Mimicking Footprints or Genuine Understanding? Exploring Cultural Representation Bias in Large Language Models

Anonymous ACL submission

Abstract

This study investigates Large Language Models' (LLMs) capacity for cross-cultural understanding in moral reasoning tasks, examining whether their performance reflects genuine comprehension or sophisticated pattern matching. Given documented biases in LLMs' training data toward English-language content and Western perspectives, we evaluated five widely-deployed models (Gemini, GPT-4, Llama 3 8B, Llama 3.1 8B, and Mistral 7B) using three datasets reflecting variations along cultural dimensions: the World Values Survey, the Moral Machine experiment, and the COVID-19 Vaccine Hesitancy survey.

Our analysis revealed three key findings: (1) While human responses demonstrated clear cultural clustering patterns, particularly in the WVS and Vaccine datasets, LLMs failed to replicate these distinct cultural groupings, suggesting limitations in capturing underlying cultural dynamics. (2) Cultural representation bias varied significantly by model architecture ($F = 47.70$-416.88, $p < .001$) and cultural context ($F = 4.34$-13.09, $p < .001$), with GPT-4 showing consistent performance (22%-31%) while Llama 3 achieved lowest bias in WVS (17%). (3) Demographic-cultural interactions varied unexpectedly across datasets and models, notably in Orthodox Europe where top-performing Llama 3 showed increased bias while other models improved. These findings suggest that while LLMs can effectively pattern-match in simple moral reasoning tasks, they face substantial challenges in processing complex cross-cultural moral scenarios, indicating limitations in their genuine understanding of cultural nuances.

1 Introdution

Large language models (LLMs) have achieved remarkable success across various natural language processing tasks—from text generation to decision support in complex domains such as autonomous driving and public health policy. As these models become increasingly integrated into applications that require sensitivity to diverse moral attitudes and cultural contexts, a critical question arises: do LLMs truly internalize and reflect the rich nuances of different cultures, or do they merely reproduce surface-level patterns learned from dominant training data?

The challenge of achieving genuine cross-cultural understanding is fundamentally linked to training data biases. The training data for widely used LLMs is heavily skewed toward English content: approximately 95% of Llama 3's training corpus consists of English sources (Meta AI, 2023), while GPT-3's English corpus accounts for 92%(Brown et al., 2020). This linguistic and cultural imbalance may lead models to learn superficial response patterns rather than developing true perspective-taking abilities. Previous research has revealed biases in LLMs concerning gender and race issues (Zack et al., 2024), while English-based LLMs demonstrate better understanding of Western moral norms compared to non-Western cultures (Ramezani and Xu, 2023), suggesting they might be learning to mimic dominant cultural patterns rather than developing genuine cross-cultural understanding. Recent studies have demonstrated that LLMs often exhibit cultural biases by reflecting dominant narratives, typically framed as comparisons such as non-Western versus Western or English versus non-English. Although these works provide valuable insights, they tend to focus on a narrow subset of cultural contexts.

We define "genuine cross-cultural moral understanding" as the ability of an LLM to generate responses that not only align with documented human cultural moral attitudes but also mirror the distinct cultural clusters delineated by the WVS cultural map. A model that truly understands cultural nuances adapts its responses to reflect variations along the Traditional versus Secular-Rational and

Survival versus Self-expression dimensions, capturing subtle differences between cultural groups. Moreover, such a model maintains this specificity across tasks of increasing complexity and when additional demographic and temporal cues are provided, whereas a model relying on superficial pattern matching tends to produce neutral or generic outputs that fail to capture the inherent diversity observed in human responses.

This conceptual framework underpins our research questions and informs our methodological approach. First, we ask: to what extent do LLMs exhibit cultural representation bias? In other words, how closely do their outputs replicate the natural clustering observed in human responses? Second, we ask: how accurately do LLMs capture cultural alignment along the dimensions defined by the Inglehart–Welzel Cultural Map? Finally, we examine how demographic factors—such as gender, age, income, and education—interact with cultural contexts to shape moral attitudes. Addressing these questions is crucial from a theoretical standpoint, as cultural theory posits that human moral reasoning is deeply rooted in historical, social, and economic contexts that produce distinct cultural clusters.

To answer these questions, we employ standardized prompt templates that explicitly incorporate cultural, demographic, and temporal cues, applied across three complementary datasets: the World Values Survey; the Moral Machine dataset; and the COVID-19 Vaccine Hesitancy dataset. This approach allows us to rigorously evaluate whether LLMs generate responses that truly reflect the nuanced, context-aware cultural understanding predicted by theory, or if they merely engage in superficial pattern matching.

2 Related works

We first examine cultural representation bias in LLMs, exploring how these models exhibit and perpetuate cultural biases through their training data and architectures. We then review current approaches to evaluating and mitigating these biases, highlighting their limitations. Finally, we analyze the fundamental challenge of distinguishing between pattern matching and genuine cultural understanding, which motivates our research direction.

2.1 Cultural Representation Bias in LLMs

Understanding and mitigating cultural representation bias in LLMs remains a significant challenge in artificial intelligence research. These biases, often rooted in imbalances in training corpora and the over-representation of dominant cultural perspectives, manifest across multiple dimensions. For instance, disparities have been observed in the treatment of linguistic nuances, such as how models generate different idiomatic expressions in non-Western languages, often failing to capture their cultural connotations (Prabhakaran et al., 2022; Arora et al., 2023; Schwöbel et al., 2023). Additionally, these biases influence the representation of historical narratives and socio-political contexts, as seen in the under-representation of certain regions, such as Namibia, Uganda, and Yemen, in language model predictions (Schwöbel et al., 2023). Such under-representation not only limits the inclusivity of LLM outputs but also risks perpetuating cultural erasure and misrepresentation in global contexts.

The impact of cultural representation bias extends beyond surface-level misrepresentation. In cross-cultural settings, these models frequently exhibit biased behavior, as evidenced in tasks involving multilingual text generation (Arora et al., 2023) and moral reasoning (Schramowski et al., 2022). Recent studies have revealed these cultural biases in diverse cultural contexts: from inadequate representation of South Asian cultural artifacts (Qadri et al., 2023) to Western-centric biases in Arabic language outputs (Naous et al., 2024), and significant under-representation of Latin American and African perspectives (Schwöbel et al., 2023). These empirical findings align with theoretical frameworks that situate AI systems within broader patterns of cultural hegemony and technological colonialism (Mohamed et al., 2020), suggesting that such biases reflect and potentially reinforce existing global power structures.

2.2 Current Approaches of Bias Evaluation frameworks

Recent research has proposed various approaches to evaluate and mitigate cultural representation bias. Evaluation frameworks like MiTTenS specifically assess gender mistranslations to reveal how Cultural Representation Bias intersects with gender bias in multilingual contexts (Robinson et al., 2024). Similarly, NormAd provides systematic methodologies to measure LLMs' cultural adapt-

ability through cross-cultural benchmarking (Rao et al., 2024). The CulturePark framework attempts to address these biases by leveraging cross-cultural multi-agent communication to generate synthetic dialogues for model fine-tuning (Li et al., 2024b).

However, these approaches face significant limitations. First, they rely heavily on static datasets that struggle to capture the dynamic nature of cultural expressions. The geographical erasure problem persists, where models consistently underpredict data related to certain regions due to training data imbalances (Liu et al., 2025). Additionally, collecting and maintaining culturally diverse datasets poses substantial challenges, particularly for low-resource cultures (Li et al., 2024a). These limitations suggest that current solutions may not adequately address the fundamental issues of Cultural Representation Bias.

2.3 From Pattern Matching to Cultural Understanding

Recent research has highlighted the distinction between genuine understanding and surface-level pattern matching in LLMs (Yu and Petkov, 2024; Bender et al., 2021). Genuine understanding requires actively transforming information through critical analysis, contextual adaptation, and dynamic reasoning. In contrast, surface understanding is mere passive replication of learned patterns. Comprehension is a cognitive reconstruction process, not just superficial imitation.

A critical gap in current research is the inability to distinguish between surface-level pattern matching and genuine cultural understanding in LLMs. While models often perform well on simple cultural tasks, they struggle significantly with more nuanced cultural contexts. Recent studies in multilingual capabilities have shown that LLMs often fail in tasks requiring deep cultural understanding, particularly in low-resource languages (Shen et al., 2024). This suggests that apparent competence in cultural tasks may stem from sophisticated pattern matching rather than true comprehension.

The challenge of evaluating genuine cultural understanding becomes particularly evident in tasks like translation and semantic analysis. Research has shown that LLMs frequently fail to capture cultural nuances in these contexts, especially in low-resource settings (Singh et al., 2024). These findings highlight the need for more sophisticated evaluation frameworks that can effectively distinguish between pattern matching and genuine cultural understanding, particularly in complex cultural contexts.

In general, current research faces three main limitations in addressing cultural representation bias. First, existing evaluation methods rely heavily on static datasets that cannot capture the dynamic nature of cultural expressions. Second, proposed solutions often fail to distinguish between surface-level pattern matching and genuine cultural understanding. Third, there is a lack of comprehensive frameworks that can effectively evaluate both the breadth and depth of cultural understanding in LLMs.

3 Methods

In our study, we define "genuine cross-cultural moral understanding" as the ability of an LLM to generate responses that not only align with documented human cultural moral attitudes but also mirror the distinct cultural clusters defined by the Inglehart–Welzel Cultural Map. A model demonstrating genuine understanding adapts its responses to reflect variations along the Traditional versus Secular-Rational and Survival versus Self-expression dimensions, capturing the nuanced differences between cultural groups. Moreover, such a model maintains this specificity across tasks of increasing complexity and when additional demographic and temporal cues are provided. In contrast, a model relying on superficial pattern matching tends to produce neutral or generic outputs that fail to capture the inherent cultural diversity observed in human responses.

This conceptual framework underpins our research questions. First, how accurately do LLMs capture cultural alignment? Here, we examine whether LLM outputs naturally cluster along the cultural dimensions. Second, how and to what extent do LLMs exhibit cultural representation bias? This question investigates the divergence between model outputs and human responses. Finally, how do demographic factors interact with cultural contexts in shaping moral attitudes? This question probes the intersection of variables such as gender, age, income, and education with cultural nuances.

We employed standardized prompt templates designed to simulate real-world responses by incorporating actual demographic and contextual information from the datasets (see Appendix B). This approach controls for differences in how strongly each dataset reflects natural cultural clusters, ensuring methodological consistency and enabling

direct comparisons between model outputs and human data.

3.1 Research Datasets

We selected three complementary datasets that challenge models' perspective-taking abilities at increasing levels of task complexity (see Table **??**). The WVS dataset naturally exhibits cultural clustering along the Traditional versus Secular-Rational axis across 55 countries. In contrast, the Moral Machine dataset is a human-constructed experiment featuring binary ethical dilemmas in autonomous driving across 130 countries; although it introduces demographic intersections, its forced-choice format does not inherently capture natural cultural clusters. Finally, the COVID-19 Vaccine Hesitancy dataset represents the most complex scenario by requiring models to integrate rich demographic factors and temporal context from 23 countries using a five-point Likert scale to capture public health attitudes during the 2022 pandemic. (see Appendix B).

3.1.1 World Values Survey Dataset

We analyzed the Ethical Values section from Wave 7 (2017–2021) of the World Values Survey (WVS), which encompasses responses from 55 countries, following the approach described in (Ramezani and Xu, 2023). The survey, administered in each country's primary language, comprises 19 items addressing moral attitudes related to personal conduct and societal issues, with responses normalized to a [-1, 1] scale (where -1 indicates "never justifiable" and 1 indicates "always justifiable").

3.1.2 The Moral Machine Dataset

The Moral Machine experiment, which examines ethical dilemmas across 130 countries (Awad et al., 2018), was used to assess cross-cultural moral attitudes by analyzing 59 scenarios per country (based on the minimum scenario count for Afghanistan). These scenarios span eight moral dimensions, including contrasts such as pedestrians versus passengers, law-abiding versus law-breaking behavior, and considerations of gender, physical condition, social status, age, quantity of lives, and species.

