

# CuMA: Aligning LLMs with Sparse Cultural Values via Demographic-Aware Mixture of Adapters

Ao Sun<sup>1</sup>, Xiaoyu Wang<sup>2</sup>, Zhe Tan<sup>1</sup>, Yu Li<sup>2</sup>, Jiachen Zhu<sup>3</sup>, Yuheng Jia<sup>2,4\*</sup>, Shu Su<sup>1\*</sup>

<sup>1</sup>School of Civil Engineering, Southeast University

<sup>2</sup>School of Computer Science and Engineering, Southeast University

<sup>3</sup>ByteDance Inc.

<sup>4</sup>Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, China  
{sunao, yhjia, sushu}@seu.edu.cn

## Abstract

As Large Language Models (LLMs) serve a global audience, alignment must transition from enforcing universal consensus to respecting cultural pluralism. We demonstrate that dense models, when forced to fit conflicting value distributions, suffer from **Mean Collapse**, converging to a generic average that fails to represent diverse groups. We attribute this to **Cultural Sparsity**, where gradient interference prevents dense parameters from spanning distinct cultural modes. To resolve this, we propose CuMA (Cultural Mixture of Adapters), a framework that frames alignment as a conditional capacity separation problem. By incorporating demographic-aware routing, CuMA internalizes a *Latent Cultural Topology* to explicitly disentangle conflicting gradients into specialized expert subspaces. Extensive evaluations on WorldValuesBench, Community Alignment, and PRISM demonstrate that CuMA achieves state-of-the-art performance, significantly outperforming both dense baselines and semantic-only MoEs. Crucially, our analysis confirms that CuMA effectively mitigates mean collapse, preserving cultural diversity. Our code is available at <https://github.com/Throll/CuMA>.

## 1 Introduction

Large Language Models (LLMs) have achieved remarkable success in general-purpose reasoning (Gao et al., 2024). To ensure these models remain helpful and harmless, alignment techniques like Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017; Ouyang et al., 2022) are widely adopted. This paradigm typically uses a monolithic reward model to capture human preferences (Frick, 2025). This approach is effective for consensus-based tasks, such as safety compliance (Xue et al., 2024), code generation (Chen

et al., 2021), and mathematical reasoning (Zhang et al., 2025c), where a globally optimal response generally exists.

However, as LLMs serve a global user base, alignment must extend to cultural resonance (Adilazuarda et al., 2024; Oh et al., 2025). In subjective domains, response utility is culturally contingent, meaning a response considered insightful in one community may be irrelevant in another (Khamassi et al., 2024). Consequently, human values are inherently pluralistic and often conflicting (Sorensen et al., 2024). Existing methods (Christiano et al., 2017; Ouyang et al., 2022; Rafailov et al., 2023; Gu et al., 2025) optimize a dense set of parameters over such data, implicitly assuming a unified value system. When minimizing error across conflicting modes, dense models gravitate towards a statistical average, leading to Mean Collapse.

This results in the model collapsing divergent values into a single dominant representation, suppressing minority perspectives and imposing a monolithic consensus (Durmus et al., 2024). Mean Collapse manifests as "mode-covering" behavior, where models output generic, diluted responses. Crucially, this average is rarely neutral. Driven by imbalances in pre-training corpora (Alkhamissi et al., 2024; Zhu et al., 2025; Öncel et al., 2024) and the homogeneity of crowd-sourced annotators (Li et al., 2025; Li, 2024), the learned "mean" often reflects Western, Educated, Industrialized, Rich, and Democratic (WEIRD) norms (Santurkar et al., 2023; Henrich et al., 2010).

We argue that this failure is rooted in gradient interference. Human values are inherently sparse (Kostina et al., 2015), clustering into distinct, conflicting modes rather than forming a continuous spectrum (Liu et al., 2025). A single dense model cannot simultaneously fit these opposing clusters (Sukiennik et al., 2025; Adilazuarda et al., 2024). Consequently, to minimize global error, it converges to a statistical average, or the "diluted

\* Corresponding author.

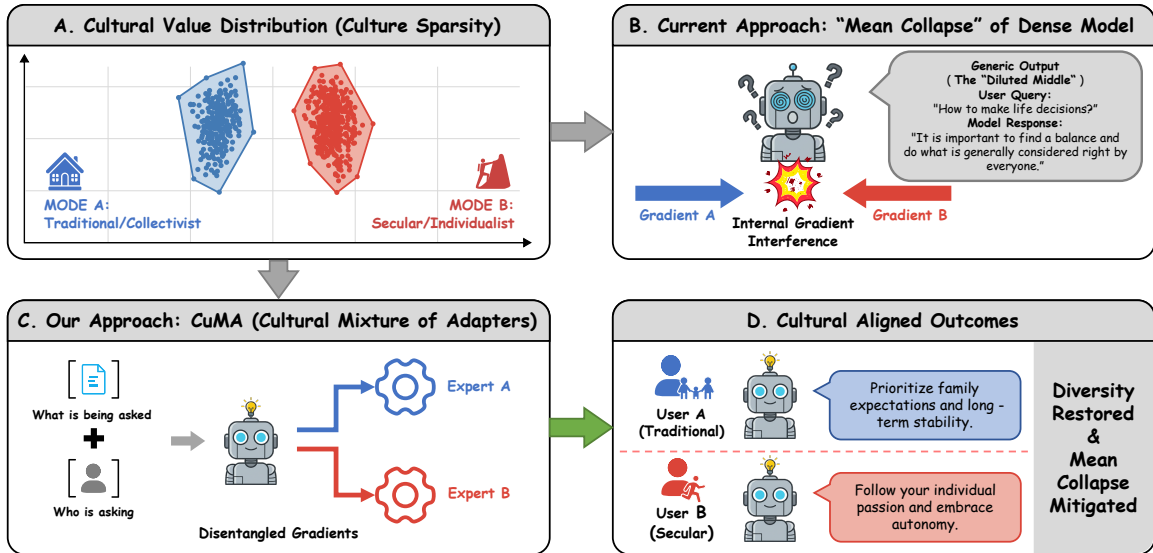


Figure 1: **Mechanism of Mean Collapse and the CUMA Solution.** (A) Human values exhibit *Cultural Sparsity*, forming distinct modes (e.g., Traditional vs. Secular). (B) Standard dense models suffer from **Gradient Interference** when optimizing for conflicting modes simultaneously. This forces the model into **Mean Collapse** (the "Diluted Middle"), producing generic responses that fail to resonate with any group. (C) **CUMA** addresses this via *Demographic-Aware Routing*, explicitly disentangling gradients into specialized experts. (D) Consequently, the model generates distinct, culturally resonant outcomes for diverse users, effectively restoring value diversity.

middle", as visualized in Figure 1.

To address this, we propose CUMA, a framework that reformulates alignment as a conditional capacity separation problem. Standard Mixture-of-Experts (MoE) route tokens based solely on internal hidden states (Zhou et al., 2022; Li and Zhou, 2024), struggling to distinguish culturally conflicting preferences within similar contexts (Wang et al., 2024a). This design is motivated by the insight that cultural differences are driven by both semantic and demographic proxies (Adilazuarda et al., 2024). Therefore, CUMA conditions expert selection on the joint representation of semantic content and the user’s demographic profile. This allows the router to learn a latent cultural topology, where parameter subspaces are specialized not just by what is being asked, but by who is asking, effectively isolating gradients and preserving cultural diversity (Fu and Lapata, 2022).

Our contributions are as follows: (1) We formally identify cultural sparsity as the geometric root of alignment failure in pluralistic settings, demonstrating that dense parameterization inevitably leads to *Mean Collapse*, a structural inability to resolve conflicting modes;

(2) We propose CUMA, a framework that implements conditional capacity separation via demographic-aware routing to explicitly disentangle conflicting gradients into specialized parameter

subspaces, allowing the model to learn a latent cultural topology that isolates interference;

(3) Extensive evaluations on WorldValuesBench, Community Alignment, and PRISM show that CUMA achieves state-of-the-art performance, significantly outperforming dense baselines. Analysis confirms that this disentanglement effectively restores generative diversity and mitigates the Mean Collapse found in standard dense models.

## 2 Problem Formulation

In this section, we establish the theoretical foundations of our framework. From a probabilistic perspective, we formulate cultural alignment as a conditional modeling task dependent on demographic context. We then characterize the geometry of pluralistic values through the lens of *Cultural Sparsity*, and analyze why dense parameterization fails to capture this geometry, leading to *Mean Collapse*.

