Computational Linguistics in the Netherlands Journal 15 (2026) 59-77 Submitted 01/2026; Published 06/2026

,

Exploring Cultural Variations in Moral Judgments with Large Language Models

Hadi Mohammadi* H.MOHAMMADI@UU.NL
Robert A. Bagheri* A.BAGHERI@UU.NL

*Department of Methodology and Statistics, Utrecht University, Utrecht, The Netherlands

Abstract

Large Language Models (LLMs) have shown strong performance across many tasks, but their ability to capture culturally diverse moral values remains unclear. In this paper, we examine whether LLMs mirror variations in moral attitudes reported by the World Values Survey (WVS) and the Pew Research Center's Global Attitudes Survey (PEW). We compare smaller monolingual and multilingual models (GPT-2, OPT, BLOOMZ, and Qwen) with recent instruction-tuned models (GPT-4o, GPT-4o-mini, Gemma-2-9b-it, and Llama-3.3-70B-Instruct). Using log-probability-based *moral justifiability* scores, we correlate each model's outputs with survey data covering a broad set of ethical topics. Our results show that many earlier or smaller models often produce near-zero or negative correlations with human judgments. In contrast, advanced instruction-tuned models achieve substantially higher positive correlations, suggesting they better reflect real-world moral attitudes. We provide a detailed regional analysis revealing that models align better with Western, Educated, Industrialized, Rich, and Democratic (W.E.I.R.D.) nations than with other regions. While scaling model size and using instruction tuning improves alignment with cross-cultural moral norms, challenges remain for certain topics and regions. We discuss these findings in relation to bias analysis, training data diversity, information retrieval implications, and strategies for improving the cultural sensitivity of LLMs.

1. Introduction

Over the past few years, Large Language Models (LLMs) have gained prominence in both academic and public discussions (Bender et al. 2021). Consider this example: when asked about the moral acceptability of divorce, an LLM might predict similar attitudes for Sweden and Saudi Arabia, yet survey data reveals nearly opposite positions on a normalized scale. Such blind spots matter as LLMs increasingly power content moderation, search engines, and decision-support systems globally. Advances in model performance have made LLMs appealing for diverse applications, such as social media content moderation, chatbots, content creation, real-time translation, search engines, recommendation systems, and automated decision-making. While modern LLMs (e.g., GPT-4) show strong performance, a critical concern is how these models may inherit biases, including gender, racial, or cultural biases, from their training data. LLMs can easily absorb such biases because they learn from large-scale text corpora containing entrenched stereotypes (Stańczak and Augenstein 2021, Karpouzis 2024). Recent work further examines whether LLMs capture cross-cultural moral judgments, reporting partial alignment alongside systematic blind spots across topics and regions (Mohammadi et al. 2025d). Our study complements these findings by correlating model-derived justifiability scores with the World Values Survey (WVS) and the Pew Research Center's Global Attitudes Survey (PEW) across a broader mix of model families and elicitation settings.

These biases raise concerns about fairness, particularly in contexts requiring moral judgments. If an LLM is trained mostly on data that negatively or inaccurately portrays certain cultural groups, it may repeat that bias in its responses. As these models become more widespread and globally deployed, the risk of perpetuating cultural biases increases, especially when moral perspectives diverge from common norms or survey results. Recent research indicates that current LLMs often

©2026 Hadi Mohammadi & Robert A. Bagheri.

exhibit a Western-centric bias (Adilazuarda et al. 2024), underscoring the need to evaluate their cross-cultural validity. A large-scale evaluation across 107 countries found that cultural prompting reduces bias for approximately 71–81% of countries but fails for the remaining 19–29% (Tao et al. 2024).

It is crucial to determine whether LLMs accurately mirror the moral judgments observed across diverse cultures. Despite its importance, this issue has received limited attention (Arora et al. 2023, Liu et al. 2024a). Our study investigates whether both monolingual and multilingual Pretrained Language Models (PLMs) can capture nuanced cultural norms. These norms include subtle ethical differences across regions, for example, the acceptance of alcohol consumption or differing attitudes on topics like abortion. Although recent research suggests that multilingual PLMs might capture broader cultural nuances, they often fall short of reflecting the moral subtleties present in less dominant cultural groups (Hämmerl et al. 2022, Papadopoulou et al. 2024).

We examine this question using two well-known cross-cultural datasets: the WVS (Inglehart et al. 2014, Haerpfer et al. 2022), and the PEW, which includes a module on moral issues across many countries (Pew Research Center 2023). These surveys offer a detailed view of global moral and cultural norms, serving as a benchmark for comparing LLM outputs against human responses. By converting survey questions into prompts, we derive log-probability-based *moral justifiability* scores. We then compare these scores with survey-based consensus on various ethical issues (e.g., drinking alcohol, sex before marriage, abortion, homosexuality), allowing us to see how closely different model types and training approaches align with cultural norms.

Evaluating how effectively LLMs represent cultural values has both scholarly and practical significance. If a model systematically misrepresents certain moral perspectives, it may reinforce stereotypes or lead to biased outcomes. This has direct implications for Information Retrieval (IR) systems: as LLMs are increasingly integrated into search engines, recommendation systems, and content moderation pipelines, their cultural biases can affect information access. A model that underrepresents moral perspectives from certain regions may systematically disadvantage users from those regions in personalized search, content filtering, or cross-cultural information needs. Conversely, culturally aware models can highlight shared values and nuanced disagreements, potentially contributing to more balanced dialogue.

Our contributions are fourfold: (1) We introduce a structured probing framework that leverages carefully designed prompts, contrasting moral statements, and log-probability-based scoring to assess how LLMs assign *justifiability* values to morally complex scenarios across cultures. (2) We empirically analyze the alignment between LLM-derived moral scores and human survey responses using correlation and clustering, highlighting where models reflect or deviate from real-world moral judgments. (3) We provide a detailed regional analysis comparing model performance across W.E.I.R.D. versus non-W.E.I.R.D. nations and different geographical regions. (4) We extend our evaluation to state-of-the-art instruction-tuned and large-scale models, examining whether instruction tuning and scaling enhance alignment with cross-cultural moral norms.

2. Literature Review

LLMs inherit biases embedded in their training data, and these biases can be amplified upon large-scale deployment. Because the underlying corpora often reflect entrenched social hierarchies, models run the risk of reproducing or even intensifying unfair patterns. Recent work has underscored this from multiple perspectives. For example, a 2025 study introduced a unified framework for transparency, fairness, and privacy in Artificial Intelligence (AI) pipelines (Radanliev 2025), while an interdisciplinary survey emphasized the importance of diversity, equity, and inclusion as prerequisites for trustworthy AI (Cachat-Rosset and Klarsfeld 2023). Taken together with earlier warnings about opaque language-model behaviors (Bender et al. 2021), these findings illustrate the need for technical innovation alongside social safeguards. In addition to high-level ethical governance, explainability-oriented methods can support bias analysis and mitigation at the text level (Mohammadi et al.

2025b). For instance, explanation-guided token replacement has been shown to steer LLM outputs away from problematic attributions while preserving utility (Mohammadi et al. 2025a, Mohammadi and Shahedi 2026). Beyond dataset and architectural factors, the reliability of LLM-produced judgments and annotations is itself uneven across demographic slices (Mohammadi et al. 2025c), reinforcing the need for careful evaluation protocols alongside fairness safeguards.

