Whose Values Prevail? Bias in Large Language Model Value Alignment

Ruoxi Qi[#]
(ruoxiqi@connect.hku.hk)
Department of Psychology,
University of Hong Kong
Pokfulam Road, Hong Kong

Gleb Papyshev[#]
(gleb@ust.hk)
Division of Social Science, Hong Kong University of Science & Technology
Clear Water Bay, Hong Kong

Kellee Tsai
(k.tsai@northeastern.edu)
Department of Political Science, Northeastern University
Boston, Massachusetts, United States

Antoni B. Chan
(abchan@cityu.edu.hk)
Department of Computer Science, City University of Hong Kong
Kowloon Tong, Hong Kong

Janet H. Hsiao
(jhhsiao@ust.hk)
Division of Social Science and Department of Computer Science & Engineering, Hong Kong University of Science & Technology
Clear Water Bay, Hong Kong

Abstract

As large language models (LLMs) are increasingly integrated into our lives, concerns have been raised about whether they are biased towards the values of particular cultures. We show that while LLMs were biased toward the values of WEIRD populations, some non-Western populations, including East Asia and Russia, were also represented relatively well. Notably, the Rich dimension was the strongest predictor of LLM's alignment instead of the most discussed Western dimension. This suggests the need to attend to less prosperous populations instead of focusing only on easily accessible populations. We also found that one source of this bias could be unbalanced training data as approximated by an Internet Freedom measure, and that prompting the model to act as individuals from different populations reduced the bias but could not eliminate it. These findings raise the importance of training process disclosure and the consideration of culture-specific models to ensure ethical usage of LLMs.

Keywords: Large Language Model (LLM), Value Alignment, WEIRD Population

Introduction

Recently, large language models (LLMs) are increasingly integrated into users' daily life and are being applied to more and more diverse tasks, including critical tasks such as medical diagnosis (Biesheuvel et al., 2024) and legal judgement (Lai et al., 2024). At the same time, there is a growing concern about potential risks if the model generates harmful output when interacting with its users or makes biased decisions in a critical system. In particular, LLMs are black boxes whose inner mechanisms are not transparent, thus making it difficult to predict when and why they would exhibit unacceptable behaviors. To address these issues, recent research has started to focus on value alignment, which aims to ensure that LLMs consistently adhere to human values (Ji et al., 2023). Various techniques can be used to align LLMs with human values, from reinforcement learning and supervised fine-tuning using human preferences (Ouyang et al., 2022) to in-context alignment that directly prompts the model to generate value-aligned content (Ganguli et al., 2023). Benchmarks have also been developed to evaluate whether LLMs indeed follow human values, including benchmarks for specific issues such as social bias (e.g., Kocielnik et al., 2023) and benchmarks that assess LLMs' inherent values more generally (e.g., Duan et al., 2024).

However, there are substantial cross-cultural differences in humans' values and judgements about the inappropriateness of LLMs' output (Davani et al., 2024). While it is generally agreed that LLMs should be aligned with human values, the question that remains is, whose values should LLMs adhere to? Currently, various studies have found that LLMs are biased toward the values of certain groups (e.g. Rozado, 2023; Santurkar et al., 2023). Specifically, it has been suggested that LLMs may be biased toward the values of Western, Educated, Industrialized, Rich, and Democratic (WEIRD) populations (Benkler et al., 2023; Atari et al., 2023), similar to how psychological theories can be biased by over-reliance on WEIRD samples (Henrich et al., 2010). Others have argued that rather than being biased toward certain groups, LLMs may have their own idiosyncratic response patterns that do not align with any groups (Boelaert et al., 2024).

The current study aimed to systematically investigate whether LLMs exhibit a bias toward WEIRD values. We used the World Values Survey (WVS) dataset (Haerpfer et al., 2022), which contains questions that assess different aspects of human values and data from large, nationally representative samples of different countries/territories. We tested an LLM on a subset of these questions and measured the distance between its responses and the responses of different human populations for comparison. We separately quantified each of the five dimensions of WEIRD-ness rather than relying on a binary distinction between WEIRD and non-WEIRD, borrowing the approaches from two studies that examined sample representativeness of user studies (Linxen et al., 2021; Septiandri et al., 2024). who coined the term WEIRD, has argued that researchers should not attempt to decompose each letter of WEIRD, given that it is a backronym created as a conscious-raising rhetorical tool

[#] These authors contributed equally and are co-first authors.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society* ©2025 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

without solid theoretical grounding. Without assuming theoretical importance behind each letter, here we tested which components of this well-known backronym actively contributed to the biases observed. Currently, the term WEIRD is often used as a synonym of Western. As a result, researchers often treat WEIRD-ness as a binary distinction and use the easily accessible East Asian populations as the non-WEIRD comparison group, despite many similarities between the two populations (Krys et al., 2024). Thus, identifying critical components of WEIRD-ness beyond Westernness can help researchers better recognize which populations are underrepresented.

