From Surveys to Narratives: Rethinking Cultural Value Adaptation in LLMs

Anonymous ACL submission

Abstract

Adapting cultural values in Large Language Models (LLMs) presents significant challenges, particularly due to biases and data limitations. Previous work aligns LLMs with different cultures using survey data, primarily from the World Values Survey (WVS). However, it remains unclear whether this approach effectively captures cultural nuances or produces distinct cultural representations for various downstream tasks. In this paper, we systematically investigate WVS-based training for cultural value adaptation and find that relying solely on survey data can homogenize cultural norms and interfere with factual knowledge. To address these issues, we propose augmenting WVS with encyclopedic and scenario-based cultural narratives from Wikipedia and NormAd. Our experiments across multiple cultures show that this approach captures more enhances differentiated cultural values and improves downstream classification performances.

1 Introduction

Recent developments in Large Language Models (LLMs) suggest LLMs align closely with the cultural values of Western, Educated, Industrialized, Rich, and Democratic (WEIRD, Henrich et al. 2010) societies without adaptations (Johnson et al., 2022; Ramezani and Xu, 2023; Cao et al., 2023, among others). Such a WEIRD-centric bias can harm specific groups, perpetuate stereotypes, and limit a model's usefulness to a diverse global audience. Indeed, culture is a distinct and vital aspect of human society, influencing behavior, norms, and worldviews (Geertz, 2017). Yet most LLMs lack robust mechanisms to adapt their outputs in ways that reflect different cultural value systems. In this work, we use the term *cultural adaptation* to refer to the process of adjusting an LLM's behavior so that it reflects the norms, attitudes, values, and beliefs of a specific culture.[1]

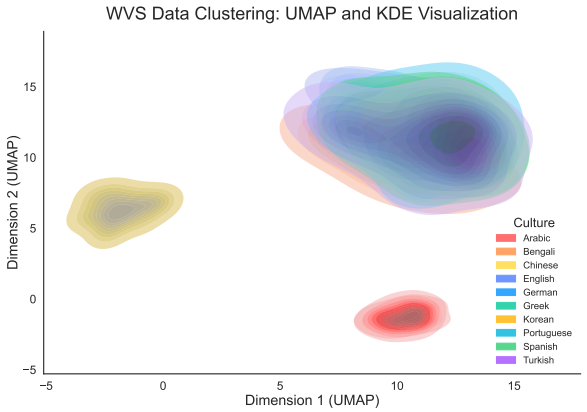


Figure 1: UMAP-KDE visualization of cultural value distributions from World Values Survey (WVS) data reveals significant homogenization. While Arabic (lower right) and Chinese (left) cultures form distinct clusters, many others converge in the upper right. This suggests that current WVS-based training may not sufficiently capture cultural nuances.

Culture is an integral aspect of human society that encompasses far more than language alone (Geertz, 2017). Indeed, two cultures may share the same primary language (e.g., English in the US vs. Australia) yet differ in social norms, historical contexts, and what is considered offensive or polite. Therefore, training on the same language data without incorporating cultural nuances can lead to suboptimal or biased outputs. Existing work often adapts LLMs to cultural values by leveraging self-reported survey data (Li et al., 2024a; Xu et al., 2024; Li et al., 2024b) such as the World Values Survey (WVS, Haerpfer et al. 2022). Although WVS offers a quantitative glimpse into cultural attitudes (e.g., *"How important is family in your life?"*

[1]For the purposes of this paper, we focus on "culture" at a linguistic-regional level (e.g., Iraq and Jordan represent **Arab** culture vs. Argentina and Mexico that represent **Spanish** culture), but we acknowledge that culture is more nuanced, including sub-cultures within a group and intersectional factors such as ethnicity and religion (Adilazuarda et al., 2024).

on a scale from 1 to 4), it remains unclear how to best translate these numeric indications into concrete behavior in downstream tasks (e.g., classification of offensiveness in different linguistic-cultural settings). Beyond survey responses on values and opinions, culture also includes social norms, historical contexts, and nuanced beliefs (Liu et al., 2024) that may not be fully captured through self-reported questionnaires. As shown in Figure 1, even WVS data for distinct cultures may converge into overlapping clusters in latent space (showing semantic similarities), potentially homogenizing nuanced cultural dimensions.

Ideally, cultural value adaptation should enhance performance within each culture. However, several challenges emerge. First, adapting multiple cultural values through data prompts may create interference similar to that seen in multilingual models (Conneau et al., 2020; Wang et al., 2020), particularly given language-culture interconnections (Adilazuarda et al., 2024; Hershcovich et al., 2022; Hovy and Yang, 2021). Second, the reliability of cultural value training data is uncertain. Studies show discrepancies between attitude and actual behavior in human (Gross and Niman, 1975; Fazio, 1981), raising concerns about the World Values Survey's (WVS) ability to capture deep cultural behavior for LLM training (Liu et al., 2024), necessitating further investigation.

In this work, we investigate these research problems by critically evaluating existing cultural value adaptation methods. Through a series of experiments, our findings highlight the limitations of WVS for cultural value adaptation in LLMs, which require more nuanced methods than current approaches (Laskar et al., 2024). While WVS provides insights into cultural values, it fails to capture deeper contextual dimensions or inform how values translate into behavior in downstream tasks. To addressing this, we propose supplementing WVS with descriptive, culturally grounded data (e.g., Wikipedia, NormAd).

To summarize, our contributions are:

- We identify ***cultural interference*** in adaptation using WVS data while enhancing downstream tasks, such as offensiveness classification, demonstrating that WVS-based adaptation homogenizes cultural behaviors rather than preserving their distinctions.
- We demonstrate ***knowledge interference*** from adaptation, where adding cultural data can degrade factual knowledge understanding (e.g., lowering MMLU scores).
- We propose ***a solution to mitigate both cultural and knowledge interference***, using context-rich sources (Wikipedia, NormAd) to enhance cultural adaptation while preserving factual knowledge.

2 Methodology

We designed a series of experiments to investigate our research problem systematically. This section outlines our methodologies for cultural adaptation and evaluation. We start with zero-shot prompting, then describe single-culture adapter fine-tuning, and outline how we diagnose potential interference using additional tasks such as MMLU (Massive Multitask Language Understanding, Hendrycks et al. 2021). We will describe datasets, models, and evaluation metrics in §3.

2.1 Zero-Shot Prompting

Zero-shot prompting leverages a pre-trained LLM without additional fine-tuning. To adapt the model for a specific target culture, we use simple instructions that reference the culture. For instance, for an OFFENSEVAL-style task, we use the following prompt in Table 1:

You are a `{country}` chatbot that understands `{country}`'s cultural context. **Question:** Is the following sentence offensive according to `{country}`'s cultural norms? **Input:** `{input_txt}` **Answer:** [Select one: 1. Offensive, 2. Not offensive]

Table 1: Zero-shot prompt template for offensiveness classification. We list the full prompts used in our study in Appendix E.

Here, the model's responses rely entirely on cultural or multilingual knowledge that was encoded during pre-training. This can create systematic biases when the training data is skewed toward dominant cultural paradigms, which may disadvantage underrepresented groups (Guo et al., 2024).

2.2 Cultural Value Adaptation via Fine-tuning

Moving beyond zero-shot prompts, we explore explicit fine-tuning with culture-specific data using **Single-Culture Adaptation**. Following Li et al. (2024a), we train a separate LoRA adapter (Hu et al., 2022) for each cultural context using data

from a single, or a combination of data sources. Each adapter is specialized to reflect the norms, attitudes, or knowledge of that specific culture. However, data sparsity and overfitting are risks, particularly for cultures with limited samples.

Using the single-culture adaptation strategy, each LoRA adapter is trained to reflect the high-level values from the chosen dataset (WVS alone, or WVS plus additional cultural knowledge). At inference time, the adapter is activated based on the test target culture specification.

3 Experimental Setup

We base our experiments on the CultureLLM (Li et al., 2024a) framework, one of the earliest popular adaptation frameworks for cultural values. We design our experimental setup to evaluate across multiple LLMs and languages. Below, we briefly describe datasets used for training and evaluation, model and training hyperparameters, and evaluation metrics.

3.1 Linguistic-Cultural Settings

We conduct experiments on ten distinct linguistic-cultural settings, here we use the ISO 693-3 code for simplicity: Arabic (ara, Iraq and Jordan), Bengali (ben, Bangladesh), Chinese (zho, China), English (eng, United States), German (deu, Germany), Greek (ell, Greece), Korean (kor, South Korea), Portuguese (por, Brazil), Spanish (spa, Argentina and Mexico), and Turkish (tur, Turkey).

