

# Cultural Palette: Pluralising Culture Alignment via Multi-agent Palette

Anonymous ACL submission

## Abstract

Large language models (LLMs) face challenges in aligning with diverse cultural values despite their remarkable performance in generation, which stems from inherent monocultural biases and difficulties in capturing nuanced cultural semantics. Existing methods struggle to adapt to unknown culture after fine-tuning. Inspired by cultural geography across five continents, we propose *Cultural Palette*, a multi-agent framework that redefines cultural alignment as an adaptive "color-blending" process for country-specific adaptation. Our approach harnesses cultural geography across five continents through three key steps: First, we synthesize the *Pentachromatic Cultural Palette Dataset* using GPT-4o, refining continental-level dialogues with Hofstede's cultural dimensions to establish foundational cultural representations. Second, five continent-level alignment agents form specialized cultural communities that generate region-specific draft responses. Third, a Meta Agent employs *Cultural MoErges* to dynamically blend these cultural "colors" through attention-gated parameter merging, akin to mixing pigments on a palette, resolving conflicts while preserving cultural nuances to produce the final culturally-aligned response. Extensive experiments across various countries demonstrate that *Cultural Palette* surpasses existing baselines in cultural alignment.

## 1 Introduction

Large Language Models (LLMs), such as GPT-4 (Achiam et al., 2023), have shown remarkable performance in aligning with human values for real-world applications by incorporating diverse perspectives (Han et al., 2024; Sorensen et al., 2024a) and preferences (Ouyang et al., 2022; Chen et al., 2024) through specific fine-tuning and alignment techniques (Hu et al., 2022; Ouyang et al., 2022). However, alignment with human values is predominantly influenced by pluralistic, multicultural factors including traditions, societal norms and more

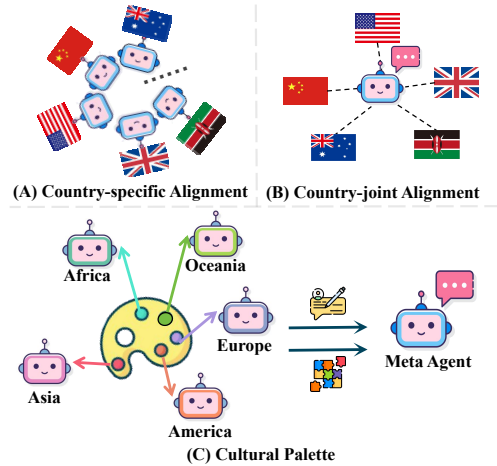


Figure 1: Comparison between (A) country-specific alignment (Li et al., 2024a,b; Feng et al., 2024), (B) country-joint alignment (Li et al., 2024c,a) and (C) Cultural Palette.

(Feng et al., 2024), especially in our multicultural society, necessitating LLMs to perceive cross-cultural differences and reflect pluralistic feedback (Sorensen et al., 2024b; Feng et al., 2024) to avoid mono-cultural bias (Shi et al., 2024) risked by mere RLHF (Ouyang et al., 2022). Early studies primarily involve collecting cultural data corpora for fine-tuning to enhance cultural-awareness in LLMs (Chan et al., 2023; Lin et al., 2024; Kirk et al., 2024), while extremely time-consuming.

Recent research has explored two primary approaches to integrating cultural elements into LLMs: (i) **multi-culture data composition for alignment**, which simulates social surveys to synthesize culture-specific corpora (Li et al., 2024b; Shi et al., 2024), cultural value preference choices (Li et al., 2024a; Xu et al., 2025) seeded from the World Values Survey (WVS) (Haerpfner et al., 2022) or PRISM dataset (Kirk et al., 2024) collecting preferences for LLMs responses across cultural regions for fine-tuning and alignment training, which is limited by underrepresented cultures and may lead to overfitting and value leakage from the WVS

benchmark (Zhou et al., 2023). (ii) **multi-culture community mechanisms**, which leverage pluralistic feedback among multi-agent (Feng et al., 2024; Lake et al., 2024) yet struggles with unseen cultures (Feng et al., 2024). In summary, pluralistic cultural alignment for LLMs encounters three major limitations: (1) Existing alignment techniques like RLHF (Ouyang et al., 2022; Hong et al., 2024) often neglect intersectional identities and marginalized viewpoints between countries instead of among community, limiting the models’ ability to provide culturally inclusive outputs. (2) Existing datasets for pluralistic cultural alignment mainly synthesize multiple-choice questions for cultural preference seeded from WVS benchmark for fine-tuning, lacking cultural dialogues needed for nuanced understanding. (3) Cultural norms evolve over time; however, current methods lack effective paradigms to adapt to this fluidity. Requiring fine-tuning on country-specific corpora (Li et al., 2024a; Xu et al., 2025) to construct cultural community (Sorensen et al., 2024b; Feng et al., 2024) leads to inflexibility and limited practicality (Xu et al., 2025).

To address above limitations, we propose *Cultural Palette*, a framework integrating the core cultural "colors" of five continents <sup>1</sup> (Crang, 2013), collectively forming a palette that symbolizes global cultural diversity via multi-agent coordination grounded in **cultural geography findings that culture differences among global countries are profoundly shaped by geographical influences within cultural evolution and spread across five continents** (Crang, 2013). As depicted in Fig. 1, *Cultural Palette* first generates cultural dialogues based on the *PRISM* dataset (Kirk et al., 2024) for purifying responses representing the cultural characteristics of five continents (Sec. 3) under Hofstede’s cultural dimensions (Insights, 2021) for perceiving culture differences (Hofstede, 2009), conceptualized as primary colors for cultural alignment (Sec. 4.1) and then coloring five continent cultures to form a multi-agent cultural community (Sec. 4.2). Our key contributions are as follows:

- To the best of our knowledge, we introduce the first **Cultural Palette** framework that models cultural alignment as **dynamic blending** via *Cultural MoErges*, outperforming four country-specific and joint alignment methods.
- We release the **Pentachromatic Cultural**

<sup>1</sup>We adopt the five-continent definition from Buschmann and Nolde (2018): Africa, America, Asia, Europe & Oceania.

**Palette Dataset**, a diverse collection of cultural dialogues synthesized using LLMs and refined using Hofstede dimensions to capture cultural semantics across five continents.

- We propose a multi-agent framework that **coordinates five continent-level drafting agents with a meta agent for country-level alignment**. Extensive experiments across 18 countries demonstrate consistent gains over five pluralistic alignment methods.

## 2 Related Work

**Pluralistic Cultural Alignment.** Existing alignment techniques, such as DPO (Rafailov et al., 2024) and ORPO (Hong et al., 2024) guide LLMs to learn from preferred-signal data but do not account for cultural factors. To enhance pluralistic culture alignment in LLMs, existing researches focus on aligning with specific countries or values (Huang et al., 2024; Xu et al., 2025) via fine-tuning (Li et al., 2024a,b; Huang et al., 2024; Xu et al., 2025) and alignment (Kirk et al., 2024; Li et al., 2024d; Guo et al., 2025), yet remains constrained by reliance on inherent cultural values (Feng et al., 2024; Vamplew et al., 2024). To alleviate the above dependencies, recent research leverages LLMs’ zero-shot capabilities (Kojima et al., 2022; Li et al., 2024b) for cultural alignment through multi-agent collaboration, introducing a cultural-aligned judger LLM to summarize community comments via Overton, steerable, and distributional modes (Sorensen et al., 2024b; Feng et al., 2024; Shetty et al., 2025), while challenges persist in interacting with unseen culture (Feng et al., 2024). To further address the complexities of cultural diversity, we propose a novel approach inspired by cultural evolution across five continents (Crang, 2013). By conceptualizing these continental cultures as fundamental colors on a palette, we introduce a multi-agent framework integrated with a *Cultural MoErges* (Goddard et al., 2024) mechanism integrating semantic relationships at both continent and country levels to dynamically blend colors adapting to judging diverse cultural alignment and ensuring nuanced and context-aware response through multi-agent collaboration.

**Data Synthesis for Cultural Alignment.** Aligning LLMs for cultural pluralism—whether through fine-tuning (Li et al., 2024a,b; Shi et al., 2024), alignment (Kirk et al., 2024; Li et al., 2024d; Guo et al., 2025), or agent-based approaches (Sorensen

et al., 2024b; Feng et al., 2024)—requires extensive, culture-specific datasets. While existing datasets like PRISM (Kirk et al., 2024) collect user feedback on LLM responses across diverse countries, they primarily focus on preference ranking rather than generating culturally nuanced dialogues, limiting their utility for training models to understand implicit cultural norms (Chen et al., 2024). Recent studies have demonstrated that self-instruct (Wang et al., 2023) or multi-agent (Li et al., 2024a) can effectively synthesize culturally nuanced data through LLM-driven multi-step generation and refinement, including expanding datasets based on the World Values Survey (WVS) (Haerpfer et al., 2022) to study cultural dominance and alignment, with benchmarks such as CultureLLM (Li et al., 2024a), CulturePark (Li et al., 2024b), and CultureSPA (Xu et al., 2025). However, LLM-driven data synthesis, seeded from the WVS for multiple-choice data pairs, may introduce biases in cultural options and lead to value leakage from the WVS benchmark (Zheng et al., 2023; Zhou et al., 2023). To overcome these limitations, we expanded Prism (Kirk et al., 2024) to cultural dialogues from five continents, creating *Pentachromatic Cultural Palette Dataset* (Sec. 3) through self-feedback contrastive aggregation of cultural differences.