In the case that such WEIRD bias indeed exists, it would be important to consider potential solutions to mitigate the bias. For instance, Atari et al. (2023) argued that this bias is likely due to WEIRD dominance in LLM's training data. Similar to the well-known aphorism of "garbage in, garbage out," where low-quality training data can lead to problematic output, biased data would also lead to biased output (i.e., WEIRD in, WEIRD out). Such biases in the training data cannot simply be reduced with increased model size or more training and would instead be further amplified (McKenzie et al., 2023). Therefore, researchers should focus more on reducing the bias in the training data if it is indeed the source of an LLM's WEIRD bias. Accordingly, the current study directly tested this hypothesis through mediation analyses, using internet usage data as proxies to assess the dominance of different countries/territories in the LLM's training data.

Another issue to consider is that if aligning with WEIRD values is undesirable, it would still be necessary to determine what kind of values an LLM should be aligned with instead. However, individuals can hold a variety of contrasting but reasonable beliefs about values (Rawls, 1999), thus making it difficult to decide on a single set of values that LLM should adhere to. Gabriel (2020) proposed some possible solutions, such as identifying values that are agreed upon globally or adding up individual views in a democratic manner, but these may not be the most optimal solutions given the current technologies, where LLMs can be steered toward different preferences (Kirk et al., 2024). While LLMs that went through a one-size-fits-all alignment process may have to take a certain set of values and thus be biased toward certain groups, steering the model to different values when interacting with different users can overcome this problem. Therefore, the current study also investigated an LLM's steerability by prompting it to simulate different identities and tested whether such prompting approach could reduce its bias toward WEIRD values.

LLM's Alignment with Different Populations and Its Steerability

We investigated whether the responses of GPT-3.5 Instruct[1] are better aligned with some populations than others, with the hypothesis that it may be biased toward the values of WEIRD populations. In addition, we tested whether alignment can be improved by prompting the LLM to simulate different identities (i.e., its steerability) and whether this approach can also reduce the bias toward WEIRD populations. Finally, we examined a potential source of this bias: dominance of WEIRD values in LLM's training data.

Methods

Materials To assess the values of different populations, we used the WVS Wave 7 dataset (Haerpfer et al., 2022), containing responses of nationally representative samples from 64 countries/territories collected from 2017 to 2022. We focused on the 47 questions in the Social Values, Attitudes, and Stereotypes bloc, since these questions capture general societal values across cultures.

For each country/territory, we quantified the different dimensions of WEIRD-ness using the same approach as previous studies (Linxen et al., 2021; Septiandri et al., 2024): (1) **Western**, binary distinction between Western and non-Western based on Huntington (2000)'s classification; (2) **Educated**, mean years of schooling based on UNDP Human Development Report (United Nations Development Programme, 2022); (3) **Industrialized**, Competitive Industrial Performance (CIP) Index from the United Nations Industrial Development Organization (UNIDO, 2022); (4) **Rich**, gross national income (GNI) per capita in thousands of international dollars (World Bank, 2021a); (5) **Democratic**, political rights ratings (Freedom House, 2024a).

An LLM's potential bias toward WEIRD values could be due to their dominance in the training data. However, while language dominance in the training data is available for some models, dominance for other demographic characteristics is more difficult to assess. Since LLM's training data tend to be mostly from internet users, here we used three metrics as proxies for estimating the dominance of different countries/territories in the training data: (1) *Internet Population*, which is the total number of internet users that are from a country (World Bank, 2021b, 2021c); (2) *Internet Percentage*, which measures the amount of internet users as a percentage of the country's population (World Bank, 2021c); (3) *Internet Freedom*, which includes ratings for obstacles to access, limits on content, and violations of user rights (Freedom House, 2024b). While Internet Population directly measures the number of individuals that can contribute to LLM's training data, Internet Percentage considers whether these individuals are representative of the country's population (i.e., less representative if they only form a small proportion). Internet Freedom further shows if the content produced by a country's users truly reflects their opinions or if it is subjected to censorship or content filtering, either by the state or by the technology companies.

[1] We used this older model since more recent models such as GPT-4 tended to output extreme probabilities that were difficult to normalize into percentages comparable to human response distributions and refused to respond more often, causing missing data that would pose a problem to the subsequent clustering analysis.

