From Surveys to Narratives: Rethinking Cultural Value Adaptation in LLMs

M. Farid Adilazuarda[1], **Chen Cecilia Liu**[2], **Iryna Gurevych**[2], **Alham Fikri Aji**[1]
[1]MBZUAI
[2]Ubiquitous Knowledge Processing Lab (UKP Lab),
Technical University of Darmstadt and
National Research Center for Applied Cybersecurity ATHENE, Germany

Abstract

Adapting cultural values in Large Language Models (LLMs) presents significant challenges, particularly due to biases and limited training data. Prior work primarily aligns LLMs with different cultural values using World Values Survey (WVS) data. 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 investigate these issues, we augment WVS with encyclopedic and scenario-based cultural narratives from Wikipedia and NormAd. While these narratives may have variable effects on downstream tasks, they consistently improve cultural distinctiveness than survey data alone. Our work highlights the inherent complexity of aligning cultural values to guide task-specific behavior. Code: `https://github.com/faridlazuarda/from-surveys-to-narratives`.

1 Introduction

Recent research in Large Language Models (LLMs) suggests 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). The WEIRD-centric bias can harm specific groups and limit the 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). However, current research lacks robust mechanisms to adapt LLMs' outputs in ways that reflect different cultural value systems (i.e., culturally adapt LLMs).[1]

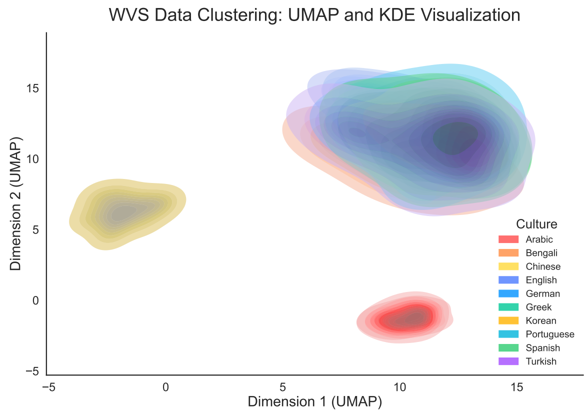


Figure 1: UMAP-KDE visualization of cultural value distributions from WVS data reveals significant homogenization. While Arabic (lower right) and Chinese (left) cultures form distinct clusters, many others converge in the upper right. This suggests that current WVS-based training may be insufficient to capture cultural nuances.

Existing work often adapts LLMs to cultural values by leveraging self-reported survey data (Li et al., 2024a; Xu et al., 2025; 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?"* 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., 2025) 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), poten-

[1]For 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).

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 18052–18079
November 4-9, 2025 ©2025 Association for Computational Linguistics

tially homogenizing nuanced cultural dimensions.

Ideally, cultural value adaptation should also enhance downstream tasks within each culture. However, several challenges emerge. First, adapting multiple cultural values may create interference similar to that seen in multilingual models (Conneau et al., 2020; Wang et al., 2020), 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 WVS's ability to accurately reflect cultural behavior for LLM training, necessitating further investigation.

In this work, we tackle these challenges through a critical evaluation of current cultural value adaptation methods. Through a series of experiments, we reveal the key limitations of using WVS as training data (§4): while WVS provides insights into cultural values, it lacks the contextual depth needed to inform value-driven behavior in downstream tasks. Given these limitations of survey data, we investigate whether augmenting WVS with richer narrative sources like encyclopedic descriptions (Wikipedia) and scenario-based norms (NormAd; Rao et al. 2025) yields more culturally distinct and effective LLM adaptations. We assess the impact on preserving cultural distinctiveness, downstream task performance, and factual knowledge.

To summarize, our contributions are: **1)** We identify ***cultural interference*** in adaptation using WVS, showing that it improves tasks like offensiveness classification but tends to homogenize cultural behaviors rather than preserve their differences. **2)** We demonstrate ***knowledge interference*** from adaptation, where adaptation can degrade factual knowledge understanding. **3)** We investigate the impact of ***augmenting WVS with context-rich narratives (Wikipedia, NormAd)*** and find that, while their effects on downstream tasks can vary, they help preserve cultural distinctiveness. Our analysis showcased the challenge of aligning cultural values to influence task-specific behavior and highlighted the need for further research into *which types of data* best support effective cultural adaptation.

2 Methodology

We systematically designed experiments to investigate our research question. This section details our methodologies for cultural adaptation and performance evaluation. We begin with zero-shot prompting, followed by single-culture adapter fine-tuning, and conclude with an analysis of cross-cultural interference using auxiliary tasks such as MMLU (Massive Multitask Language Understanding; Hendrycks et al. 2021). We 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

Beyond zero-shot prompts, we explore explicit fine-tuning with culture-specific data, referred to as *single-culture adaptation* in our paper. 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 a specific culture. However, data sparsity and overfitting are risks, particularly for cultures with limited samples.

In single-culture adaptation, each LoRA adapter is trained to reflect the high-level cultural values present in the training dataset. During inference, the appropriate adapter is activated based on the test target culture specified.

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 the datasets used for training and evaluation, the 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).

Terminology. In this paper, we use *culture* to denote a linguistic–regional grouping (e.g., "Arab" represented by Iraq and Jordan; "Spanish" by Argentina and Mexico). *Country* refers to specific national datasets used to instantiate a culture. *Language* denotes the linguistic form of inputs/outputs (e.g., ara, spa). While related, these are not interchangeable; our experiments condition adapters on *culture*, evaluate on country-specific test sets, and control for language.

3.2 Training Dataset

We established 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 target cultures. These articles can enrich the numeric survey data with a qualitative background.[2]

NormAd. NormAd (Rao et al., 2025) offers a structured collection of cultural norms and situational examples, demonstrating how abstract values are operationalized 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.

3.3 Evaluation Dataset

We use two sets 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 or Appendix F.2 for the complete list, which consists of 59 datasets). The test data contains a total of 68607 multilingual, culturally sensitive texts annotated for offensiveness.

MMLU. To evaluate the model's general knowledge retention capabilities after cultural adaptation, 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 the cultural adapter alters 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 train LoRA adapters on each model with rank 64, a batch size of 32, a learning rate of 2×10^{-4} with a dropout rate of 0.1, for 6 epochs. All adapters were trained on a single 80GB A100 GPU with bf16 and gradient checkpointing. Other details on training are in Appendix B.

We evaluate three popular open-source models, each with instruction-tuned variants, to ensure that the results generalize across distinct pretraining

[2] See Table 17 for the Wikipedia pages used.

corpora while remaining reproducible (Dubey et al., 2024; Rivière et al., 2024; Team, 2024). We focus on 7-9B parameter models, to ensure the LoRA tuning fits on a single high-memory GPU.

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. Our 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 (also referred to as cultural-task F1) as the primary metric for evaluating classification performance on both probability and freeform-based evaluations.