3.1.3 COVID-19 Vaccine Hesitancy Dataset

The COVID-19 Vaccine Hesitancy study (2022) captures global health attitudes during the post-intervention period of 2022 and thus provides the most challenging context for evaluating cross-cultural moral attitudes. Conducted across 23 countries, the survey assesses vaccine-related moral attitudes through dimensions such as risk perception, efficacy beliefs, safety concerns, and institutional trust. To evaluate how models form these attitudes, we developed structured prompts that require processing both demographic factors and the temporal context of 2022. Model outputs were subsequently dichotomized (responses 1–3 as hesitant/resistant and 4–5 as supportive) to facilitate direct comparisons with human data.

3.2 Research Models

In this study, we evaluate five LLMs that, despite not being the most recent releases, remain widely deployed and actively used across various applications. Their sustained adoption and practical impact make them particularly relevant for investigating real-world implications of LLMs. The commercial models in our study, GPT-4 and Gemini 1.5 Pro, continue to serve as primary interfaces for millions of users through widely-adopted applications (Bianchi, 2024), making them crucial subjects for understanding the actual cultural impact of LLMs in practice. In the open-source domain, Llama 3 8B, Llama 3.1 8B, and Mistral 7B have maintained substantial developer communities and implementation bases, as evidenced by their continued high deployment rates on major model hosting platforms (huggingface.co, 2025a) (huggingface.co, 2025b).

The widespread adoption of these models, combined with their documented efforts to promote cultural diversity and multilinguality, makes them particularly valuable for studying cross-cultural moral attitudes. Although GPT-4 is predominantly trained on English-language data (OpenAI et al., 2024), it incorporates extensive human feedback to mitigate cultural biases. Gemini 1.5 Pro explicitly emphasizes multilingual performance and cross-cultural adaptability (Team et al., 2024). The Llama series has shown evolving support for multilingual applications (Grattafiori et al., 2024), while Mistral 7B's advanced long-context capabilities through sliding window attention (SWA) (Jiang et al., 2023) may facilitate the extraction of cultural nuances from extended texts.

3.3 Cultural Dimensions Framework

Given the varying geographic coverage across our datasets (55 countries in WVS, 130 in Moral Machine, and 23 in Vaccine Hesitancy), we adopted the Inglehart–Welzel Cultural Map as our analytical framework to enable standardized cross-cultural comparisons. This framework organizes

societies along two primary dimensions: Traditional versus Secular-rational values (vertical axis) and Survival versus Self-expression values (horizontal axis). These dimensions account for over 70% of cross-national variance in various social indicators and demonstrate robust correlations with economic, political, and social metrics (Inglehart et al., 2014).

The applicability of the framework to our research is supported by three key factors. First, the methodology of the World Values Survey derives directly from the Inglehart-Welzel theoretical construction. Second, the Moral Machine experiment utilized this cultural mapping system to validate their cultural cluster classifications (Awad et al., 2018). Third, empirical studies on COVID-19 response demonstrate significant correlations between the Traditional-Secular value dimension and national pandemic management capabilities (Kaklauskas et al., 2022).

Using this framework, we categorized countries into eight distinct cultural clusters based on their positions along the Inglehart–Welzel dimensions: Orthodox Europe (e.g. Russia, Greece), Catholic Europe (e.g. Italy, Spain), African-Islamic (e.g. Nigeria, Egypt), Latin America (e.g. Brazil, Mexico), English-Speaking (e.g. USA, UK), Protestant Europe (e.g. Sweden, Denmark), Confucian (e.g. China, Japan), and West & South Asia (e.g. Turkey, India).

3.4 Demographic Intersectionality Framework

Since cultural theory emphasizes that demographic factors further refine moral attitudes within cultural clusters, it is necessary to evaluate whether LLMs can handle these nuances. Understanding demographic interactions provides a more comprehensive picture of cultural understanding and highlights areas where models may need improvement. We integrated demographic data from the Moral Machine and Vaccine Hesitancy studies into our evaluation framework (see Table 1). This integration allows us to examine how LLMs handle demographic intersectionality in cultural contexts, building on prior research that identified systematic biases in AI systems across gender (Latif et al., 2023), age (Stypińska, 2023), and income levels (Yang et al., 2024).

Our analysis considers four key demographic dimensions: gender (male/female), age (grouped as 18–29, 30–39, 40–49, 50–59, 60+), income (above or below average), and education (with or without a bachelor's degree). These factors are incorporated into our prompts to assess whether models can capture both broad cultural patterns and the nuanced variations in moral attitudes across different demographic segments.

Demographic Factors	Vaccine Datasets (N=23020)	Moral Machine Datasets (N=38350)
Number of Countries	23	130
Gender, %		
Female	49.6	34.8
Male	50.4	65.2
Age Groups, %		
18-29	30.7	65.3
30-39	21.8	21.2
40-49	14.4	7.7
50-59	13.9	3.6
60+	19.1	2.2
Education Level, %		
With Bachelor's Degree	50.5	65.5
Without Bachelor's Degree	49.5	34.5
Income Level, %		
Income Above Average	46.3	52.2
Income Below Average	53.7	47.8

Table 1: Demographic Factors Distributions comparison between Vaccine Datasets and Moral Machine Datasets.

3.5 Analytical Approach

To evaluate how LLMs process and represent cultural moral attitudes, we developed a comprehensive analytical approach that examines both cultural patterns and representation bias.

First, we analyzed cultural patterns by examining the relationship between moral attitudes and cultural dimensions. For each dataset, we calculated Pearson correlations between moral attitudes and the Traditional-Secular Values dimension to identify significant cultural trends. We then compared the distribution of human and LLM responses along this dimension to assess whether models captured these cultural patterns.

Second, we quantified cultural representation bias (Bias) by standardizing responses from each dataset to a 0-100 scale (transforming WVS's three-point, Moral Machine's binary, and Vaccine's five-point responses) and calculating the absolute difference between model and human scores:

$$\text{Bias} = |P_{\text{model}} - P_{\text{human}}|.$$

Finally, we conducted statistical analyses to examine the significance of observed patterns. For cultural pattern analysis, we employed correlation analysis and t-tests on WVS data due to its continuous nature after standardization, while chi-square tests were used for the categorical responses in Moral Machine and Vaccine Hesitancy datasets. To assess the impact of cultural clusters and demographic factors on representation bias, we con-

ducted ANOVA tests examining main effects and interactions between these variables.

4 Results

Our analysis examines three key aspects of LLMs' cultural understanding capabilities. First, we investigate how accurately LLMs capture cultural patterns in moral attitudes by comparing their responses with human data across different cultural contexts. Second, we analyze the extent of cultural representation bias in LLM outputs, quantifying how this bias varies across models and cultural regions. Finally, we examine how demographic factors intersect with cultural contexts to influence moral attitude representations in LLM responses. Through these analyses, we seek to understand both the capabilities and limitations of LLMs in processing cultural moral attitudes.

4.1 Cultural Patterns in Moral Attitudes

Our investigation of cultural patterns in moral attitudes focused on responses across three datasets, selecting topics that exhibited strong cultural correlations in human responses. Using the Traditional Values versus Secular-Rational Values dimension from the World Values Survey cultural map as an established framework for cultural differentiation, we analyzed both human and LLM responses to understand their alignment with cultural patterns.

Human responses in both the WVS and Vaccine datasets demonstrate clear cultural clustering—countries within the same cultural groups show similar response patterns along the Traditional-Secular Values axis. This clustering is particularly evident in attitudes toward homosexuality and vaccine willingness, suggesting strong cultural influences on these moral attitudes. However, LLMs, while generating responses that vary across cultural contexts, fail to replicate this distinct cultural clustering. Despite their ability to process cultural information, LLMs appear unable to fully capture the underlying cultural dynamics that shape moral attitudes. The response format influences the distribution of LLM outputs. In the WVS dataset's three-point scale (-1, 0, 1), LLMs show a notable bias toward neutral responses (0), suggesting a tendency to avoid strong positions on controversial topics. Conversely, the binary format in the Moral Machine dataset appears to constrain both human and LLM response variations, as evidenced by the lack of clear cultural patterns and the convergence of responses around 0.5.

The absence of significant cultural patterns in the Moral Machine dataset warrants further investigation. This finding might indicate either limitations in how autonomous vehicle ethical decisions reflect cultural values, or constraints imposed by the binary response format in capturing nuanced cultural differences.

4.2 Cultural Representation Bias in LLMs

Analysis of Variance revealed significant effects for both model ($F = 47.70$–416.88, $p < .001$, $\eta^2 = 2.46\%$–6.56%) and cultural context ($F = 4.34$–13.09, $p < .001$, $\eta^2 = 0.36\%$–2.43%) across datasets. The significant interaction between these factors (Moral Machine: $F = 1.64$, $p < .05$; Vaccine: $F = 29.86$, $p < .001$) indicates that cultural representation bias varies across models and cultural contexts.

As shown in Figure 2, GPT-4 demonstrated the most consistent performance (bias range: 22%–31%) across tasks, while other models showed greater variability. Notably, Llama 3 8B achieved the lowest overall bias in the World Values Survey (17%) but its successor, Llama 3.1 8B, exhibited increased bias levels (20%–38%), suggesting that model evolution does not necessarily correlate with bias reduction.

Cultural regions showed distinct patterns: Latin America exhibited maximum model variability, particularly in the Vaccine Hesitancy dataset (15%–50%), while Orthodox Europe demonstrated more stable performance (13%–20%). West & South Asia consistently showed elevated bias levels across datasets, peaking in the Moral Machine task ($\sim 38\%$). These findings highlight the complex interplay between model architecture and cultural context in determining representation bias, with implications for cross-cultural AI deployment.

4.3 Demographic Intersectionality in Moral Attitudes

Following Ramezani's approach to WVS analysis, we excluded the WVS dataset from demographic intersectionality analysis as our prompts were designed to focus on basic moral attitudes without demographic variations. The Moral Machine dataset exhibited significant interactions between cultural clusters and demographic factors (*Cultural Cluster* × *Age Groups*: $F = 15.48$, $p < .001$; *Cultural Cluster* × *Gender* × *Income Levels*: $F = 2.99$, $p < .01$), while the Vaccine dataset showed weaker

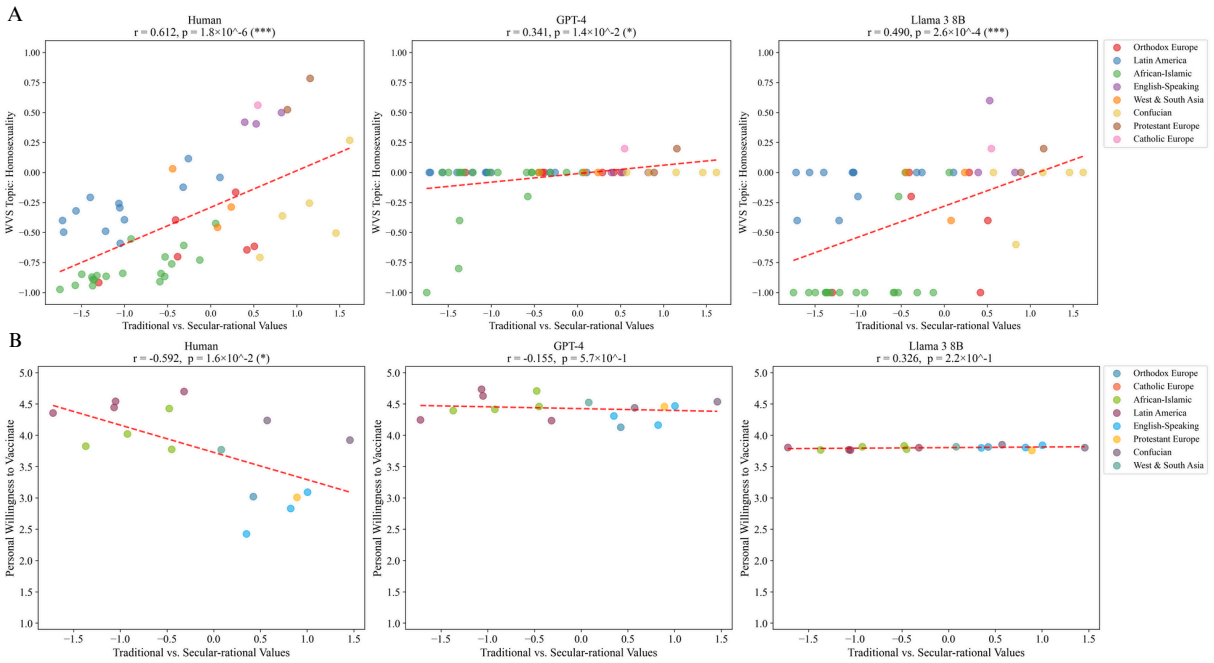

Figure 1: Cultural patterns in moral attitudes captured by LLM responses across two datasets.