**Model Merging.** LLMs trained on a single corpus often suffer from limited generalization and suboptimal decision-making, as they may converge to local optima that restrict efficiency and versatility across tasks (Lawson and Qureshi, 2024; Yang et al., 2024b). To address this, prior work explores collaborative model merging strategies—such as Task Arithmetic (Ilharco et al., 2023), Ties (Yadav et al., 2024), and Model Stock (Jang et al., 2025)—to improve performance on complex decision-making (Lawson and Qureshi, 2024) and generation tasks (Pourreza et al., 2024). More recently, MoErges (Goddard et al., 2024; Tang et al., 2024; Jiang et al., 2024) unifies model merging with Mixture-of-Experts architectures through an untrainable design, initializing MoE MLP layers with task-specific model weights. Building on these advances, we investigate the role of model merging in culturally grounded decision-making by systematically comparing merging-based methods with cultural joint training approaches (Xu et al., 2025; Li et al., 2024a). Furthermore, we ground our proposed *Cultural MoErges* in palette theory, demonstrating how continent-level cultural repre-

sentations can be dynamically composed to foster country-level cultural pluralism.

### 3 Pentachromatic Cultural Palette Dataset

We contribute a Pentachromatic Cultural Palette dataset with the diverse cultural semantics and values across five continents in social dialogues.

**Dataset Construction.** Cultural differences in evolution across five continents (Crang, 2013) shape values and semantics in different countries. While existing datasets like PRISM (Kirk et al., 2024) collect preference rankings, they lack nuanced dialogues necessary for training models to understand context-specific norms. To address this, we propose a Pentachromatic Cultural Palette dataset, built from PRISM, to capture cultural nuances essential for reflecting the full diversity of cultural semantics in social dialogues. Following (Li et al., 2024c), We enhance cultural richness through a multi-step data synthesis process powered by GPT-4o, utilizing user queries  $\{Query_j\}_{j=1}^n$  from the PRISM dataset. As shown in Fig. 2, our pipeline begins with a self-instruct (Wang et al., 2023) step to generate initial continent-aware responses. These responses then undergo a critique phase based on cross-continent feedback to identify divergent norms. Finally, in the continent-specific aggregation stage, each response is polished through an iterative self-refinement loop (Madaan et al., 2024). Within this loop, the model acts as a self-judge to iteratively enhance the response, ensuring cultural relevance while preventing cross-cultural contradictions. This process is capped at a maximum of  $Max = 3$  rounds and terminates early once a response passes predefined quality checks. And finally we get the Pentachromatic Cultural Palette Dataset <sup>2</sup>, represented as  $\{(Query_j, Response_j^{continent})\}_{j=1}^n$ , where  $continent \in \{Africa, Asia, Europe, America, Oceania\}$ .

### 4 The Cultural Palette Framework

Following Wang et al. (2025), our multi-agent cultural palette encompasses a country-specific

<sup>2</sup>Appendix A.1 details the multi-step synthesis prompts, Appendix A.2 reports dataset scale, and Fig. 3 validates the five-continent base colors. All synthesized data undergoes rigorous human ethics review for cultural fairness and safety (Kirk et al., 2024; Li et al., 2024a).

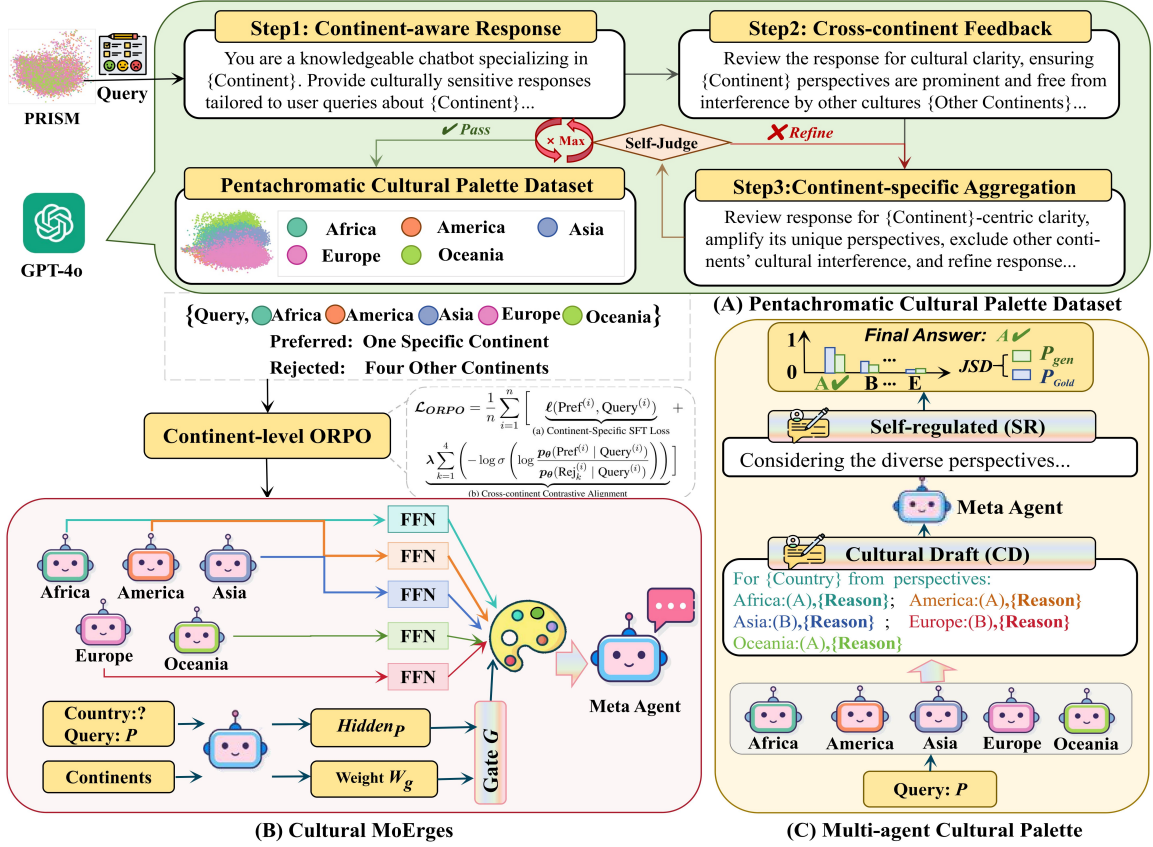


Figure 2: Architecture of Cultural Palette consisting of continent-level alignment on Pentachromatic Cultural Palette Dataset (Sec. 3), Cultural MoErges for Meta Agent (Sec. 4.1) and Multi-agent Cultural Palette (Sec. 4.2). Details of our ORPO alignment loss formulation are deferred to Appendix B for brevity.

Meta Agent (Sec. 4.1) aggregating drafts from five continent-level cultural communities (4.2).

#### 4.1 Cultural Palette for Meta Agent

**Joint/Merging Alignment.** After constructing the Pentachromatic Cultural Palette dataset, we leverage it to perform continent-level alignment for LLMs as cultural decision-maker judger aligned with users, ensuring culturally relevant outputs tailored to diverse regions. Inspired by Li et al. (2024c), we introduce two alignment strategies to highlight the benefits of model merging for overall cultural pluralism. The first strategy, joint alignment, integrates all cultural dialogue data into a single LLM. The second strategy involves creating separate models for each culture, which are then merged into a unified LLM using separately various merging techniques including Ties (Yadav et al., 2024), Model Stock (Jang et al., 2025) and MoErges (Goddard et al., 2024; Jiang et al., 2024) to demonstrate that MoErges exhibits the greatest potential for cultural alignment (see Sec. 5.2).

**Cultural MoErges (CM) for Meta Agent.** Inspired by the merging paradigm of MoErges (Goddard et al., 2024; Tang et al., 2024), we propose Cultural MoErges for Meta Agent to merge cultural knowledge of all five continent-level experts as a shared resource, enabling dynamic adaptation to country-specific contexts following Goddard et al. (2024) while acting as a global cultural reviewer. Five continent-aligned models  $\mathcal{E}_{\text{Group}} = \{E_{\text{Africa}}, E_{\text{Asia}}, E_{\text{Europe}}, E_{\text{America}}, E_{\text{Oceania}}\}$  are trained on region-specific datasets, sharing a base LLM architecture but differing in FFN components following (Feng et al., 2024), and the activation mechanism dynamically adjusts expert contributions through a continent-to-country aware gate that computes weights  $G(P) \in \mathbb{R}^5$  to prioritize expert contributions based on national cultural cues in user prompt  $P$  via:  $G(P) = \text{Softmax}(\text{hidden}_P \cdot W_g)$ , with  $W_g \in \mathbb{R}^{d \times 5}$  initialized via hidden states of continent-level system prompts (e.g., "You are a knowledge chatbot about {continent}...") and  $\text{hidden}_P$  derived from the base LLM's encoding

of  $P$  such that  $G(P)$  blend experts’ parameters into  $LLM_{\text{meta}}$ ’s architecture following MoErges paradigm (Wang et al., 2025), the Meta Agent’s parameter space  $\Theta_{\text{meta}}$  is defined as:

$$\Theta_{\text{meta}} = \Theta_{\text{shared}} \oplus \left( \sum_{C \in \mathcal{E}_{\text{Group}}} G(P) \cdot \Theta_C^{\text{FFN}} \right). \quad (1)$$

where  $\Theta_{\text{shared}} \subset \Theta_{\text{base}}$  denotes frozen shared parameters inherited from the base LLM and  $\Theta_C^{\text{FFN}}$  represents the FFN parameters of continent-level expert  $C$ , dynamically fused via Softmax-normalized weights  $G(P)$ . Empirical experiments in Sec. 5.4 demonstrate that Cultural MoErges outperforms alternative strategies in achieving both cultural pluralism and country-specific alignment.

