# Extrinsic Evaluation of Cultural Competence in Large Language Models

**Anonymous ACL submission**

## Abstract

Productive interactions between diverse users and language technologies require outputs from the latter to be culturally relevant and sensitive. Prior works have evaluated models' knowledge of cultural norms, values, and artefacts, without considering how this knowledge manifests in downstream applications. In this work, we focus on extrinsic evaluation of cultural competence in two text generation tasks, open-ended question answering and story generation. We quantitatively and qualitatively evaluate model outputs when an explicit cue of culture, specifically nationality, is perturbed in the prompts. Although we find that model outputs do vary when varying nationalities and feature culturally relevant words, we also find weak correlations between text similarity of outputs for different countries and the cultural values of these countries. Finally, we discuss important considerations in designing comprehensive evaluation of cultural competence in user-facing tasks.

## 1 Introduction

*Cultural competence* is the ability to effectively and appropriately communicate with socioculturally different audiences (Deardorff, 2009).[1] People demonstrate cultural competence by tailoring their utterances to the participants in a conversation (Bell, 1984; Hawkins et al., 2021; Wu et al., 2023a). These adaptations range from sociolinguistic variations (e.g., using 'soccer' or 'football' depending on the context) to appropriately using facts (e.g., in India, the Prime Minister is the head of the government, but in USA, the President is). Hence, for effectively serving diverse users, outputs from large language models (LLMs) need to be culturally relevant (Hovy and Yang, 2021).

Cultural competence consists of multiple components, including a person's *knowledge* of a culture, which then supplements their *skills* of effectively communicating with people from that culture (Deardorff, 2006; Fantini and Tirmizi, 2006; Alizadeh and Chavan, 2016). So, cultural competence of LLMs should also be evaluated along both these aspects. Contemporary works have largely targeted the *knowledge* component of cultural competence by evaluating LLMs' knowledge of cultural values, norms, and artefacts (§ 2.2). Such evaluation is *intrinsic* because it is decoupled from the manifestation of this knowledge in downstream applications (Jones and Galliers, 1995).

In this work, we focus on *extrinsic* evaluation of cultural competence. Extrinsic evaluation setups should closely mimic user interactions with a system (Jones and Galliers, 1995). We select the tasks of story generation and open-ended question answering (QA), both of which have high representation in user interactions with LLMs (Zhao et al., 2024). We evaluate the lexical variations in outputs of 6 LLMs for 195 nationalities, and by proxy culture, for these tasks using both qualitative and quantitative analyses. Further, recent intrinsic evaluations have heavily relied on surveys from crosscultural psychology, like Hofstede's Cultural Dimensions (Hofstede et al., 2010) and World Values Survey (Haerpfer et al., 2022), as a measure of cultural values across countries (Arora et al., 2023; Cao et al., 2023; Durmus et al., 2023; Ramezani and Xu, 2023; AlKhamissi et al., 2024; Masoud et al., 2024). Thus, we evaluate whether the text distributions of outputs correlate with the cultural values of countries, as captured by these surveys. Our three main research questions are:

**RQ1:** Do models vary outputs when explicit cues of culture are present in the input prompt?

**RQ2:** Do model outputs contain culturally relevant vocabulary?

**RQ3:** Are model outputs for countries with similar cultural values, also similar?

[1] interchangeably, the terms intercultural competence and crosscultural competence are also used.

By measuring the variance in the outputs, we find that models make non-trivial adaptations for different nationalities (§ 5.1). Next, inspecting the vocabulary of these outputs, we find that they contain culturally relevant words (§ 5.2). Finally, we find only a weak correlation between the text distributions and cultural values of countries, as measured by crosscultural psychology surveys frequently used in contemporary work (§ 5.3).

Our findings show that intrinsic and extrinsic measures of cultural competence do not correlate. This necessitates developing holistic evaluations to analyse cultural competence in tasks representative of user interactions with LLMs.

All our code and data will be open-sourced.

## 2 Related Work

### 2.1 Cultural Competence

*Cultural competence* is the ability to effectively communicate with a socioculturally different audience (Deardorff, 2009). While multiple definitions exist (Alizadeh and Chavan, 2016), agreed-upon components include (a) the *awareness* about one's positionality and attitude, (b) the *knowledge* about the language, values, beliefs, practices, symbols etc. of a culture, and (c) the *skill* of appropriately using this *knowledge* when communicating (Howard-Hamilton et al., 1998; Deardorff, 2006; Fantini and Tirmizi, 2006; Deardorff, 2009).[2]

The *knowledge* component requires understanding differences in values, beliefs, and preferences across societies. Surveys in crosscultural psychology, like Hofstede's Cultural Dimensions (HCD) (Hofstede, 2001) and World Values Survey (WVS) (Haerpfer et al., 2022) attempt to elicit these differences across cultures, proxied by nationalities, using value-based questions.[3] Survey responses from a large number of individuals are used to quantify the differences in cultural values across countries. Hofstede's theory, in particular, has been widely adopted in fields requiring cultural competence such as communication, education, business, and healthcare (Ahern et al., 2012; Burai, 2016; Chang and Wu, 2023; Singh and Kumari, 2023). [4]

### 2.2 Cultural Competence in LLMs

There is a growing body of work on ensuring that LLMs align with diverse human values (Hershcovich et al., 2022; Wu et al., 2023b; Kirk et al., 2024; Sorensen et al., 2024) and can serve socioculturally diverse users (Hovy and Yang, 2021; Hershcovich et al., 2022; Adilazuarda et al., 2024). Specifically, prior works have evaluated LLMs for:

1. *Reflection of diverse cultural values* on crosscultural psychology surveys (like HCD and WVS) using MCQs, Chain of Thought prompting, or personas (Arora et al., 2023; Cao et al., 2023; Durmus et al., 2023; Ramezani and Xu, 2023; AlKhamissi et al., 2024; Masoud et al., 2024).

2. *Knowledge about varying norms* in social settings like dining, gifting, etc., using yes-no questions (Dwivedi et al., 2023), natural language inference (Huang and Yang, 2023), red-teaming (Chiu et al., 2024), situational questions (Rao et al., 2024; Shi et al., 2024), and graphs (Acharya et al., 2020).

3. *Commonsense and figurative language understanding* using MCQs (Nguyen et al., 2023; Palta and Rudinger, 2023; Kabra et al., 2023; Kim et al., 2024; Koto et al., 2024; Wang et al., 2024), and pragmatic games (Shaikh et al., 2023).

4. *Information about cultural artefacts* like food, clothing, etc. (Li et al., 2024b; Seth et al., 2024).

These works reveal gaps in LLMs' knowledge of non-western cultures, complimenting known demographic biases in LLMs (Mishra et al., 2020; Zhou et al., 2022; Basu et al., 2023; Jha et al., 2023; Schwöbel et al., 2023; Naous et al., 2024).

These evaluations focus on the *knowledge* component of cultural competence and are *intrinsic* because they are decoupled from the manifestation of this knowledge in user-facing tasks. Our work is complementary as we evaluate cultural competence in the *extrinsic* setup of text generation.