Procedures and Analyses GPT-3.5 Instruct was prompted to answer the questions in the Social Values, Attitudes, and Stereotypes bloc of the WVS Wave 7 Survey, preceded by a context prompt that either instructed it to act as *an average person living in a certain country* without specifying the country (**default response prompt**) or specified a country/territory included in the WVS dataset (**simulated identity prompt**; e.g., to act as *an average person living in Canada*). All questions were in binary or Likert scale format that could be responded with a number. We obtained the model's next-token log probabilities for the different options and normalized them into a response distribution, where the probabilities of all the options would add up to 1.

We measured how well GPT-3.5 Instruct's responses align with different countries/territories for each prompt type and each question by calculating the similarity between model and human responses. Specifically, we aggregated the human responses on each question for each country/territory to form a response distribution with percentages of participants choosing each option and computed the Wasserstein distance (WD) between the response distributions of the model and human data (Santurkar et al., 2023). WD measures the cost of transforming one distribution into another and can account for the ordinal structure of the options. A smaller WD between model and human responses indicated higher similarity and thus better value alignment. WD was then divided by the theoretical maximum (number of options – 1) to normalize across questions with different numbers of options. Average WD across questions was calculated for each prompt type and each country/territory as the measure of value alignment.

We investigated whether GPT-3.5 Instruct exhibited a bias toward WEIRD values by examining the relationship between the WEIRD scores of the countries/territories and the model's value alignment with the countries/territories as assessed using WD. Better alignment (lower WD) for countries/territories with higher WEIRD scores would indicate a bias toward WEIRD populations. We then compared the WD for the two prompt types to assess the model's steerability, where reduced WD (improved alignment) for the simulated identity prompt would indicate that the model is steerable. We also tested whether simulated identity prompting could reduce the WEIRD bias through moderation analyses, hypothesizing that the association between the WEIRD scores and alignment would be weaker when given the simulated identity prompt than the default response prompt. Finally, we investigated whether the source of the WEIRD bias was due to the dominance of WEIRD populations in the training data, using mediation analyses to test if the association between the WEIRD scores and alignment could be accounted for by the three internet-related proxies for training data dominance.

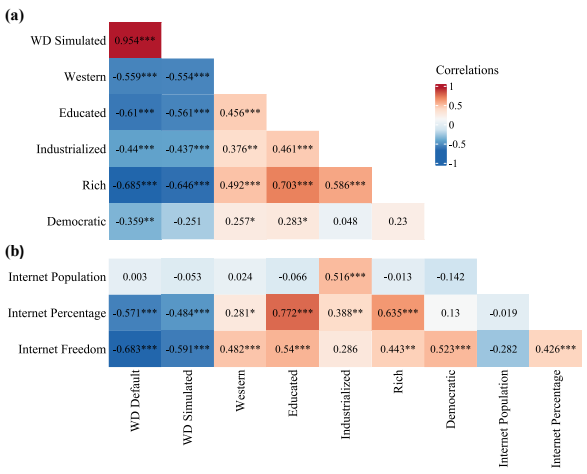

Figure 1: (a) WD correlations between the model's value alignment using two prompt types and the WEIRD scores, and (b) correlations of these variables with the proxies for country/territory dominance in training data (Pearson's r, $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$; created with the ComplexHeatmap package, Gu et al., 2016). Western is a binary variable where Western is 1 and non-Western is 0.

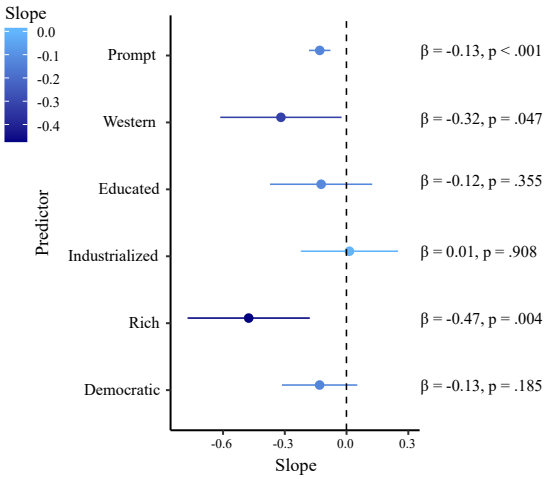

Figure 2: Standardized regression coefficients of prompt type and the WEIRD scores for predicting WD (error bars: 95% CI) [2]. Sum coding was used (with simulated identity prompt/Western coded as 1 and the other level as -1) to ensure that the other predictors' slopes were estimated while averaging across the levels of prompt and Western.

