Multilingual != Multicultural: Evaluating Gaps Between Multilingual Capabilities and Cultural Alignment in LLMs

Jonathan Rystrøm
Oxford Internet Institute
University of Oxford, UK

Hannah Rose Kirk
Oxford Internet Institute
University of Oxford, UK

Scott A. Hale
Oxford Internet Institute
University of Oxford, UK

Correspondence: jonathan.rystrom@oii.ox.ac.uk

Abstract

Large Language Models (LLMs) are becoming increasingly capable across global languages. However, the ability to communicate across languages does not necessarily translate to appropriate cultural representations. A key concern is US-centric bias, where LLMs reflect US rather than local cultural values. We propose a novel methodology that compares LLM-generated response distributions against population-level opinion data from the World Value Survey across four languages (Danish, Dutch, English, and Portuguese). Using a rigorous linear mixed-effects regression framework, we compare three families of models: Google's Gemma models (2B–27B parameters), AI2's OLMo models (7B-32B parameters), and successive iterations of OpenAI's turbo-series. Across the families of models, we find no consistent relationships between language capabilities and cultural alignment. While the Gemma models have a positive correlation between language capability and cultural alignment across all languages, the OpenAI and OLMo models are inconsistent. Our results demonstrate that achieving meaningful cultural alignment requires dedicated effort beyond improving general language capabilities.

1 Introduction

Spearheaded by accessible chat interfaces to powerful models like ChatGPT (OpenAI, 2022), LLMs are reaching hundreds of millions of users (Milmo, 2023). These models are deployed across diverse contexts: from tutoring mathematics (Khan, 2023) to building software applications (Peng et al., 2023) to assisting in legal cases (Tan et al., 2023). While most LLMs demonstrate multilingual abilities (Üstün et al., 2024), the ability to communicate across languages does not necessarily translate into appropriate cultural representations. Disentangling language capabilities and cultural alignment is crucial for understanding how LLMs should be examined and audited (Mökander et al., 2024) and for ensuring these technologies work for diverse people (D'ignazio and Klein, 2023; Weidinger et al., 2022).

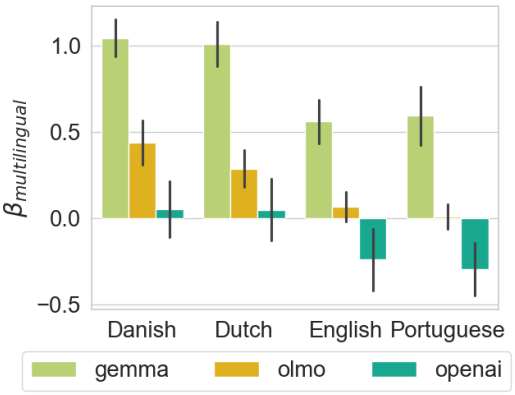

Figure 1: The relationship between multilingual capability and cultural alignment is inconsistent across LLM families, as shown by coefficients from our linear mixed-effects model ($\beta_{multilingual} = \beta_{flm}$; Eq. 3; §3.2). OpenAI and OLMo models show negative or insignificant relationships outside of Danish and Dutch, while Gemma models show positive relationships throughout ($p < .05$).

Given the Silicon Valley origins of many frontier AI labs and the prevalence of American English training data, we might expect LLMs to exhibit US-centric cultural biases despite their multilingual capabilities. These companies comprise a narrow slice of human experience, limiting the voices that contribute to critical design decisions in LLMs (D'ignazio and Klein, 2023). They typically train LLMs on massive amounts of predominantly English text and employ American crowd workers to rate and evaluate the LLMs' responses (Johnson et al., 2022; Kirk et al., 2023). Far too often, the benefits and harms of data technologies are unequally distributed, reinforcing biases and harming

Proceedings of the First Interdisciplinary Workshop on OMMM associated with RANLP 2025,
pages 74–85, Varna, Bulgaria, Sep 8-10, 2025.
https://doi.org/10.26615/978-954-452-101-1-009

already minoritized groups (Birhane, 2020; Milan and Treré, 2019; Khandelwal et al., 2024). Understanding how LLMs represent different cultures is thus paramount to establishing risks of representational harm (Rauh et al., 2022) and ensuring the technology's utility is shared across diverse communities.

Increasing diversity and cross-cultural understanding is stymied by unchecked assumptions in both alignment techniques and evaluation methodologies. First, there is an assumption that bigger and more capable LLMs trained on more data will be inherently easier to align (Zhou et al., 2023; Kundu et al., 2023), but this sidesteps the thorny question of pluralistic variation and cultural representations (Kirk et al., 2024b). Thus, it is unclear whether improvements in architecture (Fedus et al., 2022) and post-training methods (Kirk et al., 2023; Rafailov et al., 2023) translate into improvements in cultural alignment.

Although studies like the World Values Survey (WVS) have documented how values vary across cultures (EVS/WVS, 2022), it remains unclear whether more capable LLMs—through scaling or improved training—better align with these cultural differences (Bai et al., 2022; Kirk et al., 2023). While the WVS has been used in prior research on values in LLMs, these studies have focused predominantly on individual models' performance within an English-language context. (Cao et al., 2023; Arora et al., 2023; AlKhamissi et al., 2024). This paper addresses this gap by developing a methodology for assessing how well families of LLMs represent different cultural contexts across multiple languages. We compare two distinct paths to model improvement: systematic scaling of instruction-tuned models and commercial product development comprising scaling and innovation in post-training to accommodate pressures from capabilities, cost, and preferences (OpenAI et al., 2024b).