We propose a *cultural distinctiveness* metric, **C-DIST** score, to further quantify a model's ability to preserve cultural distinctiveness. 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:

$$D = \frac{1}{n} \sum_{i=1}^{n} \vec{n}_i. \tag{1}$$

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-DIST score of 1.0. A lower score suggests interference or homogenization, as illustrated in Figure 2. A diagonal-dominant matrix reflects strong cultural specialization — meaning each culture-specific model performs best on evaluations of its own culture — indicating minimal interference between cultural adaptations. This metric thus quantifies the extent to which each model preserves distinct cultural representations after adaptation.

[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 $M_{i,i}$

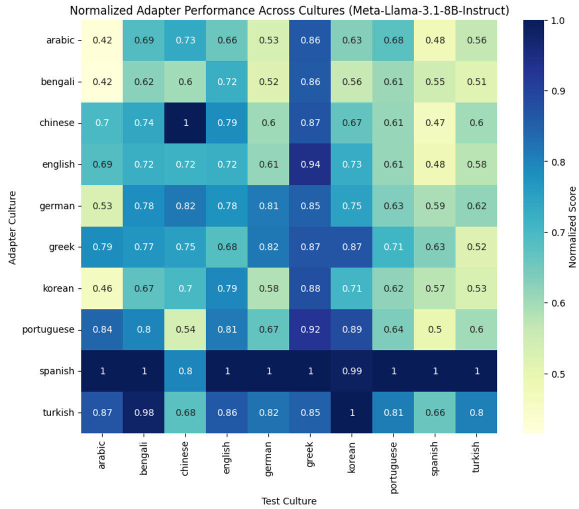


Figure 2: Single-culture adaptation using WVS data with Llama-3.1-8B-IT, evaluating cross-cultural offensiveness classification tasks. Minimal diagonal pattern is observed in this setting, with a C-DIST score of **0.76**.

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. Table 2 compares the approaches for downstream tasks using Llama-3.1-8B models: (i) zero-shot prompting, (ii) single-culture adaptation. Our results show that training using WVS is more effective in improving downstream tasks for the base model when using the single-culture adaptation strategy. Particularly, WVS training is beneficial for underrepresented cultures such as `ara` and `kor` for the base model. Surprisingly, this positive effect is not seen in the instruction-tuned model, which instead shows a decline in 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

[4]An invalid response contains nonsensical outputs, fails to follow instructions or lacks a meaningful or relevant answer to the prompt. Appendix 14 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 Ratio (%)
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
Gemma	Zero-Shot	11.75
	Single-Culture-WVS	0
Qwen	Zero-Shot	6.8
	Single-Culture-WVS	0

Table 3: Comparison of invalid response ratio across different models and scenarios. The Invalid Ratio represents the percentage of responses flagged as invalid across all test sets. We provide the complete invalid ratio table in Appendix C.2.

Appendix D.3). Compared to zero-shot prompting, both the base model and instruction-tuned model have significantly improved invalid response ratios after adaptation. This suggests that WVS fine-tuning enhances the model's general instruction-following ability but does not necessarily improve its understanding of cultural values.

The high zero-shot invalid response ratio in models shows that achieving strong performance on relevant tasks requires improvements in *both* instruction-following ability and cultural value understanding.

4.2 Observed Cultural Interference Across Models

To further investigate the effect of adaptation, we examine the single-culture adaptation results in a cross-cultural setting (i.e., training on one culture and evaluating on others). Ideally, performance should be highest when the adaptation matches the test culture, forming a diagonal pattern in a heatmap of cross-cultural evaluations. However, as shown in Figure 2, no such diagonal is observed for the instruction-tuned Llama model (with a similar pattern seen for the base model in Figure 10 in the Appendix). The cross-cultural improvements show no clear trends, and all adapters enhance performance on the Spanish test data in Figure 2. The C-DIST 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*).

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 %). **Std.**: Standard, using English WVS data. **Transl.**: Translated, using WVS data in each culture's respective language. 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.

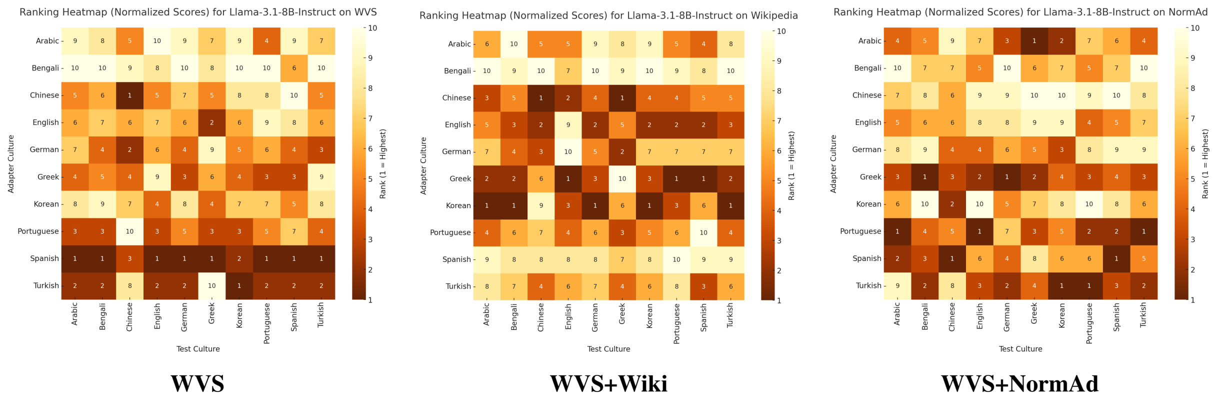


Figure 3: Heatmaps of culture-specific classification performance (Llama-3.1-8B-IT) based on the ranks of the adaptation results. Darker diagonal elements indicate stronger cultural distinctiveness and better C-DIST scores.

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 in the results show that adapting to WVS data can change factual knowledge accuracy, depending on language and dataset characteristics. Furthermore, the inconsistencies in probability-based scoring (Appendix 13) also strengthen the observation of *factual knowledge interference*. This underscores the challenge of balancing cultural distinctiveness with factual integrity with the appropriate training data.

5 Adaptation with Additional Narratives

While WVS-based training provides a useful starting point for cultural value adaptation, our results in §4 show that it rarely produces strong diagonal patterns, suggesting limited cultural specialization. This raises a key question: *what types of additional data can strengthen adaptation while preserving cultural distinctiveness?*

Social psychology studies highlight a longstanding gap between what people *think* and how they *behave* in practice (Gross and Niman, 1975; Fazio, 1981, inter alia). This raises concern about the sufficiency of self-reported values such as those in WVS for improving tasks that depend on culturally grounded behavior (§4.1). To bridge this gap, we supplement WVS with two complementary narrative sources: Wikipedia and NormAd. Unlike survey responses, these provide a more objective, context-rich source of cultural norms, institutions, and everyday practices. We hypothesize that incorporating such *narratives of culture* helps models learn both abstract values and their behavior.