Results

LLM's WEIRD Bias As shown in Figure 1a, all of the WEIRD scores were significantly and negatively correlated with the WD between human and LLM responses (i.e., positively correlated with the model's value alignment) for both prompt types (except that the democratic score was only marginally correlated with the WD given simulated identity

[2] The variance inflation factor was below 3 for all predictors.

prompt, p = .051): the higher the country/territory's WEIRD scores, the better the model's value alignment with the country/territory. This result suggested that GPT-3.5 Instruct indeed represented the values of WEIRD populations better. Since many of the WEIRD scores were also correlated with each other, we further conducted a linear mixed-effects regression with prompt type and all of the WEIRD scores as the predictors to evaluate their unique contributions for predicting the model's value alignment assessed in WD (Figure 2). The regression analysis revealed that among the WEIRD scores, only the Western and Rich dimensions were significant predictors of the model's value alignment when controlling for other variables, with Rich being the strongest predictor. Meanwhile, simulated identity prompting significantly reduced the WD/enhanced the value alignment, indicating that GPT-3.5 Instruct is indeed steerable since its value alignment can be improved by prompting it to act as different populations, although the effect size was small.

Can WEIRD Bias Be Reduced by Prompting? The results above showed that among the WEIRD scores, the Western and Rich dimensions were uniquely associated with model's value alignment, suggesting that the model's value was biased towards countries/territories that are Western (a binary variable) or have a high Rich score. To examine whether simulated identity prompting could also reduce these biases, we performed a 2 × 2 mixed ANOVA for the Western dimension and a moderated mixed-effects regression for the Rich dimension to test if the effects of these variables on value alignment assessed in WD were moderated by prompt type. The results showed that the simulated identity prompt weakened the effect of the Western dimension, $F(1, 62) = 9.73$, $p = .003$, $\eta^2_p = .32$, and the effect of the Rich dimension, $t(59) = 4.73$, $\beta = 0.21$, $p < .001$, on the model's value alignment. However, the effects remained significant even when using the simulated identity prompt for both the Western, $t(62) = –5.25$, $p < .001$, and the Rich dimensions, $t(65.34) = –5.79$, $\beta = –0.56$, $p < .001$, indicating that such prompting still cannot eliminate the bias.

Source of LLM's WEIRD Bias We then conducted mediation analyses to test whether the country/territory's dominance in the training data (as measured by the three proxies) can explain the WEIRD bias in the model's value alignment, focusing on the significant predictors, Western and Rich. Since Internet Population was not correlated with the WD value alignment measures or most of the WEIRD scores (Figure 1b), we only included Internet Percentage and Internet Freedom as the mediators. While Internet Freedom served as a significant mediator for both predictors, Internet Percentage was not a significant mediator for either (Figure 3). Meanwhile, the direct effects of the Western and Rich dimensions on WD value alignment both remained significant, indicating that the source of LLM's WEIRD bias cannot be fully explained by the mediators included.

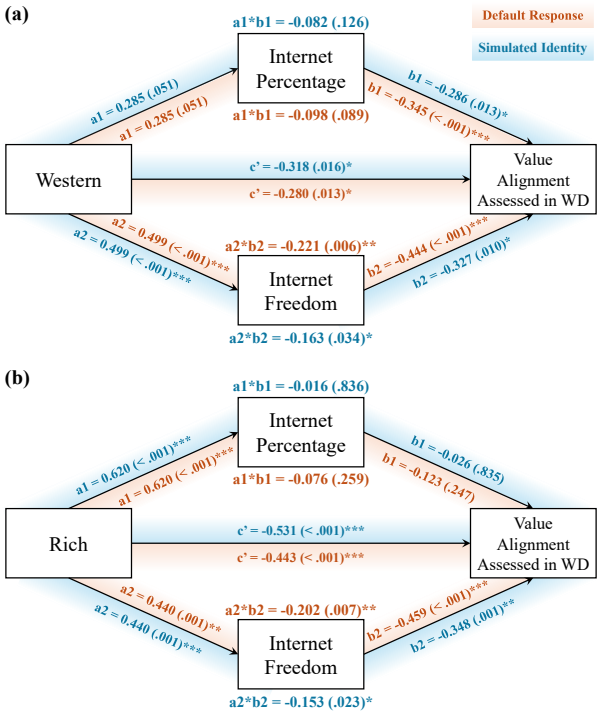

Figure 3: Indirect effect of the (a) Western and (b) Rich dimensions on value alignment assessed in WD (default response and simulated identity prompts, in different colors) through the proxies for the country/territory's dominance in training data (i.e., Internet Percentage and Internet Freedom). Path values are the standardized coefficients with p-values in parenthesis (ab: indirect effect of the WEIRD indicators on value alignment through the mediators, c': direct effect of the WEIRD indicators on value alignment, controlling for the indirect effects, $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$)