### 2.1 Cultural Alignment as Conditional Modeling

We formalize cultural alignment as a conditional modeling problem, where response validity depends on the user’s cultural context. Let  $\mathcal{X}$  denote the space of inputs (e.g., prompts),  $\mathcal{Y}$  the space of responses, and  $\mathcal{D}$  the set of demographic profiles (e.g., region, ideology) serving as observable

proxies for latent cultural values. The objective is to learn a conditional model  $P_\theta(y | x, d)$  that maximizes the likelihood of culturally resonant responses.

A crucial distinction exists between *consensus* and *non-consensus* tasks. For consensus-based tasks (e.g., safety (Lu et al., 2025; Zhao et al., 2025) or math reasoning (Ahn et al., 2024; Azerbayev et al., 2024)), an optimal response  $y^*$  is invariant to user attributes (i.e.,  $P(y|x, d) \approx P(y|x)$ ), and a single model is sufficient. However, cultural alignment is inherently a *non-consensus* problem (Tao et al., 2024): the optimal response distribution varies across  $\mathcal{D}$ , and conflicting values can be simultaneously valid. To maximize utility, the model should explicitly model the dependency on  $d$ , rather than marginalizing over it.

## 2.2 Cultural Sparsity

While distinct cultures often share universal commonalities, their preference distributions in the latent representation space typically exhibit multimodal structures, where divergent value systems form separate clusters. We term this geometric property *Cultural Sparsity*.

**Definition 2.1 (Cultural Sparsity).** Let  $P^*(y | x, d_i)$  and  $P^*(y | x, d_j)$  be the conditional value distributions for two distinct demographic profiles. Let  $\mu_k \in \mathbb{R}^m$  and  $\Sigma_k \in \mathbb{R}^{m \times m}$  denote the mean vector and covariance matrix of group  $k$  in the  $m$ -dimensional representation space (where  $m$  is the ambient dimensionality, e.g., the hidden size of the LLM). Defining the pooled covariance as  $\bar{\Sigma}_{ij} = \frac{1}{2}(\Sigma_i + \Sigma_j)$ , we categorize the distributions as *culturally sparse* if the Mahalanobis distance between their centers significantly exceeds the ambient dimension  $m$ :

$$(\mu_i - \mu_j)^\top \bar{\Sigma}_{ij}^{-1} (\mu_i - \mu_j) \gg m \quad (1)$$

This inequality implies that inter-group divergence dominates intra-group dispersion. Under such sparsity, a single dense representation is geometrically incapable of covering disjoint modes simultaneously. Consequently, the model collapses diverse values into a single expectation, failing to accurately capture distinct cultural preferences (see Appendix B.3).

## 2.3 The Failure of Dense Models: Mean Collapse

Standard alignment methods optimize a dense model  $P_\theta(y | x)$  by minimizing the forward

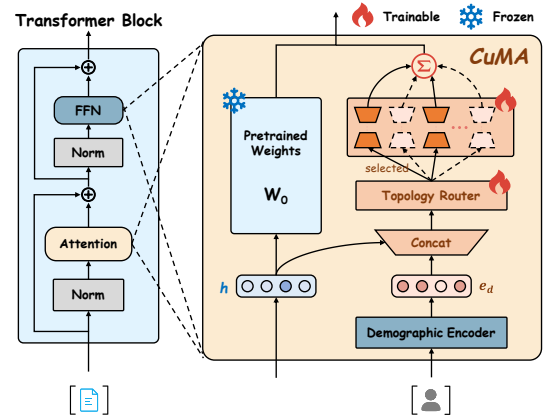


Figure 2: **Architecture of CuMA.** The framework disentangles cultural values by conditioning the routing mechanism on both semantic hidden states and demographic embeddings, effectively isolating gradients into specialized experts.

Kullback-Leibler (KL) divergence  $D_{\text{KL}}(P_{\text{data}} \parallel P_\theta)$ . While the distribution of models is theoretically complex (Machina and Mercer, 2024), the shared parameterization across conflicting groups forces the model to capture the central tendency of the aggregate gradient. We analyze this behavior using a unimodal proxy in the representation space.

**Theorem 2.1 (Mean Collapse).** Under the assumption of cultural sparsity (Eq. 1), consider a dense estimator  $P_\theta$  constrained to a single-component exponential family (e.g., a Gaussian) with mean parameter  $\mu_\theta$ . The solution minimizing the forward-KL divergence satisfies  $\mu_\theta^* = \mathbb{E}_{P_{\text{data}}}[y]$ , converging strictly to the global mixture mean. Consequently, the model exhibits *mode-covering* behavior: it centers its probability mass in the "diluted middle", a solution that is statistically optimal for minimizing global error, yet fails to capture the inherent plurality of cultural values. We provide comprehensive derivations in Appendix B: Appendix B.2 proves the mean-matching property; Appendix B.3 quantifies the exponential density decay at the collapsed mean; Appendix B.3 demonstrates the resulting variance inflation; and Appendix B.4 theoretically establishes the resolution via conditional routing.

## 3 CuMA: Modeling Latent Cultural Topology via Conditional Routing

To address Cultural Sparsity and Mean Collapse, we propose CuMA (Figure 2). Instead of using a single parameter set for conflicting values, CuMA learns a latent cultural topology and routes inputs

to specialized, demographically-aligned adapters. This design disentangles gradient interference and preserves the distinct geometry of pluralistic value distributions.

### 3.1 Demographic Encoder

To encode diverse demographic profiles and support generalization, we leverage the geometric priors in pre-trained sentence embedding models. The raw demographic profile  $d$  typically consists of structured attributes (e.g., {Country: Thailand, Religion: Buddhism, Age: 55}). We first linearize this structured set into a natural language description  $t_d$  (e.g., "A 55-year-old Buddhist resident of Thailand"). We then map  $t_d$  to a dense vector representation  $e_d \in \mathbb{R}^m$  via a frozen pre-trained embedding model  $E(\cdot)$ :

$$e_d = E(t_d) \quad (2)$$

By utilizing the frozen embedding space, we preserve the semantic topology from pre-training. Within this space, culturally related groups naturally cluster based on shared traits like geography or religion. This stable structure provides robust signals for the router to measure similarity, enabling generalization to unseen demographic groups.

### 3.2 Router as Topology Learner

The router serves as the core topological mapper. Unlike standard MoE routers that dispatch tokens based solely on internal hidden states (semantic content), our router learns the latent cultural topology by conditioning on the joint interaction between the semantic context and the demographic profile.

For a given layer input  $h \in \mathbb{R}^H$  and demographic embedding  $e_d$ , the router computes the routing logits  $s \in \mathbb{R}^N$ :

$$s = W_r \cdot [h \oplus e_d] \quad (3)$$

where  $\oplus$  denotes concatenation and  $W_r$  is the learnable routing matrix. This joint representation allows the router to disentangle what is being asked ( $h$ ) from who is asking ( $e_d$ ).

To enforce the conditional capacity separation, we activate only the Top- $k$  experts. The sparse gating weights  $g$  are computed via a softmax normalization over the selected experts:

$$g_i = \frac{\exp(s_i) \cdot \mathbb{1}[i \in \text{Top-}k(s)]}{\sum_{j=1}^N \exp(s_j) \cdot \mathbb{1}[j \in \text{Top-}k(s)]} \quad (4)$$

Guided by the latent cultural topology learned in  $W_r$ , the router directs divergent cultural modes to distinct expert subsets, thereby structurally isolating conflicting gradients and preventing interference.

### 3.3 Mixture of Cultural Adapters

To enable fine-grained adaptation while preserving general reasoning, we freeze the backbone weights  $W_0 \in \mathbb{R}^{d_{out} \times d_{in}}$  and adopt a modular parameter-efficient strategy. We instantiate the expert pool using Low-Rank Adaptation (LoRA) (Hu et al., 2021), chosen for its proven stability and efficiency in large-scale fine-tuning tasks.