3.2 Training Dataset

We established three training scenarios with data drawn from three different sources:

WVS. In this setting, we use the WVS and semantically augmented data based on Li et al. (2024a). WVS is a survey data commonly used in social sciences, as well as a proxy for cultural values in NLP (Adilazuarda et al., 2024). The dataset consists of question-and-answer pairs that provide quantitative indicators of societal beliefs and attitudes (e.g., questions on family importance, or religion).

Wikipedia. We select Wikipedia articles with detailed knowledge, region-specific norms, social practices, and historical contexts of our defined cultures. These articles can enrich the numeric survey data with qualitative background.[2]

NormAd. NormAd (Rao et al., 2024) offers a structured collection of cultural norms and situational examples, demonstrating how abstract values materialize in everyday interactions. Unlike WVS, which provides broad statistical insights, and Wikipedia, which offers descriptive knowledge, NormAd emphasizes behavioral and contextual applications of cultural principles.

[2]See Appendix 6 for the Wikipedia pages used.

3.3 Evaluation Dataset

We use two set of tasks for evaluations:

Multicultural Multilingual Offensiveness. To assess the effectiveness of adaptation in models' behavior on downstream tasks, we evaluate the adapted models using a combination of datasets (such as OffenseEval2020, Zampieri et al. 2020) following Li et al. (2024a,b, see original publications for the complete list). The evaluation setup contains a total of 68607 multilingual, culturally sensitive text annotated for offensiveness.

MMLU. To evaluate whether cultural adaptation affects the model's general knowledge capabilities, we assess each adapter's performance on factual question-answering tasks using MMLU (Mukherjee et al., 2024). The MMLU dataset focuses on factual knowledge such as mathematics, biology, chemistry etc. which contains minimal cultural sensitivity. The deviations in MMLU accuracy following cultural fine-tuning would suggest unintended interference, implying that the cultural adapter may be altering the model's underlying knowledge representations.

Using these two datasets, we enable a systematic evaluation of how effectively language models can integrate cultural perspectives into downstream tasks while preserving their factual knowledge.

3.4 Models and Training

In this work, we evaluate three variants of LLMs, including Llama-3.1-8B (base and instruction-tuned, Touvron et al. 2023; Dubey et al. 2024), Gemma-2-9B (instruction-tuned, Rivière et al. 2024), and Qwen-2.5-7B (instruction-tuned, Team 2024). In our experiments, all instruction-tuned models are suffixed with "**-IT**". We perform LoRA adaptation (Hu et al., 2022) on each model using rank-64 LoRA matrices, batch size of 32, a learning rate of 2×10^{-4}, and six training epochs. Other details on training are in Appendix B.

3.5 Evaluation Metrics

In our main paper, we evaluate each model's performance using freeform generation, assessing its ability to provide culturally relevant justifications or

context. Appendix includes additional probability-based evaluations, using token-level likelihood scores to measure the model's confidence in classifying offensive content across cultures. Further, we use F1 score as the primary metric for evaluating classification performance on both probability and freeform-based evaluations.

To further quantify a model's ability to preserve cultural distinctiveness, we propose a *cultural specificity* metric, **C-SPEC** score. For n cultures, we define a performance matrix $M \in \mathbb{R}^{n \times n}$, where $M_{i,j}$ is the F1-score when a model adapted to culture i is evaluated on test data for culture j. We compute:

1. Extract the diagonal entries[3] $\vec{d} = [M_{i,i}]_{i=1}^{n}$.
2. Normalize each $M_{i,i}$ by the maximum value in its column: $\vec{n}_i = M_{i,i} / \max_j M_{j,i}$.
3. Average these normalized diagonal entries:

$$D = \frac{1}{n} \sum_{i=1}^{n} \vec{n}_i. \quad (1)$$

In the formula above, we normalize by column (i.e., by the test culture) since each test culture set may have different difficulty and scales. This normalization also helps identify which adapter performs best for a given culture.

In an ideal scenario, the best performing adapted model for a particular culture should be based on its own culture, resulting in a C-SPEC score of 1.0. A lower score suggests interference or homogenization, as illustrated in Figure2. This metric thus quantifies the extent to which each model preserves distinct cultural representations after adaptation.

4 Adaptation with WVS: Findings and Observed Interferences

In this section, we focus on Llama-3.1-8B models (both base and instruction-tuned) to establish a clear understanding of their performance and the impact of adaptation using WVS data, including cultural and knowledge-based interferences.

4.1 Performance Gains Driven by Enhanced Instruction Following

General Observations on the Results. Using the value adaptation strategies outlined in Section 2, Table 2 compares the approaches for downstream

[3]We define "diagonal entries" as the corresponding performance of an adapter on its corresponding culture, e.g. Korean Adapter evaluated on Korean Culture test set, hence we define this as $M_{i,i}$

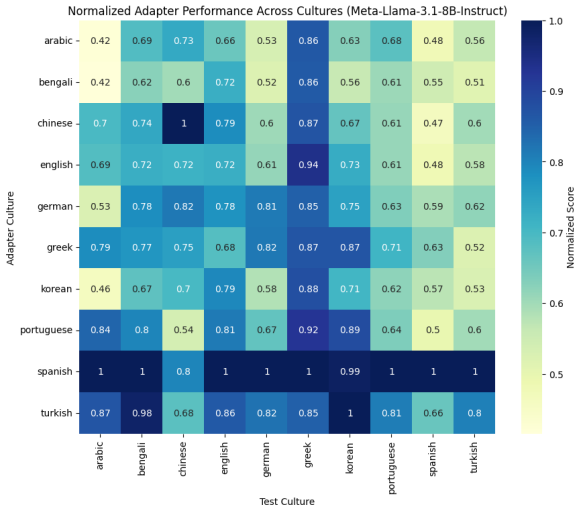


Figure 2: Single-culture adaptation using WVS data with Llama-3.1-8B-IT, evaluating cross-cultural offensiveness classification tasks. Minimal diagonal pattern is observed in this setting, with a C-SPEC score of **0.76**.

tasks using Llama-3.1 8B models: (i) zero-shot prompting, (ii) single-culture adaptation.

Observing the results in Table 2, training with the WVS survey appears more effective in improving downstream task valuation for the base model using the single-culture adaptation strategy. Particularly, WVS training is beneficial for underrepresented cultures such as `ara` and `kor`. Surprisingly, this positive effect is not observed in the instruction-tuned model, which instead shows a decline in average performance.

Performance Gain by Better Instruction Following. To understand why the instruction-tuned model did not benefit from training with WVS, we analyze its downstream task predictions by examining the ratio of invalid responses[4] before and after adaptation in Table 3 (completed results in Appendix D.3). Compared to zero-shot prompting, both base model and instruction-tuned model have significantly improved invalid response ratio after adaptation. This indicates that WVS fine-tuning enhances the model's general instruction-following ability but does not necessarily improve its understanding of cultural values.

4.2 Observed Cultural Interference Across Models

To further investigate the effect of the adaptation, we now examine the single-culture adaptation re-

[4]An invalid response contains nonsensical outputs, fails to follow instructions or lacks a meaningful or relevant answer to the prompt. Appendix 12 shows example responses.

Model	ara	ben	zho	eng	deu	ell	kor	por	spa	tur	Avg.
Zero-Shot Prompting											
Llama-3.1-8B	11.96	17.12	32.77	14.85	23.81	38.16	26.14	19.93	30.96	21.95	23.77
Llama-3.1-8B-IT	19.14	23.10	30.49	26.63	34.36	37.56	38.72	20.92	39.14	32.95	30.00
Single-Culture Adaptation - WVS											
Llama-3.1-8B	17.22	22.01	38.28	19.92	29.30	36.08	32.65	20.15	27.93	28.57	27.21
Llama-3.1-8B-IT	19.50	23.51	32.69	22.35	34.78	36.98	37.61	17.75	25.85	28.78	27.98

Table 2: Culture adaptation results (F1 scores) under three training scenarios: zero-shot prompting, single-culture adaptation training on Llama-3.1-8B models using WVS training data. The adaptation is evaluated using a multilingual offensiveness dataset (§3.3) reported with averaged F1 scores.

Methods		Invalid (%)
Llama-3.1-8B	Zero-Shot	20.12
	Single-Culture-WVS	14.68
Llama-3.1-8B-IT	Zero-Shot	21.20
	Single-Culture-WVS	10.82

Table 3: Comparison of invalid response rates across different models and scenarios. The Invalid Ratio represents the percentage of responses flagged as invalid across all culture test set.