Given these considerations, we investigate the following research questions:

RQ1 Multilingual Cultural Alignment: Does improved multilingual capability increase LLM alignment with population-specific value distributions?

RQ2 US-centric Bias: When using different languages, do LLMs align more with US values or with values from the countries where these languages are native?

We operationalise *multilingual capability* as an LLM's performance on a range of multilingual benchmarks across languages (see, e.g., Nielsen, 2023). We describe the specific benchmarks and performances in the supplementary materials.

This work makes several key contributions. First, we introduce a novel distribution-based methodology for probing cultural alignment across languages, moving beyond direct survey approaches to better capture latent cultural values (Sorensen et al., 2024). Second, we provide the first systematic comparison of how improvements in scale and post-training affect cultural alignment and US-centric bias across English, Danish, Dutch, and Portuguese through a series of robust statistical models. Third, we release a dataset of model-generated responses across multiple languages and cultural contexts as well as our code, enabling future research into cultural alignment and bias.[1] Together, these contributions advance our understanding of how LLM development choices influence cultural representation while providing tools for ongoing investigation of these critical issues.

2 Measuring Cultural Alignment

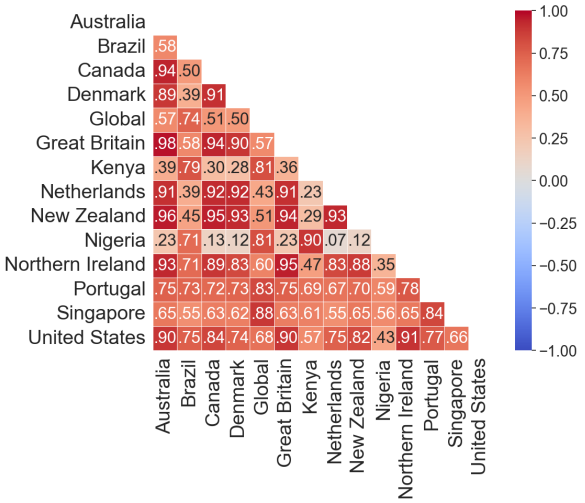

Figure 2: Pearson correlations in value polarity scores across studied countries from the World Values Survey. Value polarity scores are the fraction of the population in favour of a given topic. All correlations are positive, with most being between 0.7–0.95.

This section defines 'cultural alignment' and how to measure it in LLMs. We conceptualise cultural alignment as reproducing distributions of

[1] See github.com/jhrystrom/multicultural-alignment for code, data, and supplementary materials.

values in a particular population. Then we show how to a) get a ground-truth distribution of values using the World Values Survey (§2.1) and b) elicit value distributions from LLMs (§2.2).

Cultural alignment as value reproduction: Within a culture there will be a variety of stances to any particular topic. However, the *distribution* of stances will be characteristic among cultures. For instance, while around 8% of Danes are opposed to abortion, it is a much less contentious topic than in the US, where it's close to 40% (EVS/WVS, 2022).

We posit that cultural alignment for a specific group of people can be operationalised as how well an LLM reproduces the distribution of values over a wide range of topics (Sorensen et al., 2024). Investigating *distributions* of responses differs from previous work that directly surveys the LLMs as regular participants (e.g., Cao et al., 2023). This approach also addresses concerns raised by Khan et al. (2025) about the instability of survey-based evaluations by focusing on aggregate distributions rather than individual responses and incorporating explicit controls for response consistency. Our goal is to get more naturalistic elicitations of the underlying values whilst avoiding sycophancy and response bias (Sharma et al., 2023).

We operationalise reproduction as high correlations between *value polarity scores*: the fraction of people (or LLM responses) in favour of a topic in the population. Note, that we binarise issues to allow for simpler operationalisation. Below, we describe how we empirically estimate the value polarity score for the ground truth (§2.1) and LLMs (§2.2).

2.1 Ground Truth: World Values Survey

To get a 'ground truth' distribution of cultural values, we use the joint World Values Survey and European Values Survey (EVS; EVS/WVS, 2022). These surveys cover adults across 92 countries with samples that are nationally representative for gender, age, education, and religion. The surveys' broad coverage enables cross-cultural comparability for the many countries covered by the surveys, though some scholars note challenges in ensuring response comparability across countries (Alemán and Woods, 2016). The WVS provides both country and language identifiers for each respondent, allowing us to define populations either as citizens of a country or speakers of a language using the same underlying respondent-level data.

We select questions with binary agree/disagree or rating scale formats that allow clear classification of positive vs. negative stances, excluding questions with multiple categorical response options (see the supplementary materials for the full list of questions). These questions span environment, work, family, politics, religion, and security. We convert responses to binary indicators by determining whether each response indicates support for the measured construct, with custom coding to handle the various question formats and reverse-scored items. Finally, we calculate the value polarity score as the demographically weighted proportion of respondents with affirmative stances. Formally, we can define the value polarity score for a given population, $\mathcal{P}$ (e.g., citizens in a country or speakers of a language) and topic, q, (i.e., question within the EVS/WVS) as shown in Eq. 1:

$$\text{VPS}_{\mathcal{P},q} = \sum_{i \in \mathcal{P}} \frac{w_i}{\sum_{j \in \mathcal{P}_q} w_j} A_{i,q} \tag{1}$$

Here, $A_{i,q}$ is a binary indicator of whether participant i has a positive stance on topic q, w_i represents the survey-provided demographic weights, and $\mathcal{P}_q$ denotes respondents in population $\mathcal{P}$ who answered question q. The first term normalises the weights to account for missing responses and enables aggregation across any definition of a population (e.g., residents in a country, speakers of a language, etc.).