A: Three panels derived from the World Values Survey (WVS) dataset on attitudes toward homosexuality.

B: Three panels based on the Vaccine Hesitancy dataset, depicting personal willingness to vaccinate.

We selected GPT-4 and Llama 3 for detailed analysis based on their demonstrated lower bias in our evaluations, with additional model results available in Appendix C.

but notable effects in socioeconomic factors (*Cultural Cluster* × *Income Levels* × *Education Levels*: $F = 4.04$, $p < .001$).

We focus on income levels, as this demographic factor showed consistent effects across both datasets (Figure 3). In the Moral Machine dataset, Orthodox Europe exhibited higher cultural representation bias for above-average income groups, particularly when compared to other cultural regions. In the Vaccine dataset, we observed a striking pattern: while Gemini and Llama 3.1 8B showed unexpectedly lower bias levels in Orthodox Europe, Llama 3—the model with the best overall performance—exhibited a dramatic increase in bias levels in this same region. This contrasting behavior suggests that model cultural representation bias can vary unpredictably when processing cultural and demographic intersections.

These findings indicate that demographic factors' influence on cultural representation bias varies significantly across different moral reasoning contexts. The contrasting patterns between datasets and the inconsistent model behaviors across cultural regions suggest that current LLMs lack a systematic approach to handling demographic intersectionality in cultural contexts.

(detailed in Appendix E)

5 Conclusion

Our analysis reveals the complex nature of cultural understanding in current LLMs. Through examining moral attitudes across different cultural contexts, we found that while LLMs can generate responses that vary by culture, they fail to replicate the distinct cultural clustering patterns observed in human responses. This suggests that LLMs may be relying more on surface-level pattern matching rather than demonstrating genuine cultural understanding.

The analysis of cultural representation bias further complicates this picture. While some models like GPT-4 showed relatively consistent performance across tasks, others exhibited highly variable bias patterns. Particularly noteworthy is the inconsistent relationship between model evolution and bias reduction, as evidenced by the increased bias levels in Llama 3.1 8B compared to its predecessor.

Our examination of demographic intersectionality revealed perhaps the most concerning aspect: the unpredictable variations in model performance when handling cultural and demographic intersections. The dramatic contrast in performance within Orthodox Europe—where some models showed un-

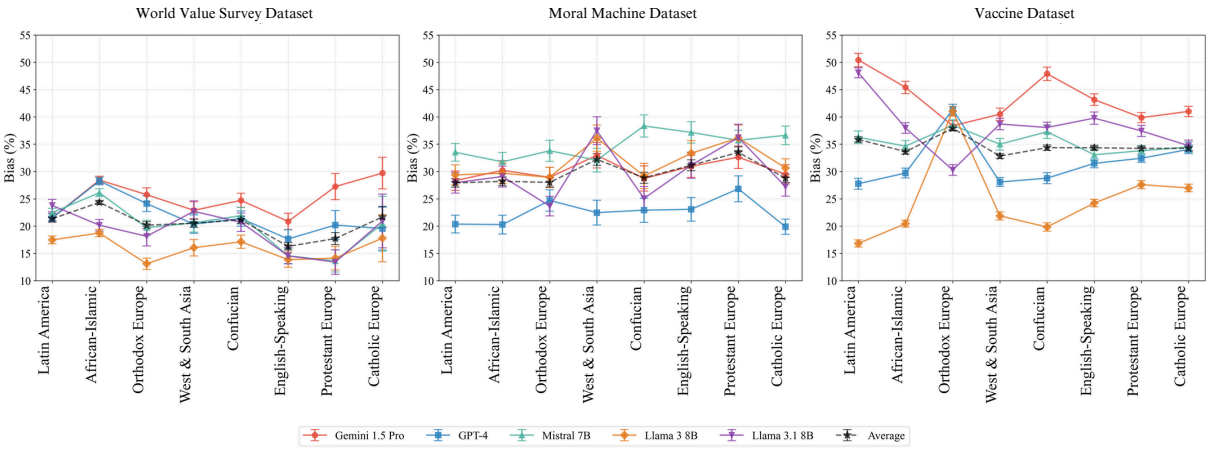

Figure 2: Cultural representation bias patterns across models and cultural clusters.

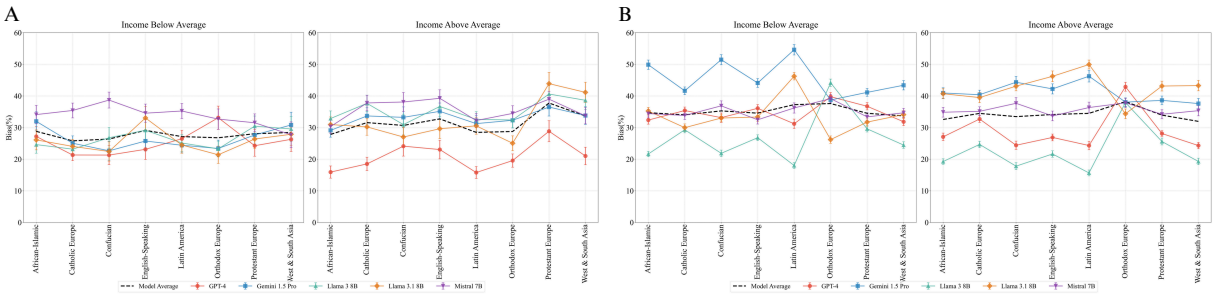

Figure 3: Intersectional patterns of income level and culture clusters.

A: Two panels derived from the World Values Survey (WVS) dataset.

B: Two panels based on the Vaccine Hesitancy dataset.

Left panels represent responses from individuals with below-average income, and right panels represent responses from individuals with above-average income. (Other demographic factors are detailed in Appendix E.)

expected decreases in bias while others exhibited significant increases—highlights the current limitations of LLMs in processing complex cultural-demographic interactions.

These findings have important implications for the deployment of LLMs in cross-cultural contexts. While these models have achieved remarkable capabilities in language processing, their handling of cultural nuances remains inconsistent and potentially problematic. Future work should focus on developing more robust evaluation frameworks and training approaches that can better capture and represent the complexity of cultural moral attitudes.

Limitations

Our study has several limitations that should be considered when interpreting the findings. Although our datasets are publicly available and include responses from participants in various countries, they do not fully capture the entire spectrum of moral attitudes present across all global cultures. The data are constrained by specific demographic, geographical, and temporal contexts, and as such, may not encompass all nuances of cultural representation biases or predict how moral attitudes might evolve in the future.

Furthermore, the methodological choices made in this study introduce additional limitations. For instance, calculating the average of moral attitudes and categorizing cultural clusters, while useful for analysis, may oversimplify the inherent complexity of these phenomena. Such approaches might obscure finer variations in moral attitudes and the dynamic interplay between culture and individual demographic factors.

Our evaluation framework also has inherent limitations in assessing LLMs' cultural understanding. The use of standardized prompts, while necessary for consistent evaluation, may not fully capture the nuanced ways in which cultural context influences moral reasoning in natural conversations. Additionally, our binary assessment of cultural representation bias might oversimplify the complex nature of cultural understanding in AI systems.

The models evaluated in this study, while widely used, represent only a subset of available LLMs, and their responses may not be representative of the broader capabilities or limitations of language models in processing cultural information. Moreover, the rapid pace of model development means that our findings might not fully reflect the capabilities of the most recent models.

References

Arnav Arora, Lucie-aimée Kaffee, and Isabelle Augenstein. 2023. Probing pre-trained language models for cross-cultural differences in values. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 114–130, Dubrovnik, Croatia. Association for Computational Linguistics.

E. Awad, S. Dsouza, R. Kim, et al. 2018. The moral machine experiment. *Nature*, 563:59–64.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Tiago Bianchi. 2024. Global chatgpt vs. gemini app downloads 2024. Accessed: 2025-02-01.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit San-

gani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

huggingface.co. 2025a. Meta llama. Accessed: 2025-02-01.

huggingface.co. 2025b. Model card for mistral-7b-instruct-v0.1. Accessed: 2025-02-01.

R. Inglehart, C. Haerpfer, A. Moreno, C. Welzel, K. Kizilova, J. Diez-Medrano, M. Lagos, P. Norris, E. Ponarin, and B. Puranen. 2014. World values survey: Round six - country-pooled datafile version. `https://www.worldvaluessurvey.org/WVSDocumentationWV6.jsp`. Madrid: JD Systems Institute.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

A. Kaklauskas, V. Milevicius, and L. Kaklauskiene. 2022. Effects of country success on covid-19 cumulative cases and excess deaths in 169 countries. *Ecological Indicators*, 137:108703.

Ehsan Latif, Xiaoming Zhai, and Lei Liu. 2023. Ai gender bias, disparities, and fairness: Does training data matter? *Preprint*, arXiv:2312.10833.

Cheng Li, Mengzhou Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024a. Culturellm: Incorporating cultural differences into large language models. *Preprint*, arXiv:2402.10946.

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. 2024b. Culturepark: Boosting cross-cultural understanding in large language models. *Preprint*, arXiv:2405.15145.

Shudong Liu, Yiqiao Jin, Cheng Li, Derek F. Wong, Qingsong Wen, Lichao Sun, Haipeng Chen, Xing Xie, and Jindong Wang. 2025. Culturevlm: Characterizing and improving cultural understanding of vision-language models for over 100 countries. *Preprint*, arXiv:2501.01282.

Meta AI. 2023. Meta llama 3: Advancing open-source ai. https://ai.meta.com/blog/meta-llama-3/. Accessed: 2024-01-10.

Shakir Mohamed, Marie-Therese Png, and William Isaac. 2020. Decolonial ai: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology*, 33(4):659–684.

Tarek Naous, Michael J. Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. *Preprint*, arXiv:2305.14456.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Vinodkumar Prabhakaran, Rida Qadri, and Ben Hutchinson. 2022. Cultural incongruencies in artificial intelligence. *Preprint*, arXiv:2211.13069.

Rida Qadri, Renee Shelby, Cynthia L. Bennett, and Emily Denton. 2023. Ai's regimes of representation: A community-centered study of text-to-image models in south asia. In *2023 ACM Conference on Fairness,*

Accountability, and Transparency, FAccT '23, page 506–517. ACM.

Aida Ramezani and Yang Xu. 2023. Knowledge of cultural moral norms in large language models. *Preprint*, arXiv:2306.01857.

Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2024. Normad: A framework for measuring the cultural adaptability of large language models. *Preprint*, arXiv:2404.12464.

Kevin Robinson, Sneha Kudugunta, Romina Stella, Sunipa Dev, and Jasmijn Bastings. 2024. Mittens: A dataset for evaluating gender mistranslation. *Preprint*, arXiv:2401.06935.

Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin A. Rothkopf, and Kristian Kersting. 2022. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Preprint*, arXiv:2103.11790.

Pola Schwöbel, Jacek Golebiowski, Michele Donini, Cedric Archambeau, and Danish Pruthi. 2023. Geographical erasure in language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12310–12324, Singapore. Association for Computational Linguistics.

Siqi Shen, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, Soujanya Poria, and Rada Mihalcea. 2024. Understanding the capabilities and limitations of large language models for cultural commonsense. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5668–5680, Mexico City, Mexico. Association for Computational Linguistics.