Discussion

Consistent with our hypothesis, GPT-3.5 Instruct indeed exhibited a bias toward the values of WEIRD populations. However, the term WEIRD is merely a conscious-raising backronym and does not guarantee that each letter has equal theoretical importance (Henrich, 2024). Indeed, our findings indicated that among the five WEIRD dimensions, only Western and Rich were found to be significant predictors for the model's value alignment after controlling for the other dimensions. In addition, while the Western dimension tended to be emphasized when discussing WEIRD populations, we found that the Rich dimension was the strongest predictor for alignment, further stressing that WEIRD-ness should not be treated as a binary distinction focused on the Western dimension.

One way to address the WEIRD bias is to investigate its source. Here we considered potential WEIRD dominance in the LLM's training data, using three proxies based on internet

usage data. Among the three proxies, Internet Freedom of a country/territory may be the most accurate estimate since it was the only significant mediator. However, the effect of the WEIRD scores on value alignment remained significant even after controlling for the mediators, indicating that the mediators could not fully explain the effect. Therefore, we either need more direct measures to assess training data dominance or we should consider other potential sources, such as biases introduced by reinforcement learning with human feedback (RLHF). Biases from RLHF may also be partially reflected by Internet Freedom (e.g., populations used to provide human feedback may be those with higher Internet Freedom). However, more information about LLMs' RLHF processes is required to draw further conclusions.

Another direction is to steer the model to different values according to different user needs. Indeed, prompting GPT-3.5 Instruct to act as an individual from the specific country/territory improved the alignment and weakened the WEIRD bias. However, the effect was small, and the bias persisted even with such prompting. While the direction is promising, the current model and prompting approach is not sufficient to eliminate the bias, similar to what previous studies observed (Boelaert et al., 2024; Santurkar et al., 2023). To further investigate the current state of the bias that persisted even after simulated identity prompting, we clustered the countries/territories based on their alignment with GPT-3.5 Instruct (measured by WD when using the simulated identity prompt) across the questions so that we could identify the countries/territories that are better represented and those that require more alignment.

Which Countries/Territories Are Better Represented by LLM?

Methods

We used the scikit-learn package (Pedregosa et al., 2011) to perform hierarchical clustering on the 64 countries/territories based on their alignment with GPT-3.5 Instruct as measured by WD when using simulated identity prompting across the 47 questions. Missing data were estimated from available data using Bayesian ridge regression through scikit-learn's IterativeImputer to ensure a complete 64 × 47 matrix for clustering. Each country/territory was initially treated as a single-item cluster. At each step, the two closest clusters would be merged into a bigger cluster until all clusters were merged into one. Distance between clusters was calculated using cosine distance and complete linkage.

Results

The hierarchical clustering analysis resulted in 4 clusters (referred to as Clusters 0 to 3; see dendrogram in Figure 4). One-way ANOVA analysis indicated that the clusters differed significantly in model alignment (WD), $F(3, 60) = 21.61$, $p < .001$, $\eta^2 = .52$. Clusters 2 and 3 had significantly better alignment than Clusters 0 and 1, while there were no significant differences between Clusters 2 and 3 or Clusters 0 and 1. Cluster 3 included English-speaking countries, such as the United States, as well as some countries in Western Europe and Latin America, while Cluster 2 contained major East Asian countries/regions and Russia.

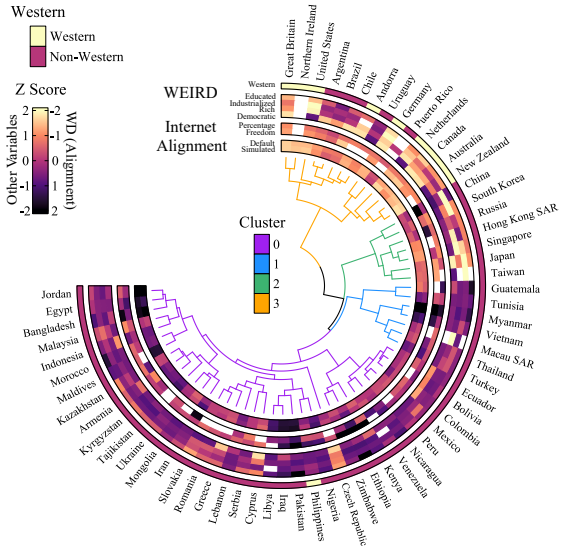

Figure 4: Circular heatmap representation (circlize package; Gu et al., 2014) of each country/territory's model alignment measured by WD, Internet Percentage and Internet Freedom, and WEIRD scores, where lighter shades indicate better alignment (smaller WD) and higher values for the other variable (see color scale, white indicates missing data). All of the values are in z-scores. The dendrogram represents the hierarchical clustering results based on model alignment (in WD) from simulated identity prompting across questions.