Here, we focus our evaluation on instruction-tuned models to better reflect real-world use and extend it beyond Llama to include Gemma and Qwen models, demonstrating the generality.

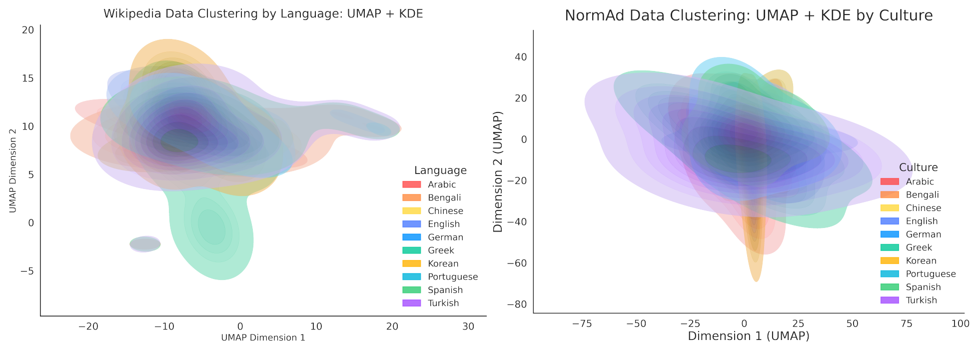
5.1 Results and Discussions

Intuitively, beyond surveys, additional sources enrich the training data. Wikipedia offers encyclopedic knowledge about institutions, historical contexts, and everyday practices, providing descriptive and factual information. In contrast, NormAd supplies procedural, scenario-based norms, including roles and obligations expressed in concrete situations, directly reflecting how abstract cultural rules are applied in everyday settings. The adaptation results are presented as follows.

Improved C-DIST with NormAd. Empirically (Table 5), the addition of WVS and NormAd (*WVS+NormAd*) often enhances cultural distinctiveness, with improvements statistically significant in most cases. Table 5 shows that integrating these datasets consistently improves C-DIST scores across all three models, indicating more culturally distinct behavior. For instance, Llama-3.1-8B-IT's C-DIST improves from 0.76 (WVS-only) to 0.89 (WVS+NormAd). Figure 3 further illustrates this

Model	Data	C-DIST	F1 Cult. (%)	F1 MMLU (%)
Llama-3.1-8B-IT	WVS	0.76 ± 0.012	29.61 ± 1.91	42.78 ± 0.84
	Wiki	0.81 ± 0.009	35.39 ± 2.77	26.33 ± 2.05
	NormAd	0.85 ± 0.014	38.42 ± 2.92	19.63 ± 1.38
	WVS+Wiki	0.78 ± 0.010	31.19 ± 1.80	49.02 ± 0.88
	WVS+NormAd	$\mathbf{0.89 \pm 0.011}$	$\mathbf{40.94 \pm 0.98}$	50.43 ± 0.93
	WVS+Wiki+NormAd	0.76 ± 0.013	38.21 ± 0.85	$\mathbf{52.61 \pm 0.97}$
Gemma-2-9B-IT	WVS	0.81 ± 0.011	39.22 ± 1.66	45.31 ± 0.78
	Wiki	0.83 ± 0.010	36.67 ± 2.61	8.23 ± 1.92
	NormAd	0.79 ± 0.013	37.10 ± 2.64	8.07 ± 1.96
	WVS+Wiki	0.80 ± 0.012	37.25 ± 1.60	47.05 ± 0.82
	WVS+NormAd	$\mathbf{0.83 \pm 0.012}$	$\mathbf{40.01 \pm 0.69}$	55.19 ± 0.86
	WVS+Wiki+NormAd	0.73 ± 0.018	37.90 ± 0.83	$\mathbf{64.94 \pm 1.01}$
Qwen2.5-7B-IT	WVS	0.92 ± 0.006	48.05 ± 1.41	68.32 ± 0.56
	Wiki	0.89 ± 0.008	44.21 ± 2.38	58.32 ± 1.62
	NormAd	0.91 ± 0.007	48.31 ± 2.42	65.57 ± 0.58
	WVS+Wiki	0.90 ± 0.007	46.00 ± 1.39	68.22 ± 1.57
	WVS+NormAd	$\mathbf{0.94 \pm 0.006}$	47.67 ± 0.40	67.51 ± 0.57
	WVS+Wiki+NormAd	0.86 ± 0.010	44.13 ± 0.37	67.33 ± 0.98

Table 5: Mean ± standard deviation across 3 runs with different seeds. **NormAd** drives the largest gains in cultural distinctiveness (C-DIST) and cultural-task F1, especially when paired with WVS (Llama/Gemma). **Wikipedia** alone yields smaller C-DIST gains, but, when combined with WVS, stabilizes or improves MMLU (knowledge retention). Variability (sd) is modest for C-DIST and F1, and slightly higher for narrative-only MMLU rows, consistent with greater instability when training without WVS grounding.



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.

shift, as the heatmaps become more diagonal and show reduced cross-cultural interference. *Incorporating additional cultural narratives retains cultural distinctiveness.*

The addition of NormAd data (*WVS+NormAd*) also leads to notable gains in offensive classifications compared to training with WVS data alone. For instance, Llama-3.1-8B-IT's performance on the 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. However, Gemma-2-9B-IT and Qwen2.5-7B-IT see a marginal change in F1 Cult., when WVS is augmented with NormAd. This highlights that while Llama-3.1-8B-IT showed clear benefits on this downstream task from narrative augmentation, the effect on tasks is model-dependent.

Role of Wikipedia and MMLU. Adding Wikipedia to WVS (*WVS+Wiki*) generally preserves or improves MMLU performance compared to training on WVS alone. Combining all three sources (*WVS+Wiki+NormAd*) yields the best MMLU results for the Llama-3.1-8B-IT and Gemma-2-9B-IT models. However, our results show anomalies, indicating the ongoing challenge of achieving robust cultural adaptation without compromising general knowledge retention.

Complementary roles of Wikipedia and NormAd. The general trend indicates that context-rich data, *when added to WVS*, effectively helps offset the knowledge interference introduced by survey data alone. In particular, this is potentially due to NormAd containing concrete, situation-based rules, so the model learns how to apply norms in decisions. Wikipedia provides a broad factual context that stabilizes general knowledge and offsets drift from survey-only training (WVS). As a result, narrative-only fine-tuning can make knowledge less stable (Appendix C.2), and WVS-only adapters tend to blur across cultures (§4.2). Mixing *WVS+NormAd* restores cultural separation, and adding Wikipedia yields the most reliable balance overall.

Overall, our findings suggest that *curated narratives are crucial for retaining the model's foundational understanding of cultural knowledge during adaptation.*

6 Further Analysis

Our empirical results suggest that adding objective cultural descriptions and context-specific examples improves cultural distinctiveness and performance on downstream tasks. We also found that pairing WVS with scenario-based narratives (WVS+NormAd; optionally +Wiki) is a strong default when aiming for culturally distinct behavior without sacrificing factual knowledge. 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 differences after training.