For example, if 80% of Danish respondents who answered the same-sex marriage question expressed support (after demographic reweighting), Denmark's value polarity score for this topic would be 0.8. Thus, a culture's values can be represented as a vector, where each element corresponds to a value polarity score for a specific topic.

2.2 Ecologically valid LLM responses

Testing cultural alignment effectively requires embedding contextual and cultural elements in ways that maintain ecological validity. At a high level, eliciting values from an LLM consist of two steps: 1) Iteratively prompting the model with the selected topics and 2) extracting the stances from each model response.

Setting prompt context: Developing ecologically valid prompts requires careful consideration. When evaluating LLM responses to value-laden topics, simply asking questions like "What proportion of people support topic X?" or "Do you support

topic X?" proves inadequate (e.g., Rozado, 2024). Such direct approaches suffer from three key limitations: they generate false positives through excessive agreement, fail to reflect realistic usage patterns, and provide insufficient variation to assess cultural alignment (Röttger et al., 2024). They also struggle to capture instance-specific harms that emerge when systems misalign with users' cultural contexts (Rauh et al., 2022).

Instead, we adopt an implicit approach by asking the model to generate responses from hypothetical respondents. For example, prompting "imagine surveying 10 random people on topic X. What are their responses?" This method reveals the model's latent opinion distribution while avoiding the limitations of direct questioning. Details for prompt construction are provided in the supplementary materials.

Seeding cultural responses: Having a method for eliciting distributions of values, the next step is to seed culture. One typical way of seeding a specific culture is to explicitly instruct the LLM either by mentioning a specific country ('imagine surveying 10 random Americans') or through describing specific personas ('Imagine surveying a 85-year-old Danish woman...'; AlKhamissi et al., 2024). The problem with these demographic prompting approaches is that they stray from actual uses of LLMs. Users are unlikely to explicitly mention their demographic information or nationality (Zheng et al., 2023a).

Instead, we use language as a proxy for cultural origin. For instance, a prompt in Danish is assumed to come from a Dane. This approach creates an intentional distinction in our analysis: we can compare 'language-level' alignment (all speakers of a language globally) with 'country-level' alignment (all people from specific nations where that language is native). As argued by Havaldar et al. (2023), users speaking a particular language would expect culturally appropriate responses in that language. For languages spoken in multiple countries, this approach is intentionally ambiguous. The ambiguity allows us to elicit the underlying 'default' alignment rather than the general ability to emulate cultures (Tao et al., 2024). We validate this approach by showing that LLM responses exhibit significantly lower self-consistency between languages compared to within languages, demonstrating that language impacts output (see the supplementary materials). To create prompts across languages, we use `gpt-3.5-turbo` to translate our original English prompts. Although previous literature has shown strong translation capabilities in LLMs (Yan et al., 2024), we nonetheless manually verify the translations.

Annotating and aggregating responses: Finally, to transform the LLMs' hypothetical survey responses into vectors of stances, we use an LLM-as-a-judge approach (Zheng et al., 2023b; Guerdan et al., 2025). Specifically, we use `gpt-4.1-mini` (OpenAI et al., 2025) to label each substatement as either 'pro', 'con', or 'null' given the context of the topic and a representative pro and con statement (generated with an LLM and validated by the authors). We then calculate the proportion of 'pro' versus 'con' responses as the LLM's value polarity score for the given statement. For instance, a response with seven 'pro', one 'con', and two 'null' statement would yield a value polarity score of 0.875 ($\frac{7}{8}$). A complete, unabridged example can be found in the supplementary materials. Formally, we label each substatement from the full set of hypothetical statements, $G_{q,g}$, for topic q and generation g as r. Furthermore, we label the classifier as $\ell(r)$. We then formalise the value polarity score for a given instance of a generation for a topic ($\text{VPS}^{\text{LLM}}_{q,g}$) as shown in Eq. 2:

$$\text{VPS}^{\text{LLM}}_{q,g} = \frac{\sum_{r \in G_{q,g}} [\, \ell(r) = \text{pro} \,]}{\sum_{r \in G_{q,g}} [\, \ell(r) \in \{\text{pro}, \text{con}\} \,]}, \quad (2)$$

These scores are then compared against the value polarity scores from the WVS. Specifically, we calculate the Spearman rank correlation to obtain a measure of similarity between the LLMs' responses and the value distributions of a given population.

To validate the LLM-as-judge, we manually annotate 200 statements. We iteratively refine the prompts and the LLM used until we reach satisfactory performance. We find a 91% agreement and a mean absolute error for value polarity of 4.5% over the dataset, ensuring consistent statistics between LLM and human annotation (Guerdan et al., 2025).

3 Experimental Setup

To investigate whether improving the multilingual capabilities of LLMs improves cultural alignment, we set up an experiment using a carefully chosen set of models and languages. We examine two

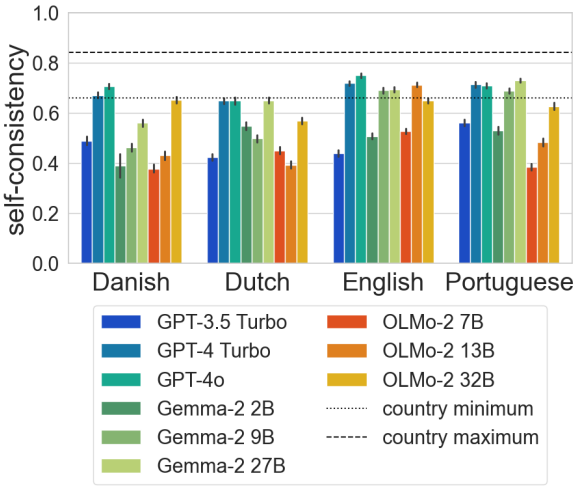

Figure 3: Self-consistency in responses for LLMs and WVS countries. LLMs have lower self-consistency than resampled WVS responses—shown by the dashed lines—particularly in non-English languages.