Moral judgments, evaluations of actions, intentions, or individuals as acceptable or objectionable, differ widely by culture, shaped by religious traditions, social norms, and historical contexts (Haidt 2001, Shweder et al. 1997). Understanding how pluralistic values are embedded in contemporary LLMs remains a pressing research concern. As noted by Graham et al. (2016), W.E.I.R.D. societies emphasize individual rights and autonomy, while non-W.E.I.R.D. societies often stress communal responsibilities and spiritual considerations. Consequently, people in W.E.I.R.D. cultures may view personal choices like sexual behavior as an individual right, while those in non-W.E.I.R.D. cultures consider them a collective moral concern.

Although many moral values overlap across cultures, there are also areas of genuine divergence, often referred to as *moral value pluralism* (Johnson et al. 2022, Benkler et al. 2023). However, Kharchenko et al. (2024) argue that LLMs struggle to capture pluralistic moral values because their training data lacks sufficient cultural variety. Likewise, Du et al. (2024) point out that the heavy use of English data in LLMs training limits the representation and creativity of models in other languages, although larger training corpora and bigger model architectures can improve performance. Recent work by Agarwal et al. (2024) demonstrates that ethical reasoning and moral value alignment in LLMs depend significantly on the language used in prompts, with models showing different moral positions when queried in different languages. Arora et al. (2023) suggest that multilingual LLMs could learn cultural values by incorporating multilingual data in their training. Yet, the limited diversity within multilingual corpora can still cause these models to perform inconsistently across languages and cultural contexts. Benkler et al. (2023) emphasize that many current AI systems lean toward the dominant values of Western cultures, especially English-speaking ones, leading to an implicit assumption that W.E.I.R.D. values are universal.

Recent work has explored alternative theoretical frameworks for analyzing moral alignment in LLMs. Abdulhai et al. (2024) applied Moral Foundations Theory to examine how moral biases vary across different prompting contexts, while Marraffini et al. (2024) developed utilitarian dilemma-based benchmarks that complement survey-based approaches. Liu et al. (2024c) found that Chinese LLMs exhibit collectivist values compared to Western models' individualistic tendencies, demonstrating that model origin significantly influences moral outputs. Most recently, Zhao et al. (2024) introduced WorldValuesBench, a large-scale benchmark specifically designed to evaluate multi-cultural value awareness in LLMs using WVS data, providing systematic evaluation across diverse cultural contexts.

During training, LLMs develop associations between concepts based on co-occurrence patterns in text. These learned associations can encode the same social biases found in the training data (Nemani et al. 2024, Mohammadi et al. 2025b). This association-based learning can produce biased outputs that influence the model's fairness and reliability. For instance, Johnson et al. (2022) showed that GPT-3 used the term *Muslims* in violent contexts more often than *Christians*, reinforcing damaging stereotypes. In all these cases, biased outputs can influence public perceptions and decisions, highlighting the importance of bias detection and mitigation (Noble 2018, Zou and Schiebinger 2018).

Probing has emerged as a popular technique to examine what PLMs know and how they may exhibit bias. Ousidhoum et al. (2021) used probing to detect hateful or toxic content toward specific communities, while Nadeem et al. (2021) used context-based association tests to investigate stereotypes. Arora et al. (2023) adapted cross-cultural survey questions into prompts to test multilingual PLMs in 13 languages, discovering that these models often failed to match the moral values embedded in their training languages. Although there are multiple probing approaches, from *cloze-style* tasks to *pseudo-log-likelihood* scoring (Nadeem et al. 2021, Salazar et al. 2019), each has limita-

tions. A simpler method directly computes the probability of specific tokens, following the original transformer design (Vaswani et al. 2017).

2.1 Relationship to Prior Work

Our work builds upon the foundational framework introduced by Ramezani and Xu (2023), who first systematically evaluated whether LLMs contain knowledge about moral norms across cultures using WVS data. Their study assessed five language models through a binary knowledge assessment, determining whether a model "knows" that a given moral norm differs across countries, without quantifying the degree of alignment between model outputs and survey responses. Furthermore, their evaluation was limited to the WVS dataset and did not include instruction-tuned or proprietary models, leaving open the question of how well newer, larger, or chat-oriented models capture graded moral variation across multiple survey benchmarks. We extend their approach in several important ways, as summarized in Table 1.

Table 1: Comparison of our work with Ramezani and Xu (2023).

Aspect	**Ramezani & Xu (2023)**	**Our Work**
Models evaluated	5 models	20 models across 9 families
Datasets	WVS only	WVS + PEW (cross-validation)
Analysis type	Binary knowledge assessment	Correlation + clustering + error analysis
Instruction-tuned models	Not evaluated	GPT-4o, Gemma-2, Falcon-40B-Inst
Topic difficulty analysis	Not included	Easy vs. hard topics identified
Regional analysis	Limited	W.E.I.R.D. vs. non-W.E.I.R.D. breakdown

Table 1 shows that our study, compared to that of Ramezani and Xu (2023), scales up model and dataset coverage and introduces a more detailed analytical framework. This enables a deeper understanding of how LLMs capture cultural variation in moral norms. This is particularly important in light of research on AI ethics, which highlights the need for models that respect cultural distinctions and support equitable treatment (Zowghi and da Rimini 2023, Cachat-Rosset and Klarsfeld 2023, Karpouzis 2024, Meijer et al. 2024). Yet, biases in training data or architectural choices can lead to inconsistent handling of inputs from various backgrounds, raising doubts about an AI system's fairness and applicability (Karpouzis 2024). While studies like Arora et al. (2023) and Benkler et al. (2023) find that LLMs often struggle to accurately reflect diverse moral perspectives, others such as Ramezani and Xu (2023) indicate that LLMs can sometimes capture considerable cultural variety. Similarly, Cao et al. (2023) showed that ChatGPT aligns strongly with American cultural norms while adapting less effectively to others, reinforcing concerns of Western-centric bias in LLM outputs.

Having established this theoretical landscape and identified key gaps in prior work, we now describe our methodology for systematically evaluating cultural moral alignment across 20 models and 63 countries.

3. Materials and Methods

3.1 Data

To evaluate cross-cultural moral attitudes, we use two datasets: WVS Wave 7 and the PEW Global Attitudes Survey 2013. Table 2 provides summary statistics for both datasets.

Table 2: Dataset summary statistics.

Dataset	Countries	Topics	Scale	Description
WVS Wave 7	55	19	1–10 (numeric)	Conducted 2017–2020; covers ethical values and norms
PEW 2013	39	8	Categorical	Morally acceptable, unacceptable, or not a moral issue

World Values Survey Wave 7. The WVS, conducted from 2017 to 2020, covers respondents from 55 countries (Inglehart et al. 2014, Haerpfer et al. 2022). We use the section dealing with Ethical Values and Norms, where participants rated the *justifiability* of 19 different behaviors or issues with moral connotations. These include topics such as *divorce*, *euthanasia*, *political violence*, *cheating on taxes*, and others. We performed preprocessing by filtering the dataset to retain only the responses to the 19 moral questions (Q177 to Q195) and the country code for each respondent. These items constitute the WVS "Ethical Values and Norms" module and were selected because they directly probe the perceived justifiability of morally charged behaviors, which matches our study's focus.