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-DIST scores.

6.1 Summary of Findings

Based on the analysis above, here we summarize our findings. Fine-tuning solely on WVS data shows clear limitations for effective cultural value adaptation, as evidenced by consistently reduced C-DIST scores, variable downstream task performance, and diminished factual knowledge retention. While overall performance may vary across tasks, augmenting survey data with more descriptive sources enables a model to *retain cultural distinctiveness* and *retain factual knowledge* better. Combining WVS survey data with NormAd situational norms consistently yields clearer cultural separation, as evidenced by improved C-DIST score (Table 5). Wikipedia data offers moderate gains through structured knowledge, but NormAd's scenario-based behavioral cues drive stronger cultural differentiation when paired with WVS.

Our findings suggest that combining scenario-based narratives (e.g., NormAd) with survey patterns (WVS) better preserves cultural distinctiveness and should be investigated further.

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., 2025; 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., 2025; Myung et al., 2025). Particularly, Jiang et al. (2025) 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). Similarly, Beck et al. (2024) introduced sensitivity tests to probe how models respond to culturally charged questions, revealing both strengths and blind spots. These evaluation-oriented studies complement adaptation methods by identifying where misalignments occur and motivating the development of more nuanced cultural adaptation strategies. While related to our work, we focus 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., 2025; Sorensen et al., 2022), which often prioritize Western or English-centric perspectives, leading to misalignment when applied to non-WEIRD cultures (Mihalcea et al., 2025). 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 investigated the limitations of using World Values Survey (WVS) data for cultural value adaptation in LLMs and explored the potential of augmenting it with scenario-based cultural narratives. Our findings reveal that relying solely on WVS can lead to homogenized cultural representations and interfere with factual knowledge. We demonstrate that incorporating encyclopedic (Wikipedia) and scenario-based (NormAd) narratives, particularly the latter, effectively enhances the cultural distinctiveness of adapted models.

While some variations in results were observed, we found that the augmentation could still improve nuanced cultural representations and preserve factual knowledge in models. Our findings reveal a complex trade-off between cultural distinctiveness, task performance, and knowledge retention, highlighting the need for further research on optimal data combinations and adaptation strategies to balance these competing objectives.

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.

In our experiments, the training and evaluation sets are imbalanced across cultures and languages (see Appendix F for details). To minimize confounding effects, we take several measures, such as reporting macro-averaged results and normalizing C-DIST scores. While our setup is intended to reflect realistic resource disparities, the results may still correlate with data availability.

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.

Acknowledgment

This work has been supported by the LOEWE Distinguished Chair "Ubiquitous Knowledge Processing", LOEWE initiative, Hesse, Germany (Grant Number: LOEWE/4a//519/05/00.002(0002)/81) and the German Federal Ministry of Research, Technology and Space and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE. We thank Yongxin Huang, German Ortiz for their feedback on a draft of this paper.

References

2019. Turkish Spam V01. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5WG7F.

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.

aimansnigdha. 2018. Bangla-abusive-comment-dataset. https://github.com/aimansnigdha/Bangla-Abusive-Comment-Dataset.

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.

Miguel Ángel Álvarez-Carmona, Estefanía Guzmán-Falcón, Manuel Montes-y-Gómez, Hugo Jair Escalante, Luis Villaseñor Pineda, Verónica Reyes-Meza, and Antonio Rico Sulayes. 2018. Overview of MEX-A3T at ibereval 2018: Authorship and aggressiveness analysis in mexican spanish tweets. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018*, volume 2150 of *CEUR Workshop Proceedings*, pages 74–96. CEUR-WS.org.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. 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.

Shiladitya Bhattacharya, Siddharth Singh, Ritesh Kumar, Akanksha Bansal, Akash Bhagat, Yogesh Dawer, Bornini Lahiri, and Atul Kr. Ojha. 2020. Developing a multilingual annotated corpus of misogyny and aggression. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 158–168, Marseille, France. European Language Resources Association (ELRA).

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.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. 2020. I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6193–6202, Marseille, France. European Language Resources Association.

Çağrı Çöltekin. 2020. A corpus of turkish offensive language on social media. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6174–6184, Marseille, France.

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. *arXiv preprint arXiv:2408.16482*.

Shammur Absar Chowdhury, Hamdy Mubarak, Ahmed Abdelali, Soon-gyo Jung, Bernard J. Jansen, and Joni Salminen. 2020. A multi-platform Arabic news comment dataset for offensive language detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6203–6212, Marseille, France. European Language Resources Association.

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.

daanVeer. 2020. Korean hatespeech dataset. https://github.com/daanVeer/HateSpeech_dataset.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pages 512–515.

Angel Felipe Magnossão de Paula and Ipek Baris Schlicht. 2021. AI-UPV at iberlef-2021 DETOXIS task: Toxicity detection in immigration-related web news comments using transformers and statistical models. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021), XXXVII International Conference of the Spanish Society for Natural Language Processing., Málaga, Spain, September, 2021*, volume 2943 of *CEUR Workshop Proceedings*, pages 547–566. CEUR-WS.org.

Rogers P. de Pelle and Viviane P. Moreira. 2017. Offensive comments in the brazilian web: a dataset and baseline results.

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. *arXiv preprint arXiv: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.

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018. Overview of the task on automatic misogyny identification at ibereval 2018. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018*, volume 2150 of *CEUR Workshop Proceedings*, pages 214–228. CEUR-WS.org.

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. *arXiv preprint arXiv: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.

HASOC. 2020. Hasoc2020. https://hasocfire.github.io/hasoc/2020/index.html.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

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*.

Fatemah Husain. 2020. OSACT4 shared task on offensive language detection: Intensive preprocessing-based approach. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 53–60, Marseille, France. European Language Resource Association.

Liwei Jiang, Taylor Sorensen, Sydney Levine, and Yejin Choi. 2025. Can language models reason about individualistic human values and preferences? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6757–6794, Vienna, Austria. Association for Computational Linguistics.

Zhuoren Jiang, Zhe Gao, Guoxiu He, Yangyang Kang, Changlong Sun, Qiong Zhang, Luo Si, and Xiaozhong Liu. 2019. Detect camouflaged spam content via StoneSkipping: Graph and text joint embedding for Chinese character variation representation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6187–6196, Hong Kong, China. Association for Computational Linguistics.

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 arXiv:2203.07785*.

Sanaa Kaddoura and Safaa Henno. 2024. Dataset of arabic spam and ham tweets. *Data in Brief*, 52(10990):4.

Kaggle. 2019. Jigsaw-multilingual-toxicity. https://www.kaggle.com/code/tarunpaparaju/jigsaw-multilingual-toxicity-eda-models.

Kaggle. 2021. 5k turkish tweets with incivil content. https://www.kaggle.com/datasets/kbulutozler/5k-turkish-tweets-with-incivil-content.