## 3 Extrinsic Evaluation of Cultural Competence

Jones and Galliers (1995) describe *extrinsic* evaluation criteria as, '*those relating to its function, i.e its role in relation to its setup's purpose*'. So, consider the two broad use cases of LLMs: (a) classification tasks, and (b) generation tasks. While

[2] For LLMs, we only rely on analogy to 'knowledge' and 'skills', and do not invoke analogies to 'awareness'.

[3] For example one of the questions in the Hofstede's survey is "In choosing an ideal job, how important would it be to you to have sufficient time for your personal or home life?".

[4] We note that defining the underpinning concept of culture itself remains elusive. Numerous works have attempted to synthesize the definitions of culture across disciplines, highlighting its complex and multi-faceted nature (Kroeber and Kluckhohn, 1952; Baldwin et al., 2006). Broadly, culture is a shared collection of knowledge, values, practices, norms, and beliefs that manifest in expression as behavioural and linguistic patterns (Kroeber and Kluckhohn, 1952).

incorporating cultural knowledge has been shown to benefit classification tasks like hate-speech detection and commonsense reasoning (Zhou et al., 2023; Li et al., 2024a; Shi et al., 2024), to the best of our knowledge there is no prior work focusing on open-ended text generation tasks.

Specifically, we obtain model outputs when nationalities in prompts are perturbed. We propose quantitative (§ 3.1) and qualitative (§ 3.2) analyses to evaluate these outputs for cultural competence.

### 3.1 Quantitative Evaluation

We evaluate outputs quantitatively in two ways:

**Lexical Variance** In order to quantify how much the generated language varies when nationalities are perturbed, we measure the variance in distance between outputs, where distance is computed according to a specific representation (§ 4.3.1).

**Correlation with Cultural Values** Prior works have relied on cultural values measured by surveys like HCD and WVS for intrinsic evaluation of cultural competence (Arora et al., 2023; Cao et al., 2023; Durmus et al., 2023; Ramezani and Xu, 2023; AlKhamissi et al., 2024; Masoud et al., 2024). So, we evaluate whether the text distributions of outputs correlate with distributions of cultural values. The intuition is to analyse whether countries with similar cultural values have similar text outputs.

We use the Kendall's $\tau_c$ rank correlation for this analysis. For each nationality (called the anchor), we rank all other countries by: (a) the similarity between their output to the output of the anchor, and (b) the difference in their cultural values and that of the anchor to the anchor. We use Kendall's $\tau_c$ to calculate the rank correlation of these rankings.

### 3.2 Qualitative Evaluation

The qualitative evaluation is intended to assess the characteristics of the outputs when the nationalities are perturbed. For this, we inspect the vocabulary of the LLM outputs by surfacing words that occur more frequently in the outputs of a particular country. We used the TF-IDF statistic to obtain words highly relevant to a particular country. The outputs were first tokenized using NLTK (Bird et al., 2009). Then, we created term frequency vocabulary of all the unigrams occurring in the outputs for each country, considering all outputs of a country as a single 'document'. We then calculate the TF-IDF score for all these unigrams and manually inspect the top 15 words for a subset of countries.

## 4 Experimental Setup

### 4.1 Tasks and Data

We select two tasks, story generation and open-ended question answering for our experiments. These were selected as they fulfil two main criteria. First, they have a sizeable representation in user interactions with LLMs (Zhao et al., 2024). And second, they represent diverse types of generation tasks with story generation on the creative end of the spectrum, while open-ended question answering being on the factual end of the spectrum.

**Open-Ended Question Answering (QA)** We created a list of 345 topics across 13 categories. We selected the categories (biology, chemistry, economics, environment, humanities, history, law, maths, physics, politics, religion, space, and world affairs) to ensure diversity in topics. Next, we curated topics for each category by referring to textbooks and encyclopedias.[5] Examples of topics include: 'elections' in 'politics', 'inertia' in 'physics', 'photosynthesis' in 'biology'. For this task, use a simple prompt template:

'Explain `{topic}` to a/an `{nationality}` person in English.'

These results in prompts like 'Explain elections to an Indian Person in English'.[6]

**Story Generation** We created a list of 35 topics for children's stories. We used online websites and children's storybooks to come up with topics. Examples include topics like moral values ('honesty', 'kindness'), characters ('farm animals', 'birds'), and places ('school', 'jungle'). Similar to QA, we use a simple prompt template:

'Write a children's story about `{topic}` for a/an `{nationality}` kid in English.'

This results in prompts like 'Write a children's story about honesty for a Japanese kid in English.'

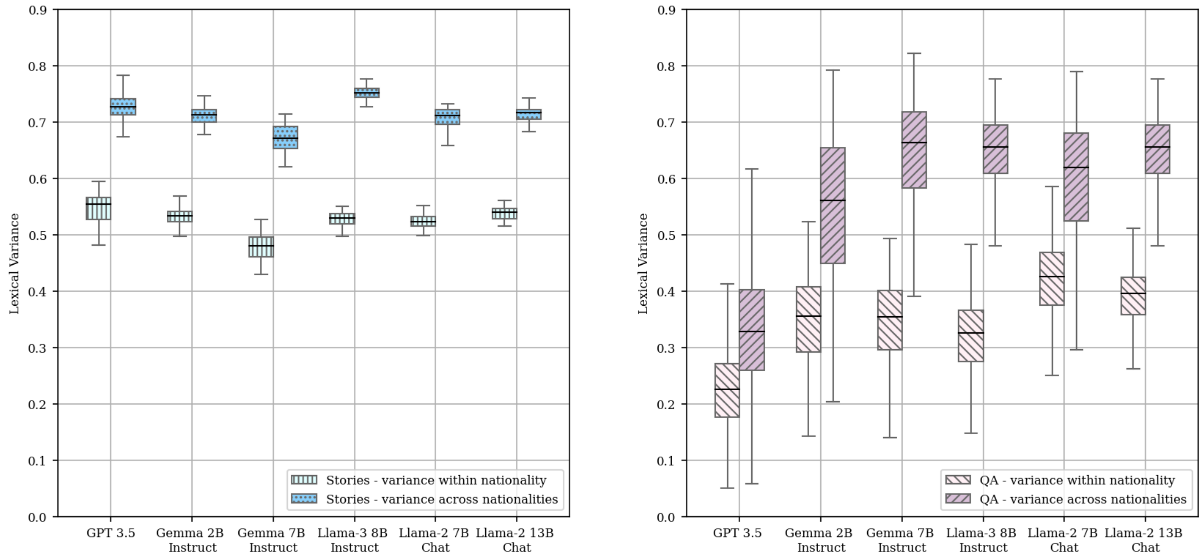
### 4.2 Models

We evaluate the following LLMs: (a) GPT 3.5 Turbo (`gpt-3.5-turbo-0125`)[7], queried via API between February 23 and March 28 2024.

[5]The datasheet (Gebru et al., 2021) is in Appendix A and curation logs will be released with the data upon publication.