## 4.2 Multi-agent Cultural Palette

As illustrated in Fig. 2, multi-agent cultural palette consists of draft from community comments aligned with five continents, self-regulated aggregation and final decision from Meta Agent.

**Stage 1: Cultural Draft (CD).** Given a country-specific question  $P$  that includes a target country  $C$  and a query  $Q$ , we engage a continent-level LMs community  $\mathcal{E}_{\text{Group}} = \{E_{\text{Africa}}, E_{\text{Asia}}, E_{\text{Europe}}, E_{\text{America}}, E_{\text{Oceania}}\}$ , each acted as a proxy for a regional community voice informed by local norms, rhetorical patterns, and sociocultural context. While  $C$  is often closely affiliated with one or more continents, all community LMs are jointly activated to span both proximate and distal cultural standpoints aligned with pluralistic roadmap (Sorensen et al., 2024b), promoting culturally diverse yet contextually coherent perspectives (Crang, 2013), forming a pluralistic Cultural Draft *Draft* via culture-aware prompting  $P_{\text{draft}}$ :

$$\text{Draft} = \sum_{E \in \mathcal{E}_{\text{Group}}} \text{Continent} \oplus E(Q, P_{\text{draft}}). \quad (2)$$

where  $\oplus$  denotes the association of the response with its continent tag *Continent*.

**Stage 2: Self-regulated Aggregation (SR).** While continent-level LMs in the community contribute diverse—and occasionally conflicting—cultural standpoints, the objective is not to reconcile these differences but to reinterpret them through the lens of the target country, thereby supporting downstream alignment with the country-

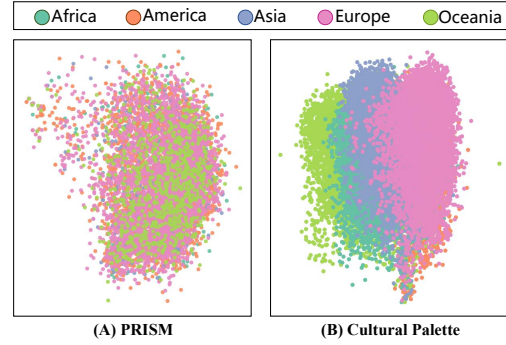


Figure 3: Comparative analysis of semantic representations on the PRISM and Pentachromatic Cultural Palette datasets using Multilingual embeddings (microsoft/Multilingual-MiniLM-L12-H384) with PCA-based dimensionality reduction. Data points are color-coded by continent, based on the country associated with each QA pair.

specific response distribution. To this end, we introduce a *Self-regulated Aggregation* mechanism, wherein  $LLM_{\text{meta}}$ , activated under country  $C$  via Cultural Merge parameter space (Sec. 4.1), recontextualizes each regional draft according to culturally encoded priors aligned with  $C$ . Rather than flattening disagreement,  $LLM_{\text{meta}}$  preserves both consonant and dissonant perspectives, reshaping them through a culturally grounded filter that centers  $C$ ’s normative worldview. Inspired by reflective adaptation principles in formative learning (Nicol and Macfarlane-Dick, 2006),  $LLM_{\text{meta}}$  serves as a cultural interpreter—transforming heterogeneous inputs into a coherent intermediate response, denoted as  $Answer_{\text{regulate}}$ , that preserves cross-cultural plurality while aligning with the socio-contextual frame of country  $C$ .

**Stage 3: Final Decision.** Conditioned on the intermediate output  $Answer_{\text{regulate}}$ , the meta agent  $LLM_{\text{meta}}$  synthesizes a culturally grounded response using the final decision prompt  $P_{\text{final}}$ . Leveraging the aggregated drafts as contextual input, it generates a country-specific answer  $Answer_{\text{final}}$  and distribution  $P_{\text{gen}}$  that reflects the normative worldview of target country  $C$ , supporting post-hoc cultural alignment evaluation by comparing  $P_{\text{gen}}$  with empirical response distributions  $P_{\text{gold}}$  (Sec. 5.2), ultimately enabling continent-to-country cultural alignment.

## 5 Experiments

### 5.1 Models and Implementation

**Data Synthesis.** We utilized the original questions from the PRISM dataset (Kirk et al., 2024),

which comprises 7,805 questions sourced from 75 birth countries. To achieve balanced representation, we augmented it by synthesizing 7,805 questions per continent, ensuring equal distribution among five continents and resulting in a total of 39,025 question-answer pairs, with each continent having 31220 preference pairs derived from the interactions between every two continents. As shown in Fig. 3, the PRISM dataset’s substantial distributional overlap across continents underscores its limitations in capturing cultural distinctions, posing challenges for modeling nuanced cultural segmentation and aligning with region-specific applications. In contrast, our Pentachromatic Cultural Palette dataset demonstrates superior clustering, with data points forming distinct, continent-specific groups that effectively model cultural uniqueness and enable more accurate, culturally aligned representations. The distribution of topics and continents are detailed in Appendix A.2.

**Cultural Alignment.** We conducted all experiments using Llama3.1-8B-Instruct (Dubey et al., 2024) and Qwen2.5-7b-Instruct (Yang et al., 2024a). Additionally, we employed five continent-specific alignments with ORPO (Hong et al., 2024), a parameter-efficient monolithic alignment method, on our Pentachromatic Cultural Palette Dataset, utilizing MS-Swift (Zhao et al., 2025) for alignment and MergeKit (Goddard et al., 2024) for merging on two NVIDIA A100 GPUs (see Appendix B for experiment setting & alignment loss details). For model training and evaluation, we partitioned our dataset was split 9:1 for training and testing, with the test set evaluating Cultural MoErges against alternative alignment methods.

## 5.2 Datasets and Tasks

**Task 1: Cultural Alignment w/ NLI evaluation.** To evaluate Cultural MoErges against alternative merging methods and alignment strategies, we measure LLMs’ ability to generate continent-level responses aligned with cultural values. Following Feng et al. (2024), we use a DeBERTa-v3-large NLI model fine-tuned on standard NLI datasets (Sileo, 2024) to compute a semantic-consistency score as the Entailment probability between gold values  $R_{gold}$  and LLM-generated responses  $R_{llm}$ :

$$S_{semantic} = P(\text{Entailment} | R_{gold}, R_{llm}) \times 100\% \in [0, 100], \quad (3)$$

where higher percentages indicate stronger alignment with the gold-standard cultural values.

**Task 2: Cultural Palette w/ GlobalOpinionQA.** Following Feng et al. (2024); Xu et al. (2025), we compute alignment scores  $S_{align}$  for 18 countries shown in Table 1 across five continents using the GlobalOpinionQA dataset (DURMUS et al., 2024), which aggregates global survey data. For each country-level survey question, LLMs are prompted to integrate national context and generate probability distributions  $P_{gen}$  over answer options, which are then compared to ground-truth survey response distribution  $P_{gold}$  via Jensen-Shannon distance to quantify cultural alignment:

$$S_{align} = 1 - \frac{1}{2} \text{KL}(P_{gen} \| M) - \frac{1}{2} \text{KL}(P_{gold} \| M). \quad (4)$$

where  $M$  is the average distribution of  $P_{gen}$  and  $P_{gold}$ , and KL denotes the Kullback-Leibler divergence. Higher  $S_{align}$  indicates better cultural alignment with real-world survey distributions.

## 5.3 Baselines

We first evaluated Cultural MoErges against four baseline alignment strategies: **(1) Prompting:** We promote cultural alignment by appending instructions via the same prompting template as culture-aware prompting following (Feng et al., 2024; Xu et al., 2025); **(2) Joint:** Following Xu et al. (2025), we implement culturally joint alignment with ORPO (Hong et al., 2024) for overall cultural LLM alignment; **(3) Tie:** Following Yadav et al. (2024), we merge LLMs among continent-specific models by resetting minor fine-tuned parameters, resolving sign conflicts, and merging parameters with consistent signs; **(4) Model Stock:** effectively integrate fine-tuned LLMs across continent-specific applications (Jang et al., 2025).

Subsequently, we compare Cultural Palette with five classical and latest baselines on Llama3.1-8B-Instruct (Dubey et al., 2024) and Qwen2.5-7b-Instruct (Yang et al., 2024a): **(1) Vanilla:** Following (Feng et al., 2024), we directly employ LLM to generate responses according to user queries without any modifications. **(2) Prompting:** We follow (Feng et al., 2024) in promoting cultural pluralism with the same instruction “Make sure your response reflects diverse values and perspectives”. **(3) Joint:** Xu et al. (2025) adopted joint cultural alignment for comprehensive LLM alignment with culture-aware prompting. **(4) Mixture-of-Experts (MoE):** User queries are matched to the best-suited community LM, which generates culturally aligned responses. **(5) Modular:** Feng et al. (2024) proposed three

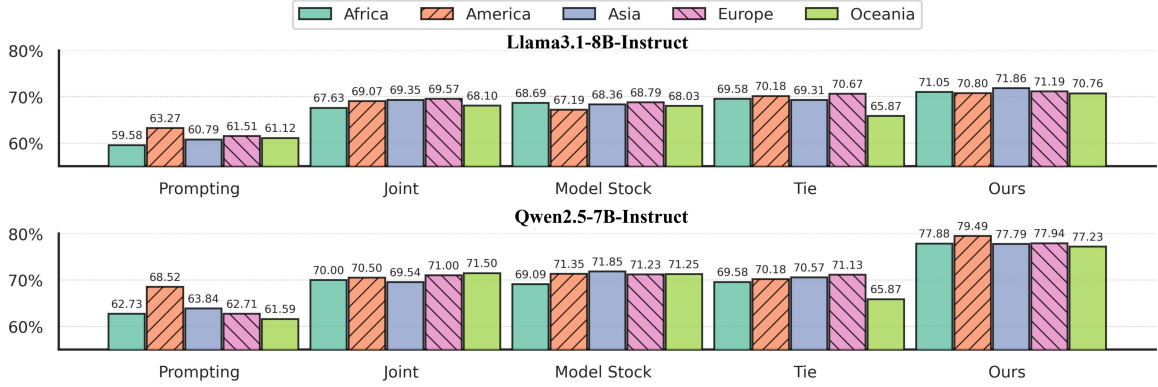


Figure 4: Comparison of semantic-consistency alignment score between our Cultural Palette and other methods (Prompting, Joint) and merging strategies (Tie and Model Stock) on Llama3.1-8B-Instruct and Qwen2.5-7B-Instruct.

forms of multi-agent collaboration, including Over-ton, steerable, and distributional modes.