Kaggle. 2022. turkish offensive language detection. https://www.kaggle.com/datasets/toygarr/turkish-offensive-language-detection.

Habibe Karayiğit, Çiğdem İnan Acı, and Ali Akdağlı. 2021. Detecting abusive instagram comments in turkish using convolutional neural network and machine learning methods. *Expert Systems with Applications*, 174:114802.

Jean Lee, Taejun Lim, Heejun Lee, Bogeun Jo, Yangsok Kim, Heegeun Yoon, and Soyeon Caren Han. 2022. K-MHaS: A multi-label hate speech detection dataset in Korean online news comment. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3530–3538, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

João Augusto Leite, Diego Silva, Kalina Bontcheva, and Carolina Scarton. 2020. Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 914–924, Suzhou, China. 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 *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. 2024b. CulturePark: Boosting cross-cultural understanding in large language models. In *Advances in Neural Information*

Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2025. Culturally aware and adapted NLP: A taxonomy and a survey of the state of the art. *Transactions of the Association for Computational Linguistics*, 13:652–689.

Rada Mihalcea, Oana Ignat, Longju Bai, Angana Borah, Luis Chiruzzo, Zhijing Jin, Claude Kwizera, Joan Nwatu, Soujanya Poria, and Thamar Solorio. 2025. Why ai is weird and shouldn't be this way: towards ai for everyone, with everyone, by everyone. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press.

Jihyung Moon, Won Ik Cho, and Junbum Lee. 2020. BEEP! Korean corpus of online news comments for toxic speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 25–31, Online. Association for Computational Linguistics.

Hamdy Mubarak, Hend Al-Khalifa, and Abdulmohsen Al-Thubaity. 2022. Overview of OSACT5 shared task on Arabic offensive language and hate speech detection. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, pages 162–166, Marseille, France. European Language Resources Association.

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.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, Nedjma Ousidhoum, Jose Camacho-Collados, and Alice Oh. 2025. BLEnD: a benchmark for llms on everyday knowledge in diverse cultures and languages. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA. Curran Associates Inc.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4675–4684, Hong Kong, China. Association for Computational Linguistics.

Juan Carlos Pereira-Kohatsu, Lara Quijano-Sánchez, Federico Liberatore, and Miguel Camacho-Collados. 2019. Detecting and monitoring hate speech in twitter. *Sensors*, 19(21):4654.

Flor Miriam Plaza-del Arco, Arturo Montejo-Ráez, L. Alfonso Ureña-López, and María-Teresa Martín-Valdivia. 2021. OffendES: A new corpus in Spanish for offensive language research. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1096–1108, Held Online. INCOMA Ltd.

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 Sukumar Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2025. NormAd: A framework for measuring the cultural adaptability of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2373–2403, Albuquerque, New Mexico. Association for Computational Linguistics.

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. *arXiv preprint arXiv:2408.00118*.

Nauros Romim, Mosahed Ahmed, Hriteshwar Talukder, and Md Saiful Islam. 2021. Hate speech detection in the bengali language: A dataset and its baseline evaluation. In *Proceedings of International Joint Conference on Advances in Computational Intelligence: IJCACI 2020*, pages 457–468. Springer.

Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov. 2021. SOLID: A large-scale semi-supervised dataset for offensive language identification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 915–928, Online. Association for Computational Linguistics.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2016. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*, volume 17 of *Bochumer Linguistische Arbeitsberichte*, pages 6–9, Bochum.

Paul Röttger, Haitham Seelawi, Debora Nozza, Zeerak Talat, and Bertie Vidgen. 2022. Multilingual HateCheck: Functional tests for multilingual hate speech detection models. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 154–169, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Omar Sharif and Mohammed Moshiul Hoque. 2022. Tackling cyber-aggression: Identification and fine-grained categorization of aggressive texts on social media using weighted ensemble of transformers. *Neurocomputing*, 490:462–481.

Hoyun Song, Soo Hyun Ryu, Huije Lee, and Jong Park. 2021. A large-scale comprehensive abusiveness detection dataset with multifaceted labels from Reddit. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 552–561, Online. Association for Computational Linguistics.

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. *arXiv preprint arXiv:2302.13971*.

Francielle Vargas, Isabelle Carvalho, Fabiana Rodrigues de Góes, Thiago Pardo, and Fabrício Benevenuto. 2022. HateBR: A large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7174–7183, Marseille, France. European Language Resources Association.

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.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the germeval 2018 shared task on the identification of offensive language.

Guohai Xu, Jiayi Liu, Ming Yan, Haotian Xu, Jinghui Si, Zhuoran Zhou, Peng Yi, Xing Gao, Jitao Sang, Rong Zhang, Ji Zhang, Chao Peng, Fei Huang, and Jingren Zhou. 2023. Cvalues: Measuring the values of chinese large language models from safety to responsibility. *arXiv preprint arXiv:2307.09705*.

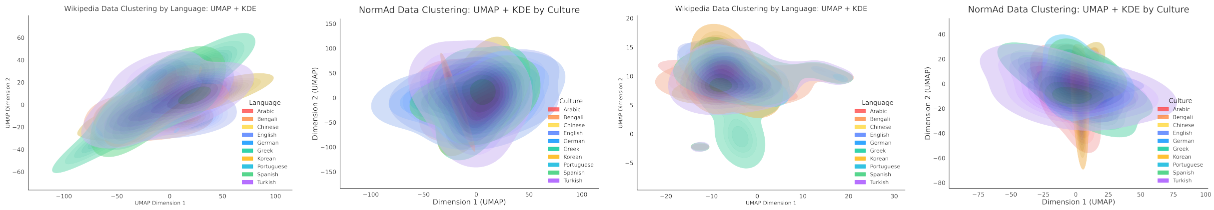
Shaoyang Xu, Yongqi Leng, Linhao Yu, and Deyi Xiong. 2025. Self-pluralising culture alignment for large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6859–6877, Albuquerque, New Mexico. Association for Computational Linguistics.

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

Jingyan Zhou, Jiawen Deng, Fei Mi, Yitong Li, Yasheng Wang, Minlie Huang, Xin Jiang, Qun Liu, and Helen Meng. 2022. Towards identifying social bias in dialog systems: Framework, dataset, and benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3576–3591, Abu Dhabi, United Arab Emirates. Association 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 Zero-Shot Prompting and Single Culture Adaptation Results

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
Gemma-2-9b-IT	17.98	50.65	20.30	46.30	50.18	45.94	60.40	38.80	27.40	46.35	40.43
Qwen2.5-7B-Instruct	45.41	58.88	25.30	38.29	60.30	48.27	53.86	54.87	45.72	60.37	49.13
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
Gemma-2-9b-IT	15.54	43.95	24.10	33.92	41.01	49.09	61.01	37.66	37.15	48.81	39.22
Qwen2.5-7B-Instruct	39.30	59.24	25.78	40.39	57.85	48.02	53.79	51.77	51.31	57.47	48.49

Table 7: Culture adaptation results (F1 scores) under three training scenarios: zero-shot prompting and single-culture adaptation (training on Llama-3.1-8B models using WVS data). Evaluation uses a multilingual offensiveness dataset (§3.3), reported as averaged F1 scores.