Model	Culture	Std.	Transl.
Llama-3.1-8B	ara	32.24	32.83
	ben	48.67	51.81
	zho	38.21	41.08
	eng	23.00	29.58
	deu	33.55	39.68
	ell	30.75	31.55
	kor	27.59	27.57
	por	**46.41**	**28.77**
	spa	35.53	35.27
	tur	19.74	18.02
	Avg.	33.57	33.62
Llama-3.1-8B-IT	ara	41.99	37.81
	ben	45.45	42.77
	zho	41.35	46.28
	eng	42.81	49.18
	deu	40.40	41.92
	ell	46.05	36.34
	kor	41.80	44.63
	por	40.11	38.08
	spa	43.77	38.60
	tur	43.93	40.46
	Avg.	42.78	41.61

Table 4: MMLU evaluation after single-culture adaptation with WVS data (F1 Score %). Performance variation is evident across cultural adapters, with observed factual knowledge retention and potential cultural biases. The zero-shot performance is **35.05** for Llama-3.1-8B and **45.38** for Llama-3.1-8B-IT.

sults cross-culturally (train in one culture, evaluate in other cultures). Ideally, the culture with matching adaptation should perform the best comparing of using other adaptations (i.e., a diagonal on a heatmap of cross-cultural evaluation). Illustrated in Figure 2, no diagonal was observed for the instruction-tuned Llama model (similar observation for the base model in Figure 10 in Appendix). The cross-cultural improvements shows no clear patterns and all adapters can improve performance on the Spanish test data in Figure 2. The C-SPEC score (introduced in §3) remains below 0.80 for both models.

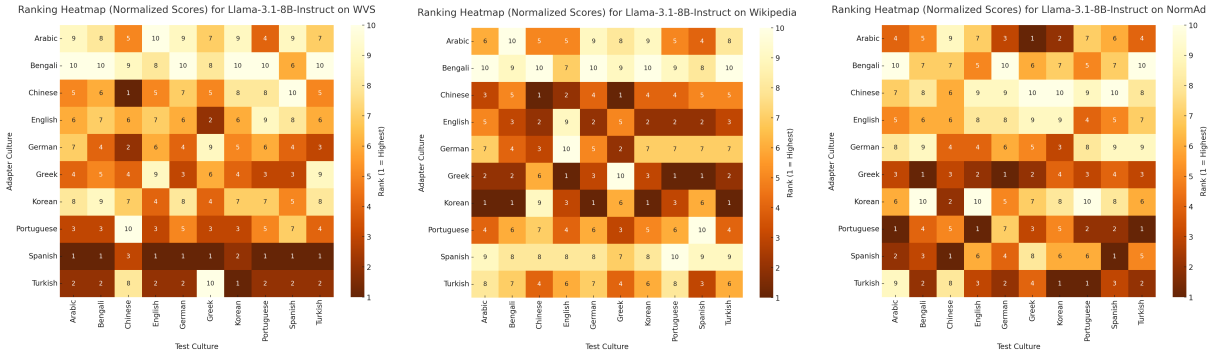
The results further suggest that WVS is not necessarily the best data source for improving cultural values, as the adapted models fail to preserve their own culture's perspectives, leading to compromised cross-cultural result improvements (i.e., *cultural interference*).

4.3 Factual Knowledge Interference

Fine-tuning improves cultural alignment but may unintentionally impact factual knowledge (Mukherjee et al., 2024). Ideally, cultural value adaptation should not affect objective QA performance.

Table 4 presents the results of single-culture adaptation on MMLU. Both Llama-3.1-8B and Llama-3.1-8B-IT exhibit significant variability when trained under two conditions: standard (using English WVS data) and translated (WVS values in their respective languages). Additionally, the base model shows a decline in performance compared to zero-shot prompting, while the instruction-tuned model shows performance improvements.

These fluctuations indicate that adapting to WVS data can change factual knowledge accuracy, depending on language and dataset characteristics. Furthermore, the inconsistencies in probability-based scoring (Appendix 11) also strengthen the observation of *factual knowledge interference*. This underscores the challenge of balancing cultural specificity with factual integrity with the appropriate training data.



WVS **WVS+Wiki** **WVS+NormAd**

Figure 3: Heatmaps of culture-specific classification performance (Llama-3.1-8B-IT) using different data sources based on the ranks of the adaptation results. Darker diagonal elements indicate stronger cultural distinctiveness and better C-SPEC scores, introduced in 5. Notably, WVS+Wiki and especially WVS+NormAd yield higher C-SPEC than WVS-only.

Model	Data	F1 Cult. (%)	C-SPEC	F1 MMLU (%)	Std.
Llama-3.1-8B-IT	WVS	29.61	0.76	42.78	1.36
	WVS+Wiki	31.19	0.78	49.02	2.66
	WVS+NormAd	40.94	0.89	50.43	1.32
Gemma-2-9B-IT	WVS	39.22	0.81	45.31	4.19
	WVS+Wiki	37.25	0.80	47.05	3.93
	WVS+NormAd	40.01	0.83	55.19	3.11
Qwen2.5-7B-IT	WVS	48.05	0.92	68.32	0.47
	WVS+Wiki	46.00	0.90	68.72	0.95
	WVS+NormAd	47.67	0.94	67.51	0.62

Table 5: Averaged performances on downstream offensiveness classifications (F1 Cult.), MMLU evaluations (F1 MMLU), and cultural specificity (C-SPEC) results for various instruction-tuned models and data configurations. Std. (Standard Deviation) measures the variations across culture adapters in MMLU.

5 Enhancing Cultural Adaptation with Different Data

While WVS-based training provides a foundation for cultural value adaptation, our results in Section 4 show that it seldom produces strong diagonal patterns, indicating limited cultural specialization. A critical question is *what additional data could enhance cultural value adaptation in downstream tasks?*

It is known in social sciences that humans exhibit gaps in what people "think" and how they "behave" (Gross and Niman, 1975; Fazio, 1981, inter alias). This suggests that self-reported value data, such as the World Values Survey (WVS), may be insufficient for adapting and improving tasks that require behavioral changes based on cultural values. Furthermore, numerical responses to questions like *"How important is family in your life? on a scale from 1 to 4"* in WVS do not directly indicate whether someone would label a sentence attacking family members as offensive (which is an issue relevant to the task evaluated in this work). Hence, we incorporate two additional cultural data sources, Wikipedia and NormAd, hypothesizing that introducing data containing more objective *narratives of culture* could enhance the model's performance and understanding of cultural values.

In this section, we now focus our evaluations on instruction-tuned models, as users in real-world applications are more likely to interact with these than with base models. Additionally, we expand our evaluations beyond Llama models to include Gemma and Qwen to show the generality of our findings.

5.1 Improved Overall Performance

The addition of Wikipedia and NormAd data leads to notable gains in overall classification performance. For instance, Llama-3.1-8B-IT's performance on the downstream offensiveness classification tasks (denoted as **F1 Cult.** in Table 5) rises from 29.61% (WVS-only) to 40.94%

(WVS+NormAd), reflecting the value of richer, context-laden cultural information. Similarly, Gemma-2-9B-IT also benefits from data augmentation, though to a lesser degree of 0.79 points in F1. Overall, WVS+NormAd also provides its largest results boost.

5.2 Improved C-SPEC with Wikipedia and NormAd

Beyond raw performance gains in downstream offensiveness classification tasks, the addition of Wikipedia and NormAd significantly enhances cultural specificity (i.e., C-SPEC). As noted in Table 5, integrating these datasets consistently increases the C-SPEC scores for all three models, indicating more culture-distinct behavior. For example, Llama-3.1-8B-IT's C-SPEC improves from 0.76 (WVS-only) to 0.89 (WVS+NormAd), and Figure 3 illustrates how heatmaps (show in the ranking of the results for better visualization) become more diagonal and less cross-cultural interference. Incorporating additional cultural narrative data from WVS and NormAd enhances culture-specific performance and highlighting unique cultural references more effectively.

5.3 Improved Factual Knowledge

Finally, Table 5 also indicates that combining WVS with Wikipedia and NormAd boosts MMLU scores. Llama-3.1-8B-IT's MMLU rises from 42.78% (WVS-only) to 50.43% (WVS+NormAd), and Gemma-2-9B-IT experiences a similar improvement (45.31% to 55.19%) under the same configuration. These findings imply that curated knowledge (e.g., encyclopedic facts in Wikipedia, normative behavior patterns in NormAd) augments the model's base understanding of cultural context. Although Qwen2.5-7B-IT shows smaller or mixed variations at higher performance levels, the general trend remains that more context-rich data increases both cultural adaptation and factual competence.