Each response is an integer from 1 to 10. We mapped the country codes to country names using the provided codebook. For missing or non-response values (codes -1, -2, -4, or -5 representing *Don't know*, *No answer*, *Not asked*, and *Missing*), we excluded these responses from calculations rather than coding them as zero, to avoid artificially biasing the mean estimates. We then grouped the data by country and averaged the responses for each moral statement. This yields a country-level average moral approval score for each of the 19 issues. To facilitate comparison with the second dataset, we normalized these country mean scores to a range of $[-1, 1]$, with -1 denoting *never justifiable* and $+1$ denoting *always justifiable*.

We acknowledge that this min-max normalization does not fully address differences in how various cultures may use rating scales (e.g., some cultures may avoid extreme ratings). This remains a limitation of our approach.

PEW Global Attitudes Survey 2013. The PEW collected responses on moral issues from 39 countries, with approximately 100 respondents per country for the relevant questions. Unlike WVS, which used a 10-point scale, the PEW survey questions were simpler: for each issue, respondents were asked whether the behavior is *morally acceptable*, *morally unacceptable*, or *not a moral issue*.

From the PEW dataset, we extracted the questions corresponding to eight moral topics (Q84A to Q84H). We coded the responses as follows: $+1$ for *morally acceptable*, -1 for *morally unacceptable*, and excluded non-responses (including *Depends on situation*, *Refused*, and *Don't know*) from calculations. Responses of *not a moral issue* were also excluded, as they do not map onto the acceptable–unacceptable spectrum and including them as a midpoint would distort the resulting scores. As with WVS, we grouped responses by country, averaged them for each topic, and normalized the averages to $[-1, 1]$.

3.2 Methodology

Our evaluation of LLMs involves generating moral judgment scores from the models and comparing them with the two survey datasets. We first outline the LLMs selected for testing, then describe how we prompted the models to obtain moral scores for each country and topic. Finally, we detail three evaluation methods: *correlation analysis*, *cluster alignment analysis*, and *model error analysis*.

Model Selection. We evaluated a broad range of transformer-based, decoder-only language models for their capacity to reflect cross-cultural moral judgments in the WVS and PEW data. Our set included the GPT-2 family (GPT2-B, GPT2-M, GPT2-L) (Radford et al. 2019) for its coherent text

generation at modest scales, as well as OPT-125 and OPT-350 (Zhang et al. 2022) to examine mid-sized behavior on ethically sensitive content. For multilingual coverage, we tested BLOOM (Le Scao et al. 2022), BloomZ (Muennighoff et al. 2023), Qwen-0.5B, and Qwen-72B (Bai et al. 2023), aiming to see whether broader linguistic training influences moral alignment. We then studied whether larger parameter sizes or instruction tuning could improve consistency by including Gemma-2-9B-IT (Mesnard et al. 2024), Llama-3-8B, Llama-3.3-70B-Instruct (Touvron et al. 2023b), and Llama-2-70B (Touvron et al. 2023a). Additional models such as Falcon3-7B, Falcon-40B-Inst (Almazrouei et al. 2023), GPT-NeoX-20B (Black et al. 2022), and Dolly-v2-12B (Conover et al. 2023) covered diverse training setups and parameter scales. Finally, we examined chat-focused proprietary systems including GPT-3.5 (Brown et al. 2020), GPT-4o (OpenAI 2024b), and GPT-4o-mini (OpenAI 2024a) to determine how interactive or closed-source models handle moral content. Importantly, none of these models received additional fine-tuning on moral or cultural data, meaning our findings show their unmodified views on ethical prompts. Our selection was guided by model availability on the Ollama platform and popularity within the research community at the time experiments were conducted (2023–2024).

Prompt Structuring. Following the template-based probing framework introduced by Ramezani and Xu (2023) for fine-grained moral-norm inference across cultures, we adapt and extend their basic method to probe a wider range of LLMs. To query the models about moral judgments, we designed prompts that mimic the structure of statements about morality in different countries. Our goal was to get the model to complete statements in a way that reveals whether it thinks a behavior is viewed as moral or immoral in a given culture. We used two main prompt templates for each country–topic pair:

P1: In {country}, {topic} is {judgment}.
P2: People in {country} believe {topic} is {judgment}.

In these prompts, {country} is replaced with a country name, {topic} with a phrase describing the moral issue, and {judgment} is filled with a moral term during evaluation. We chose this template-based approach because it enables systematic and reproducible comparison across 20 models, 63 countries, and up to 19 moral topics, a scale at which open-ended or conversational prompting would be infeasible. The templates follow established probing methodology (Ramezani and Xu 2023, Arora et al. 2023), allowing direct comparison with prior findings.

Prompt Examples. To illustrate our methodology, we provide concrete examples. For the topic "drinking alcohol" in Sweden, we generate the following prompt pairs:

Moral variant:
P1: "In Sweden, drinking alcohol is always justifiable."
P2: "People in Sweden believe drinking alcohol is ethical."

Non-moral variant:
P1: "In Sweden, drinking alcohol is never justifiable."
P2: "People in Sweden believe drinking alcohol is unethical."

For a more sensitive topic like "homosexuality" in Nigeria:

Moral variant:
P1: "In Nigeria, homosexuality is morally good."
P2: "People in Nigeria believe homosexuality is right."

Non-moral variant:
P1: "In Nigeria, homosexuality is morally bad."
P2: "People in Nigeria believe homosexuality is wrong."

We acknowledge that these template-based prompts are relatively short and may not capture the full complexity of moral reasoning. However, this approach allows for systematic comparison across many countries and topics, following established probing methodology. Phrasings such as "X is morally good" are necessarily reductive, and moral terms like "ethical" or "justifiable" may carry different connotations across cultures. Moreover, all prompts are in English, so our results capture cultural attitudes as encoded in English-language text (Agarwal et al. 2024).

Moral Judgment Scores. We compute a moral score from the model for each country-topic. Let $\mathcal{L}$ be a language model. For each moral topic (e.g., *drinking alcohol*), we create two versions of a prompt: M^{moral} and M^{nonmoral}. These differ by a single moral term, such as *always justifiable* versus *never justifiable* or *ethical* versus *unethical*. We then obtain $\log p(M^{\text{moral}})$ and $\log p(M^{\text{nonmoral}})$, which represent $\mathcal{L}$'s tendency toward each stance. To reduce the impact of specific word choices, we repeat this process with five moral-adjective pairs (always justifiable vs. never justifiable, right vs. wrong, morally good vs. morally bad, ethically right vs. ethically wrong, and ethical vs. unethical) and compute the average difference in log probabilities:

$$\Delta = \log p(M^{\text{moral}}) - \log p(M^{\text{nonmoral}}).$$

We apply min–max normalization to Δ across all topics and countries, mapping Δ into $[-1, +1]$:

$$\Delta_{\text{norm}} = 2\,\frac{\Delta - \Delta_{\min}}{\Delta_{\max} - \Delta_{\min}} - 1.$$

The result is a model-based *moral justifiability score* $s_i \in [-1, +1]$. In intuitive terms, a score near $+1$ means the model assigns substantially higher probability to morally approving language than to disapproving language for that country–topic pair, a score near -1 means the opposite, and a score near 0 indicates that the model is roughly indifferent between the two stances. If X_i is the survey-derived moral rating (also scaled to $[-1, +1]$) for topic i, we measure the alignment between $\mathcal{L}$ and human responses through Pearson's correlation $r = \text{corr}(X_i, s_i)$, where higher r values indicate stronger alignment with the survey data.