C.2 Full Invalid Ratio

Methods		Inv. Cult. (%)	Inv. MMLU (%)
Llama-3.1-8B	Zero-Shot	20.12	2.3
	WVS	14.68	0
	NormAd	15.90	**70.0**
	WVS+Wiki	14.04	0
	WVS+NormAd	13.22	0
	WVS+Wiki+NormAd	12.85	0
Llama-3.1-8B-IT	Zero-Shot	21.20	0
	WVS	10.82	0
	NormAd	11.73	**72.3**
	WVS+Wiki	9.73	0
	WVS+NormAd	8.91	0
	WVS+Wiki+NormAd	8.35	0
Gemma-2-9B-IT	Zero-Shot	13.23	0
	WVS	0	0
	NormAd	9.7	**82.7**
	WVS+Wiki	6.32	0
	WVS+NormAd	5.89	0
	WVS+Wiki+NormAd	6.21	0
Qwen2.5-7B-IT	Zero-Shot	9.4	0
	WVS	0	0
	NormAd	7.5	**10.1**
	WVS+Wiki	0	0
	WVS+NormAd	0	0
	WVS+Wiki+NormAd	0	0

Table 8: Invalid response rates on cultural evaluation sets (*Invalid Cult.*) and on MMLU (*Invalid MMLU*). All MMLU invalid ratios are lower than the 20.12 % cultural baseline of Llama-3.1-8B—*except* for the purposely inflated **NormAd**-only rows, which remain dramatically worse.

C.3 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 9: Results for Combined-Culture Adaptation on WVS.

C.4 Freeform Generation

C.4.1 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.

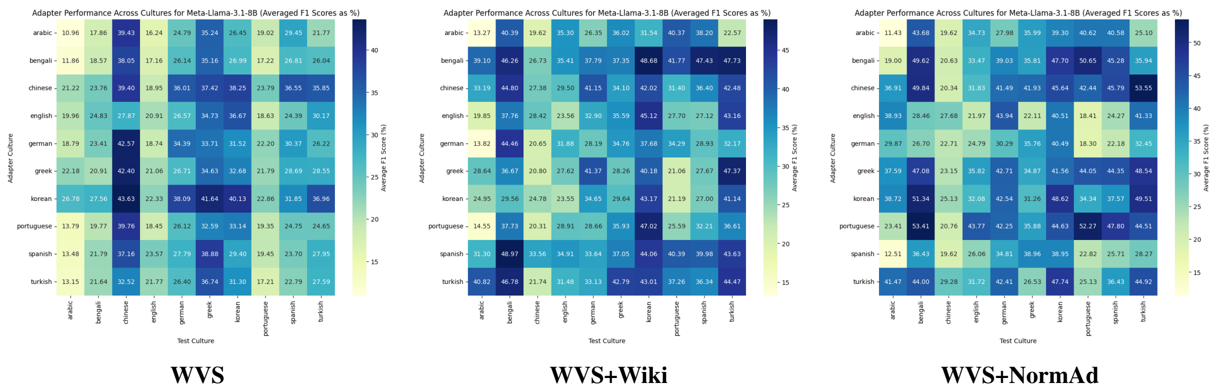


Figure 6: Heatmaps of culture-specific classification performance (Llama-3.1-8B) using different data sources based on the ranks of the adaptation results.

C.4.2 Performance Tables - Llama-3.1-8B-Instruct

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

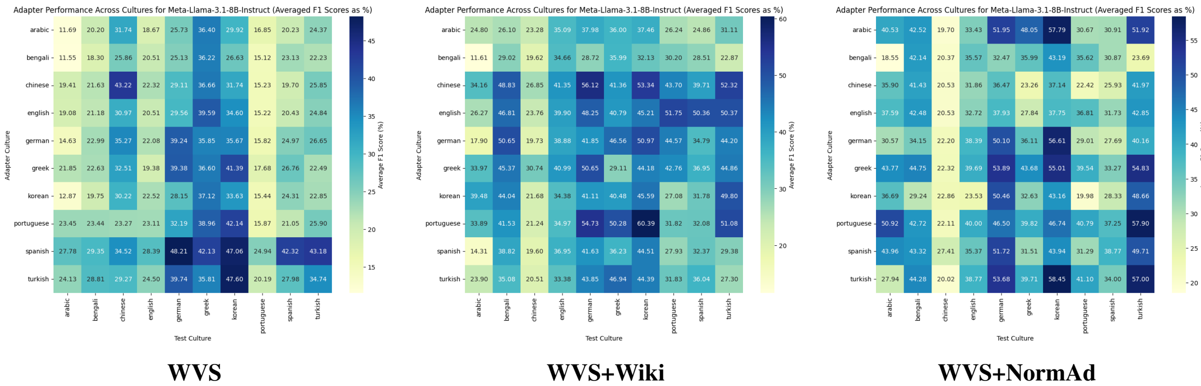


Figure 7: Heatmaps of culture-specific classification performance (Llama-3.1-8B-IT) using different data sources based on the ranks of the adaptation results.

C.4.3 Performance Tables - Qwen2.5-7B-IT

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

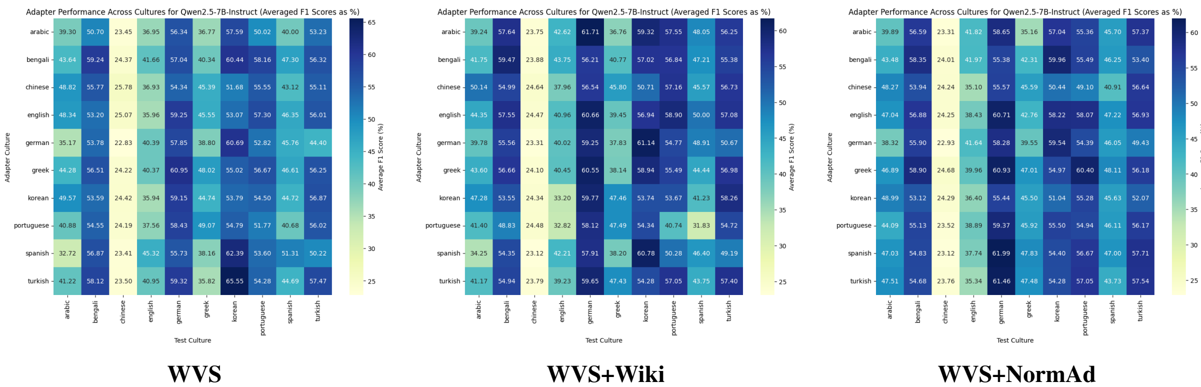


Figure 8: Heatmaps of culture-specific classification performance (Qwen2.5-7B-IT) using different data sources based on the ranks of the adaptation results.