[6]We observed that not including the phrase 'in English' in the prompt resulted in GPT 3.5's output often being in the dominant language of the particular country, for example for 'Mexican' the output is in Spanish. While this is an interesting phenomenon, analyzing this is beyond the scope of this paper.

[7]https://platform.openai.com/docs/models/gpt-3-5-turbo




(a) Story Generation

(b) Question Answering

Figure 1: Lexical Variance in outputs. The variance of outputs across nationalities is consistently higher than the variance of outputs within nationalities. Story generation has a higher median variance than QA across models.

(b) Gemma 2B instruct and 7B instruct (Team et al., 2024) (c) Llama 2 7B chat and 13B chat (Touvron et al., 2023) (d) Llama 3 8B instruct (AI@Meta, 2024) . We sample 5 responses per prompt, using a temperature of 0.3. We generate a maximum of 100 tokens for QA and 1000 tokens for stories.

## 4.3 Metrics

### 4.3.1 Text Similarity

**BLEU** BLEU (Papineni et al., 2002) calculates the precision of the n-grams present in the model-generated candidate text as compared to a gold reference text. We re-purpose this to calculate the similarity between two outputs. Because BLEU is not symmetric, we take the average of the two possible BLEU scores, one with each of the outputs as a candidate and the other as a reference.

**Word Edit Distance (WED)** WED is word-level Levenshtein distance (Levenshtein, 1966), normalized by the length of the longer text.

We picked BLEU and WED to focus on capturing the differences in lexical items between two outputs, e.g., the use of ‘soccer’ or ‘football’.[8]

[8]In early expeirments we found that semantic metrics like BERTscore (Zhang* et al., 2020) or embedding similarity might not be suitable because: (a) a lot of culturally relevant words from the outputs were converted to [UNK] tokens, (b) we did not see differences in the embeddings for outputs that were qualitatively different, especially in QA; perhaps partly because of (a) and and because, intuitively the the different words convey the same meaning.

### 4.3.2 Difference in Cultural Values

Following prior work, we rely on data from cross-cultural psychology surveys to measure the difference in cultural values among countries.

**Hofstede’s Cultural Dimensions (HCD)** Hofstede’s cultural theory quantifies the culture of a country along 6 dimensions. Using the VSM2013 version of the data available for 94 countries, we represent each country with 6 dimensions.[9]

**World Values Survey (WVS)** We use data from 64 countries and represent each country with 249 dimensions using the 249 questions from WVS[10][11]

We calculate the distance in cultural values between two countries as the magnitude of the vector distance between their HCD or WVS representations.

## 5 Results

### 5.1 Variance due to Nationality Perturbation

Our first research question was to analyse the extent of variation in outputs when nationalities are perturbed in the prompt. For this, we quantify the lexical variance (§ 3.1) in outputs, as measured by word edit distance in Figure 1. We find that model

[9]https://geerthofstede.com/research-and-vsm/dimension-data-matrix/

[10]There are additional questions that are either non-ordinal or descriptive in nature or are experimental, which we ignore.

[11]https://www.worldvaluessurvey.org/WVSDocumentationWV7.jsp

| Nationality | Top 15 highest TF-IDF scoring words for GPT 3.5's outputs of Story Generation |
|---|---|
| Afghan | amir , ali , afghanistan , ahmad , zahra , amina , rostam , babar , sara , omar , cally , farid , afghan , treehouse , bari |
| American | tommy , lily , america , jack , jake , buddy , mommy , max , town , daddy , acres , sarah , finley , assignment , surgery |
| British | oliver , england , jack , tommy , lily , willowbrook , thomas , sherwood , littleton , emily , british , jones , merlin , london , teddy |
| Canadian | liam , canada , emily , jack , alex , sarah , canadian , maple , tim , lily , beavers , smith , sammy , moose , robby |
| Chinese | li , mei , china , ming , chen , wu , wei , xiao , wukong , feather , ping , lake , bao , snowball , chinese |
| German | hans , germany , lena , anna , fritz , max , gretchen , bauer , lorelei , herr , lila , liesl , rübezahl , emma , karl |
| Indian | raj , india , rani , arjun , ravi , priya , guru , peacock , krishna , raja , meena , gupta , durga , beggar , temple |
| Nigerian | kola , nigeria , tunde , bola , kemi , ade , oya , adaeze , ayo , zuri , lagos , jide , nigerian , simba , heron |

Table 1: Top 15 highest TF-IDF scoring words for GPT 3.5's outputs of story generation for selected countries

outputs do vary with changing nationalities for both tasks across models. Moreover, these variations are non-trivial and task dependent, as described below.

**Control experiment: variance within nationality** We want to ensure that that the variance observed across nationalities are non-trivial, i.e. they do not occur because of the non-deterministic nature of generation in models. For this, we also measure the variance within multiple outputs for a particular nationality. We find that the variance for outputs within nationality is consistently lower than the variance across nationalities. We confirm this with ANOVA having a p-value of <0.05 (Appendix B.2).

**Effect of task on variance** We find that the nature of the task affects the extent of variation. The median variance for story generation is higher than the median variance for QA for every model. This might be expected as story generation, had longer outputs and being a creative task allows for more adaptations. On the other hand, the difference between the upper and lower quartiles of variance for QA is larger than that for stories. This is likely because QA consists of a wider variety of topics ranging from scientific categories, where limited variations might be expected, to topics on politics and history, that allow more variation in answers than others. For example answers while explaining 'elections' (politics) might vary more as they are operationalized differently across countries, but explaining 'inertia' (physics) might not vary as much.

### 5.2 Culturally Relevant Words in Outputs

Our second research question was to characterize the content of the outputs and understand whether they contain culturally relevant words. For this, we inspected the vocabulary of the outputs. We extracted words highly correlated to a country using TF-IDF (§ 3.2). The top 15 words from a subset of countries from outputs of GPT 3.5 for story generation and topics in the politics category from QA are presented in Table 1 and 2, respectively.

We see that story generation outputs feature different names across countries. For example, 'amir' in Afghanistan, 'raj' in India, and 'oliver' in Britain. Other culturally salient artefacts such as 'temple' and 'peacock' for Indian, 'bao' in Chinese, and 'london' for UK, etc. also show up in the list.

For the topics in the politics category of the QA task, we see words referring to senate houses and political offices of the countries, for example, 'lok sabha' and 'rajya sabha' in India, 'bundestag' for Germany, and 'meshrano jirga' and 'wolesi jirga' for Afghanistan. The list also features politically polarised issues such as 'gun' in America and 'brexit' in UK. Another common feature is the names of political parties, such as 'bjp' in India, 'apc' and 'pdp' in Nigeria, and 'ndp' in Canada.