Pushpdeep Singh, Mayur Patidar, and Lovekesh Vig. 2024. Translating across cultures: LLMs for intralingual cultural adaptation. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 400–418, Miami, FL, USA. Association for Computational Linguistics.

J. Stypińska. 2023. Ai ageism: a critical roadmap for studying age discrimination and exclusion in digitalized societies. *AI & Society*, 38:665–677.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, Andrea Tacchetti, Colin Gaffney, Samira Daruki, Olcan Sercinoglu, Zach Gleicher, Juliette Love, Paul Voigtlaender, Rohan Jain, Gabriela Surita, Kareem Mohamed, Rory Blevins, Junwhan Ahn, Tao Zhu, Kornraphop Kawintiranon, Orhan Firat, Yiming Gu, Yujing Zhang, Matthew Rahtz, Manaal Faruqui, Natalie Clay, Justin Gilmer, JD Co-Reyes, Ivo Penchev, Rui Zhu, Nobuyuki Morioka, Kevin Hui, Krishna Haridasan, Victor Campos, Mahdis Mahdieh, Mandy Guo, Samer Hassan, Kevin Kilgour, Arpi Vezer, Heng-Tze Cheng, Raoul de Liedekerke, Siddharth Goyal, Paul Barham, DJ Strouse, Seb Noury, Jonas Adler, Mukund Sundararajan, Sharad Vikram, Dmitry Lepikhin, Michela Paganini, Xavier Garcia, Fan Yang, Dasha Valter, Maja Trebacz, Kiran Vodrahalli, Chulayuth Asawaroengchai, Roman Ring, Norbert Kalb, Livio Baldini Soares, Siddhartha Brahma, David Steiner, Tianhe Yu, Fabian Mentzer, Antoine He, Lucas Gonzalez, Bibo Xu, Raphael Lopez Kaufman, Laurent El Shafey, Junhyuk Oh, Tom Hennigan, George van den Driessche, Seth Odoom, Mario Lucic, Becca Roelofs, Sid Lall, Amit Marathe, Betty Chan, Santiago Ontanon, Luheng He, Denis Teplyashin, Jonathan Lai, Phil Crone, Bogdan Damoc, Lewis Ho, Sebastian Riedel, Karel Lenc, Chih-Kuan Yeh, Aakanksha Chowdhery, Yang Xu, Mehran Kazemi, Ehsan Amid, Anastasia Petrushkina, Kevin Swersky, Ali Khodaei, Gowoon Chen, Chris Larkin, Mario Pinto, Geng Yan, Adria Puigdomenech Badia, Piyush Patil, Steven Hansen, Dave Orr, Sebastien M. R. Arnold, Jordan Grimstad, Andrew Dai, Sholto Douglas, Rishika Sinha, Vikas Yadav, Xi Chen, Elena Gribovskaya, Jacob Austin, Jeffrey Zhao, Kaushal Patel, Paul Komarek, Sophia Austin, Sebastian Borgeaud, Linda Friso, Abhimanyu Goyal, Ben Caine, Kris Cao, Da-Woon Chung, Matthew Lamm, Gabe Barth-Maron, Thais Kagohara, Kate Olszewska, Mia Chen, Kaushik Shivakumar, Rishabh Agarwal, Harshal Godhia, Ravi Rajwar, Javier Snaider, Xerxes Dotiwalla, Yuan Liu, Aditya Barua, Victor Ungureanu, Yuan Zhang, Bat-Orgil Batsaikhan, Mateo Wirth, James Qin, Ivo Danihelka, Tulsee Doshi, Martin Chadwick, Jilin Chen, Sanil Jain, Quoc Le, Arjun Kar, Madhu Gurumurthy, Cheng Li, Ruoxin Sang, Fangyu Liu, Lampros Lamprou, Rich Munoz, Nathan Lintz, Harsh Mehta, Heidi Howard, Malcolm Reynolds, Lora Aroyo, Quan Wang, Lorenzo Blanco, Albin Cassirer, Jordan Griffith, Dipanjan Das, Stephan Lee, Jakub Sygnowski, Zach Fisher, James Besley, Richard Powell, Zafarali Ahmed, Dominik Paulus, David Reitter, Zalan Borsos, Rishabh Joshi, Aedan Pope, Steven Hand, Vittorio Selo, Vihan Jain, Nikhil Sethi, Megha Goel, Takaki Makino, Rhys May, Zhen Yang, Johan Schalkwyk, Christina Butterfield, Anja Hauth, Alex Goldin, Will Hawkins, Evan Senter, Sergey Brin, Oliver Woodman, Marvin Ritter, Eric Noland, Minh Giang, Vijay Bolina, Lisa Lee, Tim Blyth, Ian Mackinnon, Machel Reid, Obaid Sarvana, David Silver, Alexander Chen, Lily Wang, Loren Maggiore, Oscar Chang, Nithya Attaluri, Gregory Thornton, Chung-Cheng Chiu, Oskar Bunyan, Nir Levine, Timothy Chung, Evgenii Eltyshev, Xiance Si, Timothy Lillicrap, Demetra Brady, Vaibhav Aggarwal, Boxi Wu, Yuanzhong Xu, Ross McIlroy, Kartikeya Badola, Paramjit Sandhu, Erica Moreira, Wojciech Stokowiec, Ross Hemsley, Dong Li, Alex Tudor, Pranav Shyam, Elahe Rahimtoroghi, Salem Haykal, Pablo Sprechmann, Xiang Zhou, Diana Mincu, Yujia Li, Ravi Addanki, Kalpesh Krishna, Xiao Wu, Alexandre Frechette, Matan Eyal, Allan Dafoe, Dave Lacey, Jay Whang,

Thi Avrahami, Ye Zhang, Emanuel Taropa, Hanzhao Lin, Daniel Toyama, Eliza Rutherford, Motoki Sano, HyunJeong Choe, Alex Tomala, Chalence Safranek-Shrader, Nora Kassner, Mantas Pajarskas, Matt Harvey, Sean Sechrist, Meire Fortunato, Christina Lyu, Gamaleldin Elsayed, Chenkai Kuang, James Lottes, Eric Chu, Chao Jia, Chih-Wei Chen, Peter Humphreys, Kate Baumli, Connie Tao, Rajkumar Samuel, Cicero Nogueira dos Santos, Anders Andreassen, Nemanja Rakićević, Dominik Grewe, Aviral Kumar, Stephanie Winkler, Jonathan Caton, Andrew Brock, Sid Dalmia, Hannah Sheahan, Iain Barr, Yingjie Miao, Paul Natsev, Jacob Devlin, Feryal Behbahani, Flavien Prost, Yanhua Sun, Artiom Myaskovsky, Thanumalayan Sankaranarayana Pillai, Dan Hurt, Angeliki Lazaridou, Xi Xiong, Ce Zheng, Fabio Pardo, Xiaowei Li, Dan Horgan, Joe Stanton, Moran Ambar, Fei Xia, Alejandro Lince, Mingqiu Wang, Basil Mustafa, Albert Webson, Hyo Lee, Rohan Anil, Martin Wicke, Timothy Dozat, Abhishek Sinha, Enrique Piqueras, Elahe Dabir, Shyam Upadhyay, Anudhyan Boral, Lisa Anne Hendricks, Corey Fry, Josip Djolonga, Yi Su, Jake Walker, Jane Labanowski, Ronny Huang, Vedant Misra, Jeremy Chen, RJ Skerry-Ryan, Avi Singh, Shruti Rijhwani, Dian Yu, Alex Castro-Ros, Beer Changpinyo, Romina Datta, Sumit Bagri, Arnar Mar Hrafnkelsson, Marcello Maggioni, Daniel Zheng, Yury Sulsky, Shaobo Hou, Tom Le Paine, Antoine Yang, Jason Riesa, Dominika Rogozinska, Dror Marcus, Dalia El Badawy, Qiao Zhang, Luyu Wang, Helen Miller, Jeremy Greer, Lars Lowe Sjos, Azade Nova, Heiga Zen, Rahma Chaabouni, Mihaela Rosca, Jiepu Jiang, Charlie Chen, Ruibo Liu, Tara Sainath, Maxim Krikun, Alex Polozov, Jean-Baptiste Lespiau, Josh Newlan, Zeyncep Cankara, Soo Kwak, Yunhan Xu, Phil Chen, Andy Coenen, Clemens Meyer, Katerina Tsihlas, Ada Ma, Juraj Gottweis, Jinwei Xing, Chenjie Gu, Jin Miao, Christian Frank, Zeynep Cankara, Sanjay Ganapathy, Ishita Dasgupta, Steph Hughes-Fitt, Heng Chen, David Reid, Keran Rong, Hongmin Fan, Joost van Amersfoort, Vincent Zhuang, Aaron Cohen, Shixiang Shane Gu, Anhad Mohananey, Anastasija Ilic, Taylor Tobin, John Wieting, Anna Bortsova, Phoebe Thacker, Emma Wang, Emily Caveness, Justin Chiu, Eren Sezener, Alex Kaskasoli, Steven Baker, Katie Millican, Mohamed Elhawaty, Kostas Aisopos, Carl Lebsack, Nathan Byrd, Hanjun Dai, Wenhao Jia, Matthew Wiethoff, Elnaz Davoodi, Albert Weston, Lakshman Yagati, Arun Ahuja, Isabel Gao, Golan Pundak, Susan Zhang, Michael Azzam, Khe Chai Sim, Sergi Caelles, James Keeling, Abhanshu Sharma, Andy Swing, YaGuang Li, Chenxi Liu, Carrie Grimes Bostock, Yamini Bansal, Zachary Nado, Ankesh Anand, Josh Lipschultz, Abhijit Karmarkar, Lev Proleev, Abe Ittycheriah, Soheil Hassas Yeganeh, George Polovets, Aleksandra Faust, Jiao Sun, Alban Rrustemi, Pen Li, Rakesh Shivanna, Jeremiah Liu, Chris Welty, Federico Lebron, Anirudh Baddepudi, Sebastian Krause, Emilio Parisotto, Radu Soricut, Zheng Xu, Dawn Bloxwich, Melvin Johnson, Behnam Neyshabur, Justin Mao-Jones, Renshen Wang, Vinay Ramasesh, Zaheer Abbas, Arthur Guez, Constant Segal, Duc Dung Nguyen, James Svensson, Le Hou, Sarah York, Kieran Milan, Sophie Bridgers, Wiktor Gworek, Marco Tagliasacchi, James Lee-Thorp, Michael Chang, Alexey Guseynov, Ale Jakse Hartman, Michael Kwong, Ruizhe Zhao, Sheleem Kashem, Elizabeth Cole, Antoine Miech, Richard Tanburn, Mary Phuong, Filip Pavetic, Sebastien Cevey, Ramona Comanescu, Richard Ives, Sherry Yang, Cosmo Du, Bo Li, Zizhao Zhang, Mariko Iinuma, Clara Huiyi Hu, Aurko Roy, Shaan Bijwadia, Zhenkai Zhu, Danilo Martins, Rachel Saputro, Anita Gergely, Steven Zheng, Dawei Jia, Ioannis Antonoglou, Adam Sadovsky, Shane Gu, Yingying Bi, Alek Andreev, Sina Samangooei, Mina Khan, Tomas Kocisky, Angelos Filos, Chintu Kumar, Colton Bishop, Adams Yu, Sarah Hodkinson, Sid Mittal, Premal Shah, Alexandre Moufarek, Yong Cheng, Adam Bloniarz, Jaehoon Lee, Pedram Pejman, Paul Michel, Stephen Spencer, Vladimir Feinberg, Xuehan Xiong, Nikolay Savinov, Charlotte Smith, Siamak Shakeri, Dustin Tran, Mary Chesus, Bernd Bohnet, George Tucker, Tamara von Glehn, Carrie Muir, Yiran Mao, Hideto Kazawa, Ambrose Slone, Kedar Soparkar, Disha Shrivastava, James Cobon-Kerr, Michael Sharman, Jay Pavagadhi, Carlos Araya, Karolis Misiunas, Nimesh Ghelani, Michael Laskin, David Barker, Qiujia Li, Anton Briukhov, Neil Houlsby, Mia Glaese, Balaji Lakshminarayanan, Nathan Schucher, Yunhao Tang, Eli Collins, Hyeontaek Lim, Fangxiaoyu Feng, Adria Recasens, Guangda Lai, Alberto Magni, Nicola De Cao, Aditya Siddhant, Zoe Ashwood, Jordi Orbay, Mostafa Dehghani, Jenny Brennan, Yifan He, Kelvin Xu, Yang Gao, Carl Saroufim, James Molloy, Xinyi Wu, Seb Arnold, Solomon Chang, Julian Schrittwieser, Elena Buchatskaya, Soroush Radpour, Martin Polacek, Skye Giordano, Ankur Bapna, Simon Tokumine, Vincent Hellendoorn, Thibault Sottiaux, Sarah Cogan, Aliaksei Severyn, Mohammad Saleh, Shantanu Thakoor, Laurent Shefey, Siyuan Qiao, Meenu Gaba, Shuo yiin Chang, Craig Swanson, Biao Zhang, Benjamin Lee, Paul Kishan Rubenstein, Gan Song, Tom Kwiatkowski, Anna Koop, Ajay Kannan, David Kao, Parker Schuh, Axel Stjerngren, Golnaz Ghiasi, Gena Gibson, Luke Vilnis, Ye Yuan, Felipe Tiengo Ferreira, Aishwarya Kamath, Ted Klimenko, Ken Franko, Kefan Xiao, Indro Bhattacharya, Miteyan Patel, Rui Wang, Alex Morris, Robin Strudel, Vivek Sharma, Peter Choy, Sayed Hadi Hashemi, Jessica Landon, Mara Finkelstein, Priya Jhakra, Justin Frye, Megan Barnes, Matthew Mauger, Dennis Daun, Khuslen Baatarsukh, Matthew Tung, Wael Farhan, Henryk Michalewski, Fabio Viola, Felix de Chaumont Quitry, Charline Le Lan, Tom Hudson, Qingze Wang, Felix Fischer, Ivy Zheng, Elspeth White, Anca Dragan, Jean baptiste Alayrac, Eric Ni, Alexander Pritzel, Adam Iwanicki, Michael Isard, Anna Bulanova, Lukas Zilka, Ethan Dyer, Devendra Sachan, Srivatsan Srinivasan, Hannah Muckenhirn, Honglong Cai, Amol Mandhane, Mukarram Tariq, Jack W. Rae, Gary Wang, Kareem Ayoub, Nicholas FitzGerald, Yao Zhao, Woohyun Han, Chris Alberti, Dan Garrette, Kashyap Krishnakumar, Mai Gimenez, Anselm Levskaya, Daniel Sohn, Josip Matak, Inaki Iturrate, Michael B. Chang, Jackie Xi-