Formally, a standard LoRA module modulates the frozen weights by learning a low-rank update  $\Delta W = BA$ , where  $B \in \mathbb{R}^{d_{out} \times r}$  and  $A \in \mathbb{R}^{r \times d_{in}}$  are trainable matrices with rank  $r \ll \min(d_{in}, d_{out})$ . We extend this formulation to a Mixture of LoRA Experts. We initialize  $N$  distinct expert modules, denoted as  $\{(A_i, B_i)\}_{i=1}^N$ . Guided by the sparse routing weights  $g$  (Eq. 4), the forward pass for a hidden state  $h$  becomes:

$$h' = W_0 h + \sum_{i=1}^N g_i \cdot \underbrace{(B_i A_i h)}_{\text{Expert } i} \quad (5)$$

CUMA constructs a demographic-aware update  $\Delta W(d) = \sum g_i(d) B_i A_i$ . This ensures that conflicting cultural values are processed by separate parameter combinations, directly preventing the gradient interference that causes mean collapse.

### 3.4 Optimization Objectives

CUMA adopts a flexible optimization strategy designed to accommodate varying data granularities. The training process establishes foundational alignment via Conditional Supervised Fine-Tuning (SFT), which can be further refined through Conditional Preference Optimization when preference annotations or group-based rewards are available. The complete training procedure, detailing the curriculum transition and objective selection, is summarized in Appendix C.

Accordingly, the generalized objective function is a weighted combination of the active task loss and an auxiliary load-balancing regularization:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_{\text{lb}} \mathcal{L}_{\text{lb}} \quad (6)$$

where  $\mathcal{L}_{\text{task}}$  corresponds to either the SFT, DPO, or GRPO objective depending on the training stage.

We provide the detailed formulations for each objective component and the full training algorithm in Appendix C.

## 4 Experimental Setup

Our experiments are designed to investigate the nature of cultural sparsity and evaluate the efficacy of conditional capacity separation. Specifically, we aim to answer the following three research questions (**RQs**):

- **RQ1:** Can CUMA achieve superior cultural alignment compared to dense baselines across diverse benchmarks, and how does it perform under varying data scales?
- **RQ2:** How does CUMA mitigate *mean collapse* to avoid the generic, uncertain response patterns of dense models, and to what extent does it preserve the intrinsic diversity of cultural value distributions?
- **RQ3:** Does the demographic-aware router successfully capture the latent cultural topology and enable generalization to unseen demographic groups?

### 4.1 Datasets and Metrics

We evaluate CUMA on three benchmarks using a 10:1 train/test split; see Appendix D.4 for detailed statistics.

**WorldValuesBench (WVB):** Derived from the World Values Survey, this benchmark evaluates value prediction across distinct cultural regions (Zhao et al., 2024). Given a demographic profile, the model predicts the value stance on a multiple-choice scale. **Metrics:** We report Accuracy and Macro-F1. Additionally, acknowledging the ordinal nature of Likert-scale responses (Zhao et al., 2024), we report the Wasserstein-1 Distance (e.g., Earth Mover’s Distance (EMD)). This metric quantifies the structural divergence between the model’s predicted probabilities and the human value distribution, where a lower distance indicates superior alignment.

**Community Alignment (CA):** This dataset (Zhang et al., 2025a) captures conflicting preferences of diverse social groups on controversial topics. We evaluate two sub-tasks: preference prediction and response generation. **Metrics:** We use

Accuracy and Macro-F1 for prediction. For generation, we employ a GPT-4o-based<sup>1</sup> judge to compute the pairwise Win-Rate (details in Appendix D.5). An expert audit on a 100-sample subset with five annotators confirms strong agreement with the automated judge (Cohen’s  $\kappa = 0.84$ ; see Appendix G). We specifically evaluate the preference-optimized models (SFT+DPO and SFT+GRPO) against the base model to assess alignment validity.

**PRISM:** PRISM (Kirk et al., 2024) links fine-grained individual profiles to open-ended, multi-turn conversations. **Metrics:** We report the Win-Rate, adopting the identical evaluation setting as the CA generation task.

### 4.2 Baselines

We compare CUMA against three categories of alignment strategies to isolate performance sources.

**Inference-Time Baselines.** These methods steer the base model without parameter updates. We consider: (1) Vanilla Baseline, the unaligned base model representing default pre-training bias; (2) Persona Prompting (Lutz et al., 2025), which prepends a demographic-specific system prompt; and (3) Prompt Steering (Miehling et al., 2025), employing  $k$ -shot ( $k = 3$ ) demonstrations retrieved from matching demographics to guide the model via analogy (see Appendix D.5).

**Dense Fine-Tuning.** These methods update a single set of global parameters on the combined multi-cultural dataset. We include: (1) Full Fine-Tuning (FFT), updating 100% of parameters; (2) P-Tuning v2 (Liu et al., 2022), which optimizes deep prompt vectors; (3) LoRA (Hu et al., 2021), standard Low-Rank Adaptation ( $r = 64$ ); and (4) DoRA (Liu et al., 2024), which decomposes weights into magnitude and direction components. These methods represent the "one-size-fits-all" parameterization, which we hypothesize is structurally prone to mean collapse.

**Sparingly Activated Adapters.** We compare against state-of-the-art MoE-LoRA architectures including (1) MixLoRA (Li et al., 2024b) and (2) HydraLoRA (Tian et al., 2024). These models utilize sparse parameter structures but route based solely on semantic hidden states. We include them

<sup>1</sup>Model version: gpt-4o-2024-11-13.

to verify whether semantic routing alone is sufficient to resolve cultural conflicts, or if explicit demographic conditioning (as in CUMA) is necessary.

### 4.3 Implementation Details

We implement CUMA on two backbones: Llama-3.1-8B-Instruct (Grattafiori et al., 2024) and Qwen3-8B (Yang et al., 2025). We utilize a frozen Qwen3-Embedding-0.6B (Zhang et al., 2025b) as the demographic encoder. All models are trained on NVIDIA RTX PRO 6000 GPUs. We employ the AdamW optimizer with a cosine decay schedule. For CUMA, we set the number of experts  $N = 8$  with Top- $k = 2$  routing, applying LoRA adapters ( $r = 8/64$ ). Detailed hyperparameters and prompt templates are provided in Appendix D.

## 5 Results and Analysis

In this section, we present empirical findings addressing our research questions. We first evaluate CUMA’s overall efficacy against baselines (RQ1), then analyze its ability to mitigate mean collapse and preserve diversity (RQ2). We further investigate the learned latent topology and its generalization capabilities (RQ3), concluding with ablation studies on key architectural components.

### 5.1 Overall Alignment Performance

Table 1 summarizes results across three benchmarks, showing consistent trends for both Llama-3.1-8B and Qwen3-8B.

**Structural Limitations of Dense Models.** Dense methods (FFT, LoRA, DoRA) show a distinct performance ceiling. On Llama-3.1 WVB, even Full Fine-Tuning (45.25% Acc) lags significantly behind CUMA (50.46% Acc). This saturation indicates a structural bottleneck: the "one-size-fits-all" parameterization suffers from gradient interference when optimizing for conflicting values, forcing convergence towards an averaged solution rather than distinct cultural modes.

**Efficiency of Demographic Conditioning.** CUMA proves that alignment depends on routing precision, not just parameter scale. The low-rank variant ( $r=8$ , 1.53% params) consistently outperforms the larger HydraLoRA (2.31% params), e.g., +2.4% Acc on Llama-3.1 WVB. This confirms that conditioning routing on demographic topology allocates capacity more effectively than

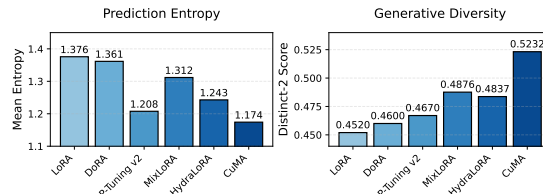


Figure 3: **Quantitative Verification of Mean Collapse.** (Left) Dense baselines (e.g., LoRA, DoRA) exhibit high prediction entropy ( $H \approx 1.38$ ), indicating probability mass dispersion typical of mean collapse. CUMA significantly reduces uncertainty ( $H \approx 1.17$ ). (Right) In open-ended generation, CUMA achieves the highest Distinct-2 score, confirming that it avoids repetitive, generic templates by accessing specialized cultural vocabularies.

semantic-only MoEs, achieving superior results with fewer parameters.

**Mitigating Semantic Stereotyping.** A critical divergence appears between Accuracy and EMD in baselines. Semantic sparse methods (MixLoRA, HydraLoRA) achieve competitive Accuracy but suffer high EMD (e.g., 0.28 vs. 0.19 for CUMA on Qwen3). This "High-Accuracy, High-EMD" pattern suggests "stereotyping": models predict the mode based on semantics but miss the nuanced probability spread. CUMA’s superior EMD indicates it successfully models the diverse shape of human value distributions rather than memorizing stereotypes.