different kinds of model improvements: scaling and commercial product development. These cases provide complementary perspectives on the effects of multilingual capabilities on cultural alignment. Scaling is the most well-studied path to improving LLMs (Kaplan et al., 2020; Ganguli et al., 2022). Commercial product development, on the other hand, comprises both scale and innovation in post-training to accommodate different pressures from capabilities, cost, and preferences (Kirk et al., 2024a). For scaling, we use the instruction-tuned Gemma models (Gemma et al., 2024) and OLMo-2 models (OLMo et al., 2025), while for product development, we use OpenAI's turbo-series models (OpenAI, 2022; OpenAI et al., 2024a,b). We provide details of these model families in §3.1. A breakdown of the computational cost is in the supplementary materials.

Languages: For the languages, we compare English with Danish, Dutch, and Portuguese. This set allows us to test multiple assumptions about cultural alignment. English represents a widely used case: it is a global language with speakers across many countries represented in the WVS (see Fig. 2). This diversity allows us to assess whether LLMs align more strongly with US values or those of other English-speaking nations.

Danish and Dutch serve as controlled test cases since they are primarily used in a single country. If cultural alignment stems from pre-training data, models should show strong Danish/Dutch cultural alignment when using these languages, despite their small share of training data (Kreutzer et al., 2022). Alternatively, if alignment emerges from post-training processes—which are predominantly English-based (Blevins and Zettlemoyer, 2022)–responses in these languages should align more with US values.

Portuguese presents an interesting case since it is an official language in several countries. We investigate whether the LLM responses are more aligned to Portugal or Brazil—two countries that show distinct value patterns in relation to each other and the US (see Fig. 2). This allows us to test whether an LLM aligns more strongly with one country's values, the aggregate values of all language users, or US values.

For each language-model pair, we collect 300 prompt-response pairs to power our statistical analysis sufficiently (see §3.2). After filtering out responses that either lacked the required hypothetical survey format or were in a language other than the prompt, we obtained between 111–299 valid responses per combination. We calculate the correlation in value polarity scores at three levels: country (e.g., US or Denmark), language (pooling all speakers of a given language), and global (weighted values from all WVS/EVS participants).

3.1 Models

We examine three model families representing different development approaches: Gemma (Gemma et al., 2024) and OLMo (OLMo et al., 2025) for improvements through scaling and OpenAI's turbo series for commercial product development, combining scaling with post-training improvements (OpenAI, 2022; OpenAI et al., 2024a,b). Other preliminary experiments included different versions of LLaMA models (Touvron et al., 2023) and Mistral models (Jiang et al., 2023). However, these models either failed to consistently follow instructions or always answered in English regardless of the prompt language. See the supplementary materials for a more thorough description of the LLMs.

3.2 RQ1: Multilingual Cultural Alignment

To statistically assess whether improving the multilingual capabilities of LLMs improves cultural alignment, we construct a linear mixed-effects regression (LMER; Luke, 2017) based on the experimental setup described above. Our LMER follows standard practices and has three core components:

- **Core coefficient:** The coefficient of interest

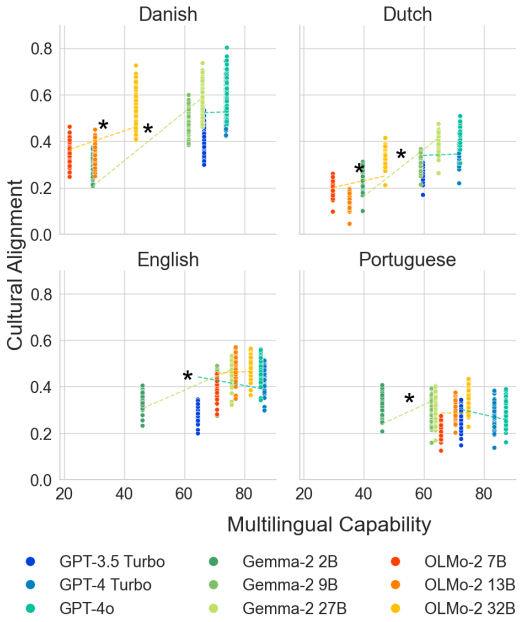

Figure 4: Language capability (x-axis) vs cultural alignment scores (y-axis) across languages. Stars indicate significance ($p < .05$) in our linear mixed-effects regression of multiple runs (See §3.2). OpenAI models (blue) and OLMo models (red) show negative/insignificant relationships outside of English, while the Gemma models (green) show positive relationships throughout ($p < .05$).

is the three-way interaction between model family, language, and multilingual capability. This tests whether the multilingual capability–alignment relationship differs by model family and response language, directly addressing **RQ1**.

- **Random effects:** We include a model-specific random intercept α_j to account for repeated measures of cultural alignment for the same LLM. This models variation between LLMs and can improve efficiency over standard linear regressions (Luke, 2017).

- **Control for self-consistency:** We include a consistency-by-language term to help ensure that higher alignment scores reflect genuine cultural adaptation rather than reduced response noise, which can inflate scores (Kahneman et al., 2021).

We calculate self-consistency as the Spearman correlation between value polarity scores (defined in §2) of repeated responses to identical topics, adjusted by the reliability of the LLM annotation (see §2.2; Charles, 2005). A score of 1.0 indicates perfect consistency; 0.0 indicates random responses. Population-level resampling of the human WVS responses yields values between 0.66 and 0.84 (see Fig.3 and the supplementary materials).