Table 1: Differences in WEIRD dimensions (except the binary variable Western), Internet Percentage, and Internet Freedom across the four clusters.

Variable	F	p	η^2
Educated	8.16	< .001	.304
Industrialized	14.81	< .001	.442
Rich	15.39	< .001	.447
Democratic	4.14	.010	.179
Internet Percentage	5.95	.001	.245
Internet Freedom	9.36	< .001	.401

We then took a closer look at the WEIRD scores and Internet Percentage and Internet Freedom of the clusters to better understand their respective characteristics (visualized in Figure 4). Chi-square test of association indicated a significant relationship between the Western dimension and cluster membership, χ^2 (3, N = 64) = 32.25, $p < .001$, with all of the Western countries belonging to Cluster 3 except for the Philippines. One-way ANOVA analyses revealed significant between-cluster differences in all of the remaining WEIRD

dimensions (Table 1). More specifically, Clusters 2 and 3 were significantly higher in the Educated and Rich scores than Clusters 0 and 1. Cluster 2 was significantly more Industrialized than all of the other clusters, while Cluster 3 was more Industrialized than the remaining clusters. Cluster 3 was significantly more Democratic than all of the other clusters. The other differences were not significant. The clusters also differed significantly in Internet Percentage and Internet Freedom (Table 1), with Clusters 2 and 3 having higher Internet Percentage than Clusters 0 and 1 and Cluster 3 having higher Internet Freedom than all other clusters.

Discussion

The clustering results confirmed that the binary distinction between WEIRD and non-WEIRD with a focus on the Western dimension is problematic when considering LLM's value bias. We found that both the typical WEIRD populations (Cluster 3) and non-Western but Educated, Industrialized, and Rich populations (Cluster 2) are well-represented by GPT-3.5 Instruct. In psychological research, East Asian samples have been commonly used as the non-WEIRD comparison group. Conclusions drawn from these WEIRD versus East Asian comparisons are often generalized more broadly as conclusions about WEIRD versus the rest of the world, even though East Asian societies are similar to WEIRD societies yet dissimilar from the other societies in many aspects (Krys et al., 2024). Similarly, research on LLM value alignment should not only focus on WEIRD and East Asian societies, especially since they are both already relatively well-represented. It is also important to note that the relatively well-represented Cluster 2 did not necessarily have higher internet freedom, although it was a mediator for the effect of WEIRD scores on model alignment. Since the mediation was only partial, other factors may better explain why Cluster 2 was relatively well-represented.

General Discussion

We investigated whether LLM's values are biased toward the values of WEIRD populations and explored two potential directions for mitigating such bias. We found that there was a general tendency for GPT-3.5 Instruct to represent countries/territories with high WEIRD scores more accurately. However, the WEIRD dimensions were highly correlated with each other, meaning that the effect of each dimension on LLM's value alignment could be amplified by others endogenously. After controlling for other dimensions, we found that only Western and Rich remained significant predictors. In particular, while the Western dimension is often emphasized, the Rich dimension was a stronger predictor of LLM's value alignment than Western. This finding was further confirmed by the clustering results, which revealed that non-Western but Rich, Educated, and Industrialized populations were also represented relatively well by the model. These populations included East Asian countries and regions, which are commonly used as non-WEIRD comparison groups in psychological studies (Krys et al., 2024). Our findings thus demonstrate the importance of considering more diverse populations beyond Western and East Asian populations in LLM value alignment research.

We also considered two potential directions for mitigating LLM's value bias toward countries/territories that are Western and high in the Rich score. One direction is to determine the source of the bias so that further research could focus on ways to remove the source. Here we used three proxies of a country/territory's dominance in LLM's training data based on internet usage data to investigate whether dominance in the training data could be a source of the bias. We found that only the proxy Internet Freedom could be a potential source. Also, it could only partially explain the effect of the Western or Rich dimensions on LLM's value alignment. One possibility is that these internet usage data may not accurately capture a country/territory's dominance in the training data as the training dataset for many LLMs are not publicly available. Moreover, even for models with open training datasets, demographic information about the people who produced the data is often unavailable. Another consideration is that many LLMs have gone through the RLHF process, where biases could be introduced by human evaluators either due to their own biases or due to the instructions given to them (Casper et al., 2023). Similar to the training datasets, information about LLMs' RLHF processes is often limited.