6 Further Analysis

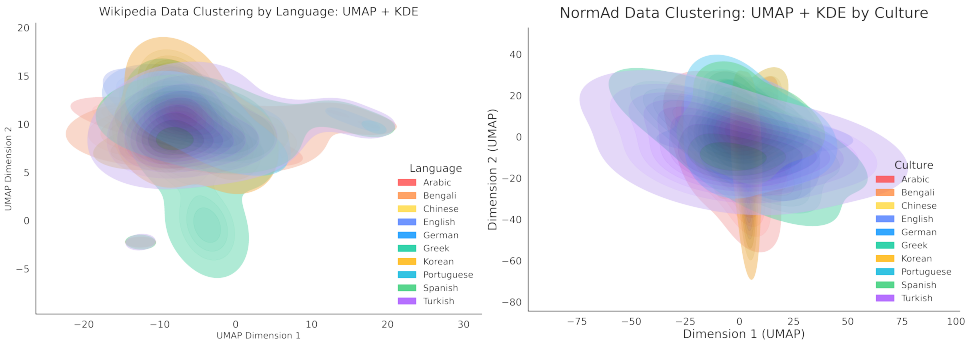
Our empirical results confirm that adding objective cultural descriptions and context-specific examples improves cultural value adaptation on downstream tasks. In this section, we analyze the data further to understand why.

Overlapping Embeddings versus Distinct Adaptations. We first embed each data source using LaBSE (Feng et al. 2022, a multilingual embedding model that compresses texts into a shared semantic space), then project the embedding with kernel density estimation (KDE). The results for WVS, Wikipedia and NormAd are shown in Figure 1 and Figure 4 respectively. It is interesting to note that there is no distinct separation between cultures within a dataset. This suggests that semantic differences in the data are not the primary factor influencing downstream tasks.

This discrepancy likely occurs because Wikipedia and NormAd differ in *how* they encode cultural details, even if their embeddings are not sharply separated (see Table 6 in Appendix for data examples). Wikipedia provides broad encyclopedic summaries, covering historical contexts and traditions, while NormAd provides scenario-specific norms that directly inform cultural behaviors (e.g., respecting elders in formal gatherings). These nuanced differences at the domain level do not necessarily create distinct embedding clusters. Nevertheless, the descriptive, scenario-based NormAd dataset enhances fine-tuning by providing more targeted cultural cues. As a result, the model can better isolate culture-specific behaviors, yielding higher C-SPEC scores.

6.1 Summary of Findings

Augmenting survey data with more descriptive sources enables a model to retain distinct cultural values: **Fine-tuning with additional cultural narratives** makes each adapter more specialized to its target culture, as evidenced by the improved diagonal in the heatmaps and higher C-SPEC scores in Table 5. While Wikipedia adds moderate benefits, NormAd's **situational norms consistently yield clearer cultural separation**. Despite similar embeddings in KDE plots, Wikipedia and NormAd influence models in fundamentally different ways than WVS. NormAd provides structured, scenario-based norms that guide models toward clearer cultural distinctions. Unlike broad textual data, these norms offer direct behavioral cues, making fine-tuning more effective in encoding culturally specific reasoning. These results show that **numeric survey data alone (WVS) is rarely sufficient for robust cultural value adaptation** requiring downstream behavioral changes. Textual resources that explicitly articulate context and behavior can substantially improve the cultural specialization.



UMAP KDE on Wikipedia

UMAP KDE on NormAd

Figure 4: Kernel Density Estimation (KDE) plots of UMAP embeddings using LaBSE (Feng et al., 2022) for Wikipedia and NormAd datasets. These visualizations show the density distributions of the data in the reduced-dimensional space.

7 Related Work

General Adaptation to Cultural Values. Several existing work approaches cultural value adaptations in LLMs through prompting (AlKhamissi et al., 2024; Wang et al., 2024; Tao et al., 2024), continual pre-training on diverse multilingual data (Wang et al., 2024; Choenni et al., 2024) or direct tuning on survey data or synthetic data based on survey (Li et al., 2024a; Xu et al., 2024; Li et al., 2024b). In particular, the basis of our investigation, CultureLLM (Li et al., 2024a), employs semantically augmented data from the World Values Survey (WVS) to represent the average opinion of a culture. In this paper, we extend the investigation using descriptive cultural principles and provide a comprehensive analysis.

Recent research also explored value prediction with In-Context Learning (ICL)-based adaptation methods (Choenni and Shutova, 2024; Jiang et al., 2024). Particularly, Jiang et al. (2024) showed a mild inconsistency when models adapted using individual data from one continent were evaluated using data from another (e.g., training data for other continents generally improves alignment to Oceania people). While related to our work, we focus specifically on the impact at the country level rather than the broader continent level.

Pluralistic Alignment. Related to cultural value adaptation, recent studies advocate for pluralistic alignment (Sorensen et al., 2024), wherein a model should reflect the values of multiple stakeholders or sub-groups. Feng et al. (2024) proposed a modular pluralistic alignment method, which primarily focuses on integrating diverse opinions. This research direction differs from typical existing cultural value adaptation work which mainly focuses on reflecting the averaged value of a culture (Li et al., 2024a,b; Tao et al., 2024; AlKhamissi et al., 2024; Choenni et al., 2024, inter alia).

Cultural Inconsistencies in LLMs. Recent work highlights the challenges LLMs face in maintaining consistent cultural values across different linguistic and social contexts (Adilazuarda et al., 2024; Beck et al., 2024). One of the reasons why these inconsistencies arise is due to biases in training data (Mihalcea et al., 2024; Sorensen et al., 2022), which often prioritize Western or English-centric perspectives, leading to misalignment when applied to non-WEIRD cultures (Mihalcea et al., 2024). Additionally, Mukherjee et al. (2024), shows that even the current LLMs are prone to a slight cultural and noncultural perturbation even on factual questions such as MMLU. This work builds upon the findings on how existing adaptation strategies address cultural disparities in downstream tasks.

8 Conclusion

In this paper, we present a comprehensive study that highlights both the potential and limitations of cultural value adaptation using WVS as training data. We find that models trained exclusively on WVS often fail to capture distinct cultural values, resulting in minimal "cultural specificity", or what we termed as C-SPEC , across cultures. By integrating WVS data with cultural narratives such as Wikipedia or NormAd dataset, we enhance cultural specialization and reduce interference in downstream tasks. Our findings highlight the importance of combining objective survey data with cultural narratives for more accurate cultural representation and improved downstream tasks performance.

Limitations

In this work, we focus on a select set of data as the source data for adaptation, including the World Values Survey (WVS), Wikipedia, and NormAd. While these datasets offer diverse cultural signals, they each come with inherent biases. For instance, WVS could be subject to self-reporting biases, Wikipedia reflects editorial biases, and NormAd consists of curated examples that may not fully represent all cultural variations.

Furthermore, our evaluation is limited to selected culturally sensitive tasks, which may not fully capture the broader range of tasks needed to assess how cultural value adaptation influences behavior. However, such an investigation requires careful task design and is beyond the scope of this work.

Ethics Statement

Our work aims to enhance cultural value adaptations in NLP systems while carefully considering potential societal impacts. While this research may help reduce Western-centric bias and improve offensive content classification by incorporating diverse cultural values, we acknowledge the risks of potential misuse, including cultural stereotyping and demographic profiling. We emphasize that our findings should be applied thoughtfully, with continuous consideration of cultural context, while being careful not to anthropomorphize LLMs by attributing to them true cultural understanding or awareness. Additionally, we encourage future research to develop more nuanced methodologies and evaluation frameworks that better represent cultural diversity in NLP systems.

References

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15763–15784, Miami, Florida, USA. Association for Computational Linguistics.

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.

Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2024. Sensitivity, performance, robustness: Deconstructing the effect of sociodemographic prompting. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2589–2615, St. Julian's, Malta. 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.

Rochelle Choenni, Anne Lauscher, and Ekaterina Shutova. 2024. The echoes of multilinguality: Tracing cultural value shifts during language model fine-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15042–15058, Bangkok, Thailand. Association for Computational Linguistics.