**Holistic Alignment across Modalities.** This distributional fidelity translates to robust generation. With DPO/GRPO, CUMA achieves dominant Win-Rates on CA (78.2%) and PRISM (76.8%) with Qwen3, surpassing dense baselines ( $\approx 65\%$ ). This verifies CUMA’s ability to map latent values into coherent, culturally aligned responses.

### 5.2 Verification of Mean Collapse

To address RQ2, we employ Prediction Entropy (WVB) and Distinct-2 scores (CA-generation/PRISM) to diagnose mean collapse.

As shown in Figure 3, dense models exhibit high entropy ( $H_{\text{mean}} \approx 1.38$ ), reflecting the "diluted middle" behavior predicted in Appendix B.3. In the context of next-token prediction, high entropy indicates a flat, indecisive distribution where the model hedges across options rather than committing to a culturally specific response, which is a symptom of variance inflation from merging conflicting modes. CUMA reduces entropy to 1.17, indicat-

Category	Method	Trainable Params	WorldValuesBench (WVB)			Community Alignment (CA)				PRISM	
			Acc $\uparrow$	Macro-F1 $\uparrow$	EMD $\downarrow$	Acc $\uparrow$	Macro-F1 $\uparrow$	Win-Rate vs Base (DPO)	Win-Rate vs Base (GRPO)	Win-Rate vs Base (DPO)	Win-Rate vs Base (GRPO)
<b>BACKBONE: LLAMA-3.1-8B</b>											
<i>Inference-Time Strategies</i>	Vanilla Baseline	0.00%	32.42	22.99	0.3967	26.70	20.79	-	-	-	-
	Persona Prompting	0.00%	37.06	23.90	0.3105	26.10	21.57	55.5%	56.2%	55.2%	55.8%
	Prompt Steering (3-shot)	0.00%	27.50	11.14	0.2507	26.80	22.74	56.8%	57.5%	56.5%	59.2%
<i>Dense Fine-Tuning</i>	Full Fine-Tuning (FFT)	100.0%	45.25	<u>30.50</u>	0.2205	45.15	32.30	63.5%	65.2%	61.5%	63.2%
	P-Tuning v2	0.94%	43.80	29.10	0.2470	43.50	30.85	57.2%	58.8%	55.5%	56.8%
	LoRA	0.37%	34.30	22.37	0.2537	38.53	30.50	60.5%	62.1%	58.8%	59.5%
	DoRA	0.38%	36.50	25.10	0.2587	39.20	31.50	61.8%	63.5%	59.5%	61.2%
<i>Sparsely Activated Adapters</i>	MixLoRA	3.01%	45.20	29.80	0.2440	46.80	34.60	66.5%	68.2%	64.2%	65.8%
	HydraLoRA	2.31%	46.50	29.90	0.2350	47.90	36.20	<u>69.8%</u>	<u>69.5%</u>	65.5%	<u>68.2%</u>
	CUMA ( $r = 8$ )	1.53%	<u>48.90</u>	<u>30.50</u>	<u>0.1903</u>	<u>50.12</u>	<u>38.50</u>	68.5%	<u>73.8%</u>	<u>68.8%</u>	67.5%
	CUMA	4.15%	<b>50.46</b>	<b>32.50</b>	<b>0.1870</b>	<b>52.45</b>	<b>40.12</b>	<b>72.2%</b>	<b>74.5%</b>	<b>71.2%</b>	<b>73.5%</b>
<b>BACKBONE: QWEN3-8B</b>											
<i>Inference-Time Strategies</i>	Vanilla Baseline	0.00%	31.68	18.92	0.3851	31.20	17.75	-	-	-	-
	Persona Prompting	0.00%	34.92	21.05	0.2864	32.80	21.00	57.1%	58.5%	56.2%	57.0%
	Prompt Steering (3-shot)	0.00%	28.08	12.36	0.2299	26.00	22.19	59.5%	60.8%	58.4%	59.5%
<i>Dense Fine-Tuning</i>	Full Fine-Tuning (FFT)	100.0%	45.54	28.21	0.2228	49.50	36.20	66.8%	68.5%	63.5%	65.2%
	P-Tuning v2	0.94%	45.04	28.17	0.2358	47.50	34.80	59.5%	61.2%	57.5%	58.8%
	LoRA	0.37%	40.06	22.02	0.2700	38.53	30.50	63.2%	65.5%	61.5%	62.2%
	DoRA	0.38%	42.78	24.73	0.2773	39.20	31.50	64.5%	66.8%	62.8%	64.1%
<i>Sparsely Activated Adapters</i>	MixLoRA	3.01%	43.50	26.44	0.2904	51.50	38.80	70.5%	72.8%	67.5%	69.2%
	HydraLoRA	2.31%	45.36	28.12	0.2793	52.80	40.20	71.5%	73.6%	68.5%	70.4%
	CUMA ( $r = 8$ )	1.53%	<u>49.02</u>	<u>29.70</u>	<u>0.1980</u>	<u>55.40</u>	<u>43.10</u>	<u>75.8%</u>	<u>76.5%</u>	<u>73.2%</u>	<u>75.5%</u>
	CUMA	4.15%	<b>50.64</b>	<b>31.50</b>	<b>0.1876</b>	<b>57.20</b>	<b>44.80</b>	<b>77.5%</b>	<b>78.2%</b>	<b>74.5%</b>	<b>76.8%</b>

Table 1: **Main Results on Cultural Alignment Benchmarks.** Comparison of CUMA against static, dense, and sparse baselines across two backbones: **Llama-3.1-8B** and **Qwen3-8B**. **Trainable Params** denotes the exact percentage of trainable parameters relative to the base model. Standard LoRA, DoRA, and CUMA imply rank  $r = 64$  unless specified otherwise ( $r = 8$ ). For Win-Rates, we report results after DPO and GRPO stages respectively. **Bold** indicates the best performance, and underline indicates the second best performance.

ing sharper, more decisive alignment for each demographic profile. Crucially, this per-profile decisiveness does not sacrifice cross-profile diversity: CUMA achieves a Distinct-2 score of 0.52, outperforming dense baselines ( $\approx 0.45$ ), confirming that it produces distinct, culturally resonant outputs rather than repetitive generic templates.

### 5.3 Latent Cultural Topology and Generalization

To address **RQ3**, we investigate the learned geometric representation and its generalization potential.

**Visualizing the Latent Topology.** Figure 4 visualizes expert activation patterns for 65 countries via t-SNE. The router spontaneously organizes demographics into clusters aligning with sociological frameworks (e.g., Inglehart–Welzel (Inglehart and Welzel, 2005)), such as the African-Islamic bloc and Confucian sphere. This confirms the construction of a latent cultural topology, where groups with shared value affinities share model capacity without explicit supervision.

### Quantitative Verification: Zero-Shot Transfer.

We validate generalization by evaluating on held-out demographic profiles (Table 2). Despite lacking supervision for these specific profiles, CUMA exhibits robust topological transfer, with an average accuracy drop of only 2.12% and minimal EMD increase (+0.0244). The English-Speaking cluster shows the smallest drop (-1.67%), benefiting from richer pre-training representation and higher sociological compactness on the Inglehart-Welzel cultural map (Inglehart and Welzel, 2005). Even the more dispersed African-Islamic cluster degrades only marginally (-2.36%).

To contextualize these results, Table 3 compares CUMA’s zero-shot performance against all baselines under the same leave-one-cluster-out protocol. CUMA achieves 48.10% zero-shot accuracy, significantly outperforming the strongest zero-shot baseline (HydraLoRA, 36.2%) and even surpassing the full-supervision accuracy of several dense baselines (e.g., FFT: 45.54%; LoRA: 40.06%).