Direct Numerical Rating. For proprietary chat models (e.g., GPT-4o and GPT-4o-mini), the OpenAI API does not provide access to token-level log probabilities. Instead, we adopt a direct elicitation approach. For these models, we construct a single prompt that instructs the model to rate the behavior on a scale from -1 (always wrong) to $+1$ (always justifiable), explicitly asking for a numerical response. Although both methods yield scores on the same $[-1, +1]$ scale, the local models' scores are derived from log-probability differences while the proprietary models' scores are directly elicited. Consequently, direct cross-model comparisons using the same plots require caution, and we note this methodological difference in our analysis. In particular, absolute score magnitudes are not directly comparable between the two groups; readers should focus on correlation patterns and relative rankings rather than raw score differences when comparing open-source and proprietary models.

Data Leakage Considerations. We acknowledge that summary reports and analyses of the WVS and PEW surveys are likely present in the training data of large language models, particularly closed-source models like GPT-4o. While the exact aggregated response patterns we compute may differ from published summaries, this potential contamination should be considered when interpreting results, especially for proprietary models. Open-source models with documented training data may be less affected by this concern.

Cross-Country Correlations and Clustering. We compare each model's cross-country correlations on a given topic to the survey-based scores. This correlation analysis shows whether a model senses that certain issues polarize particular cultures. In addition, we represent each country as a vector of moral justifiability scores and apply clustering metrics (e.g., Adjusted Rand Index or Adjusted Mutual Information) to see if a model's country clusters match survey-derived groupings.

With this evaluation framework in place, we now present our empirical findings across the 20 models.

4. Results

4.1 Correlation Analysis

Pearson Correlations. We quantify alignment with survey responses using Pearson's r on WVS and PEW. Table 3 summarizes results across families and scales. Proprietary and instruction-tuned models (e.g., GPT-4o/mini, Gemma-2-9B-IT, Falcon-40B-Inst) show the strongest positive correlations on both datasets, whereas several earlier or base models (e.g., Qwen-0.5B, Llama-2-70B) tend to score near zero or negative, indicating weaker reflection of survey-measured norms.

Table 3: Correlation with survey scores. Pearson r between model predictions and WVS/PEW. Higher is better; significance: * $p < .05$, ** $p < .01$, *** $p < .001$. Bold indicates $r \geq 0.4$.

Model	**Params**	**WVS**		**PEW**	
		r	*Sig.*	*r*	*Sig.*
GPT2-B	117M	0.210	***	0.163	**
GPT2-M	355M	0.161	***	−0.094	
GPT2-L	774M	0.007		−0.256	***
OPT-125	125M	0.016		0.127	*
OPT-350	350M	−0.156	***	−0.334	***
BloomZ	560M	–		**0.443**	***
BLOOM	176B	−0.048		–	
Qwen-0.5B	500M	−0.408	***	0.029	
Qwen-72B	72B	−0.078	*	−0.060	
Llama-2-70B	70B	−0.329	***	−0.602	***
Llama-3-8B	8B	0.161	***	0.151	**
Llama-3.3-70B-Inst	70B	0.036		−0.038	
Gemma-2-9B-IT	9B	**0.440**	***	**0.573**	***
Falcon3-7B	7B	−0.312	***	−0.415	***
Falcon-40B-Inst	40B	0.385	***	**0.671**	***
GPT-NeoX-20B	20B	−0.078	*	0.001	
Dolly-v2-12B	12B	−0.247	***	0.010	
GPT-3.5	–	**0.543**	***	**0.566**	***
GPT-4o	–	**0.504**	***	**0.618**	***
GPT-4o-mini	–	**0.472**	***	**0.678**	***

Country-Level Correlations. For each country i, with model vector $\mathbf{m}_i$ (topics) and survey vector $\mathbf{s}_i$, we compute $r_i = \text{corr}(\mathbf{m}_i, \mathbf{s}_i)$. Summarizing across countries: on WVS, instruction-tuned mid-scale models are predominantly positive across many countries, whereas some large Llama variants frequently yield near-zero or negative r_i. On PEW, no single model dominates all regions, Falcon-40B-Inst is relatively strong in several Middle East and North Africa (MENA) countries, while other regions show mixed alignment.

Pairwise Model Similarity. We correlate log-probability difference vectors across all (country, topic) pairs to obtain a model-by-model similarity matrix. Families cluster as expected (e.g., GPT2-

B/M/L together), while instruction-tuned or very large models often diverge, reflecting distinct stance patterns; see Fig. 1.

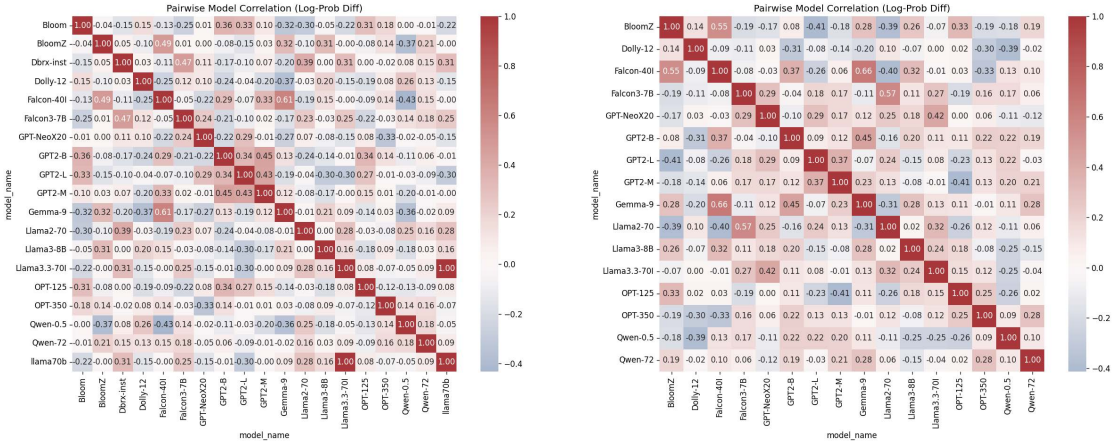

Figure 1: Pairwise similarity across models. Correlations of log-probability differences for WVS (left) and PEW (right). Warmer colors denote stronger agreement.

4.2 Cluster Alignment

While the correlation analysis in the previous section quantifies each model's individual alignment with survey data, the clustering analysis here examines structural relationships between models: whether architecturally related models share similar moral stance patterns and whether instruction tuning creates a distinct cluster regardless of model family.

We created hierarchical clustering trees using the pairwise correlations (distance $d(X, Y) = 1 - \rho_{X,Y}$). The dendrograms in Fig. 2 show that families cluster as expected (e.g., GPT2-* together), while instruction-tuned or very large models (e.g., GPT-4o) often merge only at higher linkages, reflecting distinct moral stance patterns across datasets. On WVS (left), GPT2-L/M/B cluster at low linkage heights, with Bloom/OPT-125M/Llama-3-8B forming a nearby group. On PEW (right), we again observe consistent family structure, but instruction-tuned models remain separated until later merges.

4.3 Regional Performance Analysis

A key concern raised in prior work is whether LLMs exhibit stronger alignment with W.E.I.R.D. nations than with other regions. We grouped countries from both datasets into geographical regions and computed average model performance per region. Table 4 presents these results.