Rochelle Choenni and Ekaterina Shutova. 2024. Self-alignment: Improving alignment of cultural values in llms via in-context learning. *CoRR*, abs/2408.16482.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi,

Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

RH Fazio. 1981. Direct experience and attitude behavior consistency. *Advances in experimental social psychology*, 14.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Shangbin Feng, Taylor Sorensen, Yuhan Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia Tsvetkov. 2024. Modular pluralism: Pluralistic alignment via multi-LLM collaboration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4151–4171, Miami, Florida, USA. Association for Computational Linguistics.

Clifford Geertz. 2017. *The interpretation of cultures*. Basic books.

Steven Jay Gross and C Michael Niman. 1975. Attitude-behavior consistency: A review. *Public opinion quarterly*, 39(3):358–368.

Yufei Guo, Muzhe Guo, Juntao Su, Zhou Yang, Mengqiu Zhu, Hongfei Li, Mengyang Qiu, and Shuo Shuo Liu. 2024. Bias in large language models: Origin, evaluation, and mitigation. *CoRR*, abs/2411.10915.

Christian Haerpfer, Alejandro Moreno Ronald Inglehart, Christian Welzel, Jaime Diez-Medrano Kseniya Kizilova, Milena Lagos, Pippa Norris, Eduard Ponarin, and Bianca Puranen. 2022. World values survey: Round seven.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Joseph Henrich, Steven J Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83.

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.

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.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.

Liwei Jiang, Taylor Sorensen, Sydney Levine, and Yejin Choi. 2024. Can language models reason about individualistic human values and preferences? *CoRR*, abs/2410.03868.

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. *ArXiv preprint*, abs/2203.07785.

Md Tahmid Rahman Laskar, Sawsan Alqahtani, M Saiful Bari, Mizanur Rahman, Mohammad Abdullah Matin Khan, Haidar Khan, Israt Jahan, Amran Bhuiyan, Chee Wei Tan, Md Rizwan Parvez, Enamul Hoque, Shafiq Joty, and Jimmy Huang. 2024. A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13785–13816, Miami, Florida, USA. Association for Computational Linguistics.

Cheng Li, Mengzhou Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024a. CultureLLM: Incorporating cultural differences into large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. 2024b. CulturePark: Boosting cross-cultural understanding in large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2024. Culturally aware and adapted NLP: A taxonomy and a survey of the state of the art. *CoRR*, abs/2406.03930.

Rada Mihalcea, Oana Ignat, Longju Bai, Angana Borah, Luis Chiruzzo, Zhijing Jin, Claude Kwizera, Joan Nwatu, Soujanya Poria, and Thamar Solorio. 2024. Why ai is weird and should not be this way: Towards

ai for everyone, with everyone, by everyone. *CoRR*, abs/2410.16315.

Sagnik Mukherjee, Muhammad Farid Adilazuarda, Sunayana Sitaram, Kalika Bali, Alham Fikri Aji, and Monojit Choudhury. 2024. Cultural conditioning or placebo? on the effectiveness of socio-demographic prompting. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15811–15837, Miami, Florida, USA. Association for Computational Linguistics.

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

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

Morgane Rivière, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozinska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucinska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju-yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjösund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, and Lilly McNealus. 2024. Gemma 2: Improving open language models at a practical size. *CoRR*, abs/2408.00118.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L. Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. Position: A roadmap to pluralistic alignment. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*.

Taylor Sorensen, Joshua Robinson, Christopher Rytting, Alexander Shaw, Kyle Rogers, Alexia Delorey, Mahmoud Khalil, Nancy Fulda, and David Wingate. 2022. An information-theoretic approach to prompt engineering without ground truth labels. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 819–862, Dublin, Ireland. Association for Computational Linguistics.

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.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael Lyu. 2024. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6349–6384, Bangkok, Thailand. Association for Computational Linguistics.

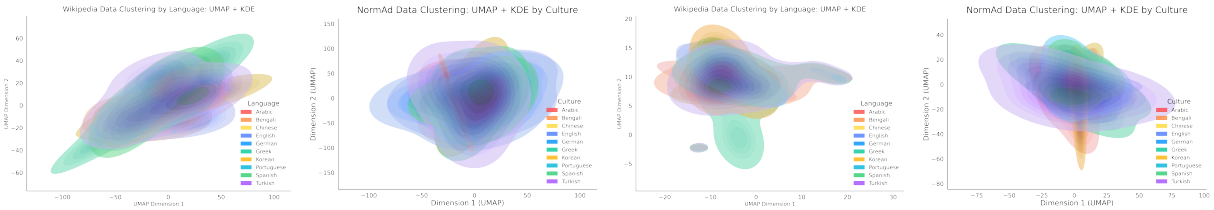
Zirui Wang, Zachary C. Lipton, and Yulia Tsvetkov. 2020. On negative interference in multilingual models: Findings and a meta-learning treatment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4438–4450, Online. Association for Computational Linguistics.

Shaoyang Xu, Yongqi Leng, Linhao Yu, and Deyi Xiong. 2024. Self-pluralising culture alignment for large language models. *CoRR*, abs/2410.12971.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çagri Çöltekin. 2020. Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation, SemEval@COLING 2020, Barcelona (online), December 12-13, 2020*, pages 1425–1447. International Committee for Computational Linguistics.

A Data Characteristics

A.1 Additional KDE Plots



t-SNE KDE on Wikipedia | t-SNE KDE on NormAd | UMAP KDE on Wikipedia | UMAP KDE on NormAd

Figure 5: Kernel Density Estimation (KDE) plots using t-SNE and UMAP projections for Wikipedia and NormAd datasets. Although projection methods vary, none of the embeddings are distinctly separable by culture, indicating shared semantic similarities of data.

A.2 Samples of WVS, Wiki, and NormAd Data

Table 6 presents a comparison of social values across different cultures by showcasing sample data from the World Values Survey (WVS), Wikipedia, and the NormAd dataset.

WVS	Wikipedia	NormAd
"topic": "SOCIAL VALUES", "q_id": "27", "q_content": "One of my main goals in life has been to make my parents proud", "option": "1. Strongly agree 2. agree 3. Disagree 4. Strongly disagree"	Arab culture is the culture of the Arabs, from the Atlantic Ocean in the west to the Arabian Sea in the east, in a region of the Middle East and North Africa known as the Arab world. The various religions the Arabs have adopted throughout their history and the various empires and kingdoms that have ruled and took lead of the civilization have contributed to the ethnogenesis and formation of modern Arab culture.	(Egypt - Background) **Basic Etiquette** - It is considered impolite to point the toe, heel or any part of the foot toward another person. Showing the sole of one's shoe is also impolite. - Modest dress and presentation is highly valued in Egyptian culture. - Greetings often occur before any form of social interaction. For example, a person joining a group is expected to greet all those present. - Generally, the younger defer to the older through showing respect.
"topic": "SOCIAL VALUES", "q_id": "28", "option": "1. Strongly agree 2. agree 3. Disagree 4. Strongly disagree", "q_content": "When a mother works for pay, the children suffer"	The culture of Bengal defines the cultural heritage of the Bengali people native to eastern regions of the Indian subcontinent, mainly what is today Bangladesh and the Indian states of West Bengal and Tripura, where they form the dominant ethnolinguistic group and the Bengali language is the official and primary language. Bengal has a recorded history of 1,400 years. After the partition, Bangladeshi culture became distinct from the mainstream Bengali culture.	(Spain - Value) Flexibility in social interactions and a relaxed approach to time and financial contributions among peers.
"topic": "SOCIAL VALUES", "q_id": "29", "option": "1. Strongly agree 2. agree 3. Disagree 4. Strongly disagree", "q_content": "On the whole, men make better political leaders than women do"	Chinese culture is one of the world's oldest cultures, originating thousands of years ago. The culture prevails across a large geographical region in East Asia with Sinosphere in whole and is extremely diverse, with customs and traditions varying greatly between counties, provinces, cities, towns. The terms 'China' and the geographical landmass of 'China' have shifted across the centuries, before the name 'China' became commonplace in modernity.	(UK - Story) Elena was planning to visit her friend Tom's house. She texted Tom to ask if it was a good time for him. Upon arrival, she made sure not to overstay by asking if he needed her to leave at a certain time. Is what Elena did socially acceptable?

Table 6: Comparison of data across different cultures from the data sources used in our paper.

B Training Procedure and Data Reformulation

Following Li et al. (2024a), our experiments employ LoRA adapters with 4-bit quantization using the BitsAndBytes configuration to optimize the memory usage. We use an alpha value of 16, a dropout rate of 0.1, and a rank of 64, specifically targeting the query (`q_proj`) and value (`v_proj`) projection matrices of the transformer architecture.