Formally, the model is specified in Eq. 3:

$$\begin{aligned} \text{CA}_i &\sim \mathcal{N}(\mu_i, \sigma^2), \\ \mu_i &= \alpha_{j[i]} + \beta_{1l}\, X_{\text{cons},i} X_{l,i} \\ &\quad + \beta_{flm}\, X_{m,i} X_{f,i} X_{l,i}, \\ \alpha_j &\sim \mathcal{N}(\mu_\alpha, \sigma_\alpha^2), \quad j = 1, \ldots, J. \end{aligned} \tag{3}$$

where i indexes responses and $j[i]$ denotes the LLM producing response i. Here $X_{\text{cons},i}$ is the self-consistency score for response i, $X_{l,i}$ is the set of language indicators, $X_{f,i}$ is the set of model-family indicators, and $X_{m,i}$ is the multilingual capability score. The residual variance σ^2 represents within-LLM variation in alignment scores not explained by the fixed effects or model-specific intercept, while σ_α^2 represents between-LLM variation in average alignment.

The above statistical model allows us to analyse the relationship between multilingual capabilities and cultural alignment in model families at the level of individual languages. For example, we might find that multilingual capabilities improve cultural alignment for Gemma models for Danish but not for Dutch or vice versa.

3.3 RQ2: US-Centric Bias

We analyse model bias by comparing cultural alignment between US and local values, where "local" refers to values in the country or countries where a given language is natively spoken. We define US-centric bias as an LLM showing higher cultural alignment with US value distributions compared to local ones. To quantify this bias, we use a linear regression model that measures the differential effect of US versus local value alignment:

$$\begin{aligned} \text{CA} &= \beta_0 + \beta_1(\text{US}) \\ &\quad + \sum_{m \in \mathcal{M}} \sum_{l \in \mathcal{L}} \beta_{ml}(m \times l) \\ &\quad + \sum_{m \in \mathcal{M}} \sum_{l \in \mathcal{L}} \beta_{ml}^{\text{US}}(\text{US} \times m \times l) + \epsilon \end{aligned} \tag{4}$$

The regression's intercept (β_0,i.e., the base case) is a baseline that produces uniformly random value polarity scores. $\mathcal{M}$ is the set of models and $\mathcal{L}$ is the set of languages. `US` is a boolean feature denoting

whether the cultural alignment is to the US (if 1) or the local values (if 0). We primarily analyse the coefficients with US (β_{ml}^{US}) since these provide the *partial* effect of US-centric bias, i.e., how much more/less a given LLM is aligned to US rather than local values. Assumption checks for the regression can be seen in the supplementary materials.

4 Results

4.1 Multilingual Cultural Alignment (RQ1)

We first examine the stability of LLMs' cultural values. For LLMs lacking stable internal values, apparent improvements in cultural alignment may reflect reduced response variance rather than genuine advances (Röttger et al., 2024; Kahneman et al., 2021). We therefore analyse both the self-consistency of LLM responses and how alignment changes with model improvements.

LLMs have low self-consistency: We find low self-consistency scores across all models and languages compared to human responses in the WVS data (Fig. 3). In contrast, LLMs show generally lower self-consistency compared to the human responses, even in English, where instruction-following capabilities are strongest due to English-dominated training data. (OpenAI et al., 2024a; Gemma et al., 2024; OLMo et al., 2025).

This lower self-consistency complicates our cultural alignment analysis (Wright et al., 2024). Drawing on Kahneman et al. (2021)'s noise framework, we recognise that inconsistent responses can be as detrimental as bias with respect to the accuracy of the analysis. To address the noise, we employ larger sample sizes and incorporate consistency controls in our regression analyses.

Multilinguality does not imply cultural alignment: The relationship between model improvements and cultural alignment varies substantially across languages and model families (Fig. 1). For Gemma, there is a strong and significant positive relationship between multilingual capabilities and cultural alignment for all languages. In contrast, the relationships for the GPT-Turbo models are either insignificant or negative. For Dutch and Danish the relationships are insignificant ($\beta_{\text{gpt,nl}} = 0.049, p = 0.589$,$\beta_{\text{gpt,da}} = 0.053, p = 0.522$), and for Portuguese and English the effect is significant and negative ($\beta_{\text{gpt,en}} = -0.24, p = 0.009$, $\beta_{\text{gpt,pt}} = -0.30, p < 0.001$). Similarly for OLMo, the relationship is positive for Danish and Dutch ($\beta_{\text{OLMo,da}} = 0.44, p < 0.001$, $\beta_{\text{OLMo,nl}} = 0.29, p < 0.001$) and insignificant for English and Portuguese ($\beta_{\text{OLMo,en}} = 0.068, p = 0.115$, $\beta_{\text{OLMo,pt}} = 0.008, p = 0.825$).

The mismatch between multilingual performance and cultural alignment could suggest a capability threshold: multilingual improvements might provide rudimentary instruction following skills (Nie et al., 2024), but beyond a point, other factors—such as the preferences of developers and annotators—dominate (Kirk et al., 2024b). This could explain the smaller open weights models' higher coefficients than the gpt-turbo models (see Fig. 4 or Fig. 1). Further work is needed to understand alignment at the sub-national level.