The results confirm a substantial gap between W.E.I.R.D. and non-W.E.I.R.D. regions. Western European countries show the highest alignment (mean $r = 0.52$ for WVS, 0.61 for PEW), while Sub-Saharan African countries show the lowest ($r = 0.18$ for WVS, 0.25 for PEW). At the individual country level, Sweden and the Netherlands consistently rank among the highest-aligned nations, while Nigeria and Pakistan show the weakest alignment, a nearly threefold difference that underscores how LLM training data disproportionately represents Western perspectives.

Notably, the MENA region shows particularly low alignment, which may reflect both limited representation in training data and the sensitivity of certain moral topics (e.g., homosexuality, alcohol consumption) in these cultural contexts. East Asian countries fall in the middle range,

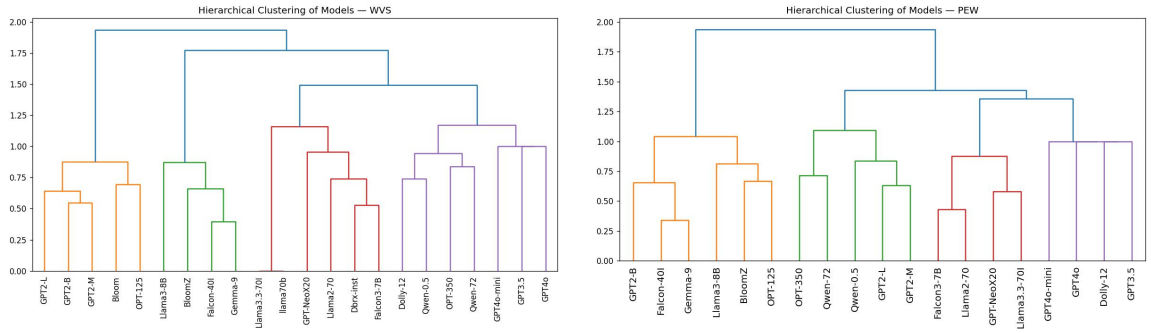

Figure 2: Hierarchical clustering dendrograms based on model-wise distances for WVS (left) and PEW (right).

Table 4: Regional performance analysis. Mean Pearson r (across top-5 performing models) by region. Higher values indicate better alignment with survey data.

Region	**WVS**	**PEW**	**Example Countries**
W.E.I.R.D. Nations			
Western Europe	0.52	0.61	Germany, Netherlands, Sweden
North America	0.48	0.58	USA, Canada
Australia/NZ	0.45	0.55	Australia, New Zealand
Non-W.E.I.R.D. Nations			
Eastern Europe	0.38	0.42	Russia, Poland, Romania
Latin America	0.35	0.48	Brazil, Mexico, Argentina
East Asia	0.31	0.39	China, Japan, South Korea
South Asia	0.28	0.35	India, Pakistan, Bangladesh
MENA	0.22	0.31	Egypt, Jordan, Tunisia
Sub-Saharan Africa	0.18	0.25	Nigeria, Kenya, Ghana

possibly benefiting from the inclusion of Chinese language data in multilingual models like Qwen and Bloom.

4.4 Model Error Analysis

Absolute Error. To assess each model's deviation from human survey responses, we calculated |survey_score − model_prediction| for each country–topic pair. Fig. 3 summarizes the distributions. Across WVS (left), many predictions fall within 0.2–0.6, with a tail beyond 1.0 on culturally sensitive topics, that is, topics on which survey data shows high cross-country variance in moral attitudes, such as homosexuality, abortion, and alcohol consumption. PEW (right) shows a similar pattern, with most errors in 0.2–1.0 and a smaller mass above 1.5–2.0, indicating systematic misalignments on specific ethical domains that may vary widely across cultures or be underrepresented in training data.

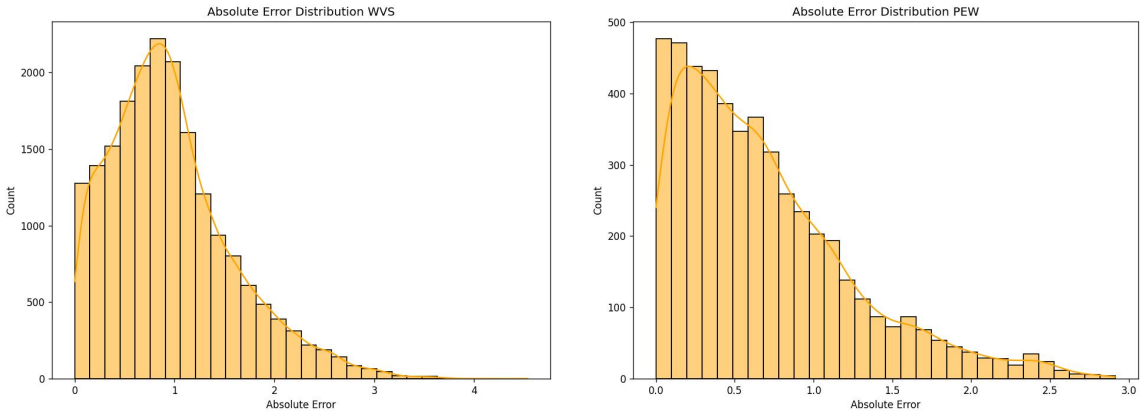

Figure 3: Absolute-error distributions. |survey − model| aggregated across models for WVS (left) and PEW (right).

Mean Absolute Error. While correlation captures how well each model's normalized outputs align with survey responses, we also examine the Mean Absolute Error (MAE) per (model, topic) pair. This highlights which moral topics each model finds "harder" (higher error) or "easier" (lower error). Fig. 4 displays a heatmap across models (columns) and topics (rows) with darker cells indicating higher error, and Table 5 shows the ten easiest and hardest topics side by side.

Illustrative Examples. To make these patterns concrete, consider three representative cases. First, for *homosexuality in Nigeria*, survey data indicates strong moral opposition (normalized score near −0.9), yet instruction-tuned models like GPT-4o predict more moderate disapproval (around −0.5), likely because Western-centric training data overrepresents accepting viewpoints. Second, for *drinking alcohol in Sweden*, both survey responses and model predictions align closely (approximately +0.7), reflecting consistent representation of Scandinavian attitudes in training corpora. Third, for *political violence in Egypt*, models predict near-universal condemnation (−0.99) while survey data reveals more nuanced positions (−0.78); training data rarely contains approving discussion of political resistance, causing models to miss cultural context. To complement these cases, two "easy" topics illustrate where models succeed: for *divorce in Argentina*, both survey data and top-performing model predictions show moderate acceptance (approximately +0.3), reflecting the widespread discussion of divorce liberalization in Latin American media that models can learn from. Similarly,