We reformulate the training data using the following formats:

1. **Standard Survey Training (WVS).** The WVS survey data is structured with clear task markers:

```
### Task: Survey Question-Answer
### Question: [question_content]
### Answer: [answer_content]
```

2. **Wikipedia.** When the Wikipedia data is used, the information is formatted as:

```
### Task: Cultural Context
### Culture: [culture_name]
### Description: [cultural_context]
```

3. **NormAd.** We integrate the data using the following prompt:

```
### Task: NormAd Cultural Context
### Culture: [culture_name]
### Country: [country_name]
### Background: [background_info]
### Rule-of-Thumb: [cultural_rule]
### Story: [narrative]
### Explanation: [detailed_explanation]
```

The training process optimizes memory usage with gradient checkpointing and uses a constant learning rate of 2×10^{-4}. The model is trained for 6 epochs with a warmup ratio of 0.03 and employs 8-bit Adam optimization with a weight decay of 0.001. For reproducibility, the process is seeded (`seed=42`) and ensures deterministic CUDA operations.

C Full Performance Tables

C.1 Combined Cultural Adaptation

Instead of learning a separate adapter per culture, we combine training data from all target cultures and produce one multi-culture adapter. This can potentially help the model recognize cross-cultural patterns or exploit data from many cultures. However, it risks "averaging out" the distinctions, possibly causing *cultural interference* (e.g., losing the unique viewpoint for each culture, akin to interference in multilinguality Conneau et al. 2020; Wang et al. 2020). While combined-culture adaptation can improve some low-resource cultures (e.g., Korean, Bengali), it could reduce performance for others, indicating cultural interference.

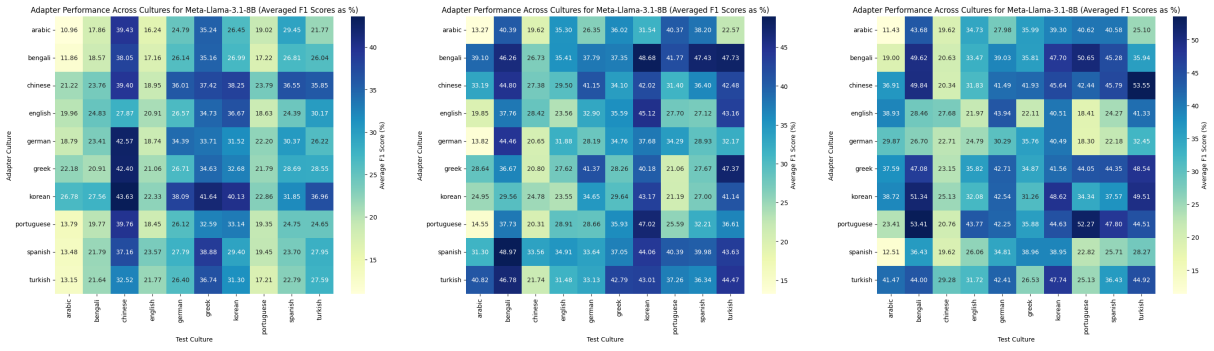
Combined-Culture Adaptation - WVS											
Model	**ara**	**ben**	**zho**	**eng**	**deu**	**ell**	**kor**	**por**	**spa**	**tur**	**Avg.**
Llama-3.1-8B	33.44	23.24	28.39	17.12	36.75	15.11	37.09	17.88	25.62	39.29	27.39
Llama-3.1-8B-IT	28.00	30.34	42.77	23.90	46.08	31.42	43.32	22.88	33.52	43.50	34.57

Table 7: Results for Combined-Culture Adaptation on WVS.

C.2 Freeform Generation

C.2.1 Performance Heatmaps - Llama-3.1-8B

Figure 6 illustrates the culture-specific classification performance of the Llama-3.1-8B model through three heatmaps corresponding to different data configurations: panel (a) uses only WVS data, panel (b) integrates cultural context from Wikipedia (WVS+Wiki), and panel (c) combines WVS with NormAd data (WVS+NormAd); in each heatmap, color gradients represent the ranks of the adaptation results, providing a visual assessment of how incorporating additional cultural sources can enhance or alter model performance across diverse cultural contexts.

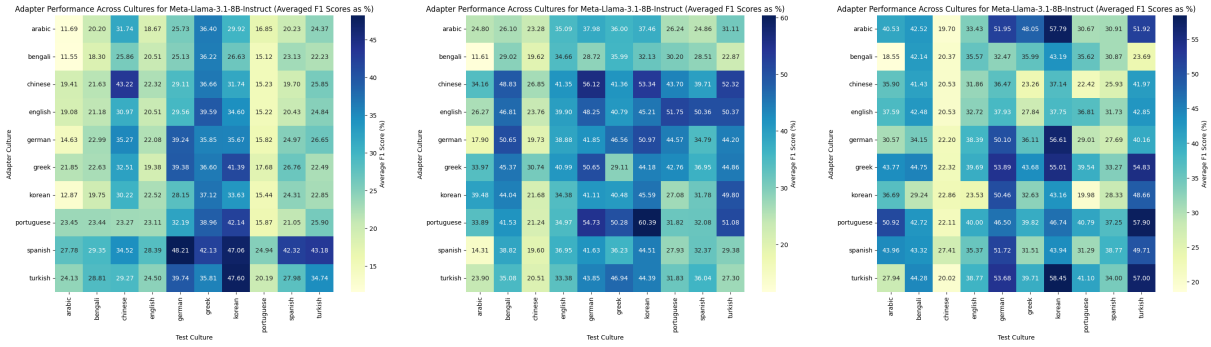


WVS **WVS+Wiki** **WVS+NormAd**

Figure 6: Heatmaps of culture-specific classification performance (Llama-3.1-8B) using different data sources based on the ranks of the adaptation results.

C.2.2 Performance Tables - Llama-3.1-8B-Instruct

Figure 7 illustrates the performance of Llama-3.1-8B-Instruct model through three heatmaps.

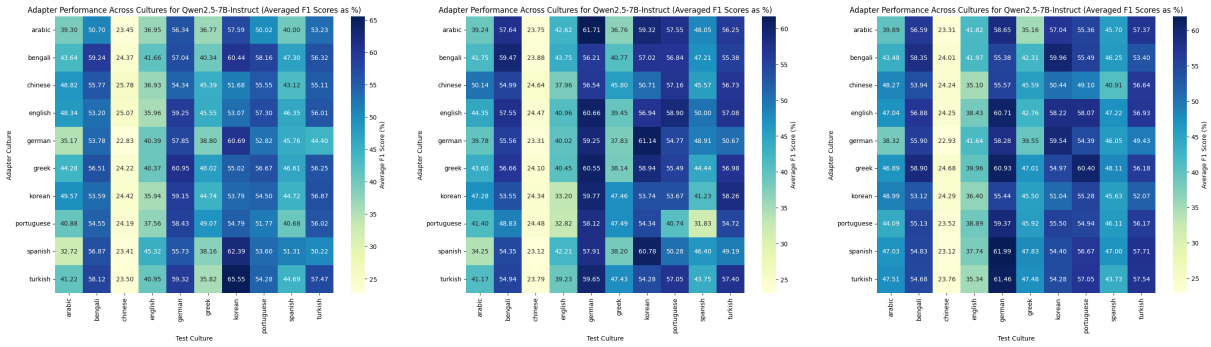


WVS **WVS+Wiki** **WVS+NormAd**

Figure 7: Heatmaps of culture-specific classification performance (Llama-3.1-8B-IT) using different data sources based on the ranks of the adaptation results.

C.2.3 Performance Tables - Qwen2.5-7B-IT

Figure 8 illustrates the performance of the Qwen2.5-7B-IT model through three heatmaps.