## 5.4 Main Result

### Cultural MoErges achieves superior alignment with continent-level values and perspectives.

As shown in Fig. 4, Cultural MoErges consistently achieves the strongest overall semantic-alignment performance across continents, outperforming both alignment-based and parameter-merging baselines. Prompting and Joint alignment methods exhibit limited regional adaptability, yielding substantially lower scores under continent-specific evaluation (e.g., 59.58% in Africa and 60.79% in Asia under Prompting on LLaMA-based models). Parameter-merging approaches, including Model Stock and Tie, provide moderate improvements but display noticeable cross-continent inconsistency (e.g., 71.13% in Europe versus 65.87% in Oceania under Tie on Qwen2.5-7B-Instruct). In contrast, Cultural MoErges enables coherent integration of continent-level reasoning, resulting in consistent gains across regions. Specifically, it achieves 71.05% in Africa and 70.76% in Oceania on LLaMA-based models, and 77.88%–79.49% on Qwen2.5-7B-Instruct, surpassing strong modular and merging baselines by approximately 1.5–3.0 percentage points. These results indicate that Cultural MoErges effectively captures continent-level cultural structure while maintaining semantic consistency. A qualitative case study for culture alignment on question-answering *w/ NLI evaluation* is detailed in Appendix C.1.

**Cultural Palette achieves superior real-world cultural alignment.** (1) **Best Performance:** As reported in Table 1, Cultural Palette consistently attains the highest cultural alignment scores across

two LLMs, outperforming all baselines in 18 countries spanning five continents. On LLaMA3.1-8B-Instruct, Cultural Palette achieves an average score of 75.90%, surpassing the strongest baseline Modular (73.82%) by 2.08 percentage points, with pronounced gains in Asia (3.73) and Africa (5.28), regions where prior methods often struggle to capture cultural nuance. On Qwen2.5-7B-Instruct, the improvement is even more significant: Cultural Palette reaches 73.64%, exceeding Modular by 4.13 percentage points, with consistent gains across America (6.35), Europe (0.48), and Asia (3.08). (2) **Alignment Fidelity:** Beyond  $S_{align}$  (Feng et al., 2024; Xu et al., 2025), we evaluate distributional fidelity using **Pearson correlation** with ground-truth survey data. Modular, which relies on country-specific supervision, generalizes poorly (Pearson  $r = -0.044$ ). In contrast, Cultural Palette, trained with continent-level priors and transferred to individual countries, achieves substantially higher correlation ( $r=0.555$ ). This demonstrates that coarse-to-fine regional transfer improves stability and generalization while outperforming direct country-level learning, which is prone to data sparsity and overfitting. Detailed case studies on GlobalOpinionQA and correlation analyses are provided in Appendix C.1 and Appendix D. (3) **GPT-4o as Cultural Teacher:** GPT-4o attains a baseline alignment score of 71.88%, and our synthetic dataset (Sec. 3) effectively enhances SLMs’ cultural alignment, surpassing GPT-4o itself in several regions.

**Ablation Study.** To quantify the contribution of each component, we perform three ablations (Table 1): (i) *w/o Cultural MoErges (CM)*: removing cross-cultural adaptation via CM; (ii) *w/o Cul-*

	America				Europe				Asia				Africa				Oceania		Avg
	USA	CAN	BOL	BRA	GBR	NLD	DEU	UKR	CHN	RUS	IND	THA	KEN	NGA	ETH	ZWE	AUS	NZL	
<b>GPT-4o</b>	72.47	70.28	72.63	69.44	71.01	69.09	73.04	72.96	72.18	74.45	72.16	72.12	73.50	72.16	72.39	74.40	70.44	69.08	71.88
<b>LLAMA3.1-8B-INSTRUCT (Dubey et al., 2024)</b>																			
<b>Vanilla</b>	67.81	65.46	66.46	66.88	59.21	64.30	63.95	59.13	62.76	62.72	66.18	63.01	67.54	66.80	67.31	58.22	64.17	55.68	63.76
<b>Prompting</b>	69.18	67.81	69.43	68.64	60.90	66.65	67.46	63.15	65.81	68.09	69.67	66.22	69.12	68.50	69.51	59.48	67.59	58.49	66.43
<b>Joint</b>	70.10	69.24	70.94	66.35	66.93	70.27	70.54	66.11	67.41	69.21	70.12	63.29	67.14	68.44	66.75	61.34	68.41	60.06	67.37
<b>MoE</b>	73.91	70.97	76.09	75.50	73.92	<u>76.69</u>	72.48	72.87	76.52	74.41	73.02	<u>75.07</u>	74.85	<b>78.76</b>	69.57	70.30	65.25	62.77	72.94
<b>Modular</b>	75.14	75.70	76.60	<b>76.29</b>	<u>76.46</u>	<b>76.71</b>	74.98	71.69	74.16	73.56	<u>74.21</u>	71.76	71.63	70.03	70.05	<u>74.80</u>	<u>72.42</u>	<u>72.57</u>	73.82
<b>Ours</b>	<b>75.85</b>	<b>77.56</b>	<b>77.88</b>	<u>75.83</u>	<b>76.48</b>	76.53	<u>76.38</u>	<b>77.69</b>	<u>77.12</u>	<b>77.67</b>	<b>74.32</b>	<b>75.54</b>	76.12	<u>77.07</u>	<b>73.65</b>	<b>74.82</b>	72.14	<b>73.50</b>	<b>75.90</b>
<b>w/o CM</b>	70.12	72.42	72.77	72.37	69.31	71.01	71.05	70.12	71.52	72.39	71.84	69.36	70.68	70.57	66.23	69.66	69.80	69.51	70.60
<b>w/o CD</b>	72.25	73.77	73.80	74.27	62.23	74.47	73.25	72.37	73.70	75.97	75.99	71.90	<b>76.19</b>	72.94	67.74	60.09	<b>73.44</b>	69.01	71.85
<b>w/o SR</b>	<u>75.54</u>	<u>77.03</u>	<u>77.39</u>	75.25	75.27	76.32	<b>76.54</b>	<u>77.19</u>	<b>77.87</b>	<u>77.42</u>	73.49	74.57	74.73	76.08	<u>71.93</u>	73.89	71.67	71.04	<u>75.18</u>
<b>QWEN2.5-7B-INSTRUCT (Yang et al., 2024a)</b>																			
<b>Vanilla</b>	61.50	60.54	62.67	63.22	63.86	63.66	61.05	59.32	63.78	60.77	60.62	62.12	59.87	63.94	63.56	57.43	54.40	58.84	61.18
<b>Prompting</b>	63.18	63.80	63.48	64.76	64.77	64.52	61.86	62.75	64.99	61.42	61.35	62.20	61.60	63.15	61.39	56.77	52.05	54.44	61.58
<b>Joint</b>	70.00	67.47	65.03	67.36	66.90	68.74	68.22	63.21	63.27	62.80	69.25	65.77	69.36	69.35	66.67	66.20	68.00	62.79	66.69
<b>MoE</b>	<u>73.56</u>	<u>72.59</u>	74.31	71.07	<b>70.96</b>	68.47	68.93	66.76	68.67	68.14	66.15	68.65	66.17	69.51	<u>70.63</u>	65.99	<u>70.51</u>	65.49	69.25
<b>Modular</b>	72.64	68.03	72.43	73.66	69.21	<b>71.21</b>	72.78	<u>75.57</u>	73.01	74.79	66.62	<u>72.19</u>	69.44	70.32	62.04	<u>70.52</u>	57.58	59.09	69.51
<b>Ours</b>	<b>75.50</b>	<b>73.92</b>	<b>78.83</b>	<b>76.71</b>	<u>69.77</u>	<u>70.89</u>	<b>73.91</b>	<b>75.58</b>	<u>73.46</u>	<b>76.09</b>	<u>72.87</u>	<b>73.02</b>	<b>76.51</b>	<b>74.51</b>	70.30	69.25	<b>74.85</b>	<b>69.57</b>	<b>73.64</b>
<b>w/o CM</b>	68.79	67.46	69.63	68.93	62.97	66.19	69.48	68.56	68.83	68.01	69.26	66.57	67.45	66.77	62.74	65.47	66.87	67.16	67.29
<b>w/o CD</b>	72.62	68.22	68.52	71.35	49.30	64.55	71.96	72.77	71.79	<u>74.86</u>	73.98	63.07	69.95	72.50	62.86	46.48	68.05	58.67	66.75
<b>w/o SR</b>	72.45	70.84	72.78	72.62	67.36	69.24	71.83	72.05	<b>73.68</b>	74.26	<b>73.06</b>	70.25	<u>72.11</u>	<u>73.30</u>	<b>72.53</b>	<b>74.24</b>	69.85	<u>68.05</u>	<u>71.69</u>

Table 1: Comparison of culture alignment score (converted to percentages) between Cultural Palette and baselines, evaluated across 18 countries grouped by region: (1) America: USA (United States), CAN (Canada), BOL (Bolivia), BRA (Brazil); (2) Europe: GBR (United Kingdom), NLD (Netherlands), DEU (Germany), UKR (Ukraine); (3) Asia: CHN (China), RUS (Russia), IND (India), THA (Thailand); (4) Africa: KEN (Kenya), NGA (Nigeria), ETH (Ethiopia), ZWE (Zimbabwe); (5) Oceania: AUS (Australia), NZL (New Zealand).