C.4.4 Performance Tables - Gemma-2-9B-IT

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

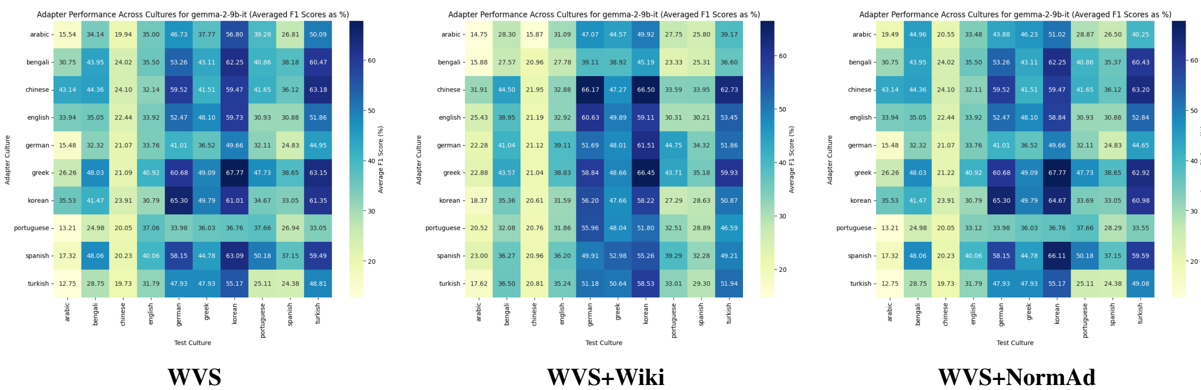


Figure 9: Heatmaps of culture-specific classification performance (Gemma-2-9B-Instruct) using different data sources based on the ranks of the adaptation results.

C.5 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 10: Normalized Scores and C-DIST on Llama-3.1-8B-IT for WVS. Rows represent the adapter culture, and columns represent the 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 11: Normalized Scores and C-DIST on Llama-3.1-8B-IT for WVS+Wikipedia. Rows represent the adapter culture, and columns represent the 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 12: Normalized Scores and C-DIST on Llama-3.1-8B-IT for WVS+NormAd. Rows represent the adapter culture, and columns represent the culture test set.

C.6 Probability-Based Generation

Table 13 shows the 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 13: 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 14 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.

<table>
<tr><th>Culture</th><th>Invalid Response Example</th></tr>
<tr><td>English</td><td>{"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}</td></tr>
<tr><td>German</td><td>{"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}</td></tr>
<tr><td>Spanish</td><td>{"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}</td></tr>
</table>

Table 14: Examples of invalid responses.

D.3 Invalid Ratio

Table 15 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
	Zero-shot	222.5849	1106	21.77%
Llama-3.1-8B	Single-Culture-WVS	143.6604	1106	14.06%
	Single-Culture-WVS+Wiki	133.6604	1106	12.97%
	Zero-shot	234.5660	1106	22.55%
Llama-3.1-8B-IT	Single-Culture-WVS	134.3774	1106	13.16%
	Single-Culture-WVS+Wiki	121.5849	1106	12.01%
	Zero-shot	11,797	58,638	20.12%
Llama-3.1-8B	Single-Culture-WVS	94,663	645,018	14.68%
	Single-Culture-WVS+Wiki	90,562	645,018	14.04%
	Zero-shot	12,432	58,638	21.20%
Llama-3.1-8B-IT	Single-Culture-WVS	69,804	645,018	10.82%
	Single-Culture-WVS+Wiki	71,019	645,018	11.01%

Table 15: Invalid count and ratio for different models and scenarios.

E Prompts

Table 16 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 16: Overview of prompts generated by getPrompt.

F Data Statistics

F.1 Training Data Statistics

Table 17 lists the data sources and URLs utilized in our experiments, encompassing the World Values Survey (WVS), Wikipedia cultural articles, and the NormAd dataset. Tables 18 and 19 provide detailed summary statistics for the Wikipedia and NormAd datasets respectively, outlining the total number of sentences, samples, and tokens per language.

Note that our training and evaluation sets are imbalanced across cultures and languages (Tables 18, 19, and 20). To minimize confounding, we (i) include per-culture scores alongside macro-averages, (ii) use the C-DIST normalization that compares each adapter against the best performer for each test culture (column-wise), and (iii) repeat all runs with three seeds and report mean $\pm$ sd. However, we avoid re-balancing due to realistic resource disparities and to preserve culture-specific signals. Hence, our results in §5 should be interpreted as gains achieved under realistic resource disparities, which may still correlate with data availability.

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 17: Data sources and URLs.

Language	Total Sentences	Total Tokens (Entire Text)	Total Tokens (Summed per Sentence)
Arabic	257	8,990	9,018
Bengali	127	4,282	4,307
Chinese	388	13,929	13,938
English	434	15,632	15,688
German	171	6,322	6,338
Greek	250	11,806	11,825
Korean	150	5,678	5,687
Portuguese	186	10,286	10,298
Spanish	76	3,662	3,666
Turkish	143	6,573	6,581

Table 18: Summary statistics for each language in our Wikipedia training dataset.

Language	Samples	Tokens
Arabic	239	102,705
Spanish	234	74,674
Chinese	134	35,988
English	209	82,144
Korean	27	6,784
German	76	21,209
Bengali	33	7,659
Portuguese	77	19,022
Greek	69	23,961
Turkish	35	15,391

Table 19: Summary statistics for each language in our NormAd training dataset.

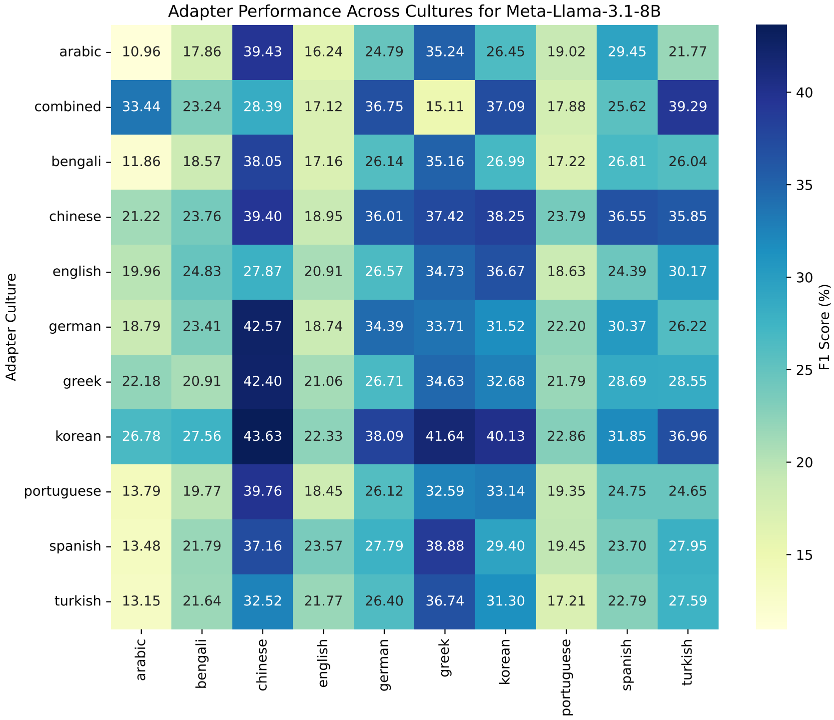
F.2 Test Data Statistics

Following Li et al. (2024a), we break down our culture test set in the table below.