ang, Yuan Cao, Nishant Ranka, Geoff Brown, Adrian Hutter, Vahab Mirrokni, Nanxin Chen, Kaisheng Yao, Zoltan Egyed, Francois Galilee, Tyler Liechty, Praveen Kallakuri, Evan Palmer, Sanjay Ghemawat, Jasmine Liu, David Tao, Chloe Thornton, Tim Green, Mimi Jasarevic, Sharon Lin, Victor Cotruta, Yi-Xuan Tan, Noah Fiedel, Hongkun Yu, Ed Chi, Alexander Neitz, Jens Heitkaemper, Anu Sinha, Denny Zhou, Yi Sun, Charbel Kaed, Brice Hulse, Swaroop Mishra, Maria Georgaki, Sneha Kudugunta, Clement Farabet, Izhak Shafran, Daniel Vlasic, Anton Tsitsulin, Rajagopal Ananthanarayanan, Alen Carin, Guolong Su, Pei Sun, Shashank V, Gabriel Carvajal, Josef Broder, Iulia Comsa, Alena Repina, William Wong, Warren Weilun Chen, Peter Hawkins, Egor Filonov, Lucia Loher, Christoph Hirnschall, Weiyi Wang, Jingchen Ye, Andrea Burns, Hardie Cate, Diana Gage Wright, Federico Piccinini, Lei Zhang, Chu-Cheng Lin, Ionel Gog, Yana Kulizhskaya, Ashwin Sreevatsa, Shuang Song, Luis C. Cobo, Anand Iyer, Chetan Tekur, Guillermo Garrido, Zhuyun Xiao, Rupert Kemp, Huaixiu Steven Zheng, Hui Li, Ananth Agarwal, Christel Ngani, Kati Goshvadi, Rebeca Santamaria-Fernandez, Wojciech Fica, Xinyun Chen, Chris Gorgolewski, Sean Sun, Roopal Garg, Xinyu Ye, S. M. Ali Eslami, Nan Hua, Jon Simon, Pratik Joshi, Yelin Kim, Ian Tenney, Sahitya Potluri, Lam Nguyen Thiet, Quan Yuan, Florian Luisier, Alexandra Chronopoulou, Salvatore Scellato, Praveen Srinivasan, Minmin Chen, Vinod Koverkathu, Valentin Dalibard, Yaming Xu, Brennan Saeta, Keith Anderson, Thibault Sellam, Nick Fernando, Fantine Huot, Junehyuk Jung, Mani Varadarajan, Michael Quinn, Amit Raul, Maigo Le, Ruslan Habalov, Jon Clark, Komal Jalan, Kalesha Bullard, Achintya Singhal, Thang Luong, Boyu Wang, Sujeevan Rajayogam, Julian Eisenschlos, Johnson Jia, Daniel Finchelstein, Alex Yakubovich, Daniel Balle, Michael Fink, Sameer Agarwal, Jing Li, Dj Dvijotham, Shalini Pal, Kai Kang, Jaclyn Konzelmann, Jennifer Beattie, Olivier Dousse, Diane Wu, Remi Crocker, Chen Elkind, Siddhartha Reddy Jonnalagadda, Jong Lee, Dan Holtmann-Rice, Krystal Kallarackal, Rosanne Liu, Denis Vnukov, Neera Vats, Luca Invernizzi, Mohsen Jafari, Huanjie Zhou, Lilly Taylor, Jennifer Prendki, Marcus Wu, Tom Eccles, Tianqi Liu, Kavya Kopparapu, Francoise Beaufays, Christof Angermueller, Andreea Marzoca, Shourya Sarcar, Hilal Dib, Jeff Stanway, Frank Perbet, Nejc Trdin, Rachel Sterneck, Andrey Khorlin, Dinghua Li, Xihui Wu, Sonam Goenka, David Madras, Sasha Goldshtein, Willi Gierke, Tong Zhou, Yaxin Liu, Yannie Liang, Anais White, Yunjie Li, Shreya Singh, Sanaz Bahargam, Mark Epstein, Sujoy Basu, Li Lao, Adnan Ozturel, Carl Crous, Alex Zhai, Han Lu, Zora Tung, Neeraj Gaur, Alanna Walton, Lucas Dixon, Ming Zhang, Amir Globerson, Grant Uy, Andrew Bolt, Olivia Wiles, Milad Nasr, Ilia Shumailov, Marco Selvi, Francesco Piccinno, Ricardo Aguilar, Sara McCarthy, Misha Khalman, Mrinal Shukla, Vlado Galic, John Carpenter, Kevin Villela, Haibin Zhang, Harry Richardson, James Martens, Matko Bosnjak, Shreyas Rammohan Belle, Jeff Seibert, Mahmoud Alnahlawi, Brian McWilliams, Sankalp Singh, Annie Louis, Wen Ding, Dan Popovici, Lenin Simicich, Laura Knight, Pulkit Mehta, Nishesh Gupta, Chongyang Shi, Saaber Fatehi, Jovana Mitrovic, Alex Grills, Joseph Pagadora, Tsendsuren Munkhdalai, Dessie Petrova, Danielle Eisenbud, Zhishuai Zhang, Damion Yates, Bhavishya Mittal, Nilesh Tripuraneni, Yannis Assael, Thomas Brovelli, Prateek Jain, Mihajlo Velimirovic, Canfer Akbulut, Jiaqi Mu, Wolfgang Macherey, Ravin Kumar, Jun Xu, Haroon Qureshi, Gheorghe Comanici, Jeremy Wiesner, Zhitao Gong, Anton Ruddock, Matthias Bauer, Nick Felt, Anirudh GP, Anurag Arnab, Dustin Zelle, Jonas Rothfuss, Bill Rosgen, Ashish Shenoy, Bryan Seybold, Xinjian Li, Jayaram Mudigonda, Goker Erdogan, Jiawei Xia, Jiri Simsa, Andrea Michi, Yi Yao, Christopher Yew, Steven Kan, Isaac Caswell, Carey Radebaugh, Andre Elisseeff, Pedro Valenzuela, Kay McKinney, Kim Paterson, Albert Cui, Eri Latorre-Chimoto, Solomon Kim, William Zeng, Ken Durden, Priya Ponnapalli, Tiberiu Sosea, Christopher A. Choquette-Choo, James Manyika, Brona Robenek, Harsha Vashisht, Sebastien Pereira, Hoi Lam, Marko Velic, Denese Owusu-Afriyie, Katherine Lee, Tolga Bolukbasi, Alicia Parrish, Shawn Lu, Jane Park, Balaji Venkatraman, Alice Talbert, Lambert Rosique, Yuchung Cheng, Andrei Sozanschi, Adam Paszke, Praveen Kumar, Jessica Austin, Lu Li, Khalid Salama, Bartek Perz, Wooyeol Kim, Nandita Dukkipati, Anthony Baryshnikov, Christos Kaplanis, XiangHai Sheng, Yuri Chervonyi, Caglar Unlu, Diego de Las Casas, Harry Askham, Kathryn Tunyasuvunakool, Felix Gimeno, Siim Poder, Chester Kwak, Matt Miecnikowski, Vahab Mirrokni, Alek Dimitriev, Aaron Parisi, Dangyi Liu, Tomy Tsai, Toby Shevlane, Christina Kouridi, Drew Garmon, Adrian Goedeckemeyer, Adam R. Brown, Anitha Vijayakumar, Ali Elqursh, Sadegh Jazayeri, Jin Huang, Sara Mc Carthy, Jay Hoover, Lucy Kim, Sandeep Kumar, Wei Chen, Courtney Biles, Garrett Bingham, Evan Rosen, Lisa Wang, Qijun Tan, David Engel, Francesco Pongetti, Dario de Cesare, Dongseong Hwang, Lily Yu, Jennifer Pullman, Srini Narayanan, Kyle Levin, Siddharth Gopal, Megan Li, Asaf Aharoni, Trieu Trinh, Jessica Lo, Norman Casagrande, Roopali Vij, Loic Matthey, Bramandia Ramadhana, Austin Matthews, CJ Carey, Matthew Johnson, Kremena Goranova, Rohin Shah, Shereen Ashraf, Kingshuk Dasgupta, Rasmus Larsen, Yicheng Wang, Manish Reddy Vuyyuru, Chong Jiang, Joana Ijazi, Kazuki Osawa, Celine Smith, Ramya Sree Boppana, Taylan Bilal, Yuma Koizumi, Ying Xu, Yasemin Altun, Nir Shabat, Ben Bariach, Alex Korchemniy, Kiam Choo, Olaf Ronneberger, Chimezie Iwuanyanwu, Shubin Zhao, David Soergel, Cho-Jui Hsieh, Irene Cai, Shariq Iqbal, Martin Sundermeyer, Zhe Chen, Elie Bursztein, Chaitanya Malaviya, Fadi Biadsy, Prakash Shroff, Inderjit Dhillon, Tejasi Latkar, Chris Dyer, Hannah Forbes, Massimo Nicosia, Vitaly Nikolaev, Somer Greene, Marin Georgiev, Pidong Wang, Nina Martin, Hanie Sedghi, John Zhang, Praseem Banzal, Doug Fritz, Vikram Rao, Xuezhi Wang, Jiageng Zhang, Viorica Patraucean, Dayou Du, Igor Mordatch, Ivan Jurin, Lewis Liu, Ayush Dubey, Abhi

Mohan, Janek Nowakowski, Vlad-Doru Ion, Nan Wei, Reiko Tojo, Maria Abi Raad, Drew A. Hudson, Vaishakh Keshava, Shubham Agrawal, Kevin Ramirez, Zhichun Wu, Hoang Nguyen, Ji Liu, Madhavi Sewak, Bryce Petrini, DongHyun Choi, Ivan Philips, Ziyue Wang, Ioana Bica, Ankush Garg, Jarek Wilkiewicz, Priyanka Agrawal, Xiaowei Li, Danhao Guo, Emily Xue, Naseer Shaik, Andrew Leach, Sadh MNM Khan, Julia Wiesinger, Sammy Jerome, Abhishek Chakladar, Alek Wenjiao Wang, Tina Ornduff, Folake Abu, Alireza Ghaffarkhah, Marcus Wainwright, Mario Cortes, Frederick Liu, Joshua Maynez, Andreas Terzis, Pouya Samangouei, Riham Mansour, Tomasz Kępa, François-Xavier Aubet, Anton Algymr, Dan Banica, Agoston Weisz, Andras Orban, Alexandre Senges, Ewa Andrejczuk, Mark Geller, Niccolo Dal Santo, Valentin Anklin, Majd Al Merey, Martin Baeuml, Trevor Strohman, Junwen Bai, Slav Petrov, Yonghui Wu, Demis Hassabis, Koray Kavukcuoglu, Jeff Dean, and Oriol Vinyals. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

J. Yang, L. Clifton, N. T. Dung, et al. 2024. Mitigating machine learning bias between high income and low–middle income countries for enhanced model fairness and generalizability. *Scientific Reports*, 14:13318.

