human

Behaviour which offers detailed methodological recommendations for such applications (Feuerriegel et al., 2026).

We hope this framing accurately conveys the nature of the contribution to Nature Human Behaviour's interdisciplinary readership.

Reviewer 3's Comments

1. Thank you for your responsive review. Using MSE rather than correlations makes the paper much more convincing and I am happy to support its' publication. Congratulations.

Response: We sincerely thank the reviewer for the constructive suggestions in the previous round and for this encouraging assessment of our revision. The reviewer's recommendation to move beyond correlations and examine prediction errors directly motivated us to adopt MSE as the primary metric, which we agree has made the evaluation more informative. We are grateful for the reviewer's support for publication.

2. One final suggestion is that it might be worth citing a recent chapter from an upcoming volume on Public Opinion and AI by the American Political Science Association to help further contextualizing the contribution of the paper alongside a larger set of considerations about the use of AI in the study of public opinion and the implications of replacing human opinion with synthetic opinion.

Clinton, Joshua et al. 2026. "Public Opinion in the Age of AI." APSA Preprints. doi: [10.33774/apsa-2026-1rg21-v3](https://doi.org/10.33774/apsa-2026-1rg21-v3).

Response: We thank the reviewer for pointing us to this timely work. We agree that research at the intersection of public opinion and AI is developing rapidly and that our evaluation of LLMs as synthetic respondents speaks directly to the broader considerations raised in that volume about replacing human opinion with synthetic opinion. We now cite Clinton et al. (2026) in the Introduction, where we discuss concerns about using LLM-simulated responses in place of human data (ref. 20).

Reviewer's report for "LLMs as Synthetic Respondents for Economic Preferences: A Global Audit"

This manuscript presents a large-scale "virtual replication" of the Global Preference Survey (GPS), which measured economic preferences (risk-taking, patience, reciprocity, altruism, and trust) in nationally representative samples across 74 countries and 42 languages, alongside demographic covariates (e.g., age, gender, region/country, and subjective cognitive skills). The authors construct LLM "personas" by conditioning prompts on these demographic profiles (via system prompts) and administer the GPS items to multiple LLMs using deterministic decoding (temperature = 0). They evaluate correspondence between LLM-generated and human responses at multiple levels of aggregation.

The main empirical conclusion is that LLM responses show statistically detectable but generally weak alignment with human preferences at the individual level. Performance varies across domains (risk-taking is typically better reproduced than trust), across model families, and across sources of heterogeneity. Synthetic responses also appear more responsive to demographic cues (gender, age, cognitive ability) than to country or cultural context, suggesting that model "preference" variation may be driven disproportionately by demographic regularities rather than country-specific structure. The overall message is appropriately cautious: while LLMs can generate plausible survey responses conditioned on demographics, the variability and limited individual-level fidelity raise concerns about using such models as substitutes for human respondents, especially for cross-cultural inference.

I find the behavioral evaluation comprehensive and the dataset-scale benchmarking valuable. The paper offers a useful snapshot of how current LLMs behave under a straightforward persona-conditioning protocol. That said, I see two limitations that reduce the manuscript's scientific contribution in its current form.

First, the analysis is almost entirely behavioral. While this is an essential starting point, the paper does not leverage what makes LLMs unique as "synthetic respondents": the possibility of more diagnostic, mechanism-oriented analyses. In particular, the study stops short of clarifying why some domains (e.g., trust) fail, why demographic gradients are amplified, or why country cues appear weak. For open-weight models, the authors could go beyond outcome comparisons and add diagnostic analyses (e.g., prompt sensitivity, representational similarity, or other probes tied to internal activations). Without at least one such analysis, the work reads more as an evaluation benchmark than as an explanation of the sources of alignment and misalignment.

Second, the behavioral results are likely to be model- and time-dependent. Given rapid model updates and deployment changes, especially for closed-source systems, some findings may only hold for specific model snapshots and collection windows. API-mediated behavior can also vary across providers and deployments (e.g., hidden updates or serving optimizations), complicating cross-model comparisons and the durability of reported effect sizes. The paper would benefit from clearer framing and/or robustness checks separating conclusions likely to

generalize (e.g., “demographic gradients > country effects”) from those that are likely contingent on the specific models and time periods studied.

Overall, I view this as a valuable benchmark documenting current limits of demographic-conditioned LLM personas for simulating economic preferences at global scale. However, for an interdisciplinary venue such as *Nature Human Behaviour*, the present version may be limited in conceptual depth. A stronger contribution would either (i) add at least one mechanistic/diagnostic analysis to explain the observed failures and successes, or (ii) strengthen claims about generalization of the behavioral results via synchronized sampling, design variations (e.g., stricter response-format constraints), or targeted robustness tests.

Detailed suggestions:

1. Add diagnostic analyses beyond regressions.

At present, even the “sources of misalignment” section relies mainly on regressions over LLM responses. If neural-level analysis is feasible for open-weight models (e.g., DeepSeek), the authors could probe internal representations to test how persona attributes and country labels are encoded and which components causally drive responses. If this is not feasible, a simpler but still informative alternative is prompt-component ablations (system persona vs country label vs individual demographic fields), combined with sensitivity analyses to quantify which prompt elements drive preference shifts beyond correlational attribution.

2. Engage the data-contamination issue more directly.

Because GPS dates to 2012, it is plausible that some items, weights, or summaries entered LLM training corpora. Data contamination would typically bias toward higher apparent alignment, yet the paper reports limited alignment. This tension is worth addressing with at least one additional analysis. For example, the authors could test memorization vs inference by paraphrasing or perturbing item wording and checking whether inferred preferences are stable or brittle. Relatedly, it would be valuable to discuss (even conceptually, if new experiments are infeasible) what interventions might improve alignment (e.g., targeted finetuning or calibration) since practitioners often care less about whether off-the-shelf models align and more about whether alignment can be improved in a controlled way.

Minor comments:

Appendix Line 54-56 (GPT-4o-mini programming error): The manuscript notes a bug where “0” responses were treated as missing and resampled until non-zero, and that this was corrected in later collection rounds. Please clarify whether the buggy data remain in the analyses. If so, a simple robustness check excluding the affected GPT-4o-mini subset (or excluding the affected items) would increase confidence that headline conclusions are not driven by mixed “buggy + corrected” data. If recollection is infeasible, explicitly reporting sensitivity would help.

Clarify use-cases: substitution vs hypothesis generation. The paper is appropriately skeptical about using LLMs as substitutes for survey inference, while also suggesting they may be useful for hypothesis generation or scenario testing. It would help to sharpen this distinction and avoid over-generalizing from a single operationalization. Even one concrete example illustrating when/why the hypothesis-generation use-case remains viable (despite weak macro/cultural validity) would strengthen the narrative.