Finally, we note that the cultural relevance of all the words on the lists is not obvious (e.g 'notably' in German in Table 2). Moreover, not all topics in the QA setting surface such interpretable lists of culturally relevant words. Especially lexicon from scientific topics did not reveal interesting

| Nationality | Top 15 highest TF-IDF scoring words for GPT 3.5's outputs for 'Politics' in QA |
|---|---|
| Afghan | afghanistan, jirga , ballot , wolesi , meshrano , elders , afghan , tribal , partners , box , strategies , target , stake , exploited , dynamics |
| American | united , states , basis , four , american , expanded , gun , fundraising , accent , congress , qualifications , residency , requirements , allowed , register |
| British | uk , british , mps , commons , reach , becomes , five , earlier , lords , scottish , brexit , kingdom , evolved , socioeconomic , previously |
| Canadian | provincial , municipal , federal , age , levels , grassroots, shapes, riding, sector, aggression, canadian, guaranteed, ndp quebec, ontario |
| Chinese | royalty , enacted , self-interests, solving , achieving, something , channels , box , health , directing , self-governing , capable , prosperous , citizenship , accumulation , accomplish |
| German | bundestag , totalitarian , he , argued , precedence , opposed , germany , upholds , notably , tourism , showcase , transition , mixed , emerged , europe |
| Indian | india , sabha , lok , rajya , linguistic , lacking , flexibility , chance , violent , anarch , hindu , bharatiya , janata , bjp , indian |
| Nigerian | nigeria, guarantees, figureheads, progressives, apc, pdp, purely, senators, problem, finances, identification, evenly, leave, lawlessness, governors |

Table 2: Top 15 highest TF-IDF scoring words for GPT 3.5's outputs for 'politics' in QA for selected countries

examples when inspecting the top-scoring TF-IDF words. This further compliments our earlier finding of output variations being different across tasks.

### 5.3 Correlation in Outputs & Cultural Values

Our third research question is analysing whether the outputs for countries with similar cultural values are similar. We report the Kendall's $\tau_c$ rank correlation (§ 3.1), averaged across countries, between BLEU text similarity and distance in cultural values measured by HCD and WVS in Figure 2.

**Effect of measure of cultural value used** When HCD is used as the measure of difference in cultural values (Figure 2a), we find that median correlation across the board[12] is greater than 0. This implies a small but positive correlation between the text distribution and cultural values of countries as measured by HCD. However, when WVS data is used, we find a small and negative correlation between text distribution and cultural values as measured by WVS (Figure 2b). All the rank correlation values were statistically significant within a significance interval of 95% in a two-sided p-test.

**Correlation for different countries** Next, we analyse the Kendall's $\tau_c$ rank correlation for different countries. Figure 3, shows two example plots for GPT 3.5 for story generation. We find that the correlation for USA, Canada, and India (in HCD) is negative, while that of Russia, China, Japan, and Australia is positive. South American, African, Southeast Asian and European countries are split between positive and negative values. This is interesting as prior work has found gaps in models' knowledge of non-western cultures (for example AlKhamissi et al. (2024); Masoud et al. (2024)), but we do not see a similar trend. Overall, the trend for each country is similar for HCD and WVS.

[12]except QA for Gemma 2B Instruct

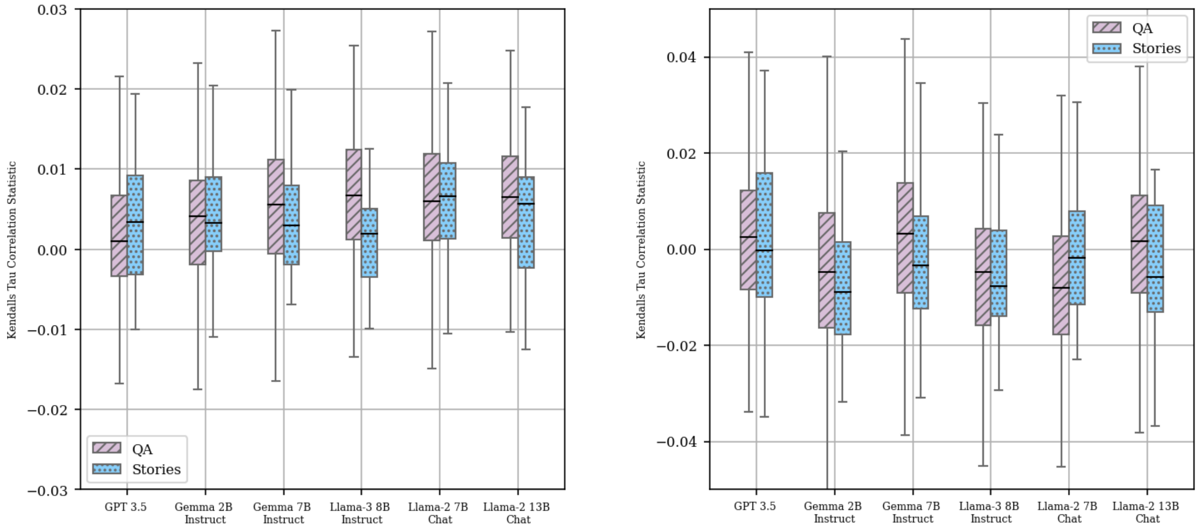
## 6 Discussion

**Correlation between Intrinsic and Extrinsic Metrics of Cultural Competence**

Together the findings for RQ2 (§ 5.2) and RQ3 (§ 5.3) suggest that intrinsic and extrinsic measures of cultural competence are not correlated. On the one hand, model outputs from our extrinsic setup feature culturally relevant words (§ 5.2). On the other hand, the text distributions are only weakly correlated with measures of cultural values widely used in intrinsic evaluations of cultural competence (§ 5.3). Thus, even if an LLM reflects the values of every country perfectly (as prior work measures by Hofstede's Cultural Dimensions or World Values Survey), this ability may not be reflective of cultural competence in downstream tasks.[13]

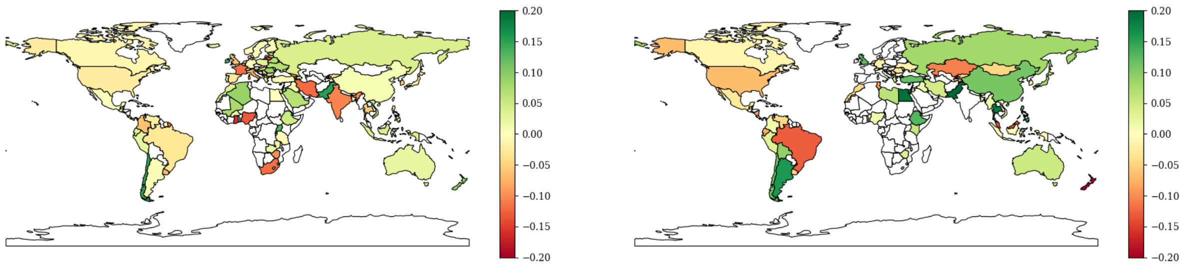
These findings underscore the importance of extrinsic evaluation of cultural competence. We thus

[13]Complementary facets of intrinsic and extrinsic evaluation have been observed in multiple settings. For example, there is limited correlation between intrinsic and extrinsic fairness metrics (Gonen and Goldberg, 2019; Goldfarb-Tarrant et al., 2021; Cao et al., 2022), and in intrinsic metrics of language model quality (like perplexity) and downstream task performance(Faruqui et al., 2016; Dudy and Bedrick, 2020).