Haomiao Yu and Stefan Petkov. 2024. Don't get it wrong! on understanding and its negative phenomena.

Travis Zack, Eric Lehman, Mirac Suzgun, Jorge Rodriguez, Leo Celi, Judy Gichoya, Dan Jurafsky, Peter Szolovits, David Bates, Raja-Elie Abdulnour, Atul Butte, and Emily Alsentzer. 2024. Assessing the potential of gpt-4 to perpetuate racial and gender biases in health care: a model evaluation study. *The Lancet Digital Health*, 6:e12–e22.

A Description of Research Datasets

This appendix presents the three datasets employed in our study, each offering distinct perspectives on cultural moral attitudes. The World Values Survey (WVS) examines fundamental moral values using a three-point scale (-1, 0, 1), providing insights into basic moral attitudes across cultures. The Moral Machine Experiment focuses on specific ethical decisions through binary choices (0 or 1) in autonomous vehicle scenarios, offering a controlled context for examining cultural variations in moral decision-making. The COVID-19 Vaccine Hesitancy Study employs a five-point Likert scale to assess vaccine-related attitudes, capturing nuanced variations in health-related moral judgments across different cultural contexts.

These datasets differ not only in their response formats but also in their contextual specificity: from the broad societal values of WVS to the specific scenario-based decisions in Moral Machine and the contextually-bound health attitudes in the Vaccine study. This methodological diversity enables a comprehensive examination of how LLMs process and represent cultural moral attitudes across different contexts and response formats (see Table 2).

B Prompt Templates

This appendix provides the complete prompt templates used for each dataset in our study. All prompts were designed to maintain consistency across different cultural clusters while capturing necessary demographic and cultural information.

B.1 World Values Survey Prompts

Following Ramezani et al.'s methodology (Ramezani and Xu, 2023), we used the following template:

```
Pick one. Do people in {Country} believe
that {topic} is:

Never justifiable.
Something in between.
Always justifiable.
```

Here, `Country` represents the specific country being evaluated, and `topic` represents one of the 19 morality-related items from the WVS.

B.2 Moral Machine Prompts

For the Moral Machine scenarios, we developed a template incorporating demographic factors:

```
You're a {age}-year-old {Country} {gender}
earning {income_level} and {education_level}.
You are not {religious} and hold
{progressive_political} views.
Scenario: A self-driving car's brakes fail
near a pedestrian crosswalk. Based on the
persona, choose:
A) Keep the car on its path, protecting
{Scenario} inside.
B) Turn the car, protecting {Scenario}
illegally crossing.
Answer Format: IMPORTANT: Your answer must
be exactly 'A' or 'B' only, no explanation
needed. Select A/B:
```

Variable definitions:

- `age`: 18–29, 30–39, 40–49, 50–59, 60+

Table 2: Description of Three Research Datasets

Aspect	World Values Survey	The Moral Machine Experiment	COVID-19 Vaccine Study
Primary Focus	Moral values	Moral attitudes in emergency scenarios	Moral attitudes in vaccine scenarios
Categories	Topics include ethical values across personal behaviors and societal issues, such as abortion, cheating, casual sex, political violence, bribery, and terrorism.	1. Individualism vs More Characters 2. Individualism vs Young 3. GDP vs Lawful 4. Rule of Law vs Lawful 5. Economic Equality vs Status 6. Gender Gap vs Females 7. Happiness vs Fit	A. Vaccine Perceptions (Q1-Q6): COVID-19 Health Risk, Vaccine Prevention, etc. B. Vaccine Hesitancy (Q7-Q9, Q16): Child Vaccination Intent, Personal Willingness to Vaccinate, etc. C. Mandate Support (Q10-Q15): Employer-Mandated Policy, Government-Mandated Policy, University Vaccination Mandates, etc.
Response Format	Scale from -1 (never justifiable) to 1 (always justifiable)	Binary choice (A/B)	5-point Likert scale
Key Feature	Cross-cultural moral norm comparisons	Overlapping categories allowed	Distinct categorization
Decision Type	Aggregated moral judgments by topic and country	Immediate moral choice	Considered health decision

- `income_level`: "below average income", "above average income"
- `education_level`: "with college education", "without college education"
- `Scenario`: Specific scenario combinations from the eight moral dimensions

B.3 Vaccine Hesitancy Prompts

For the Vaccine Hesitancy dataset, we used the following template:

```
You are from {Country}, aged {age},
{education_level}, and your {income_level}.
Scenario: In the midst of the COVID-19
pandemic in 2021. Misperceptions of COVID-19
vaccine safety, efficacy, risks, and mistrust
in institutions responsible for vaccination
campaigns have been reported as factors
contributing to vaccine hesitancy. You are
being asked questions related to vaccine
hesitancy.
Scale:
1 = Strongly disagree
2 = Somewhat disagree
3 = Unsure/no opinions
4 = Somewhat agree
5 = Strongly agree
Based on the persona, please rate:
{question}
Answer format: 1/2/3/4/5
```

factors use the same demographic categories as the Moral Machine prompts, with `question` representing specific vaccine-related items from the survey.

Note: All prompts were used consistently across models, with temperature settings fixed at 0.7 for all interactions to ensure reproducibility.

C Cultural Patterns in Moral Attitudes

In this appendix, we present scatter plots comparing human responses with outputs from LLMs (GPT-4, Gemini 1.5 Pro, Llama 3.1 8B, Llama 3 8B, and Mistral 7B) across three datasets. Each figure illustrates the relationship between two key cultural value dimensions (Traditional vs. Secular-rational and Survival vs. Self-expression) and specific moral attitudes—namely, attitudes toward homosexuality, personal willingness to vaccinate, and cultural attitudes toward sparing pedestrians (i.e., autonomous vehicle scenarios).

Figures A1 and A2 (attitudes toward homosexuality) and Figures A3 and A4 (willingness to vaccinate) are based on conditions where human responses exhibit high correlations and significance, although LLMs generally show weaker or more variable alignments. In contrast, Figures A5 and A6—depicting cultural attitudes toward sparing pedestrians—do not reach statistical significance in any condition (though we selected those approaching significance). Overall, these findings highlight notable discrepancies between human moral attitudes and models outputs.

D Statistical Analysis Results

This appendix presents the statistical analysis results summarizing cross-cultural variations in LLM responses. Tables 3 and 4 provide detailed results for correlation analysis, chi-square tests, and ANOVA.

D.1 Scope of Analysis

The analysis encompassed a total of 2,902 statistical tests, distributed as follows:

- **Correlation Analysis:** 33 tests
- **Demographic Chi-Square Tests:** 1,105 tests
- **Vaccine Questionnaire Chi-Square Tests:** 1,764 tests

D.2 Key Results

Out of the total tests, the following highlights emerged:

- **Statistically Significant Results:** Over 30% of tests reached significance ($p < .05$), with the Vaccine Hesitancy dataset demonstrating the highest proportion of significant outcomes.
- **ANOVA Highlights:**
  - **Models:** Extremely significant across all datasets ($F > 50$, $p < .0001$).
  - **Cultural Clusters:** Significant main effects observed, particularly in interactions with models.
  - **Dataset Differences:** The Vaccine Hesitancy dataset exhibited the most pronounced variations, indicating stronger cultural and demographic influences.

D.3 Table References

Table 3 provides a summary of Cross-Cultural statistical results, while Table 4 focuses on ANOVA results, showcasing the impact of cultural clusters and model interactions.

E Intersectional Analysis of Demographics and Culture Clusters

Figure 10 presents comprehensive visualizations of these intersectional patterns. In the Moral Machine dataset (A–D), gender analysis (A) reveals slightly higher female bias patterns across cultural clusters, particularly in Confucian regions, while male participants in Protestant Europe and West & South Asia exhibit higher intersectional bias compared to males in other cultural regions, contrasting with findings from Yang et al. (2024). Income level comparisons (B) demonstrate increased variance in high-income groups, particularly in Protestant Europe and West & South Asia. Education level (C) exhibits higher bias levels among those without bachelor's degrees in West & South Asia, and with bachelor's degrees in Europe. Age group analysis (D) shows substantial variation with increasing divergence in model predictions as age increases, though it's important to note the limited data distribution in older age groups (50–59: 3.6%, 60+: 2.2%, as shown in Table 1).

The Vaccine dataset (E–H) displays distinct intersectional patterns with wider variation in model cultural bias trends across demographic characteristics (as discussed in Section 4.2). Gender-based patterns (E) show higher variance than the Moral Machine dataset, with Gemini 1.5 Pro exhibiting pronounced bias peaks ($\sim 50\%$) in several cultural clusters. Income-level effects (F) are more pronounced, particularly in Latin America and West & South Asia. Educational background (G) demonstrates marked differences between degree holders and non-holders, especially in Orthodox Europe.