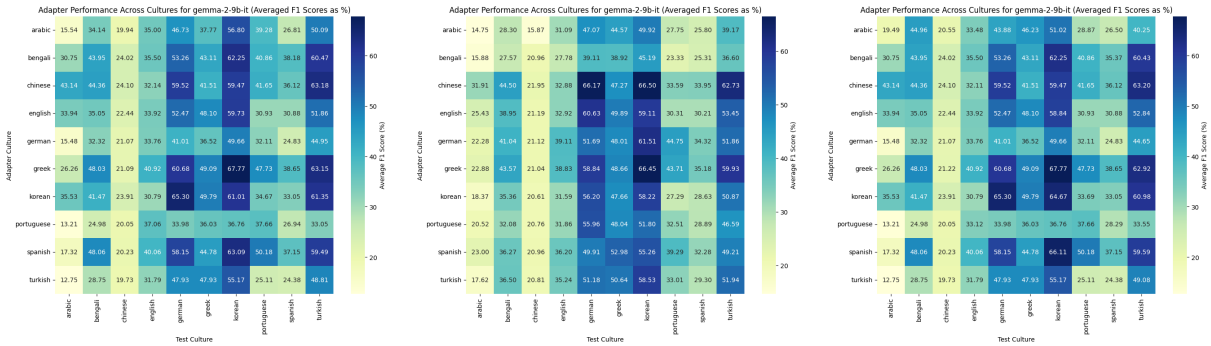


WVS **WVS+Wiki** **WVS+NormAd**

Figure 8: Heatmaps of culture-specific classification performance (Qwen2.5-7B-IT) using different data sources based on the ranks of the adaptation results.

C.2.4 Performance Tables - Gemma-2-9B-IT

Figure 9 illustrates the performance of the Gemma-2-9B-IT model through three heatmaps.



WVS **WVS+Wiki** **WVS+NormAd**

Figure 9: Heatmaps of culture-specific classification performance (Gemma-2-9B-Instruct) using different data sources based on the ranks of the adaptation results.

C.3 Normalized Scores Tables

Adapter Cult.	ara	ben	zho	eng	deu	ell	kor	por	spa	tur
ara	0.4209	0.6882	0.7343	0.6578	0.5337	0.8640	0.6284	0.6758	0.4780	0.5645
ben	0.4156	0.6237	0.5984	0.7223	0.5213	0.8598	0.5595	0.6062	0.5466	0.5148
zho	0.6986	0.7371	1.0000	0.7862	0.6038	0.8703	0.6667	0.6107	0.4654	0.5985
eng	0.6867	0.7216	0.7166	0.7225	0.6131	0.9398	0.7268	0.6103	0.4828	0.5751
deu	0.5266	0.7835	0.8161	0.7779	0.8139	0.8509	0.7493	0.6345	0.5899	0.6172
ell	0.7865	0.7711	0.7522	0.6827	0.8168	0.8688	0.8695	0.7089	0.6324	0.5208
kor	0.4633	0.6728	0.6991	0.7933	0.5838	0.8810	0.7065	0.6193	0.5745	0.5292
por	0.8442	0.7987	0.5384	0.8142	0.6676	0.9248	0.8853	0.6364	0.4975	0.5997
spa	1.0000	1.0000	0.7987	1.0000	1.0000	1.0000	0.9886	1.0000	1.0000	1.0000
tur	0.8685	0.9817	0.6772	0.8628	0.8242	0.8501	1.0000	0.8094	0.6610	0.8045

Table 8: Normalized Scores and diagonality on Llama-3.1-8B-IT for WVS. Rows represent adapter culture and columns represent culture test set.

Adapter Cult.	ara	ben	zho	eng	deu	ell	kor	por	spa	tur
ara	0.7255	0.5862	0.7980	0.8510	0.6329	0.7875	0.6219	0.7635	0.9012	0.5731
ben	0.3320	0.6027	0.4640	0.8319	0.5354	0.7861	0.5575	0.5934	0.7311	0.4903
zho	0.8268	0.7872	1.0000	0.9636	0.8755	1.0000	0.8753	0.8413	0.8521	0.7687
eng	0.7514	0.8592	0.9779	0.7852	0.9733	0.8209	0.9034	0.9299	0.9792	0.8828
deu	0.5986	0.8016	0.9445	0.7760	0.8604	0.9679	0.8233	0.7221	0.7729	0.6408
ell	0.9031	0.9440	0.7137	1.0000	0.9152	0.7502	0.8970	1.0000	1.0000	0.9678
kor	1.0000	1.0000	0.5369	0.8979	1.0000	0.8037	1.0000	0.8637	0.8274	1.0000
por	0.7863	0.7632	0.5586	0.8940	0.8065	0.9270	0.8570	0.7430	0.6613	0.7746
spa	0.4076	0.6871	0.5581	0.8136	0.6525	0.7973	0.7152	0.5486	0.6715	0.5138
tur	0.5835	0.6960	0.9223	0.8341	0.7417	0.8859	0.8456	0.7119	0.9690	0.6794

Table 9: Normalized Scores and diagonality on Llama-3.1-8B-IT for WVS+Wikipedia. Rows represent adapter culture and columns represent culture test set.

Adapter Cult.	ara	ben	zho	eng	deu	ell	kor	por	spa	tur
ara	0.7961	0.8685	0.7190	0.8358	0.9640	1.0000	0.9533	0.7462	0.7974	0.8966
ben	0.3643	0.8608	0.7432	0.8893	0.6026	0.7490	0.7124	0.8666	0.7963	0.4092
zho	0.7051	0.8463	0.7493	0.7967	0.6767	0.4841	0.6127	0.5454	0.6689	0.7248
eng	0.7383	0.8678	0.7493	0.8180	0.7038	0.5794	0.6227	0.8956	0.8185	0.7400
deu	0.6004	0.6975	0.8100	0.9597	0.9297	0.7515	0.9337	0.7058	0.7142	0.6936
ell	0.8597	0.9141	0.8144	0.9923	1.0000	0.9091	0.9074	0.9620	0.8582	0.9469
kor	0.7207	0.5973	0.8340	0.5882	0.9363	0.6791	0.7118	0.4862	0.7307	0.8404
por	1.0000	0.8727	0.8067	1.0000	0.8628	0.8287	0.7709	0.9925	0.9607	1.0000
spa	0.8634	0.8849	1.0000	0.8843	0.9596	0.6558	0.7248	0.7613	1.0000	0.8585
tur	0.5487	0.9045	0.7305	0.9694	0.9960	0.8265	0.9640	1.0000	0.8771	0.9844

Table 10: Normalized Scores and diagonality on Llama-3.1-8B-IT for WVS+NormAd. Rows represent adapter culture and columns represent culture test set.

C.4 Probability-Based Generation

Table 11 shows normalized F1 score for probability-based generation evaluations.

Language	Baseline		Translated	
	Llama-3.1-8B	Llama-3.1-8B-IT	Llama-3.1-8B	Llama-3.1-8B-IT
ara	30.52	28.83	33.24	37.81
ben	22.53	45.45	29.70	42.77
zho	28.84	41.35	35.77	46.28
eng	28.37	42.81	30.21	49.18
deu	32.53	40.40	28.80	41.92
ell	30.77	46.05	32.11	36.34
kor	30.28	41.80	34.33	44.63
por	29.24	40.11	27.55	38.08
spa	28.96	43.77	23.32	38.60
tur	30.44	43.93	30.24	40.46

Table 11: Performance on MMLU when training each adapter with different WVS cultural data. Baseline refers to fine-tuning using English-language cultural value data with the *Llama-3.1-8B* and *Llama-3.1-8B-IT* models. Translated represents training with WVS cultural values translated into the respective target language, using the *Llama-3.1-8B* and *Llama-3.1-8B-IT* models. The zero-shot performance for Arabic is 0.35 with *Llama-3.1-8B* and 0.45 with *Llama-3.1-8B-IT*.

D Invalid Answer Check

D.1 Code for Invalid Answer Filtering

We process the generated response to determine whether it contains a valid answer using the following function. If the expected answer format is not detected, a default value is assigned.

```
# Generate response from model
def getResponse(prompt, base_model=None, base_tokenizer=None):
    input_ids = base_tokenizer(prompt, return_tensors="pt").input_ids.to("cuda")

    base_model.eval()
    with torch.no_grad():
        output = base_model.generate(
            input_ids,
            max_new_tokens=25,
            temperature=0.0,
            num_beams=1,
            do_sample=False,
            eos_token_id=base_tokenizer.eos_token_id,
            pad_token_id=base_tokenizer.pad_token_id,
        )

    output_txt = base_tokenizer.decode(output[0], skip_special_tokens=True)

    # Extract the answer (e.g., "A")
    match = re.search(r'### Answer:\s*([A-D])', output_txt)
    if match:
        prediction = match.group(1)
        invalid_response = False
    else:
        prediction = "A"  # Default or handle as appropriate
        invalid_response = True

    return output_txt, prediction, invalid_response
```

Listing 1: Generating Response and Invalid Response filtering from Model.