Furthermore, the strong effect of self-consistency ($0.405 < \beta_{\text{consistency}} < 0.723, p \ll 0.001$) compared to multilingual capability suggests that noise remains a major limiting factor in analysing cultural alignment. This aligns with broader findings about the instability of LLM value elicitation (Röttger et al., 2024; Khan et al., 2025). Moreover, even the highest observed alignment scores (around 0.7; see Fig 4) indicate substantial room for improvement in how well LLMs match human cultural values and behaviours.

In conclusion, our analysis reveals a complex relationship between model improvements and cultural alignment. Although some languages show progressive improvements in cultural alignment from model scaling or iterative commercial development, others show minimal or inconsistent improvements. These findings, combined with the relatively low self-consistency of LLM responses, demonstrate that improved multilingual capability does not guarantee better cultural alignment.

4.2 US-centric Bias (RQ2)

Here, we answer RQ2 by examining US bias across languages. Specifically, we investigate relative alignment between local and US values (Fig. 5).

Our analysis reveals distinct patterns of US-centric bias across both languages and model families (Fig. 5). Languages show different susceptibilities to US bias: only one of nine LLMs exhibits US-centric bias in Danish, all in English, all in Portuguese, and none in Dutch. Note that for English, these results mean that the LLM, on average, is relatively more aligned to US values compared to other English-speaking countries like Kenya or the United Kingdom. See the supplementary materials

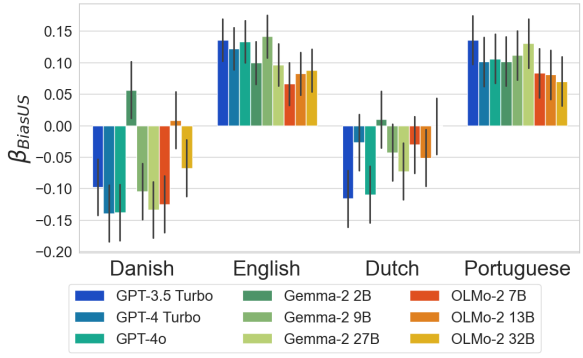

Figure 5: US-centric bias coefficients across LLMs and languages (β_{BiasUS}); see Eq. 4). Error bars are standard errors from the regression. Positive values indicate the presence of US-centric bias.

for detailed results.

The overarching pattern is that languages spoken *across* countries (English and Portuguese) show US-centric bias, whereas languages spoken in only *one* country (Danish and Dutch) show less US-centric bias. This supports the hypothesis that homogeneity in the training data can counteract US-centric bias—at least for medium-resourced, Western-European languages.

For LLMs, some specific LLMs seem more prone to bias across languages. Specifically, the small `gemma-2-2b-it` exhibits higher US-centric bias across every language except Dutch. Beyond that, we see no clear progressions in US-centric bias within any family.

In conclusion, language seems a stronger indicator of US-centric bias in LLMs compared to LLM development. Monocultural languages show insignificant to negative bias, while English and Portuguese show significant US-centric bias. Within each LLM family, we find no consistent nor significant change in US-centric bias across LLM versions. These findings underscore the complex relationship between multilingual capability and alignment.

5 Related Work

Recent work emphasizes the need for systematic auditing of LLMs' cultural alignment, particularly as these models are deployed globally (Kirk et al., 2024a; Mökander et al., 2024; Kirk et al., 2024b). Prior empirical approaches have primarily taken two paths: using transformations based on Hofstede's cultural dimensions framework or directly comparing against survey responses. Studies using Hofstede's dimensions (Masoud et al., 2025; Cao et al., 2023) provide structured cross-cultural comparisons through latent variable analysis. However, these studies assume that LLMs' latent dimensions map directly onto human dimensions, since they use formulas calibrated for humans—an assumption that warrants scrutiny (Shanahan, 2024; Schröder et al., 2025).

Recent work has explored using LLMs to simulate responses for assessing cultural alignment (Tao et al., 2024; AlKhamissi et al., 2024; Havaldar et al., 2023). Similarly to our work, these works show that LLMs struggle to represent underrepresented personas (AlKhamissi et al., 2024) and emotions (Havaldar et al., 2023) for non-English languages. Prior approaches focused on individual-level responses. In contrast, our method generates distributions of opinions across hypothetical survey participants, enabling direct comparison with population-level statistics. This distribution-based approach offers three key advantages. First, it better captures the inherent variation in cultural values within populations, paving the way for investigating distributional alignment (Sorensen et al., 2024). Second, it enables principled statistical comparison against large-scale survey data like the World Values Survey (EVS/WVS, 2022). Finally, the framework is easy to extend to new languages by automatically translating the prompts. We detail our quantitative framework for measuring alignment with observed population distributions in §2.

There is also an increasing body of work investigating political biases in LLMs (Röttger et al., 2024, 2025; Rozado, 2024). Much of this work also relies on human political surveys like the Political Compass Test. However, recent work has called for increased attention to how the randomness inherent in LLM decoding at non-zero temperatures can create instability in attributes (Röttger et al., 2024; Wright et al., 2024; Khan et al., 2025). We expand on this work by including multilingual perspectives and constructing prompts with a wide range of variations (see §2). These prompt variations, combined with statistically accounting for self-consistency in our statistical analysis (see §3.2), allow us to get a more robust measure of cultural alignment.

The relationship between model capabilities and cultural alignment remains understudied. Unlike general performance metrics that follow predictable scaling laws (Kaplan et al., 2020), cultural alignment may not improve systematically with model capabilities. This aligns with research show-

ing micro-level capabilities can be discontinuous with scale (Ganguli et al., 2022). The challenge is compounded in multilingual settings (Hoffmann et al., 2022), where static benchmarks with single correct answers fail to capture how cultural values are distributed across different topics and contexts.