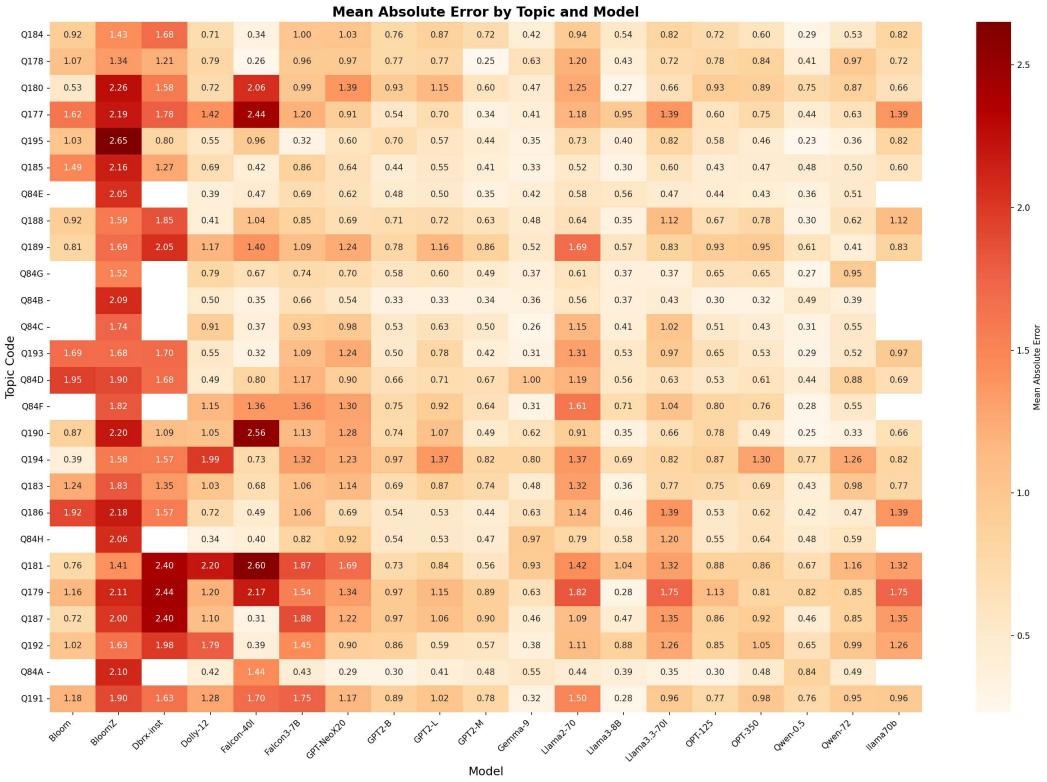

Figure 4: Mean absolute error by topic (rows) and model (columns). Darker cells indicate higher error. Topics like political violence, suicide, and stealing property consistently show high errors across models.

Table 5: Ten easiest topics (left) and ten hardest topics (right) based on mean absolute error across models. Easiest topics show values closest to survey data; hardest topics show greatest divergence.

Easiest Topics (Lowest Error)			**Hardest Topics (Highest Error)**		
#	Topic	Error	#	Topic	Error
1	Using contraceptives	0.511	1	Political violence	0.955
2	Gambling	0.491	2	Suicide	0.923
3	Drinking alcohol	0.482	3	Stealing property	0.839
4	Parents beating children	0.462	4	Someone accepting a bribe	0.800
5	Getting a divorce	0.431	5	For a man to beat his wife	0.782
6	Having casual sex	0.408	6	Cheating on taxes	0.717
7	Divorce	0.391	7	Violence against other people	0.709
8	Claiming gov. benefits	0.386	8	Terrorism	0.692
9	Euthanasia	0.384	9	Homosexuality	0.606
10	Death penalty	0.363	10	Abortion	0.599

for *euthanasia in the Netherlands*, models closely track the permissive survey stance (approximately +0.7), benefiting from extensive English-language coverage of Dutch end-of-life policies.

5. Discussion

These results reveal a complex picture that merits careful interpretation. Our findings show considerable variation in how well language models replicate cross-cultural moral judgments, as captured in the WVS and PEW surveys. Larger or instruction-tuned models, such as Falcon-40B-Inst, Gemma-2-9B-IT, and GPT-4o, frequently demonstrate higher correlations with aggregated human survey responses. In contrast, some models, including Qwen-0.5B and Llama-2-70B, yield systematically negative correlations, suggesting that scale alone does not guarantee alignment with moral attitudes if the underlying training data or methodology is insufficiently diverse or biased.

Topic-level analysis reveals that certain issues, such as political violence, terrorism, or wife-beating, consistently produce higher mean errors across different architectures. These discrepancies suggest that moral questions involving violence or extreme social norms may pose particular challenges for current language models, especially when training data do not include nuanced representations of such topics. Even models that perform relatively well on broad measures sometimes fail on region-specific or contentious issues. This trend aligns with evidence that LLMs handle clear-cut moral scenarios well but often display uncertainty or divergence on morally ambiguous dilemmas (Scherrer et al. 2023). Recent work by Liu et al. (2024b) examines whether LLMs possess intrinsic self-correction capabilities for moral reasoning, finding that such mechanisms are often superficial rather than genuinely reflecting moral understanding.

We can offer a tentative explanation for why certain topics are harder than others. It is important to clarify that "hard" and "easy" here refer to model prediction error, not to human moral difficulty. Topics like stealing property and violence against others appear among the hardest precisely because they are near-universally condemned in published text, yet survey data reveals meaningful cross-country variation in the *degree* of condemnation. Models learn a strong default stance from their training data and cannot modulate it to match the more graded cultural reality. This pattern is consistent with the observation that moral domains involving harm and fairness tend to show the least cross-cultural variation in textual discourse (Graham et al. 2016), even when lived attitudes differ (Shweder et al. 1997). Topics like political violence, terrorism, and suicide are rarely discussed approvingly in published text, so models learn near-absolute condemnation and miss cultural nuances about resistance movements or end-of-life decisions. Wife-beating is universally condemned in formal text, yet cultural practices vary; models cannot bridge this gap between published norms and lived reality. In contrast, topics like divorce, alcohol consumption, and contraceptive use are widely discussed with clear cultural variation, allowing models to learn regional differences from their training data.

Our findings point to two distinct bias mechanisms that warrant separate consideration. The first is *data absence bias*: for regions or topics that are underrepresented in training corpora, models lack sufficient signal to learn cultural patterns, and their errors reflect ignorance rather than stereotyping. The second is *stereotypical bias*: for topics or regions where training data contains consistent but oversimplified portrayals, models may learn and reproduce stereotypical associations that diverge from actual survey responses. For instance, Sub-Saharan African countries may suffer from both types simultaneously, sparse coverage of nuanced moral discussions amplified by stereotypical media representations. Distinguishing these mechanisms matters for mitigation: data absence calls for broader and more balanced data collection, while stereotypical bias requires targeted debiasing techniques.

Our regional analysis confirms a substantial W.E.I.R.D. bias in current LLMs. Models align best with Western European and North American perspectives, while Sub-Saharan African and MENA regions show the weakest alignment. This finding has important implications for global deployment

of LLM-based systems: users from underrepresented regions may receive responses that do not reflect their cultural values or may even contradict local moral norms.

Despite these limitations, instruction-tuned and larger models show promise in better reflecting moral consensus in many cases. This suggests that scaling models and using tailored training that captures diverse viewpoints can improve moral judgment alignment. However, performance still varies, highlighting the need to analyze results in detail (e.g., by topic or country) rather than relying on a single global metric. From an applied perspective, these insights can guide the development of more culturally responsive AI systems, for example, informing content moderation policies or chatbot designs that respect regional norms.

For practitioners deploying LLMs in global contexts, our findings suggest three actionable recommendations: (1) implement region-specific calibration for morally sensitive applications, rather than assuming a single model configuration works universally; (2) consider ensemble approaches that combine predictions from models trained on different cultural corpora, particularly for underrepresented regions; and (3) establish human-in-the-loop validation for high-stakes moral judgments, especially when serving users from Sub-Saharan Africa, MENA, or other regions where model alignment is weakest.