Culture	Country & Territory	Task & Dataset	#Sample
Arabic (METHOD-Ar)	Middle East	*Offensive language detection:* OffensEval2020(2000) (Zampieri et al., 2020), OSACT4(1000) (Husain, 2020), Multi-Platform(1000) (Chowdhury et al., 2020), and OSACT5(2541) (Mubarak et al., 2022). *Hate detection:* OSACT4(1000) (Husain, 2020), Multi-Platform(675) (Chowdhury et al., 2020), OSACT5(2541) (Mubarak et al., 2022), and OSACT5_finegrained(2541) (Mubarak et al., 2022). *Spam detection:* ASHT(1000) (Kaddoura and Henno, 2024). *Vulgar detection:* Multi-Platform(675) (Chowdhury et al., 2020)	14,973
Bangli (METHOD-Bn)	Bangladesh	*Offensive language detection:* TRAC2020 Task1(1000) (Bhattacharya et al., 2020), TRAC2020 Task2(1000) (Bhattacharya et al., 2020), BAD(1000) (Sharif and Hoque, 2022). *Hate detection:* Hate Speech(1000) (Romim et al., 2021). *Threat detection:* BACD(1000) (aimansnigdha, 2018). *Bias detection:* BACD(1000) (aimansnigdha, 2018).	6,000
Chinese (METHOD-Zh)	China	*Spam detection:* CCS(1000) (Jiang et al., 2019). *Bias detection:* CDial-Bias(1000) (Zhou et al., 2022). *Stance detection:* CValues(1712) (Xu et al., 2023).	3,712
English (METHOD-En)	United States	*Offensive language detection:* SOLID(1000) (Rosenthal et al., 2021). *Hate detection:* MLMA(1000) (Ousidhoum et al., 2019) and HOF(1000) (Davidson et al., 2017). *Threat detection:* CValuesJMT(1000) (Kaggle, 2019). *Toxicity detection:* MLMA(1000) (Ousidhoum et al., 2019) and JMT(1000) (Kaggle, 2019).	6,000
German (METHOD-De)	Germany and parts of Europe	*Offensive language detection:* GermEval2018(3531) (Wiegand et al., 2018). *Hate detection:* IWG_1(469) (Ross et al., 2016), IWG_2(469) (Ross et al., 2016), HASOC2020(850) (HASOC, 2020), and multilingual-hatecheck(1000) (Röttger et al., 2022).	6,319
Korean (METHOD-Ko)	South Korea	*Hate detection:* K-MHaS(1000) (Lee et al., 2022), hateSpeech(1000) (Moon et al., 2020), and HateSpeech2(1000) (daanVeer, 2020). *Abusive detection:* AbuseEval(1000) (Caselli et al., 2020), CADD(1000) (Song et al., 2021), and Waseem(1000) (Waseem and Hovy, 2016).	5,000
Portuguese (METHOD-Pt)	Brazil and parts of Latin America	*Offensive language detection:* OffComBR(1250) (de Pelle and Moreira, 2017), and HateBR(1000) (Vargas et al., 2022). *Bias detection:* ToLD-Br-homophobia(1000) (Leite et al., 2020), and ToLD-Br-misogyny(1000) (Leite et al., 2020). *Abusive detection:* ToLD-Br-insult(1000) (Leite et al., 2020).	16,250
Spanish (METHOD-Es)	Argentina, Mexico, and parts of Latin America	*Offensive language detection:* AMI(1000) (Fersini et al., 2018), MEX-A3T(1000) (Álvarez-Carmona et al., 2018), and OffendES(1000) (Plaza-del Arco et al., 2021). *Hate detection:* HatEval 2019(1000) (Basile et al., 2019), and HaterNet(1000) (Pereira-Kohatsu et al., 2019). *Bias detection:* DETOXIS_stereotype(1000) (de Paula and Schlicht, 2021), and DETOXIS_improper(1000) (de Paula and Schlicht, 2021). *Abusive detection:* DETOXIS_abusive(1000) (de Paula and Schlicht, 2021), DETOXIS_mockery(1000) (de Paula and Schlicht, 2021). *Aggressiveness detection:* DETOXIS_aggressiveness(1000) (de Paula and Schlicht, 2021). *Stance detection:* DETOXIS_stance(1000) (de Paula and Schlicht, 2021).	11,000
Turkish (METHOD-Tr)	Turkey	*Offensive language detection:* SemEval-2020(3528) (Zampieri et al., 2020), offenseCorpus(1000) (Çöltekin, 2020), offenseKaggle(1000) (Kaggle, 2021), and offenseKaggle_2(1000) (Kaggle, 2022). *Abusive detection:* ATC(1000) (Karayiğit et al., 2021). *Spam detection:* Turkish Spam(825) (mis, 2019). *Fine-grained offensive detection:* offenseCorpus(1000) (Çöltekin, 2020).	10,353

Table 20: Overview of the eight evaluation tasks and the 59 datasets used, including dataset names and their corresponding test sample sizes. For example, "OffensEval2020(2000) (Zampieri et al., 2020)" indicates that the OffensEval2020 dataset contains 2,000 test samples.

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



Adapter Culture	arabic	bengali	chinese	english	german	greek	korean	portuguese	spanish	turkish
arabic	10.96	17.86	39.43	16.24	24.79	35.24	26.45	19.02	29.45	21.77
combined	33.44	23.24	28.39	17.12	36.75	15.11	37.09	17.88	25.62	39.29
bengali	11.86	18.57	38.05	17.16	26.14	35.16	26.99	17.22	26.81	26.04
chinese	21.22	23.76	39.40	18.95	36.01	37.42	38.25	23.79	36.55	35.85
english	19.96	24.83	27.87	20.91	26.57	34.73	36.67	18.63	24.39	30.17
german	18.79	23.41	42.57	18.74	34.39	33.71	31.52	22.20	30.37	26.22
greek	22.18	20.91	42.40	21.06	26.71	34.63	32.68	21.79	28.69	28.55
korean	26.78	27.56	43.63	22.33	38.09	41.64	40.13	22.86	31.85	36.96
portuguese	13.79	19.77	39.76	18.45	26.12	32.59	33.14	19.35	24.75	24.65
spanish	13.48	21.79	37.16	23.57	27.79	38.88	29.40	19.45	23.70	27.95
turkish	13.15	21.64	32.52	21.77	26.40	36.74	31.30	17.21	22.79	27.59

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