D.2 Invalid Examples

Table 12 illustrates examples of invalid responses from our evaluation datasets (English, German, and Spanish). The model failed to properly complete its task of detecting toxic content, showing truncated or incomplete outputs across different languages.

Culture	Invalid Response Example
English	{"input": "@user @user legit nilas retarded idk", "output": "@user @user legit nilas retarded idk what to do @user @user @user", "extracted_output": "", "prediction": "", "label": "0", "invalid_response": true}
German	{"input": "@FobiMania @ManuelaSchwesig @MartinSchulz Ich scheisse auf die SPD. |LBR| SCHEINHEILGSTE PARTEI DEUTSCHLANDS !!!", "output": "### Question: Is the following sentence offensive: @FobiMania @ManuelaSchwesig @MartinSchulz Ich scheisse auf die SPD. |LBR| SCHEINHEILGSTE PARTEI DEUTSCHLANDS!!! \"1. Offensive\", \"2. Not offensive\" without explanation. ### Answer: @FobiMania @ManuelaSchw", "extracted_output": "@FobiMania @ManuelaSchw", "prediction": "@fobimania @manuelaschw", "label": "OFF", "invalid_response": true}
Spanish	{"input": "A ver si le dan un Goya al youtuber que se llevó la hostia por subnormal.", "output": "### Question: Does the following sentence contain hate speech: A ver si le dan un Goya al youtuber que se llevó la hostia por subnormal. \"1. Hatespeech\", \"2. Not Hatespeech\" without explanation. ### Answer: A ver si le dan un Goya al y", "extracted_output": "A ver si le dan un Goya al y", "prediction": "a ver si le dan un goya al y", "label": "HS", "invalid_response": true}

Table 12: Examples of invalid responses.

D.3 Invalid Rate

Table 13 shows our findings in evaluating the zero-shot model, compared to the finetuned models on WVS (Single-Culture-WVS) and WVS+Wikipedia (Single-Culture-WVS+Wiki). This table shows that the zero-shot performance of the models still suffers from the lack of instruction capability, even on the instruct variant. We further see a notable decrease in the invalid response ratio after finetuning.

Model	Scenario	Avg. Invalid Count	Total Samples	Invalid Ratio
Llama-3.1-8B	Zero-shot	222.5849	1106	21.77%
	Single-Culture-WVS	143.6604	1106	14.06%
	Single-Culture-WVS+Wiki	133.6604	1106	12.97%
Llama-3.1-8B-IT	Zero-shot	234.5660	1106	22.55%
	Single-Culture-WVS	134.3774	1106	13.16%
	Single-Culture-WVS+Wiki	121.5849	1106	12.01%
Llama-3.1-8B	Zero-shot	11,797	58,638	20.12%
	Single-Culture-WVS	94,663	645,018	14.68%
	Single-Culture-WVS+Wiki	90,562	645,018	14.04%
Llama-3.1-8B-IT	Zero-shot	12,432	58,638	21.20%
	Single-Culture-WVS	69,804	645,018	10.82%
	Single-Culture-WVS+Wiki	71,019	645,018	11.01%

Table 13: Invalid count and ratio for different models and scenarios.

E Prompts

Table 14 summarizes the prompt templates used by the `getPrompt` function for various tasks.

Task(s)	Prompt Format
offensive_detect, abusive_detect	### Question: is the following sentence offensive: {input_txt}. "1. Offensive", "2. Not offensive" without explanation. ### Answer:
hate_detect (excluding hate_detect_fine-grained)	### Question: does the following sentence contain hate speech: {input_txt}. "1. Hatespeech", "2. Not Hatespeech" without explanation. ### Answer:
vulgar_detect_mp	### Question: does the following sentence contain vulgar speech: {input_txt}. "1. Vulgar", "2. Not Vulgar" without explanation. ### Answer:
spam_detect	### Question: is the following sentence a spam tweet: {input_txt}. "1. Spam", "2. Not Spam" without explanation. ### Answer:
hate_detect_fine-grained	### Question: Does the following sentence contain hate speech? {input_txt} Please choose one of the following options without explanation: 1. Not Hatespeech, 2. Race, 3. Religion, 4. Ideology, 5. Disability, 6. Social Class, 7. Gender, ### Answer:
offensive_detect finegrained	### Question: Does the following sentence contain offensive speech? {input_txt} Please choose one of the following options without explanation: 1. Not hatespeech 2. Profanity, or non-targeted offense 3. Offense towards a group 4. Offense towards an individual 5. Offense towards an other (non-human) entity ### Answer:
hate_off_detect	### Question: does the following sentence contain hate speech or offensive content: {input_txt}. "1. Hate or Offensive", "2. Not Hate or Offensive" without explanation. ### Answer:
stereotype_detect, mockery_detect, insult_detect, improper_detect, aggressiveness_detect, toxicity_detect, negative_stance_detect, homophobia_detect, racism_detect, misogyny_detect, threat_detect, hostility_directness_detect	### Question: does the following sentence contain {entity}: {input_txt}. "0. No", "1. Yes" without explanation. ### Answer: (*Note:* {entity} is derived from the task name, e.g., bias_on_gender_detect → *gender bias*, etc.)
hate_offens_detect	### Question: does the following sentence contain hate speech: {input_txt}. "0. No", "1. Yes" without explanation. ### Answer:

Table 14: Overview of prompts generated by `getPrompt`.

F Source WVS, Wikipedia, NormAd Data

Table 15 contains the source of data used in our experiments. The first column lists datasets, and the second column provides clickable hyperlinks.

Source	URL
World Values Survey (WVS)	WVS
Wikipedia (Arab Culture)	Arab Culture
Wikipedia (Bengal Culture)	Culture of Bengal
Wikipedia (Chinese Culture)	Chinese Culture
Wikipedia (English Culture)	Culture of England
Wikipedia (German Culture)	Culture of Germany
Wikipedia (Greek Culture)	Culture of Greece
Wikipedia (Korean Culture)	Culture of Korea
Wikipedia (Portuguese Culture)	Culture of Portugal
Wikipedia (Spanish Culture)	Culture of Spain
Wikipedia (Turkish Culture)	Culture of Turkey
NormAd Dataset	NormAd

Table 15: Data sources and URLs.

G Cross-Cultural Confusion Matrix on Llama-3.1-8B

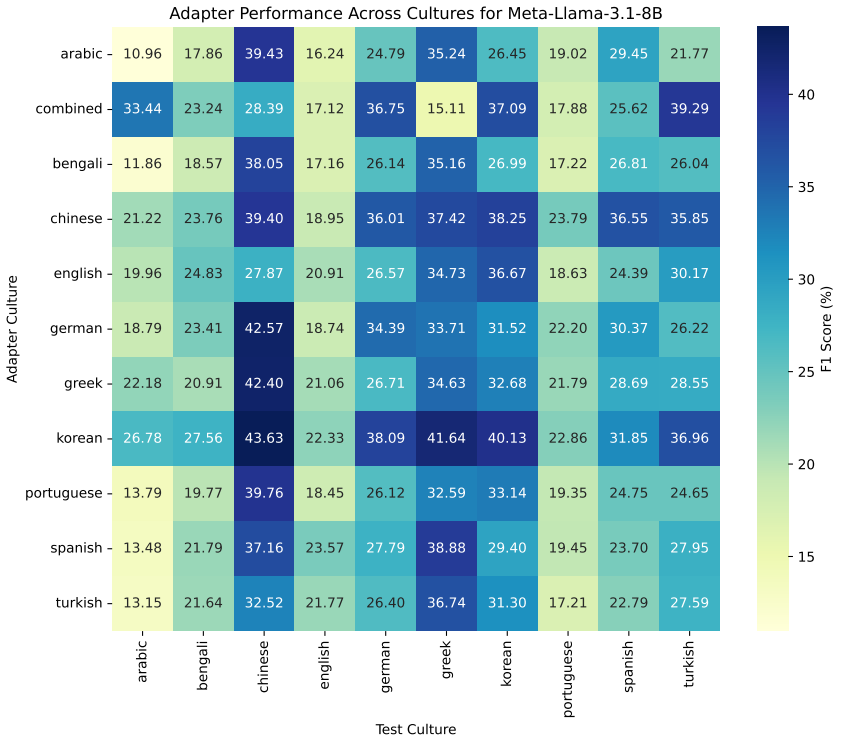


Figure 10: Cross-culture confusion matrix for the WVS-only baseline on Llama-3.1-8B (8B, base). The C-SPEC score is ≈ 0.78, reflecting substantial overlap in predictions across cultures.