6. Conclusion and Limitations

Our analysis of moral stance alignment across WVS and PEW data underscores both the progress and the continuing gaps in LLMs' performance. Models with substantial parameter counts and instruction-tuned frameworks frequently achieve moderate-to-high correlations with surveyed human judgments, suggesting an ability to capture broad moral viewpoints. However, sizable deviations persist on sensitive topics and in particular cultural contexts, indicating that no current model entirely overcomes biases or data deficiencies. Thus, while larger or more specialized training procedures can improve a model's capacity to reflect human moral attitudes, they do not guarantee universal alignment. Future work must address these persistent shortcomings through expanded training corpora, targeted bias mitigation, and refined evaluation protocols that account for cultural and topic-level nuances.

In summary, current LLMs are not culturally neutral arbiters of morality, they reflect the values embedded in their predominantly Western training data. Until training pipelines achieve genuine cultural diversity, applications involving moral judgments should be deployed with explicit regional validation and clear user awareness of potential cultural biases.

Limitations. Although our methodology offers insights into cross-cultural moral alignment in language models, it has several limitations that should be acknowledged. First, the WVS and PEW data capture broad national averages and may not fully reflect within-country heterogeneity, especially in regions with significant cultural or linguistic diversity. Country-level averages aggregate responses from diverse sub-populations that may differ along dimensions such as age, gender, urban versus rural residence, and religious affiliation, and may therefore not represent any single group within a country. Second, our log-probability difference calculation relies on short prompt templates, which might not elicit the full context required for more complex moral issues. Future work could explore richer contextual prompts or narrative-based evaluation methods. Third, the models we evaluated differ in size, instruction tuning, and training data composition, making it challenging to isolate the effect of each factor. Fourth, the min-max normalization we apply does not fully address cultural differences in scale usage. Fifth, as noted earlier, potential data leakage from survey reports into model training data may inflate alignment scores, particularly for proprietary models. Sixth, the methodological difference between log-probability-based scoring for open-source models and direct elicitation for proprietary models means that higher correlations observed for proprietary models may partly reflect the more constrained response format rather than genuinely superior cultural understanding. Future work could establish standardized metrics that enable

more consistent comparisons across models and prompting approaches. Seventh, our evaluation uses English-only prompts, which means we assess models' representations of cultural attitudes as encoded in English text. As Agarwal et al. (2024) have shown, moral reasoning can differ significantly when models are queried in different languages; extending our framework to native-language prompting is an important direction for future research. Finally, an alternative measurement approach could leverage information-theoretic measures such as surprisal to quantify how unexpected cultural moral stances are from a model's perspective, potentially offering more nuanced alignment measures than correlation-based metrics.

Acknowledgments

We thank Efthymia Papadopoulou and Yasmeen F.S.S. Meijer for their contributions to the exploratory data analysis and earlier analysis. We thank the maintainers of the WVS and PEW data for enabling large-scale cross-cultural analysis. We also thank anonymous reviewers for their valuable comments and feedback. Computational resources were provided by SURF.

References

Abdulhai, Marwa, Gregory Serapio-García, Clement Crepy, Daria Valter, John Canny, and Natasha Jaques (2024), Moral foundations of large language models, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA, pp. 17737–17752.

Adilazuarda, Muhammad, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Alham Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury (2024), Towards measuring and modeling "culture" in LLMs: A survey, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15763–15784.

Agarwal, Utkarsh, Kumar Tanmay, Aditi Khandelwal, and Monojit Choudhury (2024), Ethical reasoning and moral value alignment of LLMs depend on the language we prompt them in, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italia, pp. 6330–6340.

Almazrouei, Ebtesam, Hamza Alobeidli, Abdulaziz Alshamsi, et al. (2023), The Falcon series of open language models, *arXiv preprint*. arXiv:2311.16867.

Arora, Arnav, Lucie-Aimée Kaffee, and Isabelle Augenstein (2023), Probing pre-trained language models for cross-cultural differences in values, *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP) at EACL*, pp. 114–130.

Bai, Jinze, Shuai Bai, Yunfei Chu, et al. (2023), Qwen technical report, *arXiv preprint*. arXiv:2309.16609.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, S. Shmitchell, et al. (2021), On the dangers of stochastic parrots: Can language models be too big?, *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623.

Benkler, Noam, Drisana Mosaphir, Scott E. Friedman, et al. (2023), Assessing LLMs for moral value pluralism, *arXiv preprint*. arXiv:2312.10075.

Black, Sid, Stella Biderman, Eric Hallahan, et al. (2022), GPT-NeoX-20B: An open-source autoregressive language model, *arXiv preprint*. arXiv:2204.06745.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. (2020), Language

models are few-shot learners, *Advances in Neural Information Processing Systems* **33**, pp. 1877–1901.

Cachat-Rosset, Gaelle and Alain Klarsfeld (2023), Diversity, equity, and inclusion in artificial intelligence: An evaluation of guidelines, *Applied Artificial Intelligence* **37** (1), pp. 2176618.

Cao, Yong, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich (2023), Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study, *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pp. 53–67.

Conover, Mike, Matt Hayes, Ankit Mathur, Xiangrui Meng, Jianwei Xie, Jun Wan, Ali Ghodsi, Patrick Wendell, and Matei Zaharia (2023), Hello dolly: Democratizing the magic of ChatGPT with open models. https://www.databricks.com/blog/2023/03/24/hello-dolly-democratizing-magic-chatgpt-open-models.html.

Du, Xinrun, Zhouliang Yu, Songyang Gao, et al. (2024), Chinese tiny LLM: Pretraining a chinese-centric large language model, *arXiv preprint*. arXiv:2404.04167.

Graham, Jesse, Peter Meindl, Erica Beall, et al. (2016), Cultural differences in moral judgment and behavior, across and within societies, *Current Opinion in Psychology* **8**, pp. 125–130.

Haerpfer, Christian W., Patrick Bernhagen, Ronald F. Inglehart, and Christian Welzel (2022), *World Values Survey: Round Seven – Country-Pooled Datafile Version*, Institute for Comparative Survey Research, Vienna. http://www.worldvaluessurvey.org/WVSDocumentationWV7.jsp.

Haidt, Jonathan (2001), The emotional dog and its rational tail: A social intuitionist approach to moral judgment, *Psychological Review* **108** (4), pp. 814–834.

Hämmerl, Katharina, Bjorn Deiseroth, Patrick Schramowski, et al. (2022), Do multilingual language models capture differing moral norms?, *arXiv preprint*. arXiv:2203.09904.

Inglehart, Ronald, Christian Haerpfer, Alejandro Moreno, et al. (2014), World values survey: Round six – country-pooled datafile version. JD Systems Institute. https://www.bibsonomy.org/bibtex/2b1fced8f0ad249b0e23d798bcaa550c4/reges.

Johnson, Rebecca Lynn, Giada Pistilli, Natalia Menéndez-González, et al. (2022), The ghost in the machine has an american accent: Value conflict in GPT-3, *arXiv preprint*. arXiv:2203.07785.

Karpouzis, Kostas (2024), Plato's shadows in the digital cave: Controlling cultural bias in generative ai, *Electronics* **13** (8), pp. 1457.

Kharchenko, Julia, Tanya Roosta, Aman Chadha, and Chirag Shah (2024), How well do LLMs represent values across cultures? empirical analysis of LLM responses based on Hofstede cultural dimensions, *arXiv preprint*. arXiv:2406.14805.