**tural Draft (CD)**: excluding continent-level cultural explanations; and (iii) **w/o Self-Regulate (SR)**: eliminating the self-regulation mechanism that integrates multiple drafts. Removing **CM** causes the largest performance drop: on Qwen2.5-7B-Instruct, the average score decreases from 73.64% to 67.29%, with substantial declines in Asia (4.03 percentage points) and Oceania (4.40 percentage points). This highlights **CM**'s key role in grounding country-level alignment through continent-level priors, which provide semantic regularization and enable cross-national transfer under data scarcity or noisy cultural signals. Ablating **CD** also results in notable losses, particularly in semantically diverse regions such as the Americas (3.70) and Europe (6.91), indicating that **CD** expands the model's hypothesis space with grounded alternatives that guide downstream reasoning. Removing **SR** produces smaller but consistent drops, with an average decrease of 1.95 percentage points, emphasizing its role in integrating multiple drafts into a coherent cultural response. These findings validate our multi-agent alignment strategy: **CD** enriches semantic priors with diverse perspectives, **CM** anchors them through coarse-to-fine structural transfer, and **SR** fuses divergent outputs into a culturally coherent response, together enhancing and stabiliz-

ing culture alignment.

## 6 Conclusion

In this work, we propose Cultural Palette, a novel multi-agent framework for culturally pluralistic alignment in LLMs, modeling five continental cultures as primary colors. Powered by the Pentachromatic Cultural Palette Dataset and a self-regulating *Cultural MoErges* mechanism, Cultural Palette dynamically aggregates continent-level agents to adapt across cultural contexts. Empirical results reveal three key findings: (1) Cultural MoErges enables dynamic composition of continent-level agents, surpassing four modular and training-based baselines in semantic consistency and cultural adaptability; (2) Synthesizing diverse continental data alone suffices to achieve fine-grained continent-to-country cultural alignment, enabling practical adaptation to country-level nuances without the need for country-specific data collection; and (3) Cultural Palette consistently outperforms five pluralistic alignment baselines across both semantic and cultural alignment metrics. Our research may inspire future studies on modeling pluralistic alignment by leveraging continent-level agents to adapt to specific countries and their unique cultural contexts.

## 604 Limitations

605 Our Cultural Palette integrates the core cultural  
606 "colors" of five continents with a Meta Agent that  
607 blends these base colors into a unified multi-agent  
608 framework, symbolizing global cultural diversity  
609 (Crang, 2013). Extensive experiments and case  
610 studies show that our multi-agent framework en-  
611 hances the cultural sensitivity and adaptability of  
612 LLMs. While the multi-agent setup introduces  
613 some computational complexity, as noted in (Feng  
614 et al., 2024), we have addressed this by implement-  
615 ing optimization techniques such as efficient align-  
616 ment (Dao, 2024; Zhao et al., 2025) and inference  
617 (Kwon et al., 2023), effectively reducing compu-  
618 tational burdens. As a result, our framework re-  
619 mains competitive with other cultural alignment  
620 techniques while ensuring better scalability. Addi-  
621 tionally, we envision ongoing efforts to distill the  
622 framework’s capabilities into smaller LLM vari-  
623 ants, enabling more efficient deployment across a  
624 broader range of applications and devices.

625 Finally, as highlighted by (Xu et al., 2025; Feng  
626 et al., 2024), culture definition and cultural map  
627 division are inherently diverse and multifaceted.  
628 We validate our Cultural Palette with continent-to-  
629 country alignment based on continental cultural  
630 geography (Crang, 2013), offering a flexible frame-  
631 work that can be adapted to other cultural maps  
632 (Lechner and Boli, 2008; Inglehart and Welzel,  
633 2010) and encouraging exploration of alternative  
634 cultural divisions to expand the applicability of our  
635 Cultural Palette 🌍.

## 636 Ethical Considerations

637 This work builds on the public PRISM dataset (Kirk  
638 et al., 2024) to construct the Pentachromatic Cul-  
639 tural Palette Dataset for cultural alignment, in full  
640 compliance with its usage policies. While aiming  
641 to promote inclusivity and reduce bias, we recog-  
642 nize potential risks of stereotyping or misrepresen-  
643 tation. To address this, all generated data undergo  
644 both model-in-the-loop and post-hoc human ethics  
645 reviews for cultural sensitivity following Li et al.  
646 (2024a). Our framework advances pluralistic, eq-  
647 uitable representation via multi-agent alignment  
648 and encourages continued expansion of culturally  
649 diverse perspectives in alignment research.

## References

- 650  
651 Josh Achiam, Steven Adler, Sandhini Agarwal, Lama  
652 Ahmad, Ilge Akkaya, Florencia Leoni Aleman,  
653 Diogo Almeida, Janko Altenschmidt, Sam Altman,  
654 Shyamal Anadkat, et al. 2023. Gpt-4 technical report.  
655 *arXiv preprint arXiv:2303.08774*.
- 656 Rainer F Buschmann and Lance Nolde. 2018. *The*  
657 *World’s Oceans: Geography, History, and Environ-*  
658 *ment*. Bloomsbury Publishing USA.
- 659 Alex J Chan, José Luis Redondo García, Fabrizio Sil-  
660 vestri, Colm O’Donnel, and Konstantina Palla. 2023.  
661 Harmonizing global voices: Culturally-aware mod-  
662 els for enhanced content moderation. *arXiv preprint*  
663 *arXiv:2312.02401*.
- 664 Daiwei Chen, Yi Chen, Aniket Rege, and Ramya Ko-  
665 rlakai Vinayak. 2024. Pal: Pluralistic alignment  
666 framework for learning from heterogeneous prefer-  
667 ences. In *NeurIPS 2024 Workshop on Fine-Tuning in*  
668 *Modern Machine Learning: Principles and Scalabil-*  
669 *ity*.
- 670 Mike Crang. 2013. *Cultural geography*. Routledge.
- 671 Tri Dao. 2024. Flashattention-2: Faster attention with  
672 better parallelism and work partitioning. In *The*  
673 *Twelfth International Conference on Learning Repre-*  
674 *sentations*.
- 675 Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey,  
676 Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman,  
677 Akhil Mathur, Alan Schelten, Amy Yang, Angela  
678 Fan, et al. 2024. The llama 3 herd of models. *arXiv*  
679 *preprint arXiv:2407.21783*.
- 680 Esin DURMUS, Karina Nguyen, Thomas Liao,  
681 Nicholas Schiefer, Amanda Askill, Anton Bakhtin,  
682 Carol Chen, Zac Hatfield-Dodds, Danny Hernandez,  
683 Nicholas Joseph, et al. 2024. Towards measuring  
684 the representation of subjective global opinions in  
685 language models. In *First Conference on Language*  
686 *Modeling*.
- 687 Shangbin Feng, Taylor Sorensen, Yuhan Liu, Jillian  
688 Fisher, Chan Young Park, Yejin Choi, and Yulia  
689 Tsvetkov. 2024. Modular pluralism: Pluralistic align-  
690 ment via multi-llm collaboration. In *Proceedings of*  
691 *the 2024 Conference on Empirical Methods in Natu-*  
692 *ral Language Processing*, pages 4151–4171.
- 693 Charles Goddard, Shamane Siriwardhana, Malikeh  
694 Ehghaghi, Luke Meyers, Vladimir Karpukhin, Brian  
695 Benedict, Mark McQuade, and Jacob Solawetz. 2024.  
696 Arcee’s mergekit: A toolkit for merging large lan-  
697 guage models. In *Proceedings of the 2024 Confer-*  
698 *ence on Empirical Methods in Natural Language*  
699 *Processing: Industry Track*, pages 477–485.
- 700 Geyang Guo, Tarek Naous, Hiromi Wakaki, Yukiko  
701 Nishimura, Yuki Mitsufuji, Alan Ritter, and Wei  
702 Xu. 2025. Care: Aligning language models  
703 for regional cultural awareness. *arXiv preprint*  
704 *arXiv:2504.05154*.