(a) Correlation with Hofstede's Cultural Dimensions (HCD) (b) Correlation with World Values Survey (WVS)

Figure 2: Kendall's $\tau_c$ rank correlation between text distribution and cultural closeness of countries. For both plots, text similarity is measured using **BLEU**. For HCD correlation statistic values are greater than 0, implying a small but positive correlation (2a). However, for WVS, most correlations are less than 0, indicating small and negative correlation (2b). There are no clear trends among different models or tasks.




(a) Correlation with Hofstede's Cultural Dimensions (HCD) (b) Correlation with World Values Survey (WVS)

Figure 3: Kendall's $\tau_c$ rank correlation between cultural closeness and text outputs of **story generation** for GPT 3.5. For both plots, text similarity is measured using **BLEU**. There is a mix of positive (green) and negative (red) correlation. Russia, China, and Australia have positive correlations while India, USA, and Canada have negative correlations. European, South American, and African countries are split between positive and negative correlations.

believe that future work on advancing cultural competence should focus on tasks reflective of user interactions with language technologies.

#### Need for Comprehensive Human Evaluation

Our results show that models adapt to explicit cues of culture with culturally relevant words (§ 5.2. But, it is unclear how this will affect user experience. In prior work, Lucy et al. (2023) found mixed reactions from users when an email auto-reply system adapted to cues of their identities. Moreover, we do not consider any implicit cues of culture, like dialect or topical differences in queries (Kirk et al., 2024). Thus, understanding whether model adaptations triggered by implicit and explicit cues of culture are useful or desired by users remains open.

Further, as the qualitative evaluation shows, the output contains names that are typically associated with the ethnic majorities of the country. This is reflective of biases of the models, which can also lead to potentially offensive, and hurtful generations. While user-facing LLMs might have some, albeit imperfect, safeguards against generating outright toxic content, they might still generate stereotypical text for marginalized groups and cause representational harms (Gadiraju et al., 2023).

Thus, the design of extrinsic evaluation of cultural competence should be task-grounded and user-centred. Future work should look into designing human evaluation that considers context (when are adaptations useful?), user agency (do users want adaptations?), and representational harms (who is depicted and how?) in a holistic manner.

### Accounting for the Multi-faceted, Intersectional, and Dynamic Nature of Culture

We find that the correlation between text similarity and cultural values is affected by the measure of the cultural values (§ 5.3). One of the reasons for this might be that measures of cultural values like HCD and WVS are imperfect and incomplete. This is because are ample disagreements on the very definition of culture (Baldwin et al., 2006). In fact, Hofstede's Cultural Dimension Theory has been widely criticized for its static nature and over-simplification of culture (Signorini et al., 2009). Even so, evaluating cultural competence in LLMs heavily relies on these measures of culture, inheriting these flaws. Future work should consider diverse and complimentary measures of culture.

Further, like the Hofstede's theory, most evaluations of cultural competence are also done using static benchmarks. However, the world is an evolving place where cultural norms and values are not static. They change and develop through complex interactions among societies. Future work should focus on incorporating evaluation methods like dynamic benchmarking (Kiela et al., 2021) or dealing with disagreements (Davani et al., 2022), among others to account for the evolving nature of culture.

Finally, in our work, we use nationality as a proxy for culture. Our choice was motivated by the availability of data for cultural values for countries and by similar operationalization in prior work. However, culture cannot be anchored by nationalities alone. Moreover, countries are not monoliths and comprise of many and diverse communities. Calls for inclusive evaluations of fairness in language technologies (Bhatt et al., 2022) have led to important recent work on building fairness resources with participatory design (Dev et al., 2023b,a). We believe that methods of evaluation of cultural competence should also similarly embrace participatory and intersectional design.

Overall, the holistic evaluation of cultural competence should account for the multi-faceted, intersectional, and dynamic nature of culture.

## 7 Limitations

While our work serves as a starting point and a call to focus on the extrinsic evaluation of cultural competence, it is not free of limitations.

First, we perform limited qualitative evaluation, and we do not perform any comprehensive human evaluation of the outputs. We describe considerations for comprehensive human evaluation in § 6.

Secondly, our work is anchored on nationalities and relies on imperfect measures of cultural values. However, as we describe in detail in § 6, evaluation of cultural competence demands participatory and intersectional approaches, in addition to accounting for imperfect and static measures of cultures.

Further, our evaluation of the outputs does not reflect their pragmatic correctness. In other words, have not evaluated whether a model's adaptations for a particular question (eg. 'Explain elections...') correctly reflect how the topic is operationalized in the country. Such evaluation needs either expert knowledge or a comparison with verified sources.

Moreover, in measuring the characteristics of the text distributions, we focus only on vocabulary. This provides a starting point for cultural competence. However, culturally sensitive text will need to be evaluated for further characteristics also, for example adhering to the tonality, formality, or other stylistic expectations that might vary culturally.

Finally, in our evaluation, we prompt the model with the nationality explicitly and in English. However, there might be other implicit cues of culture that trigger adaptations such as the language and dialect of interaction, and topical differences in queries which we do not account for in this work.

We hope that future work can address these limitations to holisitcally evaluate LLMs for cultural competence in user-facing tasks.

## 8 Conclusion

In this work, we evaluated cultural competence in two tasks, story generation and open-ended question answering. Our data contributions include a hand-curated list of 345 diverse question-answering topics and 35 story generation topics. We also obtain model outputs for 6 models and 195 nationalities which we will make available for further analysis. Our methodological contributions include conceiving two quantitative and one qualitative analyses for evaluation of LLM outputs for cultural competence. Using these methods, we find that models do vary their outputs with varying nationalities (§ 5.1), outputs contain culturally relevant artefacts (§ 5.2), and model outputs weakly correlate with cultural values (§ 5.3). Our findings underscore the importance of comprehensive extrinsic evaluation of cultural competence.

## Ethical Considerations

**Broader implications and Social Impact** We do not study any sensitive content in this paper, but we note that the outputs of the models could have potentially sensitive and offensive content. Further, the cultural competence of LLMs (or lack of thereof) can lead to varying experiences for users from different demographic backgrounds. We discuss the importance of considering user agency and representational harms in this context in § 6.

**Author Positionality Statement** *Anonymized for peer review*

## References


Anurag Acharya, Kartik Talamadupula, and Mark A Finlayson. 2020. Towards an atlas of cultural commonsense for machine reasoning.

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Ashutosh Dwivedi, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in llms: A survey.

Kenneth Robinson Ahern, Daniele Daminelli, and Cesare Fracassi. 2012. Lost in translation? the effect of cultural values on mergers around the world. *The Stephen M. Ross School of Business at the University of Michigan Research Paper Series*.

AI@Meta. 2024. Llama 3 model card.

Somayeh Alizadeh and Meena Chavan. 2016. Cultural competence dimensions and outcomes: a systematic review of the literature. *Health & Social Care in the Community*, 24(6):e117–e130.

Badr AlKhamissi, Muhammad ElNokrashy, Mai AlKhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models.

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.

John Baldwin, Sandra L. Faulkner, and Michael L. Hecht. 2006. A moving target: The illusive definition of culture.