Previous work has focused primarily on English-language performance (Tao et al., 2024) or individual LLMs (Arora et al., 2023; Cao et al., 2023). Our work extends this by examining how cultural alignment systematically varies within model families and across languages, providing insight into how different development approaches—scaling and commercial product development—influence cultural representation capabilities.

There is already progress on improving the cross-cultural participation in alignment data. Two notable projects are PRISM and AYA (Kirk et al., 2024b; Üstün et al., 2024). PRISM is a large dataset of conversational preferences from a diverse participant pool. While the data is predominantly in English, it could be an important resource for better understanding and modelling diverse cultural preferences. The AYA dataset is a massively multilingual instruction fine-tuning dataset. AYA could provide further means of realising the demonstrated benefits of multilingual training (Nie et al., 2024).

6 Conclusion

Increased multilingual capabilities do not guarantee improved cultural alignment in Large Language Models. Through systematic comparison of three model families—Gemma, OLMo, and OpenAI's GPTs—we find that the relationship between improvements in multilingual capability and cultural alignment is complex. While some languages show clear improvements in alignment with increased model capabilities (e.g., Danish), others exhibit inconsistent patterns, suggesting that cultural alignment does not automatically follow gains in multilingual capabilities. Our distribution-matching methodology using World Values Survey data enabled the detection of these nuanced patterns across languages and cultural contexts.

We also find that, contrary to popular discourse, LLMs do not exhibit US-centric bias across all languages; in Danish and Dutch, they align more closely with the values of Denmark and the Netherlands, respectively, than with the US. This fits with the hypothesis that more culturally uniform data leads to less US-centric bias. Both English and Portuguese are spoken in multiple countries, whereas Dutch and Danish are predominantly spoken in one. To further validate this claim, future work could include other multi-cultural languages (like Spanish or Swahili) and monocultural languages (like Japanese)—especially with a wider geographical reach to preclude European bias.

Our findings highlight that improving cultural alignment requires dedicated effort beyond general capability scaling. Future work should focus on developing techniques that can better handle alignment with distributions of cultural values rather than single points, while ensuring meaningful participation from diverse communities in LLM development. As these models continue to reach wider audiences spanning many geographic and cultural regions, achieving robust cultural alignment becomes increasingly crucial for equitable deployment.

Acknowledgements

We are thankful for the helpful feedback from the anonymous reviewers. We also thank Shiri Dori-Hacohen, Daniel Hershcovich, and others for helpful discussions throughout the project. For compute support, the project used the Microsoft Azure Accelerating Foundation Model Research Grant. This work was supported in part by the Engineering and Physical Sciences Research Council [grant number EP/X028909/1].

References

José Alemán and Dwayne Woods. 2016. Value Orientations From the World Values Survey: How Comparable Are They Cross-Nationally? *Comparative Political Studies*, 49(8):1039–1067.

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.

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.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, and 22 others. 2022.

Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback.

Abeba Birhane. 2020. Algorithmic colonization of Africa. *SCRIPTed*, 17:389.

Terra Blevins and Luke Zettlemoyer. 2022. Language contamination helps explains the cross-lingual capabilities of english pretrained models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3563–3574, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Eric P. Charles. 2005. The correction for attenuation due to measurement error: Clarifying concepts and creating confidence sets. *Psychological Methods*, 10(2):206–226.

Catherine D'ignazio and Lauren F. Klein. 2023. *Data Feminism*. MIT press.

EVS/WVS. 2022. Joint EVS/WVS 2017-2022 Dataset.

William Fedus, Jeff Dean, and Barret Zoph. 2022. A review of sparse expert models in deep learning.

Deep Ganguli, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova Dassarma, Dawn Drain, and Nelson Elhage. 2022. Predictability and surprise in large generative models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1747–1764.

Team Gemma, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Bobak Shahriari, Alexandre Ramé, Johan Ferret, and 187 others. 2024. Gemma 2: Improving open language models at a practical size.

Luke Guerdan, Solon Barocas, Kenneth Holstein, Hanna Wallach, Zhiwei Steven Wu, and Alexandra Chouldechova. 2025. Validating LLM-as-a-judge systems in the absence of gold labels.

Shreya Havaldar, Bhumika Singhal, Sunny Rai, Langchen Liu, Sharath Chandra Guntuku, and Lyle Ungar. 2023. Multilingual language models are not multicultural: A case study in emotion. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 202–214, Toronto, Canada. Association for Computational Linguistics.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, and 13 others. 2022. An empirical analysis of compute-optimal large language model training. *Advances in Neural Information Processing Systems*, 35:30016–30030.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, and 9 others. 2023. Mistral 7B.

Rebecca L. Johnson, Giada Pistilli, Natalia Menédez-González, Leslye Denisse Dias Duran, Enrico Panai, Julija Kalpokiene, and Donald Jay Bertulfo. 2022. The Ghost in the Machine has an American accent: Value conflict in GPT-3.

Daniel Kahneman, Olivier Sibony, and Cass R. Sunstein. 2021. *Noise: A Flaw in Human Judgment*. Little, Brown.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. *arXiv:2001.08361 [cs, stat]*.

Ariba Khan, Stephen Casper, and Dylan Hadfield-Menell. 2025. Randomness, not representation: The unreliability of evaluating cultural alignment in LLMs. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, pages 2151–2165, New York, NY, USA. Association for Computing Machinery.

Sal Khan. 2023. Harnessing GPT-4 so that all students benefit. A nonprofit approach for equal access!