Another means for tackling LLM's value bias is to simply steer the model to different values depending on user needs. The current study found that prompting the model to simulate a given identity improved its value alignment with the corresponding country/territory, but the effect was limited and value bias persisted. Even with such prompting, only the WEIRD and East Asian populations were represented relatively well. Future studies may consider other methods to steer the model's values according to user needs. For instance, Li et al. (2024) found that LLMs fine-tuned to adapt to specific cultures performed better on culture-related language tasks than the unified model, which was fine-tuned with data from multiple cultures. Therefore, developing culture-specific models through fine-tuning may be a more effective strategy than expecting a unified model to adapt to different cultures only by in-context learning.

In conclusion, we found that while the GPT-3.5 Instruct LLM currently better represents the values of WEIRD populations, it also captures the values of some non-Western populations (e.g., East Asia) relatively well. While many researchers focus on the comparison between Western and non-Western populations, the Rich dimension was the strongest predictor of LLM's value bias. Therefore, future studies should attend to less prosperous populations instead of only considering easily accessible samples. We also found that one source of the value bias could be unbalanced training data as approximated by Internet Freedom, and that simulated identity prompting reduced the value bias but did not completely remove it. These findings point to the importance of training process disclosure by model developers and the consideration of developing culture-specific models through fine-tuning to ensure fair and ethical usage of LLMs.

Acknowledgements

This study was funded by Research Grant Council of Hong Kong, Area of Excellence (Project number AoE/E-601/24-N). We would like to thank Zixuan Wang for help with acquiring data from GPT-3.5 Instruct and Jinhan Zhang for advice on implementing hierarchical clustering. We are also grateful for the helpful comments and suggestions from Linus Huang and the anonymous reviewers.

References

Atari, M., Xue, M. J., Park, P. S., Blasi, D., & Henrich, J. (2023). *Which Humans?* OSF. https://doi.org/10.31234/osf.io/5b26t

Benkler, N., Mosaphir, D., Friedman, S., Smart, A., & Schmer-Galunder, S. (2023). *Assessing LLMs for Moral Value Pluralism*. arXiv. https://arxiv.org/abs/2312.10075

Biesheuvel, L. A., Workum, J. D., Reuland, M., van Genderen, M. E., Thoral, P., Dongelmans, D., & Elbers, P. (2024). Large language models in critical care. *Journal of Intensive Medicine*. Advance Online Publication.

Boelaert, J., Coavoux, S., Ollion, E., Petev, I. D., & Präg, P. (2024). *How do Generative Language Models Answer Opinion Polls?* OSF. https://doi.org/10.31235/osf.io/r2pnb

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., Wang, T., Marks, S., Segerie, C.-R., Carroll, M., Peng, A., Christoffersen, P., Damani, M., Slocum, S., Anwar, U., … Hadfield-Menell, D. (2023). *Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback*. arXiv. https://arxiv.org/abs/2307.15217

Davani, A., Díaz, M., Baker, D., & Prabhakaran, V. (2024). Disentangling Perceptions of Offensiveness: Cultural and Moral Correlates. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2007–2021.

Duan, S., Yi, X., Zhang, P., Lu, T., Xie, X., & Gu, N. (2024). *DENEVIL: Towards Deciphering and Navigating the Ethical Values of Large Language Models via Instruction Learning*. arXiv. https://arxiv.org/abs/2310.11053

Freedom House. (2024a). Freedom in the World. https://freedomhouse.org/report/freedom-world

Freedom House. (2024b). Freedom on the Net. https://freedomhouse.org/report/freedom-net

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines*, *30*(3), 411–437.

Ganguli, D., Askell, A., Schiefer, N., Liao, T. I., Lukošiūtė, K., Chen, A., Goldie, A., Mirhoseini, A., Olsson, C., Hernandez, D., Drain, D., Li, D., Tran-Johnson, E., Perez, E., Kernion, J., Kerr, J., Mueller, J., Landau, J., Ndousse, K., … Kaplan, J. (2023). *The Capacity for Moral Self-Correction in Large Language Models*. arXiv. https://arxiv.org/abs/2302.07459

Gu, Z., Eils, R., & Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, *32*(18), 2847–2849.