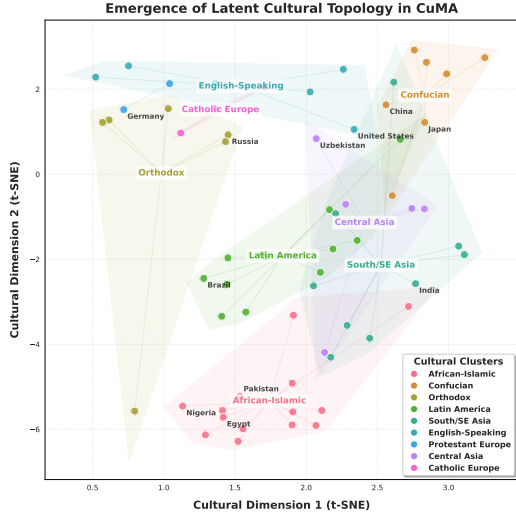


Figure 4: **Emergence of Latent Cultural Topology.** t-SNE projection of expert activation patterns across 65 nations. Without explicit supervision, the router spontaneously organizes demographic profiles into coherent clusters that align with sociological frameworks (e.g., the *African-Islamic* and *Confucian* spheres). This geometric structure facilitates zero-shot generalization by routing unseen demographic profiles to experts trained on culturally proximate groups. Details on the visualization protocol are provided in Appendix E.

#### 5.4 Ablation Studies

We validate CuMA’s components on Qwen3-8B by ablating the demographic routing branch ( $e_d$ ), semantic routing ( $h$ ), and auxiliary load balancing loss ( $\mathcal{L}_{aux}$ ). Table 4 summarizes the results.

Results demonstrate the synergy between semantic and demographic signals. Removing demographic routing ( $w/o$  Demo.) acts as a standard semantic MoE, dropping accuracy by 3.56%. This confirms that resolving cultural conflict requires explicit demographic conditioning. Conversely, in the  $w/o$  Semantic Routing setting, we replace the demographic-specific prompt with a generic instruction ("You are a helpful assistant that answers survey questions honestly"), forcing reliance solely on demographic embeddings. This causes a larger accuracy drop (-6.38%), yet still significantly outperforms the random baseline (Full Cancellation), proving that the router successfully captures latent value priors solely from the demographic topology. Finally, removing the auxiliary balancing loss ( $w/o$  Demo. & Bal. Loss) spikes EMD (0.1876  $\rightarrow$  0.2657), indicating that structural regularization is critical for preventing mode collapse and ensuring effective expert utilization. We further analyze the impact of the routing strategy (e.g., Soft vs. Top-k

Cultural Cluster	Full Sup.		Zero-Shot		Gap ( $\Delta$ )	
	Acc $\uparrow$	EMD $\downarrow$	Acc $\uparrow$	EMD $\downarrow$	Acc	EMD
African-Islamic	53.81	0.2244	51.45	0.2510	-2.36	+0.0266
Catholic Europe	47.98	0.2110	45.82	0.2350	-2.16	+0.0240
Central Asia	49.72	0.2808	47.55	0.3090	-2.17	+0.0282
Confucian	48.71	0.2387	46.60	0.2640	-2.11	+0.0253
English-Speaking	50.82	0.1970	49.15	0.2150	-1.67	+0.0180
Latin America	49.77	0.2568	47.65	0.2810	-2.12	+0.0242
Orthodox	49.87	0.2368	47.90	0.2610	-1.97	+0.0242
Protestant Europe	50.57	0.2182	48.35	0.2410	-2.22	+0.0228
South/SE Asia	50.39	0.2316	48.10	0.2580	-2.29	+0.0264
Average	50.18	0.2328	48.06	0.2572	-2.12	+0.0244

Table 2: **Zero-Shot Cross-Cultural Generalization.** Results of the zero-shot generalization experiment across 9 cultural clusters. **Full Sup.** indicates standard training, while **Zero-Shot** evaluates performance on held-out demographic profiles excluded during training. The results are aggregated by the cultural cluster of the unseen profiles. **Gap ( $\Delta$ )** denotes the performance difference between Full Supervision and Zero-Shot. The minimal degradation (Avg  $\Delta_{Acc} \approx -2.1\%$ ) confirms that CuMA effectively generalizes to unseen cultures by leveraging the latent topology. See Appendix E for experimental details.

Method	Full Sup.		Zero-Shot		Gap ( $\Delta$ )	
	Acc $\uparrow$	EMD $\downarrow$	Acc $\uparrow$	EMD $\downarrow$	Acc	EMD
LoRA	40.06	0.2700	32.40	0.3250	-7.66	+0.0550
DoRA	42.78	0.2773	33.12	0.3180	-9.66	+0.0407
P-Tuning v2	45.04	0.2358	30.50	0.3320	-14.54	+0.0962
MixLoRA	43.50	0.2904	35.80	0.2820	-7.70	-0.0084
HydraLoRA	45.36	0.2793	36.24	0.2750	-9.12	-0.0043
CuMA (ours)	50.64	0.1876	48.10	0.2100	-2.54	+0.0224

Table 3: **Zero-Shot Generalization vs. Baselines on WVB (Qwen3-8B).** All methods are evaluated under the same leave-one-cluster-out protocol. CuMA’s zero-shot accuracy (48.10%) surpasses the full-supervision accuracy of most dense baselines, demonstrating the effectiveness of the learned latent topology.

Routing) in Appendix F, finding that strict capacity separation (Top-k) is essential for resolving cultural interference.

## 6 Related Work

Existing alignment paradigms typically prioritize universal attributes (Ouyang et al., 2022; Rafailov et al., 2023), often leading to "Algorithmic Monoculture" (Zhang et al., 2025a). While recent pluralistic alignment methods (Li et al., 2024a; Xu et al., 2025; Kirk et al., 2024; Wang et al., 2024b) attempt to incorporate diverse values, they largely rely on dense parameterizations. Even when utilizing parameter-efficient variations such as LoRA (Hu et al., 2021) and DoRA (Liu et al., 2024), these

Method	Acc $\uparrow$	Macro-F1 $\uparrow$	EMD $\downarrow$
<b>CUMA (Full)</b>	<b>50.64</b>	<b>31.50</b>	<b>0.1876</b>
w/o Demographic Routing	47.08	29.98	0.1965
w/o Demo. & Bal. Loss	45.26	27.49	0.2657
w/o Semantic Routing	44.26	22.99	0.3060
Full Cancellation	32.15	19.25	0.3518

Table 4: **Ablation Studies on WVB (Qwen3-8B)**. We evaluate the impact of removing demographic routing, semantic routing, and the load balancing loss. "w/o Demo. & Bal. Loss" represents a naive semantic MoE without auxiliary loss. "Full Cancellation" denotes the removal of all routing mechanisms and demographic prompts.

methods remain fundamentally "dense" by sharing a unified weight space, which we argue renders them structurally vulnerable to gradient interference and mean collapse.

To address this, we draw upon Mixture-of-Experts (MoE) architectures (Shazeer et al., 2017). Unlike recent PEFT-MoE approaches (Li et al., 2024b; Tian et al., 2024) that rely on semantic or task-specific routing to enhance multi-task competence, CUMA re-purposes MoE for conditional capacity separation. By conditioning routing on demographic topology, we isolate conflicting cultural gradients, preventing the homogenization of distinct cultural values. A comprehensive review of related work is provided in Appendix A.

## 7 Conclusion

We introduced CUMA, a framework that reformulates cultural alignment as a conditional capacity separation problem. By using demographic-aware routing, CUMA learns a *Latent Cultural Topology* to disentangle conflicting gradients and resolve *Mean Collapse*. Results across three benchmarks show significant gains: CUMA reduces distributional divergence (EMD) to 0.1876 and outperforms dense baselines by over 5% in accuracy. It also achieves dominant Win-Rates on Community Alignment (78.2%) and PRISM (76.8%). These findings suggest that respecting the sparsity of cultural values is key to building truly pluralistic LLMs.

## Limitations

While CUMA demonstrates significant improvements in cultural alignment, several limitations remain. First, the framework relies on explicit demographic profiles to guide the routing mechanism.

In real-world scenarios, such information may be incomplete, inaccurate, or unavailable due to privacy constraints. Second, our experiments utilized a fixed number of experts ( $N = 8$ ). While this capacity proved sufficient for the benchmarks studied, capturing the full complexity of global cultural diversity may require more granular expert pools or hierarchical routing structures. Third, although CUMA generalizes well to unseen demographic groups, its performance is still bounded by the coverage and potential biases of the underlying training datasets (WVB, CA, and PRISM). Finally, the MoE-based architecture introduces modest inference overhead (see Appendix H for detailed profiling). Future work will explore implicit demographic inference and dynamic expert allocation to further enhance the flexibility of pluralistic alignment.

## Acknowledgments

Yuheng Jia was supported by the National Natural Science Foundation of China (Grant U24A20322, Grant 62576094). Shu Su was supported by the National Natural Science Foundation of China (Grant 72371072). This research work was also supported by the Big Data Computing Center of Southeast University.