Khyati Khandelwal, Manuel Tonneau, Andrew M. Bean, Hannah Rose Kirk, and Scott A. Hale. 2024. IndianBhED: A dataset for measuring india-centric biases in large language models. In *Proceedings of the 2024 International Conference on Information Technology for Social Good*, pages 231–239, Bremen Germany. ACM.

Hannah Rose Kirk, Andrew M. Bean, Bertie Vidgen, Paul Röttger, and Scott A. Hale. 2023. The past, present and better future of feedback learning in large language models for subjective human preferences and values. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2409–2430, Singapore. Association for Computational Linguistics.

Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A. Hale. 2024a. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*, 6(4):383–392.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Michael Bean, Katerina Margatina, Rafael Mosquera, Juan Manuel Ciro, Max Bartolo, Adina Williams, and 3 others. 2024b. The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language

models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, and 43 others. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Sandipan Kundu, Yuntao Bai, Saurav Kadavath, Amanda Askell, Andrew Callahan, Anna Chen, Anna Goldie, Avital Balwit, Azalia Mirhoseini, and 27 others. 2023. Specific versus General Principles for Constitutional AI.

Steven G. Luke. 2017. Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*, 49(4):1494–1502.

Reem Masoud, Ziquan Liu, Martin Ferianc, Philip C. Treleaven, and Miguel Rodrigues Rodrigues. 2025. Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8474–8503, Abu Dhabi, UAE. Association for Computational Linguistics.

Stefania Milan and Emiliano Treré. 2019. Big data from the south(s): Beyond data universalism. *Television & New Media*, 20(4):319–335.

Dan Milmo. 2023. ChatGPT reaches 100 million users two months after launch. *The Guardian*.

Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. 2024. Auditing large language models: A three-layered approach. *AI and Ethics*, 4(4):1085–1115.

Shangrui Nie, Michael Fromm, Charles Welch, Rebekka Görge, Akbar Karimi, Joan Plepi, Nazia Mowmita, Nicolas Flores-Herr, Mehdi Ali, and Lucie Flek. 2024. Do multilingual large language models mitigate stereotype bias? In *Proceedings of the 2nd Workshop on Cross-cultural Considerations in NLP*, pages 65–83, Bangkok, Thailand. Association for Computational Linguistics.

Dan Nielsen. 2023. ScandEval: A benchmark for scandinavian natural language processing. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 185–201, Tórshavn, Faroe Islands. University of Tartu Library.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, and 31 others. 2025. 2 OLMo 2 furious.

OpenAI. 2022. ChatGPT: Optimizing Language Models for Dialogue.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, and 272 others. 2024a. GPT-4 technical report.

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, A. J. Ostrow, Akila Welihinda, and 410 others. 2024b. GPT-4o system card.

OpenAI, Ananya Kumar, Jiahui Yu, John Hallman, and Michelle Pokrass. 2025. Introducing GPT-4.1.

Sida Peng, Eirini Kalliamvakou, Peter Cihon, and Mert Demirer. 2023. The impact of AI on developer productivity: Evidence from GitHub copilot.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

Maribeth Rauh, John Mellor, Jonathan Uesato, Po-Sen Huang, Johannes Welbl, Laura Weidinger, Sumanth Dathathri, Amelia Glaese, Geoffrey Irving, and 3 others. 2022. Characteristics of harmful text: Towards rigorous benchmarking of language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, pages 24720–24739, Red Hook, NY, USA. Curran Associates Inc.

Paul Röttger, Musashi Hinck, Valentin Hofmann, Kobi Hackenburg, Valentina Pyatkin, Faeze Brahman, and Dirk Hovy. 2025. IssueBench: Millions of realistic prompts for measuring issue bias in LLM writing assistance.

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Kirk, Hinrich Schuetze, and Dirk Hovy. 2024. Political compass or spinning arrow? Towards more meaningful evaluations for values and opinions in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15295–15311, Bangkok, Thailand. Association for Computational Linguistics.

David Rozado. 2024. The political preferences of LLMs. *PLOS One*, 19(7):e0306621.

Sarah Schröder, Thekla Morgenroth, Ulrike Kuhl, Valerie Vaquet, and Benjamin Paaßen. 2025. Large language models do not simulate human psychology.

Murray Shanahan. 2024. Talking about large language models. *Commun. ACM*, 67(2):68–79.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, and 9 others. 2023. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L. Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, and 3 others. 2024. Position: A roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, pages 46280–46302. PMLR.

Jinzhe Tan, Hannes Westermann, and Karim Benyekhlef. 2023. Chatgpt as an artificial lawyer? In *Ai4aj@ Icail*.

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

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, and 5 others. 2023. LLaMA: Open and Efficient Foundation Language Models.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, and 8 others. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, and 14 others. 2022. Taxonomy of Risks posed by Language Models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 214–229, Seoul Republic of Korea. ACM.

Dustin Wright, Arnav Arora, Nadav Borenstein, Srishti Yadav, Serge Belongie, and Isabelle Augenstein. 2024. LLM tropes: Revealing fine-grained values and opinions in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 17085–17112, Miami, Florida, USA. Association for Computational Linguistics.

Jianhao Yan, Pingchuan Yan, Yulong Chen, Judy Li, Xianchao Zhu, and Yue Zhang. 2024. GPT-4 vs. Human translators: A comprehensive evaluation of translation quality across languages, domains, and expertise levels. *Corr*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, and 4 others. 2023a. LMSYS-chat-1M: A large-scale real-world LLM conversation dataset. In *The Twelfth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, and 4 others. 2023b. Judging LLM-as-a-judge with MT-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, and 6 others. 2023. LIMA: Less is more for alignment. In *Thirty-Seventh Conference on Neural Information Processing Systems*.