705	Christian Haerper, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, and Bjorn Puranen. 2022. World values survey: Round seven-country-pooled datafile version 5.0. <i>Madrid, Spain &amp; Vienna, Austria: JD Systems Institute &amp; WVSA Secretariat</i> , 12(10):8.	models. <i>Advances in Neural Information Processing Systems</i> , 37:105236–105344.	760 761
712	Seungwook Han, Idan Shenfeld, Akash Srivastava, Yoon Kim, and Pulkit Agrawal. 2024. Value augmented sampling for language model alignment and personalization. <i>arXiv preprint arXiv:2405.06639</i> .	Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. <i>Advances in neural information processing systems</i> , 35:22199–22213.	762 763 764 765 766
716	Geert Hofstede. 2009. Geert hofstede cultural dimensions.	Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In <i>Proceedings of the 29th Symposium on Operating Systems Principles</i> , pages 611–626.	767 768 769 770 771 772 773
718	Jiwoo Hong, Noah Lee, and James Thorne. 2024. Orpo: Monolithic preference optimization without reference model. In <i>Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing</i> , pages 11170–11189.	Thom Lake, Eunsol Choi, and Greg Durrett. 2024. From distributional to overton pluralism: Investigating large language model alignment. In <i>Pluralistic Alignment Workshop at NeurIPS 2024</i> .	774 775 776 777
723	Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. In <i>International Conference on Learning Representations</i> .	Daniel Lawson and Ahmed H Qureshi. 2024. Merging decision transformers: Weight averaging for forming multi-task policies. In <i>2024 IEEE International Conference on Robotics and Automation (ICRA)</i> , pages 12942–12948. IEEE.	778 779 780 781 782
728	Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Song Dingjie, Zhihong Chen, Mosen Alharthi, Bang An, Juncai He, et al. 2024. Acegpt, localizing large language models in arabic. In <i>Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)</i> , pages 8132–8156.	Frank J Lechner and John Boli. 2008. <i>World culture: Origins and consequences</i> . John Wiley & Sons.	783 784
736	Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In <i>The Eleventh International Conference on Learning Representations</i> .	Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024a. Culturellm: Incorporating cultural differences into large language models. <i>Advances in Neural Information Processing Systems</i> , 37:84799–84838.	785 786 787 788 789
741	Ronald Inglehart and Chris Welzel. 2010. The wvs cultural map of the world. <i>World Values Survey</i> .	Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. 2024b. Culturepark: Boosting cross-cultural understanding in large language models. <i>Advances in Neural Information Processing Systems</i> , 37:65183–65216.	790 791 792 793 794
743	Hofstede Insights. 2021. Hofstede cultural dimensions.	Junlong Li, Jinyuan Wang, Zhuosheng Zhang, and Hai Zhao. 2024c. Self-prompting large language models for zero-shot open-domain qa. In <i>Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)</i> , pages 296–310.	795 796 797 798 799 800 801
744	Dong-Hwan Jang, Sangdoo Yun, and Dongyoon Han. 2025. Model stock: All we need is just a few fine-tuned models. In <i>European Conference on Computer Vision</i> , pages 207–223. Springer.	Xinyu Li, Ruiyang Zhou, Zachary C Lipton, and Liu Leqi. 2024d. Personalized language modeling from personalized human feedback. <i>arXiv preprint arXiv:2402.05133</i> .	802 803 804 805
748	Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. <i>arXiv preprint arXiv:2401.04088</i> .	Peiqin Lin, Shaoxiong Ji, Jörg Tiedemann, André FT Martins, and Hinrich Schütze. 2024. Mala-500: Massive language adaptation of large language models. <i>arXiv preprint arXiv:2401.13303</i> .	806 807 808 809
753	Hannah Rose Kirk, Alexander Whitefield, Paul Rottger, Andrew M Bean, Katerina Margatina, Rafael Mosquera-Gomez, Juan Ciro, Max Bartolo, Adina Williams, He He, et al. 2024. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language	Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2024. Self-refine: Iterative refinement with self-feedback. <i>Advances in Neural Information Processing Systems</i> , 36.	810 811 812 813 814 815

816	David J Nicol and Debra Macfarlane-Dick. 2006. Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. <i>Studies in higher education</i> , 31(2):199–218.	873
817		874
818		
819		
820	Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. <i>Advances in neural information processing systems</i> , 35:27730–27744.	875
821		876
822		877
823		878
824		879
825		
826	Mohammadreza Pourreza, Ruoxi Sun, Hailong Li, Lesly Miculicich, Tomas Pfister, and Sercan O Arik. 2024. Sql-gen: Bridging the dialect gap for text-to-sql via synthetic data and model merging. <i>arXiv preprint arXiv:2408.12733</i> .	880
827		881
828		882
829		883
830		884
831	Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. <i>Advances in Neural Information Processing Systems</i> , 36.	885
832		886
833		887
834		888
835		889
836	Anudeex Shetty, Amin Beheshti, Mark Dras, and Usman Naseem. 2025. VITAL: A new dataset for benchmarking pluralistic alignment in healthcare. In <i>Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 22954–22974, Vienna, Austria. Association for Computational Linguistics.	890
837		891
838		892
839		893
840		894
841		895
842		896
843	Weiyang Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Sunny Yu, Raya Horesh, Rogério Abreu De Paula, and Diyi Yang. 2024. Culturebank: An online community-driven knowledge base towards culturally aware language technologies. In <i>Findings of the Association for Computational Linguistics: EMNLP 2024</i> , pages 4996–5025.	897
844		898
845		899
846		900
847		901
848		902
849		903
850	Damien Sileo. 2024. tasksource: A large collection of NLP tasks with a structured dataset preprocessing framework. In <i>Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)</i> , pages 15655–15684, Torino, Italia. ELRA and ICCL.	904
851		905
852		906
853		907
854		908
855		909
856		910
857	Taylor Sorensen, Liwei Jiang, Jena D Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, et al. 2024a. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties. In <i>Proceedings of the AAAI Conference on Artificial Intelligence</i> , volume 38, pages 19937–19947.	911
858		912
859		913
860		914
861		915
862		916
863		917
864	Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Miresghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. 2024b. Position: a roadmap to pluralistic alignment. In <i>Proceedings of the 41st International Conference on Machine Learning</i> , pages 46280–46302.	918
865		919
866		920
867		921
868		922
869		923
870		924
871	Anke Tang, Li Shen, Yong Luo, Nan Yin, Lefei Zhang, and Dacheng Tao. 2024. Merging multi-task models	925
872		926
		927
	via weight-ensembling mixture of experts. In <i>Forty-first International Conference on Machine Learning</i> .	
	Peter Vamplew, Conor F Hayes, Cameron Foale, Richard Dazeley, and Hadassah Harland. 2024. Multi-objective reinforcement learning: A tool for pluralistic alignment. In <i>Pluralistic Alignment Workshop at NeurIPS 2024</i> .	
	Junlin Wang, WANG Jue, Ben Athiwaratkun, Ce Zhang, and James Zou. 2025. Mixture-of-agents enhances large language model capabilities. In <i>The Thirteenth International Conference on Learning Representations</i> .	
	Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khoshdel, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In <i>Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 13484–13508.	
	Shaoyang Xu, Yongqi Leng, Linhao Yu, and Deyi Xiong. 2025. Self-pluralising culture alignment for large language models. In <i>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)</i> , pages 6859–6877.	
	Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. 2024. Ties-merging: Resolving interference when merging models. <i>Advances in Neural Information Processing Systems</i> , 36.	
	An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024a. Qwen2 technical report. <i>arXiv preprint arXiv:2407.10671</i> .	
	Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. 2024b. Model merging in llms, mllms, and beyond: Methods, theories, applications and opportunities. <i>arXiv preprint arXiv:2408.07666</i> .	
	Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, et al. 2025. Swift: a scalable lightweight infrastructure for fine-tuning. In <i>Proceedings of the AAAI Conference on Artificial Intelligence</i> , volume 39, pages 29733–29735.	
	Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023. Large language models are not robust multiple choice selectors. In <i>The Twelfth International Conference on Learning Representations</i> .	
	Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. 2023. Don’t make your llm an evaluation benchmark cheater. <i>arXiv preprint arXiv:2311.01964</i> .	

## Appendix of Cultural Palette

This supplementary material provides supplementary materials: Appendix A describes the construction details of Pentachromatic Cultural Palette dataset, Appendix B details the experimental setup and hyperparameters, Appendix C presents qualitative case studies on cultural alignment across *w/ NLI evaluation (C.1)* and *w/ GlobalOpinionQA (C.2)*, and Appendix D provides an additional quantitative continent-to-country correlation analysis with Modular Pluralism (Feng et al., 2024), further validating Cultural Palette’s effectiveness in bridging continent-level reasoning with country-level cultural alignment.

### A Details for Pentachromatic Cultural Palette Dataset

#### A.1 Prompt for Pentachromatic Cultural Palette Dataset

##### Step1: Continent-aware Responses

###### **Instruction**

You are a knowledgeable chatbot specializing in {Continent}. Provide concise, accurate, and culturally sensitive responses tailored to user queries about {Continent}’s geography, culture, history, and current affairs. Ensure your responses demonstrate respect and understanding of the diverse cultures within {Continent}.

###### **1. Understand Query Context:**

Identify the topic and cultural relevance of the user’s query (e.g., geography, culture, history, or current events).

###### **2. Generate Continent-aware Response:**

Show respect for the diverse cultures, avoiding assumptions or biased statements. Highlight the uniqueness and importance of specific cultural aspects when appropriate.

###### **Examples:**

Question: {Question} Answer: {Answer}

**Input:** {Question}

**Output:** {Continent-aware Response}

Figure 5: Prompt for continent-aware responses

##### Step2: Cross-continent Feedback

###### **Instruction**

Review the response for cultural clarity, ensuring {Continent} perspectives are prominent and free from interference by other cultures {Other Continents}. If confusion arises, provide improvement suggestions using Hofstede’s cultural dimensions:

###### **1. Analyze Query Context:**

Assess whether the response fully aligns with the cultural, historical, and geographical relevance of {Continent}. Identify any unnecessary elements or overlaps with other continents’ cultural markers.

###### **2. Refine Response for {Continent} Perspective:**

Emphasize key cultural, historical, or environmental aspects unique to {Continent}. Remove any elements that might introduce cultural interference or bias. Ensure the response highlights community, tradition, and respect for local customs.

###### **3. Apply Hofstede’s Cultural Dimensions Theory for Feedback:**

- **Power Distance (PDI):** Does the response reflect the balance between leadership (e.g., village chiefs) and community roles?

- **Individualism vs. Collectivism (IDV):** Does it highlight collective well-being and community-driven decisions?

- **Masculinity vs. Femininity (MAS):** Does it prioritize care, cooperation, and harmony over competition?

- **Uncertainty Avoidance (UAI):** Does it address Oceania’s adaptability to change while respecting traditions?

- **Long-Term vs. Short-Term Orientation (LTO):** Does it balance preservation of traditions with immediate adaptability?