## References

- Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O’Neill, Ashutosh Modi, and Monojit Choudhury. 2024. [Towards measuring and modeling “culture” in LLMs: A survey](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15763–15784, Miami, Florida, USA. Association for Computational Linguistics.
- Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. [Large language models for mathematical reasoning: Progresses and challenges](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 225–237, St. Julian’s, Malta. Association for Computational Linguistics.
- Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. [Investigating cultural alignment of large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.

- Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen McAleer, Albert Q. Jiang, Jia Deng, Stella Biderman, and Sean Welleck. 2024. [Llemma: An open language model for mathematics](#). *Preprint*, arXiv:2310.10631.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, and 39 others. 2021. [Evaluating large language models trained on code](#). *Preprint*, arXiv:2107.03374.
- Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17*, page 4302–4310, Red Hook, NY, USA. Curran Associates Inc.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. [Training verifiers to solve math word problems](#). *Preprint*, arXiv:2110.14168.
- Esin Durmus, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2024. [Towards measuring the representation of subjective global opinions in language models](#). *Preprint*, arXiv:2306.16388.
- Evan Frick. 2025. [Reward modeling for human preferences](#). Master's thesis, EECS Department, University of California, Berkeley, May.
- Yao Fu and Mirella Lapata. 2022. [Latent topology induction for understanding contextualized representations](#). *Preprint*, arXiv:2206.01512.
- Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2024. [Large language models empowered agent-based modeling and simulation: a survey and perspectives](#). *Humanities and Social Sciences Communications*, 11(1):1259.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Yanggan Gu, Yuanyi Wang, Zhaoyi Yan, Yiming Zhang, Qi Zhou, Fei Wu, and Hongxia Yang. 2025. [Infifpo](#)
- [Implicit model fusion via preference optimization in large language models](#). *Preprint*, arXiv:2505.13878.
- Joseph Henrich, Steven J. Heine, and Ara Norenzayan. 2010. [The weirdest people in the world?](#) *Behavioral and Brain Sciences*, 33(2–3):61–83.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. [Lora: Low-rank adaptation of large language models](#). *Preprint*, arXiv:2106.09685.
- Ronald Inglehart and Christian Welzel. 2005. *Modernization, Cultural Change, and Democracy: The Human Development Sequence*. Cambridge University Press.
- 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)*, Vienna, Austria. Association for Computational Linguistics.
- Chengyi Ju, Weijie Shi, Chengzhong Liu, Jiaming Ji, Jipeng Zhang, Ruiyuan Zhang, Jiajie Xu, Yaodong Yang, Sirui Han, and Yike Guo. 2025. [Benchmarking multi-national value alignment for large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20042–20058.
- Mehdi Khamassi, Marceau Nahon, and Raja Chatila. 2024. [Strong and weak alignment of large language models with human values](#). *Scientific Reports*, 14(1):19399.
- Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. 2024. [The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 105236–105344. Curran Associates, Inc.
- Ekaterina Kostina, Larisa Kretova, Raisa Teleshova, Anna Tsepkova, and Timur Vezirov. 2015. [Universal human values: Cross-cultural comparative analysis](#). *Procedia - Social and Behavioral Sciences*, 214:1019–1028. Worldwide trends in the development of education and academic research, Sofia, Bulgaria, 15-18 June, 2015.
- Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024a. [Culturellm: Incorporating cultural differences into large language models](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 84799–84838. Curran Associates, Inc.
- Dengchun Li, Yingzi Ma, Naizheng Wang, Zhengmao Ye, Zhiyuan Cheng, Yinghao Tang, Yan Zhang, Lei Duan, Jie Zuo, Cal Yang, and Mingjie Tang. 2024b.

- Mixlor: Enhancing large language models fine-tuning with lora-based mixture of experts. *Preprint*, arXiv:2404.15159.
- Jiyi Li. 2024. A comparative study on annotation quality of crowdsourcing and llm via label aggregation. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6525–6529.
- Tianyi Li, Divya Sree, and Tatiana Ringenber. 2025. Assessing crowdsourced annotations with LLMs: Linguistic certainty as a proxy for trustworthiness. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 191–201, Albuquerque, USA. Association for Computational Linguistics.
- Ziyue Li and Tianyi Zhou. 2024. Your mixture-of-experts llm is secretly an embedding model for free. *Preprint*, arXiv:2410.10814.
- Haijiang Liu, Yong Cao, Xun Wu, Chen Qiu, Jinguang Gu, Maofu Liu, and Daniel Hershcovich. 2025. Towards realistic evaluation of cultural value alignment in large language models: Diversity enhancement for survey response simulation. *Information Processing & Management*, 62(4):104099.
- Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. Dora: weight-decomposed low-rank adaptation. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.
- Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-tuning v2: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68. Association for Computational Linguistics.
- Haoran Lu, Luyang Fang, Ruidong Zhang, Xinliang Li, Jiazhang Cai, Huimin Cheng, Lin Tang, Ziyu Liu, Zeliang Sun, Tao Wang, Yingchuan Zhang, Arif Hassan Zidan, Jinwen Xu, Jincheng Yu, Meizhi Yu, Hanqi Jiang, Xilin Gong, Weidi Luo, Bolun Sun, and 31 others. 2025. Alignment and safety in large language models: Safety mechanisms, training paradigms, and emerging challenges. *Preprint*, arXiv:2507.19672.
- Marlene Lutz, Indira Sen, Georg Ahnert, Elisa Rogers, and Markus Strohmaier. 2025. The prompt makes the person(a): A systematic evaluation of sociodemographic persona prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23212–23237, Suzhou, China. Association for Computational Linguistics.
- Anemily Machina and Robert Mercer. 2024. Anisotropy is not inherent to transformers. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Mexico City, Mexico. Association for Computational Linguistics.
- Erik Miebling, Michael Desmond, Karthikeyan Natesan Ramamurthy, Elizabeth M. Daly, Pierre Dognin, Jesus Rios, Djallel Bouneffouf, and Miao Liu. 2025. Evaluating the prompt steerability of large language models. *Preprint*, arXiv:2411.12405.
- Juhyun Oh, Inha Cha, Michael Saxon, Hyunseung Lim, Shaily Bhatt, and Alice Oh. 2025. Culture is everywhere: A call for intentionally cultural evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 19156–19168, Suzhou, China. Association for Computational Linguistics.
- Fırat Öncel, Matthias Bethge, Beyza Ermis, Mirco Ravanelli, Cem Subakan, and Çağatay Yıldız. 2024. Adaptation odyssey in LLMs: Why does additional pretraining sometimes fail to improve? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19834–19843, Miami, Florida, USA. Association for Computational Linguistics.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.
- Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cino Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR.
- Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.
- Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. A roadmap to pluralistic alignment. *Preprint*, arXiv:2402.05070.
- Nicholas Sukiennik, Chen Gao, Fengli Xu, and Yong Li. 2025. An evaluation of cultural value alignment in llm. *Preprint*, arXiv:2504.08863.
- 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.