Abhipsa Basu, R. Venkatesh Babu, and Danish Pruthi. 2023. Inspecting the geographical representativeness of images from text-to-image models.

Allan Bell. 1984. Language style as audience design. *Language in Society*, 13(2):145–204.

Shaily Bhatt, Sunipa Dev, Partha Talukdar, Shachi Dave, and Vinodkumar Prabhakaran. 2022. Re-contextualizing fairness in NLP: The case of India. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 727–740, Online only. Association for Computational Linguistics.

Steven Bird, Edward Loper, and Ewan Klein. 2009. Natural language processing with python: Analyzing text with the natural language toolkit.

Renata Burai. 2016. Substructure of an intercultural curriculum. comparison of selected features of the croatian culture among high school students and their teachers according to the hofstede model.

Yang Trista Cao, Yada Pruksachatkun, Kai-Wei Chang, Rahul Gupta, Varun Kumar, Jwala Dhamala, and Aram Galstyan. 2022. On the intrinsic and extrinsic fairness evaluation metrics for contextualized language representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 561–570, Dublin, Ireland. 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.

Ling-Hsing Chang and Sheng Wu. 2023. The correlation between hofstede's cultural dimensions and covid-19 data in the early stage of the covid-19 pandemic period. *Healthcare*, 11.

Yu Ying Chiu, Liwei Jiang, Maria Antoniak, Chan Young Park, Shuyue Stella Li, Mehar Bhatia, Sahithya Ravi, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2024. Culturalteaming: Ai-assisted interactive red-teaming for challenging llms' (lack of) multicultural knowledge.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Darla K. Deardorff. 2006. Identification and Assessment of Intercultural Competence as a Student Outcome of Internationalization. *Journal of Studies in International Education*, 10(3):241–266.

Darla K. Deardorff. 2009. *The SAGE Handbook of Intercultural Competence*. SAGE Publications, Inc, 2455 Teller Road,Thousand Oaks California 91320.

Sunipa Dev, Jaya Goyal, Dinesh Tewari, Shachi Dave, and Vinodkumar Prabhakaran. 2023a. Building socio-culturally inclusive stereotype resources with community engagement.

Sunipa Dev, Akshita Jha, Jaya Goyal, Dinesh Tewari, Shachi Dave, and Vinodkumar Prabhakaran. 2023b. Building stereotype repositories with complementary approaches for scale and depth. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 84–90, Dubrovnik, Croatia. Association for Computational Linguistics.

Shiran Dudy and Steven Bedrick. 2020. Are some words worth more than others? In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 131–142, Online. Association for Computational Linguistics.

Esin Durmus, Karina Nyugen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2023. Towards measuring the representation of subjective global opinions in language models.

Ashutosh Dwivedi, Pradhyumna Lavania, and Ashutosh Modi. 2023. EtiCor: Corpus for analyzing LLMs for etiquettes. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6921–6931, Singapore. Association for Computational Linguistics.

Alvino Fantini and Aqeel Tirmizi. 2006. Exploring and Assessing Intercultural Competence. Final Report of a Research Project to Explore and Assess Intercultural Outcomes in Service Program Participants Worldwide, Federation of The Experiment in International Living with funding support from the Center for Social Development at Washington University, St. Louis, Missouri, World Learning Publications.

Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. 2016. Problems with evaluation of word embeddings using word similarity tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 30–35, Berlin, Germany. Association for Computational Linguistics.

Vinitha Gadiraju, Shaun Kane, Sunipa Dev, Alex Taylor, Ding Wang, Emily Denton, and Robin Brewer. 2023. " i wouldn't say offensive but...": Disability-centered perspectives on large language models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 205–216.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III au2, and Kate Crawford. 2021. Datasheets for datasets.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. In *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)*, pages 1926–1940, Online. Association for Computational Linguistics.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, and Bjorn Puranen, editors. 2022. *World Values Survey: Round Seven - Country-Pooled Datafile Version 5.0*. JD Systems Institute & WVSA Secretariat, Madrid, Spain & Vienna, Austria.

Robert Hawkins, Irina Liu, Adele Goldberg, and Tom Griffiths. 2021. Respect the code: Speakers expect novel conventions to generalize within but not across social group boundaries. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 43(43).

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Geert Hofstede. 2001. Culture's consequences: Comparing values, behaviors, institutions and organizations across nations.

Gert Jan Hofstede, Gert Jan Hofstede, Michael Minkov, and McGraw-Hill New. 2010. Cultures and organizations: Software of the mind, 3rd ed.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.

Mary F. Howard-Hamilton, Brenda J. Richardson, and Bettina C. Shuford. 1998. Promoting multicultural education: A holistic approach. *The College Student Affairs Journal*, 18:5–17.

Jing Huang and Diyi Yang. 2023. Culturally aware natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7591–7609, Singapore. Association for Computational Linguistics.

Akshita Jha, Aida Mostafazadeh Davani, Chandan K Reddy, Shachi Dave, Vinodkumar Prabhakaran, and Sunipa Dev. 2023. SeeGULL: A stereotype benchmark with broad geo-cultural coverage leveraging generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9851–9870, Toronto, Canada. Association for Computational Linguistics.

Karen Sparck Jones and Julia R. Galliers. 1995. *Evaluating Natural Language Processing Systems: An Analysis and Review*, pages 3–63. Springer Berlin Heidelberg, Berlin, Heidelberg.

Anubha Kabra, Emmy Liu, Simran Khanuja, Alham Fikri Aji, Genta Indra Winata, Samuel Cahyawijaya, Anuoluwapo Aremu, Perez Ogayo, and Graham Neubig. 2023. Multi-lingual and multi-cultural figurative language understanding.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

Eunsu Kim, Juyoung Suk, Philhoon Oh, Haneul Yoo, James Thorne, and Alice Oh. 2024. Click: A benchmark dataset of cultural and linguistic intelligence in korean.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. 2024. The prism alignment project: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models.

Fajri Koto, Rahmad Mahendra, Nurul Aisyah, and Timothy Baldwin. 2024. Indoculture: Exploring geographically-influenced cultural commonsense reasoning across eleven indonesian provinces.

A. L. Kroeber and Clyde Kluckhohn. 1952. *Culture: A Critical Review of Concepts and Definitions*. Peabody Museum Press, Cambridge, Massachusetts.

V. I. Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707.

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

Huihan Li, Liwei Jiang, Jena D. Huang, Hyunwoo Kim, Sebastin Santy, Taylor Sorensen, Bill Yuchen Lin, Nouha Dziri, Xiang Ren, and Yejin Choi. 2024b. Culture-gen: Revealing global cultural perception in language models through natural language prompting.

Li Lucy, Su Lin Blodgett, Milad Shokouhi, Hanna M. Wallach, and Alexandra Olteanu. 2023. "one-size-fits-all"? examining expectations around what constitute"fair"or"good"nlg system behaviors.

Reem I. Masoud, Ziquan Liu, Martin Ferianc, Philip Treleaven, and Miguel Rodrigues. 2024. Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions.

Shubhanshu Mishra, Sijun He, and Luca Belli. 2020. Assessing demographic bias in named entity recognition.