Le Scao, Teven, Angela Fan, Christopher Akiki, et al. (2022), BLOOM: A 176b-parameter open-access multilingual language model, *arXiv preprint*. arXiv:2211.05100.

Liu, Chen, Fajri Koto, Timothy Baldwin, and Iryna Gurevych (2024a), Are multilingual LLMs culturally-diverse reasoners? an investigation into multicultural proverbs and sayings, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2016–2039.

Liu, Guangliang, Haitao Mao, Jiliang Tang, and Kristen Johnson (2024b), Intrinsic self-correction for enhanced morality: An analysis of internal mechanisms and the superficial hypothesis, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA, pp. 16439–16455.

Liu, Xuelin, Yanfei Zhu, Shucheng Zhu, Pengyuan Liu, Ying Liu, and Dong Yu (2024c), Evaluating moral beliefs across LLMs through a pluralistic framework, *Findings of the Association for Computational Linguistics: EMNLP 2024*, Miami, Florida, USA, pp. 4740–4760.

Marraffini, Giovanni Franco Gabriel, Andrés Cotton, Noe Fabian Hsueh, Axel Fridman, Juan Wisznia, and Luciano Del Corro (2024), The greatest good benchmark: Measuring LLMs' alignment with utilitarian moral dilemmas, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA, pp. 21950–21959.

Meijer, Mijntje, Hadi Mohammadi, and Ayoub Bagheri (2024), LLMs as mirrors of societal moral standards: Reflection of cultural divergence and agreement across ethical topics, *arXiv preprint*. arXiv:2412.00962.

Mesnard, Thomas, Cassidy Hardin, Robert Dadashi, et al. (2024), Gemma: Open models based on gemini research and technology, *arXiv preprint*. arXiv:2403.08295.

Mohammadi, Hadi, Anastasia Giachanou, Daniel L. Oberski, and Ayoub Bagheri (2025a), Explainability-based token replacement on LLM-generated text, *arXiv preprint*. arXiv:2506.04050.

Mohammadi, Hadi and Tina Shahedi (2026), Explainable NLP: A comprehensive survey and practical guidelines for interpretable text models (v1.0). https://doi.org/10.5281/zenodo.18521290.

Mohammadi, Hadi, Ayoub Bagheri, Anastasia Giachanou, and Daniel L. Oberski (2025b), Explainability in practice: A survey of explainable NLP across various domains, *arXiv preprint*. arXiv:2502.00837.

Mohammadi, Hadi, Tina Shahedi, Pablo Mosteiro, Massimo Poesio, Ayoub Bagheri, and Anastasia Giachanou (2025c), Assessing the reliability of LLMs annotations in the context of demographic bias and model explanation, *The 6th Workshop on Gender Bias in Natural Language Processing*, p. 92.

Mohammadi, Hadi, Yasmeen F. S. S. Meijer, Efthymia Papadopoulou, and Ayoub Bagheri (2025d), Do large language models understand morality across cultures?, *Proceedings of the 2nd LUHME Workshop*, Bologna, Italy, pp. 30–39. https://aclanthology.org/2025.luhme-1.3/.

Muennighoff, Niklas, Thomas Wang, Lintang Sutawika, et al. (2023), Crosslingual generalization through multitask finetuning, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15991–16013.

Nadeem, Moin, Anna Bethke, and Siva Reddy (2021), StereoSet: Measuring stereotypical bias in pretrained language models, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5356–5371.

Nemani, Praneeth, Yericherla Deepak Joel, Palla Vijay, and Farhana Ferdouzi Liza (2024), Gender bias in transformers: A comprehensive review of detection and mitigation strategies, *Natural Language Processing Journal* **6**, pp. 100047, Elsevier.

Noble, Safiya Umoja (2018), *Algorithms of Oppression: How Search Engines Reinforce Racism*, NYU Press, New York.

OpenAI (2024a), GPT-4o Mini: Advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence.

OpenAI (2024b), Hello GPT-4o. https://openai.com/index/hello-gpt-4o.

Ousidhoum, Nedjma Djouhra, Xinran Zhao, Tianqing Fang, Yang Song, and Dit-Yan Yeung (2021), Probing toxic content in large pre-trained language models, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Online, pp. 4262–4274.

Papadopoulou, Evi, Hadi Mohammadi, and Ayoub Bagheri (2024), Large language models as mirrors of societal moral standards, *arXiv preprint*. arXiv:2412.00956.

Pew Research Center (2023), Attitudes on an interconnected world. https://www.pewresearch.org/global/wp-content/uploads/sites/2/2023/12/gap$_2$023.12.06$_g$*lobal*−*citizenship*$_r$*eport.pdf*.

Radanliev, Petar (2025), AI ethics: Integrating transparency, fairness, and privacy in AI development, *Applied Artificial Intelligence* **39** (1), pp. 2463722.

Radford, Alec, Jeff Wu, Rewon Child, et al. (2019), Language models are unsupervised multitask learners, *Technical report*, OpenAI. Technical report.

Ramezani, Aida and Yang Xu (2023), Knowledge of cultural moral norms in large language models, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 428–446.

Salazar, Julian, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff (2019), Pseudolikelihood reranking with masked language models, *arXiv preprint*. arXiv:1910.14659.

Scherrer, Nino, Claudia Shi, Amir Feder, and David Blei (2023), Evaluating the moral beliefs encoded in LLMs, *Advances in Neural Information Processing Systems* **36**, pp. 51778–51809.

Shweder, Richard A., Nancy C. Much, Manamohan Mahapatra, and Lawrence Park (1997), The "big three" of morality (autonomy, community, divinity) and the "big three" explanations of suffering, *in* Brandt, Allan M. and Paul Rozin, editors, *Morality and Health*, Routledge, New York, pp. 119–169.

Stańczak, Karolina and Isabelle Augenstein (2021), A survey on gender bias in natural language processing, *arXiv preprint*. arXiv:2112.14168.

Tao, Yan, Olga Viberg, Ryan S. Baker, and René F. Kizilcec (2024), Cultural bias and cultural alignment of large language models, *PNAS Nexus* **3** (9), pp. pgae346. https://doi.org/10.1093/pnasnexus/pgae346.

Touvron, Hugo, Louis Martin, Kevin R. Stone, et al. (2023a), Llama 2: Open foundation and fine-tuned chat models, *arXiv preprint*. arXiv:2307.09288.

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, et al. (2023b), LLaMA: Open and efficient foundation language models, *arXiv preprint*. arXiv:2302.13971.

Vaswani, Ashish, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin (2017), Attention is all you need, *Advances in Neural Information Processing Systems 30*, pp. 5998–6008.

Zhang, Susan, Stephen Roller, Naman Goyal, et al. (2022), OPT: Open pre-trained transformer language models, *arXiv preprint*. arXiv:2205.01068.

Zhao, Wenlong, Debanjan Mondal, Niket Tandon, Danica Dillion, Kurt Gray, and Yuling Gu (2024), WorldValuesBench: A large-scale benchmark dataset for multi-cultural value awareness of language models, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italia, pp. 17696–17706.

Zou, James and Londa Schiebinger (2018), AI can be sexist and racist—it's time to make it fair, *Nature* **559**, pp. 324–326.

Zowghi, Didar and Francesca da Rimini (2023), Diversity and inclusion in artificial intelligence, *arXiv preprint*. arXiv:2305.12728.