- Chunlin Tian, Zhan Shi, Zhijiang Guo, Li Li, and Chengzhong Xu. 2024. [Hydralora: An asymmetric lora architecture for efficient fine-tuning](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 9565–9584. Curran Associates, Inc.
- Siqi Wang, Zhengyu Chen, Bei Li, Keqing He, Min Zhang, and Jingang Wang. 2024a. [Scaling laws across model architectures: A comparative analysis of dense and MoE models in large language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5583–5595, Miami, Florida, USA. Association for Computational Linguistics.
- Yuhang Wang, Yanxu Zhu, Chao Kong, Shuyu Wei, Xiaoyuan Yi, Xing Xie, and Jitao Sang. 2024b. [CDEval: A benchmark for measuring the cultural dimensions of large language models](#). In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 1–16, Bangkok, Thailand. Association for Computational Linguistics.
- 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.
- Wanqi Xue, Bo An, Shuicheng Yan, and Zhongwen Xu. 2024. [Reinforcement learning from diverse human preferences](#). In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 5298–5306. International Joint Conferences on Artificial Intelligence Organization. Main Track.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Lily Hong Zhang, Smitha Milli, Karen Jusko, Jonathan Smith, Brandon Amos, Wassim Bouaziz, Manon Revel, Jack Kussman, Yasha Sheynin, Lisa Titus, Bhaktipriya Radharapu, Jane Yu, Vidya Sarma, Kris Rose, and Maximilian Nickel. 2025a. [Cultivating pluralism in algorithmic monoculture: The community alignment dataset](#). *Preprint*, arXiv:2507.09650.
- Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025b. [Qwen3 embedding: Advancing text embedding and reranking through foundation models](#). *Preprint*, arXiv:2506.05176.
- Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2025c. [The lessons of developing process reward models in mathematical reasoning](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10495–10516, Vienna, Austria. Association for Computational Linguistics.
- Wenlong Zhao, Debanjan Mondal, Niket Tandon, Danica Dillion, Kurt Gray, and Yuling Gu. 2024. [World-ValuesBench: A large-scale benchmark dataset for multi-cultural value awareness of language models](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17696–17706, Torino, Italia. ELRA and ICCL.
- Xuandong Zhao, Will Cai, Tianneng Shi, David Huang, Licong Lin, Song Mei, and Dawn Song. 2025. [Improving llm safety alignment with dual-objective optimization](#). In *Proceedings of the 42nd International Conference on Machine Learning, ICML'25*. JMLR.org.
- Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Y. Zhao, Andrew Dai, Zhifeng Chen, Quoc Le, and James Laudon. 2022. [Mixture-of-experts with expert choice routing](#). In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22*, Red Hook, NY, USA. Curran Associates Inc.
- Xiaoxuan Zhu, Zhouhong Gu, Baiqian Wu, Suhang Zheng, Tao Wang, Tianyu Li, Hongwei Feng, and Yanghua Xiao. 2025. [Toremi: Topic-aware data reweighting for dynamic pre-training data selection](#). *Preprint*, arXiv:2504.00695.

## A Extended Related Work

### A.1 From Universal to Pluralistic Alignment

The dominant paradigm in LLM alignment has prioritized universal attributes such as helpfulness and safety, typically optimized via Reinforcement Learning from Human Feedback (RLHF) or Direct Preference Optimization (DPO) (Ouyang et al., 2022; Rafailov et al., 2023). While effective for objective tasks, this "one-size-fits-all" approach fails to encompass the normative diversity of global users, often collapsing into a specific Western-centric value system, a phenomenon termed "Algorithmic Monoculture" (Zhang et al., 2025a).

In response, recent research has pivoted towards *pluralistic alignment*. This transition is supported by emerging evaluation frameworks: CDEval (Wang et al., 2024b) and NaVAB (Ju et al., 2025) assess cultural knowledge and bias, while PRISM (Kirk et al., 2024) links fine-grained sociodemographics to interactive preferences. On

the methodological front, approaches like CultureLLM (Li et al., 2024a) utilize semantic data augmentation, and CultureSPA (Xu et al., 2025) employs contrastive learning to distinguish cultural norms. Others have explored personalization, predicting individual value judgments from historical context (Jiang et al., 2025).

However, a critical structural gap remains. Most existing methods treat cultural alignment as a data scale or prompting problem, attempting to inject pluralistic cultural values into a dense model. They overlook the inherent conflict arising from this multiplicity: since these values are often mutually exclusive, forcing a single set of parameters to represent them leads to gradient interference. Without structural separation, these methods remain vulnerable to mean collapse.

## A.2 Parameter-Efficient MoE for Value Disentanglement

To address parameter interference, Mixture-of-Experts (MoE) architectures have seen renewed interest, particularly when combined with Parameter-Efficient Fine-Tuning (PEFT). LoRA (Hu et al., 2021) provides a lightweight adaptation mechanism, while MoE scales capacity via conditional computation (Shazeer et al., 2017).

Recent innovations like MixLoRA (Li et al., 2024b) and HydraLoRA (Tian et al., 2024) integrate these paradigms, composing multiple LoRA adapters to handle diverse downstream tasks. While structurally similar to our approach, these methods employ experts as functional components to maximize multi-task competence. In contrast, CUMA re-purposes the MoE framework for structural value separation. We conceptualize experts not merely as skill specialists, but as culturally specialized parameter spaces that isolate conflicting cultural gradients. By conditioning routing on demographic topology rather than just semantic complexity, CUMA prevents the homogenization of distinct cultural perspectives, mitigating a key limitation in pluralistic alignment.

## B Derivations of Mean Collapse and Its Resolution

In Section 2.3, we qualitatively defined *Mean Collapse* as the convergence of a dense model to the statistical average of conflicting modes. In this appendix, we provide the rigorous mathematical derivation of this phenomenon under *Cultural Spar-*

*sity* and theoretically demonstrate how CUMA’s conditional routing resolves this structural limitation.

### B.1 Setup: The Mixture Problem

Let the true distribution of human values  $P_{\text{data}}(y)$  be a mixture of  $K$  distinct cultural modes. For analytical tractability, we approximate these modes as Gaussians. Consider a simplified case with two conflicting groups ( $K = 2$ ) with proportions  $\pi_1, \pi_2$  (where  $\pi_1 + \pi_2 = 1$ ):

$$P_{\text{data}}(y) = \pi_1 \mathcal{N}(y; \mu_1, \Sigma) + \pi_2 \mathcal{N}(y; \mu_2, \Sigma) \quad (7)$$

where  $\mu_1, \mu_2$  represent conflicting value centers in the feature space.

A standard dense model  $P_{\theta}(y|x, d)$  utilizes a monolithic parameter set  $\theta$  for all groups. Consequently, conflicting gradients from diverse groups interfere within the shared capacity. To analyze this structural tendency, we approximate the dense estimator as a single Gaussian  $\mathcal{N}(y; \mu_{\theta}, \Sigma_{\theta})$  optimized via the Forward Kullback-Leibler (KL) divergence:

$$\begin{aligned} \min_{\theta} D_{\text{KL}}(P_{\text{data}} \| P_{\theta}) \\ \iff \min_{\theta} \mathbb{E}_{y \sim P_{\text{data}}} [-\log P_{\theta}(y)] \end{aligned} \quad (8)$$

### B.2 Optimization Dynamics of Dense Models

We determine the optimal location parameter  $\mu_{\theta}^*$  by minimizing  $\mathcal{J}(\mu_{\theta}) = \mathbb{E}_{y \sim P_{\text{data}}} [-\log P_{\theta}(y)]$ . Substituting the Gaussian log-likelihood (ignoring constant terms):

$$\mathcal{J}(\mu_{\theta}) = \int P_{\text{data}}(y) \left[ \frac{1}{2} (y - \mu_{\theta})^{\top} \Sigma_{\theta}^{-1} (y - \mu_{\theta}) \right] dy \quad (9)$$

Taking the gradient with respect to  $\mu_{\theta}$  and setting it to zero:

$$\nabla_{\mu_{\theta}} \mathcal{J} = -\Sigma_{\theta}^{-1} \left( \int P_{\text{data}}(y) y dy - \mu_{\theta} \right) = 0 \quad (10)$$

Since  $\Sigma_{\theta}^{-1}$  is positive definite and  $\int P_{\text{data}}(y) dy = 1$ , we obtain:

$$\mu_{\theta}^* = \int P_{\text{data}}(y) y dy = \mathbb{E}_{P_{\text{data}}}[y] \quad (11)$$

Expanding the expectation over the mixture components:

$$\mu_{\theta}^* = \pi_1 \mu_1 + \pi_2 \mu_2 \quad (12)$$

□

This proves that the dense model strictly converges to the linearly weighted average of the modes. Regardless of the semantic distance between cultural groups, the single set of parameters is mathematically forced to the geometric center.

### B.3 Geometric Consequences under Cultural Sparsity

We now analyze the implications of this convergence when the data satisfies the *Cultural Sparsity* condition (large separation  $\delta = \|\mu_1 - \mu_2\|$ ).

**1. Probability Density Gap.** Assume a symmetric conflict where  $\pi_1 = \pi_2 = 0.5$  and  $\Sigma = I$ . The optimal dense mean lies at  $\mu_\theta^* = (\mu_1 + \mu_2)/2$ . The distance from this collapsed mean to a true mode is  $\|\mu_\theta^* - \mu_1\| = \delta/2$ .