- **Indulgence vs. Restraint (IVR):** Does it reflect the expressive and celebratory aspects of Oceanian cultures?

**Input:** {Continent-aware Responses}

**Output:** {Cross-continent Feedback}

Figure 6: Prompt for cross-continent feedback

### Step3: Continent-specific Aggregation

#### **Instruction**

You are tasked with improving the following response for the query {query} by incorporating the cross-continent review feedback.

#### **1. Review Context Response:**

Carefully read and analyze the Base Response and Review Feedback provided. Identify elements in the base response that lack focus on {Continent} or introduce cultural elements from other continents.

#### **2. Apply Review Feedback:**

Incorporate all suggestions from the review feedback into the base response. Ensure the revised response highlights {Continent}-specific culture, history, and values and focus on {Continent} Perspectives, avoiding any overlap with other cultural contexts.

#### **Input:**

Base Response: {Base response}

Review Feedback: {Review feedback}

**Output:** {Continentfocused Response}

950

Figure 7: Prompt for Continent-specific Aggregation

### Step4: Self-Judge Refinement

#### **Instruction**

You are now a cultural self-judge. Your task is to assess whether the aggregated response reflects accurate, respectful, and continent-specific perspectives for the user query. Critically evaluate the revised response based on feedback, and decide whether further improvement is needed.

#### **1. Evaluate Cultural Fidelity:**

Check if the revised response aligns with the core values, traditions, and worldview of {Continent}. Is the response focused, respectful, and specific? Does it avoid generalizations or unrelated cross-cultural references?

#### **2. Verify Feedback Integration:**

Determine whether the suggestions from cross-continent feedback have been fully addressed. Are the earlier critiques resolved in this version?

#### **3. Quality Criteria for Acceptance:**

Approve the response only if it meets all the following: - Emphasizes cultural clarity and salience - Integrates prior feedback - Avoids cultural interference - Is concise, coherent, and informative

#### **Output:**

If the response is acceptable, return: [Approved]

If the response requires revision, return: [Revise] And explain what needs to be changed.

952

Figure 8: Prompt for self-judge refinement to ensure cultural alignment

953

## A.2 Distribution in Pentachromatic Cultural Palette Dataset

954

955

Figure 9 illustrates the distribution of topics and continent over our Pentachromatic Cultural Palette Dataset. The topics were categorized using a topic modeling llm<sup>3</sup> which classifies the questions into distinct thematic clusters. The distribution of continents, on the other hand, is evenly balanced, with an equal number of questions synthesized for each continent: Africa, America, Asia, Europe, and Oceania.

956

957

958

959

960

961

962

963

964

<sup>3</sup><https://huggingface.co/classla/multilingual-IPTC-news-topic-classifier>, License: cc-by-sa-4.0

951

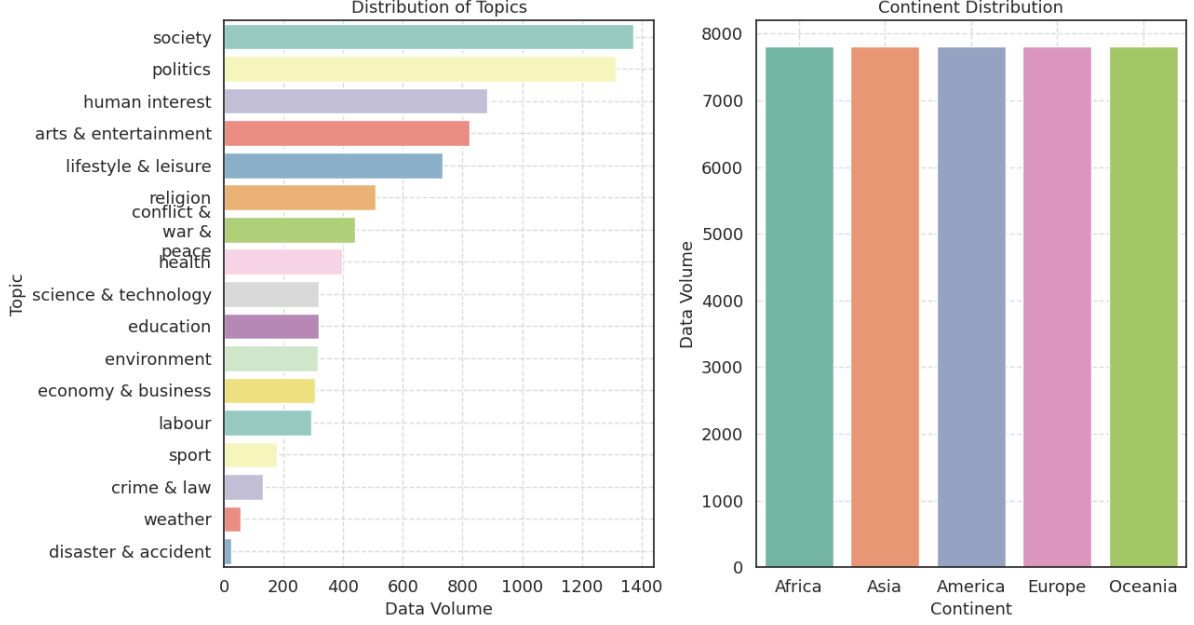


Figure 9: Distribution of topics and continents in Pentachromatic Cultural Palette Dataset.

## B Experiment Details

**Experiment Setting.** We adopt the ORPO method (Hong et al., 2024) with default hyperparameters for Llama3.1-8B-Instruct (Dubey et al., 2024) and Qwen2.5-7B-Instruct (Yang et al., 2024a), using a learning rate of 5e-5 for 2 epochs. We select the average result from five runs as the experimental results.

**Alignment Implement.** We align continent-specific LLMs using ORPO (Hong et al., 2024), optimizing the following loss,  $\mathcal{L}_{\text{ORPO}}$ :

$$\mathcal{L}_{\text{ORPO}} = \frac{1}{n} \sum_{i=1}^n \left[ \underbrace{\ell(\text{Pref}^{(i)}, \text{Query}^{(i)})}_{\text{(a) Continent-specific SFT Loss}} + \underbrace{\lambda \sum_{k=1}^4 \left( -\log \sigma \left( \log \frac{p_{\theta}(\text{Pref}^{(i)} | \text{Query}^{(i)})}{p_{\theta}(\text{Rej}_k^{(i)} | \text{Query}^{(i)})} \right) \right)}_{\text{(b) Cross-continent Contrastive Alignment}} \right] \quad (5)$$

where  $\ell(\text{Pref}^{(i)}, \text{Query}^{(i)})$  denotes the continent-specific supervised fine-tuning (SFT) loss, aligning the LLM to generate the preferred response  $\text{Pref}^{(i)}$  given query  $\text{Query}^{(i)}$ . The second term enforces cross-continent contrastive alignment by penalizing less-preferred responses  $\text{Rej}_k^{(i)}$  from other continents that diverge from the preference of the target region. We set the odds ratio  $\lambda = 0.1$ , following MS-Swift (Zhao et al., 2025), to balance in-continent preference alignment with intercontinental contrastive separation.

**Licenses.** The licenses for the codes used in our work are listed in Table 2.

988  
989

Resources	Type	Link	License
MergeKit (Goddard et al., 2024)	Code	<a href="https://github.com/arcee-ai/mergekit">https://github.com/arcee-ai/mergekit</a>	LGPL-3.0 license
MS-Swift (Zhao et al., 2025)	Code	<a href="https://github.com/modelscope/ms-swift">https://github.com/modelscope/ms-swift</a>	Apache-2.0 license

Table 2: Licenses for the software libraries utilized in our work.

## 990 C Case Study

### 991 C.1 Case Study for Cultural Alignment w/ 992 *NLI evaluation*

993 To validate our Cultural Palette’s efficacy, we con-  
994 ducted a case study (Tables 3 & 4) addressing the  
995 question, "What happens when we die?" , to demon-  
996 strate how our Cultural Palette generates diverse re-  
997 sponses across continents, reflecting each region’s  
998 unique perspective on death. In Oceania, the re-  
999 union of souls with ancestors is emphasized, high-  
1000 lighting the cultural significance of these traditions.  
1001 In Africa, the continuity of the soul and its con-  
1002 nection to ancestors reinforce communal bonds. In  
1003 Asia, beliefs such as reincarnation, enlightenment,  
1004 and ancestor worship illustrate varied views of the  
1005 afterlife. In America, a blend of perspectives is  
1006 presented, including concepts of the afterlife and  
1007 views on the cessation of consciousness. Finally,  
1008 Europe encompasses various death-related tradi-  
1009 tions, showcasing the diversity of beliefs about life  
1010 after death. This analysis underscores the model’s  
1011 ability to capture cultural contexts and reveals dis-  
1012 tinct views on death across continents.

### 1013 C.2 Case Study for Cultural Palette w/ 1014 *GlobalOpinionQA*

1015 We provide a case study for Cultural Palette w/  
1016 *GlobalOpinionQA* to demonstrate how our Cultural  
1017 Palette works, as detailed in Table 5.