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

Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. 2023. Extracting cultural commonsense knowledge at scale. In *Proceedings of the ACM Web Conference 2023*, WWW '23. ACM.

Shramay Palta and Rachel Rudinger. 2023. FORK: A bite-sized test set for probing culinary cultural biases in commonsense reasoning models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9952–9962, Toronto, Canada. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Aida Ramezani and Yang Xu. 2023. Knowledge of cultural moral norms in large language models.

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

Pola Schwöbel, Jacek Golebiowski, Michele Donini, Cédric Archambeau, and Danish Pruthi. 2023. Geographical erasure in language generation.

Agrima Seth, Sanchit Ahuja, Kalika Bali, and Sunayana Sitaram. 2024. Dosa: A dataset of social artifacts from different indian geographical subcultures.

Omar Shaikh, Caleb Ziems, William Held, Aryan Pariani, Fred Morstatter, and Diyi Yang. 2023. Modeling cross-cultural pragmatic inference with codenames duet. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6550–6569, Toronto, Canada. Association for Computational Linguistics.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Chunhua yu, Raya Horesh, Rogério Abreu de Paula, and Diyi Yang. 2024. Culturebank: An online community-driven knowledge base towards culturally aware language technologies.

Paola Signorini, Rolf Wiesemes, and Roger J. L. Murphy. 2009. Developing alternative frameworks for exploring intercultural learning: a critique of hofstede's cultural difference model. *Teaching in Higher Education*, 14:253 – 264.

Harman Preet Singh and Radha Karuna Kumari. 2023. Digital technologies in healthcare management: A study of influence of national culture for adoption of electronic health records in india and australia. *Archives of Business Research*.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. A roadmap to pluralistic alignment.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, AiTi Aw, and Nancy F. Chen. 2024. Seaeval for multilingual foundation models: From cross-lingual alignment to cultural reasoning.

Charley M. Wu, Rick Dale, and Robert D. Hawkins. 2023a. Group coordination catalyzes individual and cultural intelligence.

Winston Wu, Lu Wang, and Rada Mihalcea. 2023b. Cross-cultural analysis of human values, morals, and biases in folk tales. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5113–5125, Singapore. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. Wildchat: 1m chatgpt interaction logs in the wild.

Kaitlyn Zhou, Kawin Ethayarajh, and Dan Jurafsky. 2022. Richer countries and richer representations. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2074–2085, Dublin, Ireland. Association for Computational Linguistics.

Li Zhou, Antonia Karamolegkou, Wenyu Chen, and Daniel Hershcovich. 2023. Cultural compass: Predicting transfer learning success in offensive language detection with cultural features. In *Findings* *of the Association for Computational Linguistics: EMNLP 2023*, pages 12684–12702, Singapore. Association for Computational Linguistics.

# A Datasheet

This document is based on *Datasheets for Datasets* by Gebru et al. (2021). The latex template is based on this github repo

## A.1 Motivation

***For what purpose was the dataset created?*** *Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*

This dataset has two parts. First is a list of topics to prompt models with for two tasks, question answering and story generation to analyse difference in model outputs across nationalities. Second are the model responses for these prompts.

***Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?***

Anonymized for peer review

***What support was needed to make this dataset?*** *(e.g.who funded the creation of the dataset? If there is an associated grant, provide the name of the grantor and the grant name and number, or if it was supported by a company or government agency, give those details.)*

Anonymized for peer review

## A.2 Composition

***What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?*** *Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

The data consists of a list of topics. The model outputs contain text generated by a LLMs.

***How many instances are there in total (of each type, if appropriate)?***

35 topics for story generation and 345 topics for QA. For model outputs, each topic leads to 195 prompts (for 195 nationalities) and 5 responses are sampled for every prompt from 6 LLMs. This leads to 2018250 model outputs for QA and 175500 model outputs for stories.

***Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?*** *If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).*

This is a hand curated list of data. It is not exhaustively representative of all possible story generation topics or QA topics. For story generation in particular, we only focus on children's stories. For QA, we attempt to include diverse topics and categories. But we note that these are open-ended tasks and thus the range of topics is very wide to measure exhaustiveness.

***What data does each instance consist of?*** *"Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.*

Each instance in the topic list is simply a phrase (unigram or bigram) that is used to create a prompt for Question answering or story generation. Each instance of model output is a paragraph with maximum 100 tokens in case of QA and 1000 tokens in case of story generation.

***Is there a label or target associated with each instance?*** *If so, please provide a description.*

There are no labels

***Is any information missing from individual instances?*** *If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.*

No

***Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?*** *If so, please describe how these relationships are made explicit.*

No

***Are there recommended data splits (e.g., training, development/validation, testing)?*** *If so, please provide a description of these splits, explaining the rationale behind them.*

All of the data is intended for evaluation, we do not anticipate needing any training or validation splits.

***Are there any errors, sources of noise, or redundancies in the dataset?*** *If so, please provide a description.*

No

***Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., web-***

***sites, tweets, other datasets)?*** *If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

It is self-contained.

***Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)?*** *If so, please provide a description.*

No

***Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?*** *If so, please describe why.*

No

***Does the dataset relate to people?*** *If not, you may skip the remaining questions in this section.*

No

***Does the dataset identify any subpopulations (e.g., by age, gender)?*** *If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

For collecting model outputs, the prompt that we use explicitly mention a nationality. This is because we want to study the perturbation of the model outputs when nationalities are perturbed in the prompts. Because of this model outputs in the data are likely to contain text that refer to respective nationalities.

***Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?*** *If so, please describe how.*

No

***Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?*** *If so, please provide a description.*

No

***Any other comments?***

No

### A.3 Collection

***How was the data associated with each instance acquired?*** *Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

The topics were obtained by hand-curation. The authors first created a broad list of 13 categories that were of interest in the evaluation: biology, chemistry, environment, economics, history, humanities, law, maths, physics, politics, space, religion, world affairs. This categories were selected as intuitive categories of questions in which differences in model outputs might be observed. The authors then referred to textbooks and encyclopedia index to sample topics within these categories leading to a total 345 topics. For stories, the authors first similarly selected three broad categories on which children's stories can be written: moral values, stories with specific characters, stories with specific settings. They then used online websites and children's story books to come up with topics with these areas creating a list of 35 topics. This is the topic lists. Next, these were then used in a simple template 'Explain {topic} to a / an {nationality} person.' for QA and 'Write a story about {topic} for a / an {nationality} kid.' in story. The resulting prompts were input into 6 LLMs listed in 4.2 to obtain model outputs. 5 responses were generated for every output.

***Over what timeframe was the data collected?*** *Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created. Finally, list when the dataset was first published.*

The topic list was curated between November 2023 and January 2024. Model outputs were collected between February and April 2024.

***What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sen-***

***sor, manual human curation, software program, software API)?*** *How were these mechanisms or procedures validated?*

The entire data of topic list is human curated. The model outputs are LLMs generated. Some characteristics of the model outputs are evaluated in the paper.