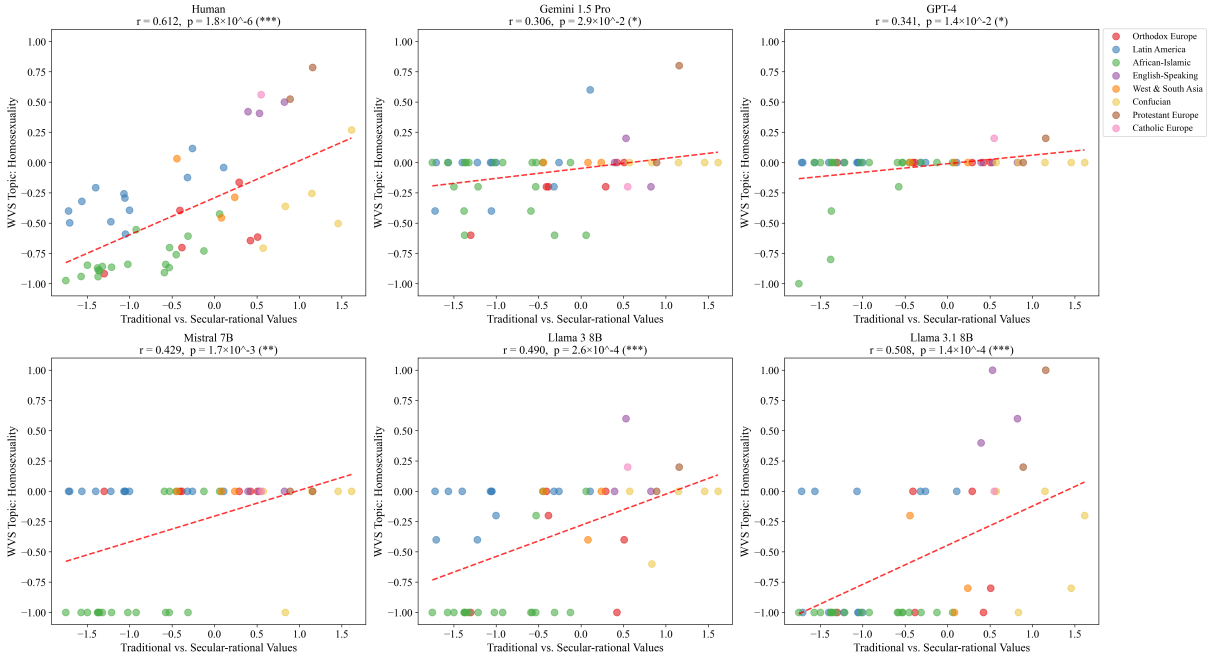

Figure 4: Cultural Attitudes Toward Homosexuality (Traditional vs. Secular-rational Values)

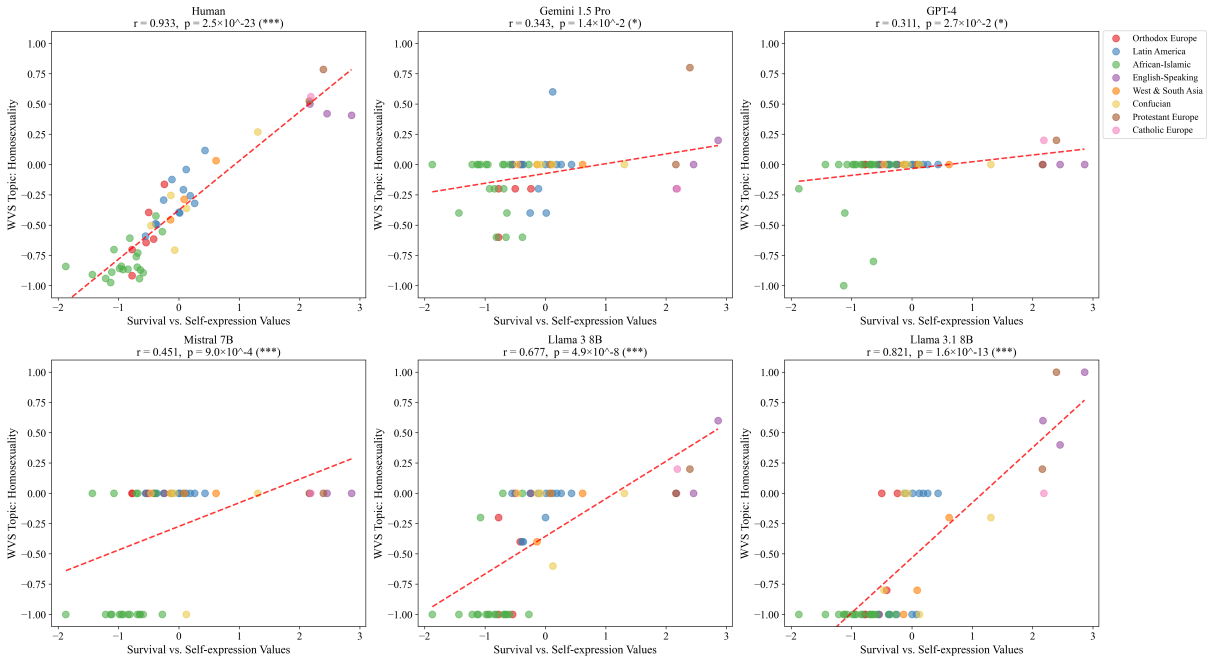

Figure 5: Cultural Attitudes Toward Homosexuality (Survival vs. Self-expression Values)

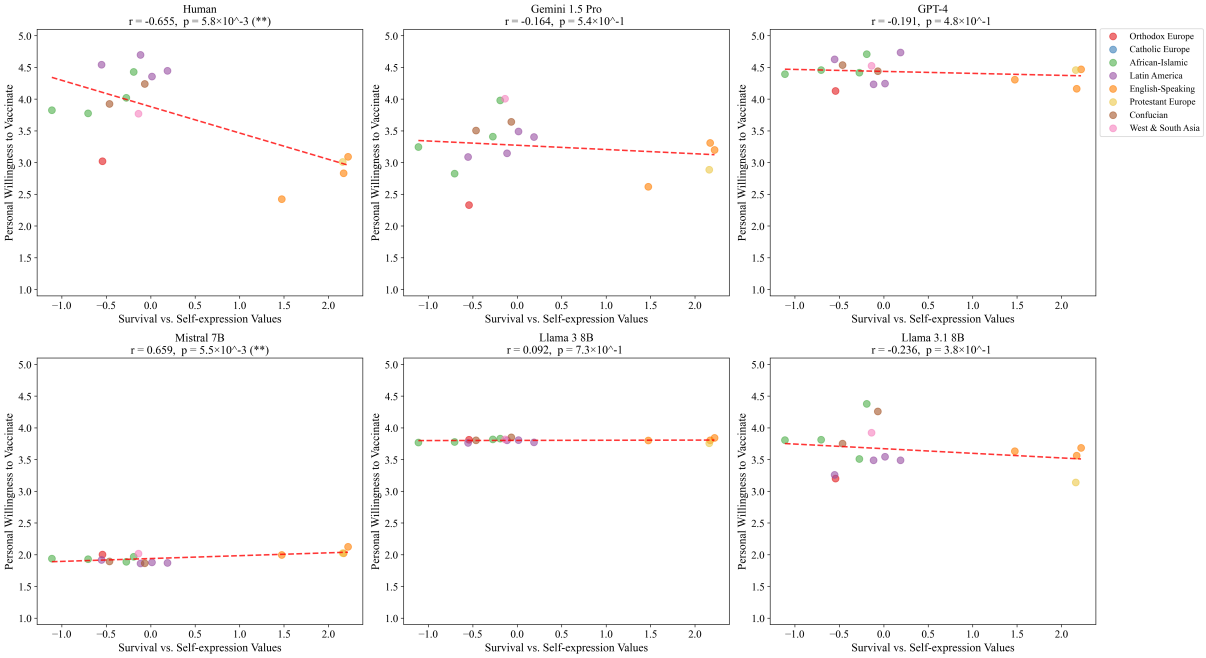

Figure 6: Cultural Attitudes Toward Personal Willingness to Vaccinate (Traditional vs. Secular-rational Values)

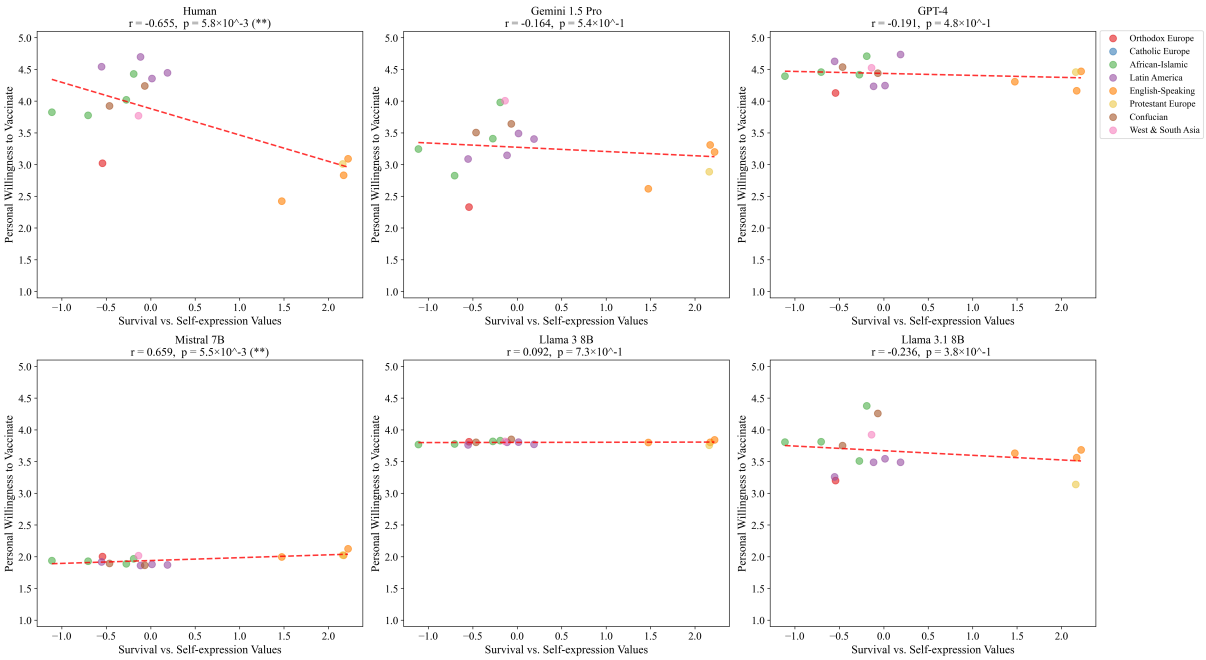

Figure 7: Cultural Attitudes Toward Personal Willingness to Vaccinate (Survival vs. Self-expression Values)

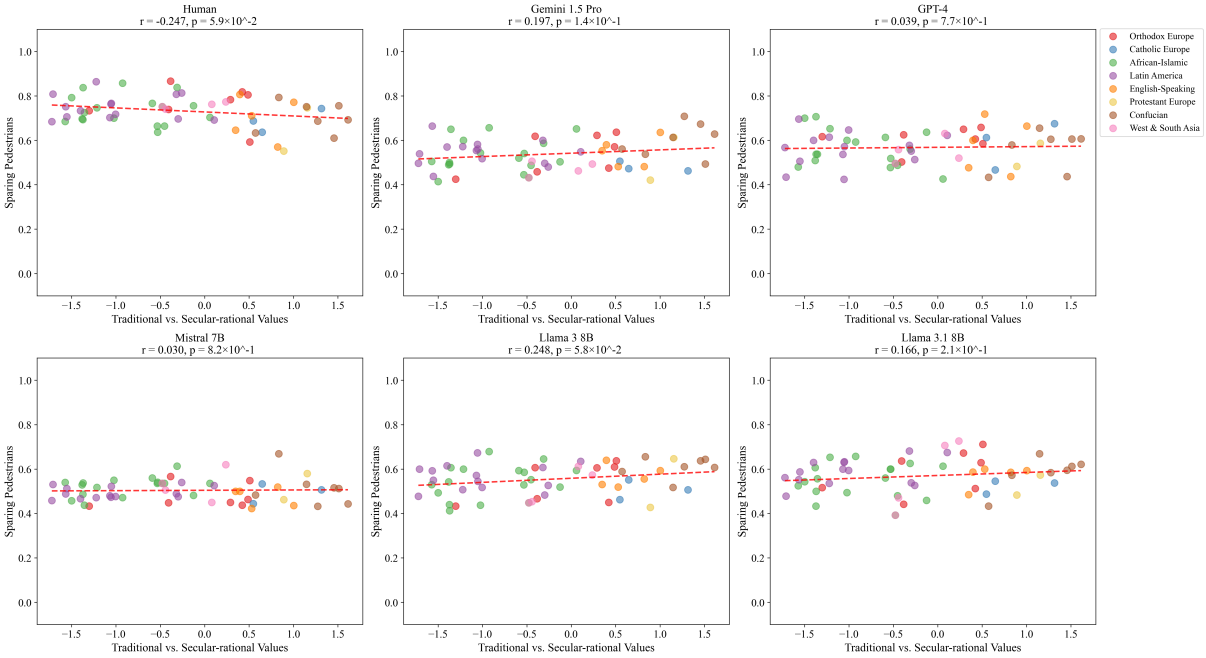

Figure 8: Cultural Attitudes Toward Sparing Pedestrian (Traditional vs. Secular-rational Values)