Continent	Input: What happens when we die?	Ground Truth	$S_{semantic}$
Oceania	<p>In Oceania, the concept of death and the afterlife varies among different cultures and communities. <b>Many indigenous cultures in Oceania believe in a spiritual realm where the deceased continue to exist and play an active role in the lives of the living.</b> For example, in some Pacific Island cultures, <b>the spirits of ancestors are believed to guide and protect their descendants, while in other cultures, the deceased are thought to journey to a spiritual realm where they are judged based on their actions in life.</b></p>	<p>In Oceania, the beliefs surrounding what occurs after death are rich and diverse. Among the indigenous communities, <b>there is a profound connection to an afterlife where spirits are reunited with their ancestors in a sacred land.</b> Additionally, some cultures embrace the idea of <b>reincarnation or embarking on a spiritual voyage to another realm beyond our earthly existence.</b> It is paramount to approach these beliefs with deep respect and understanding, honoring the profound significance they hold within Oceanian cultures.</p>	0.8927
Africa	<p>In many African cultures, <b>the concept of death is deeply intertwined with spirituality and the afterlife.</b> The belief in an afterlife varies among different African societies, but it is often seen as <b>a continuation of life in a spiritual realm.</b> The deceased are believed to journey to the spirit world, where they are judged based on their deeds in life. <b>The spirits of the ancestors are revered and believed to play a significant role in guiding the living.</b></p>	<p>In various African cultures, there is a profound belief in the continuity of life beyond death. <b>Ancestral spirits play a significant role, where the departed are thought to watch over and guide their living relatives.</b> This connection with ancestors is cherished, and rituals are performed to honor and communicate with them. Additionally, some traditions believe in <b>reincarnation, suggesting that the soul returns in a new form after death.</b> The respect for these diverse beliefs reflects the communal values deeply rooted in many <b>African societies</b>, honoring the interconnectedness between the living and the deceased.</p>	0.7671
Asia	<p>In many Asian cultures, the concept of death is deeply <b>intertwined with spiritual beliefs and traditions.</b> The idea of what happens after death varies across different cultures and religions in Asia. For example, in Buddhism, the belief is that <b>the soul is reborn into a new existence based on past actions, known as karma.</b> In Hinduism, the soul is believed to be <b>reincarnated into a new life cycle, influenced by the law of karma.</b> In many Asian cultures, the deceased are often believed to <b>continue their journey in the afterlife, where they may be reunited with ancestors or other loved ones.</b> It's important to note that these beliefs can vary widely across different cultures and individuals in Asia.</p>	<p>In Asian cultures, diverse beliefs shape perspectives on death and the afterlife. Hinduism teaches that <b>souls undergo reincarnation driven by karma, seeking spiritual growth.</b> Buddhism emphasizes <b>breaking free from the cycle of rebirth through enlightenment.</b> In Japan, <b>ancestor veneration and a belief in spirits existing in the afterlife are prominent practices.</b> The rich tapestry of Asian beliefs offers varied insights into what lies beyond our earthly existence.</p>	0.5847

Table 3: Case Study for Cultural Platette on Cultural Alignment w/ *NLI evaluation* (A).

Continent	Input: What happens when we die?	Ground Truth	$S_{semantic}$
America	<p>In American culture, the concept of what happens after death is <b>deeply rooted in various beliefs and traditions</b>. Many Americans find <b>solace in the idea of an afterlife, where the soul continues to exist in a spiritual realm</b>. This belief is often <b>tied to religious teachings, such as Christianity, Judaism, Islam, and others</b>. Some believe in <b>reincarnation</b>, while others may <b>view death as the end of physical existence</b>. It's important to acknowledge that these perspectives vary widely among individuals and communities in America.</p>	<p>In American culture, beliefs about what happens after death can vary widely. Some Americans <b>adhere to religious beliefs that include an afterlife in heaven or hell</b>, while others believe in concepts like <b>reincarnation</b>. Additionally, some Americans <b>view death from a more secular perspective where consciousness simply ceases to exist</b>.</p>	0.7557
Europe	<p>In Europe, the concept of what happens after death is deeply rooted in various cultural, religious, and philosophical beliefs. <b>The idea of an afterlife, where the soul or spirit continues to exist beyond physical death, is a common thread in many European cultures</b>. For example, <b>in Christianity, the belief in heaven and hell is prevalent</b>, while in Islam, <b>the concept of Jannah (paradise) and Jahannam (hell) is central</b>. Additionally, some European cultures <b>believe in reincarnation or the transmigration of the soul</b>. It's important to note that these beliefs can vary significantly across different European countries and communities.</p>	<p>In Europe, the beliefs regarding what happens when we die are deeply rooted in diverse cultural and religious traditions. For instance, <b>Christian belief in heaven and hell has been significant throughout European history, shaping perspectives on the afterlife</b>. Similarly, <b>Nordic cultures have a rich tradition of Valhalla, a majestic hall where warriors go after death</b>. Furthermore, the <b>Greek belief in the underworld, ruled by gods like Hades and Persephone, adds another layer to European perceptions of death</b>. These varied beliefs reflect the intricate tapestry of European culture and highlight the importance of understanding and respecting different perspectives on life's end.</p>	0.7085

Table 4: Case Study for Cultural Platette on Cultural Alignment w/ *NLI evaluation* (B).

Cultural Palette	Step
Question	Question: Most people consider both freedom and equality to be important, but if you had to choose between them, which one would you consider more important? Options: <b>A.</b> Freedom <b>B.</b> Equality <b>C.</b> Don't know <b>D.</b> No answer
Draft	<p><b>African Culture perspective:</b> From an African perspective, the concept of equality is deeply rooted in many African cultures, where community and collective well-being are highly valued. In this context, prioritizing equality over freedom is often seen as a way to promote social cohesion, reduce inequality, and ensure that everyone has a fair chance to succeed ... This philosophy suggests that individual freedom is not as important as the collective well-being of the community. <b>B. Equality</b></p> <p><b>America Culture perspective:</b> From an American perspective, considering New Zealand's unique culture, traditions, and history, one might argue that equality is more important. ... Their focus on freedom reflects a commitment to individual rights and autonomy, which are core values in New Zealand's society. While equality is undoubtedly crucial, the context of New Zealand's history and culture emphasizes the importance of freedom in fostering a just and inclusive society. <b>A. Freedom</b></p> <p><b>Oceania Culture perspective:</b> From an Oceania perspective, New Zealand's unique culture, traditions, and history emphasize the importance of freedom. ... In the Pacific region, the concept of "manaakitanga" (respect and care for others) is deeply rooted in many cultures, but it is often balanced with the importance of individual freedom and autonomy. New Zealand's emphasis on freedom reflects this balance, prioritizing the rights and freedoms of individuals while also promoting social cohesion and community. <b>A. Freedom</b></p> <p><b>Europe Culture perspective:</b> From a European perspective, New Zealand's unique culture, traditions, and history highlight a strong emphasis on both freedom and equality, reflecting its Māori and European influences. ... However, given the options provided, the most appropriate answer would be <b>C. Don't know</b>, as the question presents a false dichotomy. In New Zealand's context, both freedom and equality are considered paramount, and the country's history and culture do not necessitate a choice between them. <b>C. Don't know</b></p> <p><b>Asia Culture perspective:</b> From an Asian perspective, New Zealand's emphasis on equality is particularly relevant in its multicultural society ... In this context, equality is considered more important than freedom, as it ensures that all individuals have equal opportunities and are treated with dignity and respect. This aligns with the values of many Asian cultures, which prioritize harmony and social cohesion over individual freedom. <b>B. Equality</b></p>
Regulate	Considering the diverse cultural perspectives, New Zealand's cultural attributes can be summarized as community and collective well-being, freedom, and multiculturalism... Many cultures, including African, Asian, and Pacific Islander, prioritize community and collective well-being, while others, such as American and European, value individual freedom and autonomy... the Oceania perspective emphasizes the importance of freedom... Individual freedom is essential to promoting social cohesion and community, and New Zealand's emphasis on freedom reflects a commitment to individual rights and autonomy. ... <b>A. Freedom</b>
Answer	<p><b>A. Freedom</b> ✓</p> <p><b>Gold Distribution:</b> [0.6709999999999999, 0.242, 0.061, 0.0, 0.026000000000000002]</p> <p><b>Pred Distribution:</b> [0.6049056212210604, 0.3230594648058141, 0.060349139731253465, 0.0116857742428727]</p>

Table 5: Cultural Alignment on Decision-Making in Household Finances

1018  
1019

## D Comparative Analysis via Pearson Correlation Across Countries

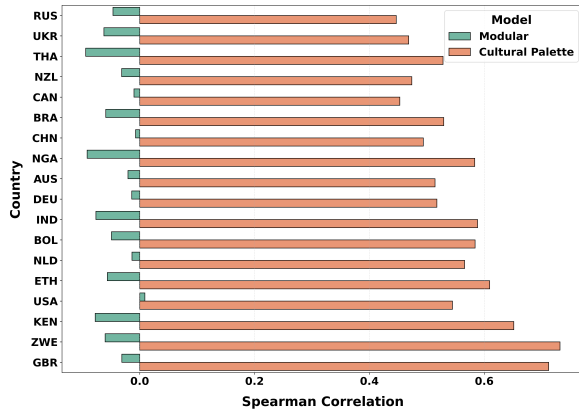


Figure 10: Comparative Analysis of Modular Pluralism (Feng et al., 2024) and Cultural Palette Models via Pearson Correlation Across Countries

1020  
1021  
1022  
1023  
1024  
1025  
1026  
1027  
1028  
1029  
1030  
1031  
1032  
1033  
1034  
1035  
1036  
1037  
1038

The comparison detailed in Fig. 10 illustrates that the Cultural Palette model consistently captures cultural and regional nuances more effectively across diverse countries such as Ethiopia, India, and Kenya, where it achieves notably higher Pearson correlations. This suggests strong adaptability to local contexts and complex sociocultural patterns. In contrast, Modular Pluralism (Feng et al., 2024) struggles in these countries, often producing low or near-zero correlations, highlighting its limited ability to model cultural heterogeneity within and across nations. Additionally, even in countries like Russia, China, and Ukraine, the Cultural Palette maintains a clear advantage, reinforcing the importance of culturally aware modeling. These qualitative insights emphasize that incorporating country-specific cultural factors is critical for building robust predictive models in global, multicultural environments.