Gu, Z., Gu, L., Eils, R., Schlesner, M., & Brors, B. (2014). *circlize* implements and enhances visualization in R. *Bioinformatics*, *30*(19), 2811–2812.

Haerpfer, C., Inglehart, R., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano J., Lagos, M., Norris, P., Ponarin, E. & Puranen, B. (eds.). (2022). World Values Survey: Round Seven – Country-Pooled Datafile Version 6.0. JD Systems Institute & WVSA Secretariat. https://www.worldvaluessurvey.org/WVSDocumentationWV7.jsp

Henrich, J. (2024). WEIRD. In M. C. Frank & A. Majid (Eds.), *Open Encyclopedia of Cognitive Science*. MIT Press.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *The Behavioral and Brain Sciences*, *33*(2–3), 61–135.

Huntington, S. P. (2000). The Clash of Civilizations? In Crothers, L. & Lockhart, C. (Eds.), *Culture and Politics*. Palgrave Macmillan.

Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., Duan, Y., He, Z., Zhou, J., Zhang, Z., Zeng, F., Ng, K. Y., Dai, J., Pan, X., O'Gara, A., Lei, Y., Xu, H., Tse, B., Fu, J., … Gao, W. (2023). *AI Alignment: A Comprehensive Survey*. arXiv. https://arxiv.org/abs/2310.19852

Kirk, H. R., Vidgen, B., Röttger, P., & Hale, S. A. (2024). The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*, *6*(4), 383–392.

Kocielnik, R., Prabhumoye, S., Zhang, V., Jiang, R., Alvarez, R. M., & Anandkumar, A. (2023). *BiasTestGPT: Using ChatGPT for Social Bias Testing of Language Models* arXiv. https://arxiv.org/abs/2302.07371

Krys, K., de Almeida, I., Wasiel, A., & Vignoles, V. L. (2024). WEIRD–Confucian comparisons: Ongoing cultural biases in psychology's evidence base and some recommendations for improving global representation. *American Psychologist.* Advance online publication.

Lai, J., Gan, W., Wu, J., Qi, Z., & Yu, P. S. (2024). Large language models in law: A survey. *AI Open*, *5*, 181–196.

Li, C., Chen, M., Wang, J., Sitaram, S., & Xie, X. (2024). *CultureLLM: Incorporating Cultural Differences into Large Language Models*. arXiv. https://arxiv.org/abs/2402.10946

Linxen, S., Sturm, C., Brühlmann, F., Cassau, V., Opwis, K., & Reinecke, K. (2021). How WEIRD is CHI? In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM.

McKenzie, I. R., Lyzhov, A., Pieler, M. M., Parrish, A., Mueller, A., Prabhu, A., McLean, E., Shen, X., Cavanagh, J., Gritsevskiy, A. G., Kauffman, D., Kirtland, A. T., Zhou, Z., Zhang, Y., Huang, S., Wurgaft, D., Weiss, M., Ross, A., Recchia, G., … Perez, E. (2023). Inverse Scaling: When Bigger Isn't Better. *Transactions on Machine Learning Research*.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., & Lowe, R. (2022). Training language models to follow

instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in Neural Information Processing Systems* (Vol. 35, pp. 27730–27744). Curran Associates, Inc.

Rawls, J. (1993). The Law of Peoples. *Critical Inquiry*, *20*(1), 36–68.

Rozado, D. (2023). The Political Biases of ChatGPT. *Social Sciences*, *12*(3), Article 3.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). Whose Opinions Do Language Models Reflect? In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *Proceedings of the 40th International Conference on Machine Learning* (Vol. 202, pp. 29971–30004). PMLR.

Septiandri, A. A., Constantinides, M., & Quercia, D. (2024). *WEIRD ICWSM: How Western, Educated, Industrialized, Rich, and Democratic is Social Computing Research?* arXiv. https://arxiv.org/abs/2406.02090

UNIDO (2022). Competitive Industrial Performance (CIP) Index. https://stat.unido.org/data/download?dataset=cip

United Nations Development Programme. (2022). Human Development Index. https://hdr.undp.org/data-center/human-development-index#/indicies/HDI

World Bank. (2021a). GNI per capita, PPP (current international $). https://data.worldbank.org/indicator/NY.GNP.PCAP.PP.CD

World Bank. (2021b). Population, total. https://data.worldbank.org/indicator/SP.POP.TOTL

World Bank. (2021c). Individuals using the Internet (% of population). https://data.worldbank.org/indicator/IT.NET.USER.ZS