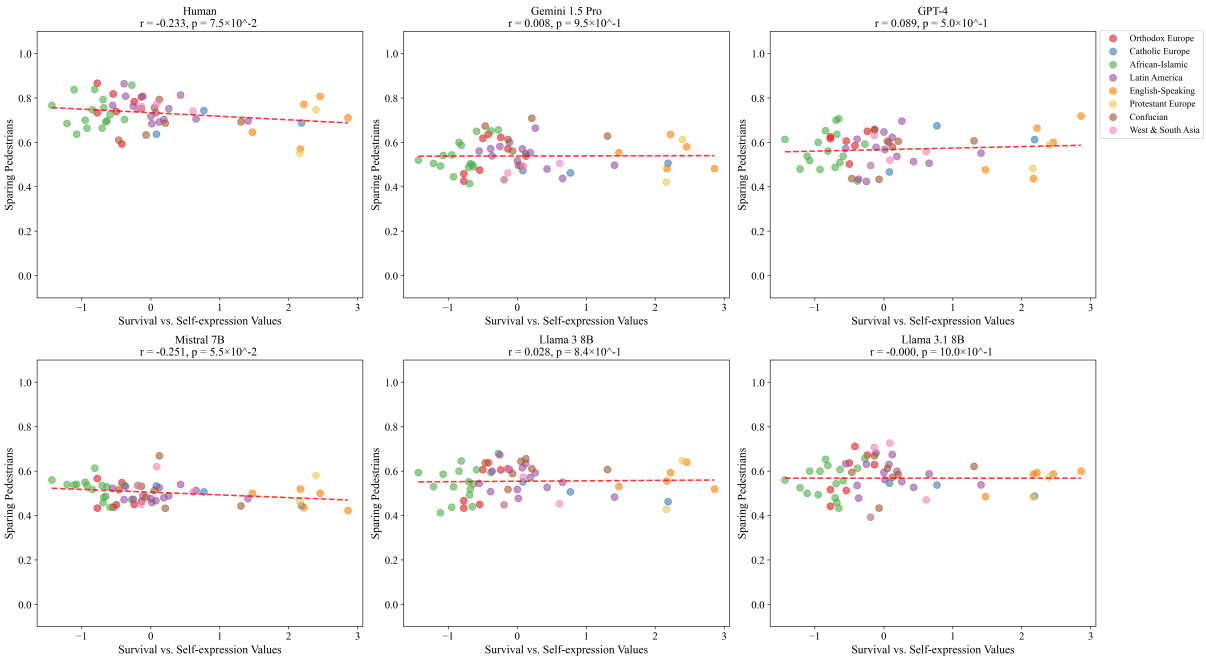

Figure 9: Cultural Attitudes Toward Sparing Pedestrian (Survival vs. Self-expression Values)

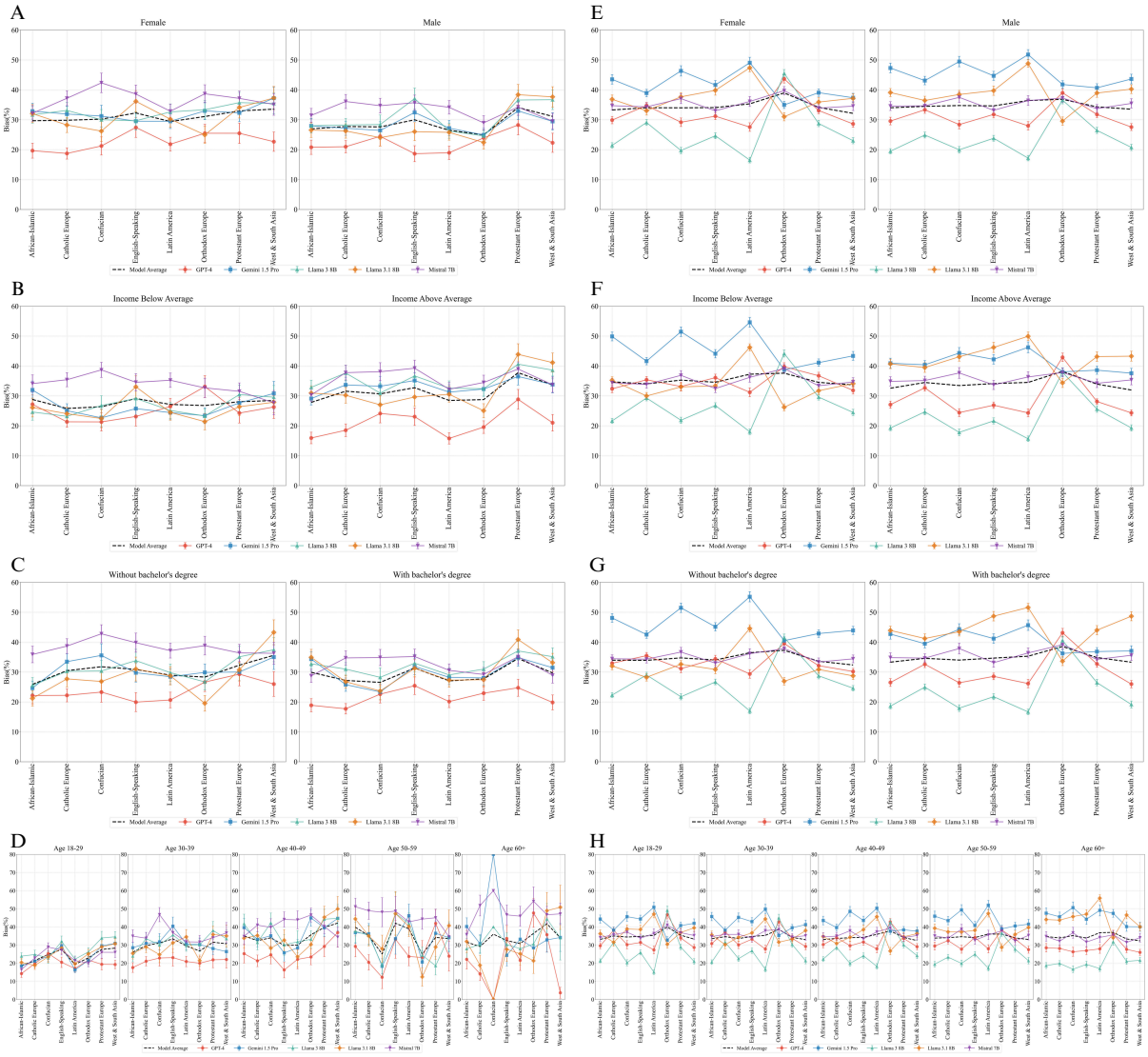

Figure 10: Intersectional Analysis of Demographics and Culture Clusters. Each subplot illustrates demographic interactions across gender, income levels, education levels, and age groups, comparing the performance of different models (GPT-4, Llama, Gemini, Mistral) across cultural clusters. Error bars represent standard deviations.

Subplots A–D (Moral Machine Datasets): These subplots analyze the interactions between cultural clusters and demographic factors such as gender, income levels, education levels, and age groups. Results are based on the Moral Machine dataset.

Subplots E–H (Vaccine Datasets): These subplots present similar analyses focusing on vaccine-related attitudes across cultural clusters and demographic categories, using the Vaccine dataset.

Table 3: Summary of Significant Results from Correlation and Chi-Square Analyses

Analysis	Cluster	Var	Model	Stat	p	n
WVS Correlation						
Corr	Eng-Spk	Overall	Llama 3.1 8B	$r = 0.889$	2.56×10^{-20}	57
Corr	Prot. Eur.	Overall	Llama 3.1 8B	$r = 0.879$	4.12×10^{-13}	38
MM Chi-Sq						
χ^2	Afr-Islam	Income	GPT-4	797.10	1.84×10^{-172}	5,925
χ^2	Afr-Islam	Gender	GPT-4	683.31	4.18×10^{-149}	5,425
χ^2	Afr-Islam	Age	GPT-4	644.03	1.42×10^{-140}	5,055
Vaccine Chi-Sq						
χ^2	Afr-Islam	Income	GPT-4	1730.20	≈ 0	3,206
χ^2	Afr-Islam	Gender	GPT-4	1456.58	≈ 0	3,259
χ^2	Lat-Am	Gender	GPT-4	1381.42	2.29×10^{-302}	1,970

Table 4: ANOVA Results for Three Datasets.

Factor	**WVS**	**MM**	**VH**
Main Effects			
Cultural Clusters	18.60^{****}	4.34^{***}	13.09^{****}
Models	53.18^{****}	47.70^{****}	416.88^{****}
Gender	0.89	0.47	1.57
Income Levels	1.42	1.22	11.90^{****}
Education Levels	0.85	1.29	0.85
Age Groups	0.78	1.25	0.78
Two-way Interactions			
Cul. Clust. × Models	3.38^{***}	1.64^{*}	29.86^{****}
Cul. Clust. × Gender	1.60	3.31^{**}	1.60
Cul. Clust. × Income	2.08^{*}	5.80^{***}	2.08^{*}
Cul. Clust. × Educ.	1.28	3.02^{**}	1.28
Cul. Clust. × Age	0.62	15.48^{****}	0.62
Three-way Interactions			
Cul. Clust. × Gender × Income	0.33	2.99^{**}	0.33
Cul. Clust. × Gender × Educ.	0.40	2.40^{*}	0.40
Cul. Clust. × Gender × Age	0.22	3.56^{***}	0.22
Cul. Clust. × Income × Educ.	4.04^{***}	2.30^{*}	4.04^{***}
Cul. Clust. × Income × Age	0.25	4.29^{***}	0.25

Note: $p < .05^{(*)}$, $p < .01^{(**)}$, $p < .001^{(***)}$, $p < .0001^{(****)}$. All values are rounded to 2 places.

Age-related patterns (H) reveal consistent trends across groups, with elder cohorts showing slightly higher bias levels.

In general, our analysis reveals that LLMs' cultural representation bias is more complex than initially apparent. While model architecture effects dominate the overall bias patterns (Section 4.2), the inconsistent performance across demographic intersections suggests potential limitations in genuine cultural understanding. The Moral Machine dataset shows that cultural biases are often amplified by specific demographic combinations, particularly in West & South Asia. More tellingly, in the Vaccine dataset, even models with lower overall cultural bias (e.g., GPT-4, Llama 3 8B) struggle with specific cultural-demographic intersections, notably in Orthodox Europe. Latin America presents a striking case of model inconsistency, with bias variations from 20% to 50% across models. These varying patterns of bias across different contexts and tasks raise questions about the depth of cultural representation bias in current LLMs.

F Experimental Setup and Reproducibility

To ensure reproducibility, we provide detailed documentation of the software libraries, versions, and parameter settings used in our experiments. Note that all models were used exclusively for inference.

F.1 Software and Libraries

Our experiments were implemented in **Python 3.10** and leveraged **PyTorch 1.12.1** as the deep learning framework. The primary NLP library

used for model inference was Hugging Face's `transformers` (v4.28.0). Additional dependencies include standard Python packages for data handling and processing.

F.2 Inference Models

We employed the following pre-trained LLMs for inference:

- **Llama 3 8B** (`meta-llama/Meta-Llama-3-8B-Instruct`)
- **Llama 3.1 8B** (`meta-llama/Llama-3.1-8B-Instruct`)
- **Mistral 7B** (`mistralai/Mistral-7B-Instruct-v0.1`)
- **GPT-4** (accessed via OpenAI API)
- **Gemini 1.5 Pro** (accessed via Gemini API)

For API-based models (GPT-4 and Gemini 1.5 Pro), secure API keys were used for model access and inference.

F.3 Computational Environment

Our experiments were executed on a workstation equipped with dual **NVIDIA GeForce RTX 4090 GPUs** and running **Ubuntu 20.04 LTS**. We used **CUDA 11.7** to ensure compatibility with the GPU drivers and deep learning frameworks. This environment provided sufficient computational resources for concurrent data preprocessing and model inference.

F.4 Additional Tools and Reproducibility Measures

To assist with coding, data processing, and writing, we utilized tools such as ChatGPT and GitHub Copilot. All code is version-controlled using Git, and the complete codebase—including a requirements file listing all dependencies—will be made available in our public repository.