***What was the resource cost of collecting the data?*** *(e.g. what were the required computational resources, and the associated financial costs, and energy consumption - estimate the carbon footprint. See Strubell et al.(Strubell et al., 2019) for approaches in this area.)*

The cost of hand-curating topic lists was about 10 researcher hours. For getting model outputs, A6000 GPUs was used for hosting the LLM to run inference for obtaining model outputs. The total inference cost was about 45 GPU hours. Model outputs from GPT 3.5 cost about 125 USD.

***If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?***

We did not sample.

***Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?***

The data was hand curated by the author and the author queried LLMs for model outputs. No additional personnel was involved.

***Were any ethical review processes conducted (e.g., by an institutional review board)?*** *If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

No human subjects or crowd workers were involved hence we did not conduct any IRB.

***Does the dataset relate to people?*** *If not, you may skip the remainder of the questions in this section.*

No

***Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?***

NA

***Were the individuals in question notified about the data collection?*** *If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

NA

***Did the individuals in question consent to the collection and use of their data?*** *If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

NA

***If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?*** *If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate)*

NA

***Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis)been conducted?*** *If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

NA

***Any other comments?***

NA

### A.4 Preprocessing / Labelling / Cleaning

***Was any preprocessing/cleaning/labeling of the data done(e.g.,discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?*** *If so, please provide a description. If not, you may skip the remainder of the questions in this section.*

No

***Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?*** *If so, please provide a link or other access point to the "raw" data.*

No cleaning was performed

***Is the software used to preprocess/clean/label the instances available?*** *If so, please provide a link or other access point.*

NA

***Any other comments?***

NA

### A.5 Uses

***Has the dataset been used for any tasks already?*** *If so, please provide a description.*

Yes, the data was used to evaluate the variations in model outputs for varying nationalities in the input prompts for two tasks in order to evaluate cultural competence.

***Is there a repository that links to any or all papers or systems that use the dataset?*** *If so, please provide a link or other access point.*

Yes. The data, paper, and code will be open sourced after peer review.

***What (other) tasks could the dataset be used for?***

The list of topics could be used for a different task evaluation. The model outputs could be further used to characterize model behaviour in these settings, such as qualitative analysis of outputs, analysis for prescence of biases and so on.

***Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?*** *For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?*

No. We do note though that the model outputs are generated content from LLMs and might content toxic, offensive, and stereotypical texts against marginalized communties. We advise discretion on part of users who choose to further utlize this data for analysis.

***Are there tasks for which the dataset should not be used?*** *If so, please provide a description.*

The topic lists should not be treated an exhaustive list of topics to evaluate cultural competence. The model outputs should not be used as gold standard answers for the particular questions or story generation tasks.

***Any other comments?***

No

### A.6 Distribution

***Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?*** *If so, please provide a description.*

The data will be open-sourced

***How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?*** *Does the dataset have a digital object identifier (DOI)?*

The data will be open-sourced onto a github repo or huggingface after publication.

***When will the dataset be distributed?***

The data will be open-sourced after publication.

***Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?*** *If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

The data will be open-sourced.

***Have any third parties imposed IP-based or other restrictions on the data associated with the instances?*** *If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*

No

***Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?*** *If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*

No

***Any other comments?***

YOUR ANSWER HERE

### A.7 Maintenance

***Who is supporting/hosting/maintaining the dataset?***

Anonymized for peer review.

***How can the owner/curator/manager of the dataset be contacted (e.g., email address)?***

Anonymized for peer review.

***Is there an erratum?*** *If so, please provide a link or other access point.*

No. The authors can be contacted via email.

***Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?*** *If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?*

This data is unlikely to be updated.

***If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a***

***fixed period of time and then deleted)?*** *If so, please describe these limits and explain how they will be enforced.*

NA

***Will older versions of the dataset continue to be supported/hosted/maintained?*** *If so, please describe how. If not, please describe how its obsolescence will be communicated to users.*

We do not intend to have multiple version.

***If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?*** *If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.*

TBD

***Any other comments?***

NA

# B Lexical Variance

## B.1 Calculation Details

The Variance between two discrete random variables can be defined as:

$$\mathrm{Var}(X) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{2} \left(x_i - x_j\right)^2$$

Within this equation $x_i - x_j$ essentially represents distance between the two points, which we replace with lexical distance or the Word Edit Distance (WED). Thus, repurposing the above variance equation, lexical variance in outputs across nationalities for a concept can be calculated as:

$$\frac{1}{|\mathcal{N}|^2} \sum_{n \in \mathcal{N}} \sum_{n' \in \mathcal{N}} \frac{1}{2} (\mathrm{WED}(O_n, O_{n'}))^2$$

Where: $\mathcal{N}$ = Set of all Nationalities, $n$ = nationality, $O_n$ = output for nationality $n$

## B.2 ANOVA results on within and across nationality lexical variance

H0 = $\mu_{within} = \mu_{across}$

H1 = they are different

The p-values are in table 3

# C Kendalls Tau Rank Correlation

## C.1 Choice of Kendalls Tau variant

We use the c variant in particular because before the ranking both the rank list have been generated by metrics that have different scales.

## C.2 An example calculation

This is a brief example of how Kendall's $\tau_c$ was calculated. Suppose there are 4 nationalities: A, B, C, D. We first take one nationality as an anchor, let's say A, and create two rank lists. The first rank list is of similarity of text outputs to A, let's say this is [B, D, C] and the second is using distance between cultural values representation (we reverse the raw rank list we get from distance in vector representation of cultural values, because this is distance while the other one similarity), let's say this is [D, C, B]. For A, the rank correlation between these two ranklist is calculated using Kendall's $\tau_c$ . We use sklearn to calculte Kendall's $\tau_c$ with default parameters. Finally, for a particular concept, we take average of Kendall's $\tau_c$ across all nationalities.

| task | model | F-statistic | p-value | Reject H0 |
|---|---|---|---|---|
| stories | llama2_7B_chat | 2255.3456 | 7.043450519806435e-54 | Yes |
| stories | llama2_13B_chat | 2821.1494 | 4.248487694091554e-57 | Yes |
| stories | llama3_8B_instruct | 3610.5356 | 1.1491538258492085e-60 | Yes |
| stories | gemma2B_it | 874.1556 | 1.5311181671386628e-40 | Yes |
| stories | gemma7B_it | 1721.6872 | 5.048199426344931e-50 | Yes |
| stories | gpt_3-58 | 1055.6979 | 3.80594818701481e-43 | Yes |
| QA | llama2_7B_chat | 911.4229 | 3.7297913016341677e-128 | Yes |
| QA | llama2_13B_chat | 1444.7691 | 3.4753916948315105e-171 | Yes |
| QA | llama3_8B_instruct | 2585.3423 | 3.2168642758230666e-235 | Yes |
| QA | gemma2B_it | 550.97 | 5.966032578765733e-90 | Yes |
| QA | gemma7B_it | 1335.7818 | 2.4154064759769195e-163 | Yes |
| QA | gpt_3-5 | 233.4199 | 1.3687492031148516e-45 | Yes |

Table 3: One Way ANOVA for within and across nationalities. All p-values suggest that H0 (same means) can be rejected.