The true probability density at the collapsed mean is:

$$\begin{aligned} P_{\text{data}}(\mu_\theta^*) &= \frac{1}{2}\mathcal{N}(\mu_\theta^*; \mu_1, I) + \frac{1}{2}\mathcal{N}(\mu_\theta^*; \mu_2, I) \\ &\propto \exp\left(-\frac{1}{2}\left\|\frac{\delta}{2}\right\|^2\right) = \exp\left(-\frac{\delta^2}{8}\right) \end{aligned} \quad (13)$$

In contrast, the density at a true mode (e.g.,  $\mu_1$ ) is dominated by the first component:

$$P_{\text{data}}(\mu_1) \approx \frac{1}{2}\mathcal{N}(\mu_1; \mu_1, I) \propto \frac{1}{2}\exp(0) = \frac{1}{2} \quad (14)$$

The likelihood ratio of the "average" response versus a culturally specific response decays exponentially:

$$\frac{P_{\text{data}}(\mu_\theta^*)}{P_{\text{data}}(\mu_1)} \approx 2 \exp\left(-\frac{\delta^2}{8}\right) \quad (15)$$

Under Cultural Sparsity (Eq. 2), where  $\delta$  significantly exceeds the ambient dimension ( $\delta^2 \gg m$ ), this ratio vanishes. The dense model effectively hallucinates a "safe middle" that corresponds to a low-density void in the cultural manifold.

**2. Variance Inflation.** Mean collapse also implies a loss of precision. By the law of total variance, the optimal covariance  $\Sigma_\theta^*$  for the dense model decomposes into two terms:

$$\begin{aligned} \Sigma_\theta^* &= \text{Var}_{P_{\text{data}}}[y] \\ &= \sum_k \pi_k \Sigma_k + \sum_k \pi_k (\mu_k - \mu_\theta^*)(\mu_k - \mu_\theta^*)^\top \end{aligned} \quad (16)$$

The second term scales quadratically with  $\delta$ . This forces the dense model to expand its probability mass to span distant modes, exhibiting Maximum Entropy behavior, generating generic, non-committal responses.

### B.4 Resolution via Conditional Routing

CUMA resolves this dilemma by introducing a conditioning variable  $d$  (demographics). The routing mechanism  $g(d)$  partitions the parameter space, modeling the conditional density:

$$P_{\text{CuMA}}(y|x, d) \approx \sum_i g_i(d) \mathcal{N}(y; \mu_i, \Sigma_i) \quad (17)$$

If the router successfully learns the topology (i.e.,  $g_k(d) \approx \mathbb{1}[d \in \text{Group}_k]$ ), the objective function decomposes into separate objectives for each expert. This allows each expert to converge to the true mode  $\mu_k$  and intrinsic covariance  $\Sigma_k$  of its respective group.

Crucially, the resulting variance for CUMA becomes:

$$\Sigma_{\text{CuMA}}^* \approx \sum_k \pi_k \Sigma_k \quad (18)$$

Comparing this to Eq. 16, CUMA explicitly eliminates the structural uncertainty term ( $\sum \pi_k (\mu_k - \mu_\theta^*)^2$ ). By removing this variance inflation, CUMA avoids the exponential density decay and maintains high fidelity to distinct cultural modes.

## C Detailed Optimization Objectives

In this section, we provide the detailed formulations for the optimization objectives. The complete training procedure is summarized in Algorithm 1.

**1. Conditional SFT.** For standard instruction following and knowledge injection, we minimize the negative log-likelihood conditioned on the demographic profile  $d$ :

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(x,y,d) \sim \mathcal{D}_{\text{SFT}}} [\log P_\theta(y | x, d)] \quad (19)$$

**2. Conditional Preference Optimization.** To sharpen the decision boundaries between cultural modes and explicitly penalize mean collapse, we align the model with human preferences. Depending on the available data format, we employ one of the following objectives:

**Option A: Conditional DPO.** When pairwise preference data ( $y_w, y_l$ ) is available, we apply Direct Preference Optimization (DPO). Our objective

contrasts a chosen response  $y_w$  against a rejected response  $y_l$  under the *same* demographic profile  $d$ :

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E} \left[ \log \sigma \left( \beta \log \frac{P_\theta(y_w|x, d)}{P_{\text{ref}}(y_w|x, d)} - \beta \log \frac{P_\theta(y_l|x, d)}{P_{\text{ref}}(y_l|x, d)} \right) \right] \quad (20)$$

Crucially, the rejected response  $y_l$  often represents a "neutral" or "mode-covering" output. Optimizing this margin forces CUMA to separate the conditional distributions, pushing the router to activate distinct experts for conflicting values.

**Option B: Conditional GRPO.** For scenarios allowing multiple valid outputs or reasoning paths, we employ Group Relative Policy Optimization (GRPO). For each input  $(x, d)$ , GRPO samples a group of outputs  $\{y_1, \dots, y_G\}$  and optimizes the policy based on group-relative advantages without a value function critic. The objective is:

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\frac{1}{G} \sum_{i=1}^G \left[ \min(\rho_i A_i, \text{clip}(\rho_i, 1-\epsilon, 1+\epsilon) A_i) - \beta D_{\text{KL}}(P_\theta || P_{\text{ref}}) \right] \quad (21)$$

where  $\rho_i = \frac{P_\theta(y_i|x, d)}{P_{\text{old}}(y_i|x, d)}$  is the importance sampling ratio, and the advantage  $A_i$  is computed by normalizing the rewards within the group:  $A_i = \frac{r_i - \text{mean}(\{r_1 \dots r_G\})}{\text{std}(\{r_1 \dots r_G\})}$ . GRPO is particularly effective in stabilizing the router by using the group mean as a dynamic baseline.

**3. Load Balancing Loss.** To prevent router collapse, we incorporate an auxiliary load balancing loss  $\mathcal{L}_{\text{lb}}$ , defined as the scaled dot-product between expert selection frequency  $f$  and average routing probability  $P$ :

$$\mathcal{L}_{\text{lb}} = N \sum_{i=1}^N f_i \cdot P_i \quad (22)$$

This regularization ensures that the latent cultural topology is mapped across the full capacity of the expert pool.

## D Implementation Details

### D.1 Model Architectures

**Backbone Models.** We evaluate CUMA using two state-of-the-art open-source backbones:

---

### Algorithm 1 CUMA Training Procedure

---

**Input:** Dataset  $\mathcal{D}$ , Pre-trained LLM  $\theta_{\text{LLM}}$ , Demographic Encoder  $E(\cdot)$

**Output:** Optimized Parameters  $\theta_r^*$ ,  $\{A_i^*, B_i^*\}_{i=1}^N$

- 1: **Initialization:** Freeze  $\theta_{\text{LLM}}$  and  $E(\cdot)$ . Initialize router  $\theta_r$  and  $N$  LoRA experts with random weights.
  - 2: *// Stage 1: Conditional SFT*
  - 3: **for** each batch  $\mathcal{B} = \{(x, d, y)\} \in \mathcal{D}_{\text{SFT}}$  **do**
  - 4:   Encode demographics:  $e_d \leftarrow E(d)$
  - 5:   Forward pass to compute  $P_\theta(y|x, d)$  via sparse routing (Eq. 4)
  - 6:   Compute Loss:  $\mathcal{L} = \mathcal{L}_{\text{SFT}} + \lambda \mathcal{L}_{\text{lb}}$
  - 7:   Update  $\theta_r, A_i, B_i \leftarrow \text{AdamW}(\nabla \mathcal{L})$
  - 8: **end for**
  - 9: *// Stage 2: Conditional Preference Optimization (DPO or GRPO)*
  - 10: **for** each batch  $\mathcal{B} \in \mathcal{D}_{\text{Pref}}$  **do**
  - 11:   Encode demographics:  $e_d \leftarrow E(d)$
  - 12:   **if** method is **DPO** **then**
  - 13:     Input batch pairs  $\{(x, d, y_w, y_l)\}$
  - 14:     Compute implied rewards relative to reference model  $\pi_{\text{ref}}$ :
  - 15:      $r_w \leftarrow \beta \log(P_\theta(y_w|x, d)/P_{\text{ref}}(y_w|x, d))$
  - 16:      $r_l \leftarrow \beta \log(P_\theta(y_l|x, d)/P_{\text{ref}}(y_l|x, d))$
  - 17:      $\mathcal{L}_{\text{task}} = -\log \sigma(r_w - r_l)$
  - 18:   **else if** method is **GRPO** **then**
  - 19:     Input batch  $\{(x, d)\}$ . Sample group outputs  $\{y_1, \dots, y_G\}$  from  $P_{\text{old}}$ .
  - 20:     Compute rewards  $\{r_1, \dots, r_G\}$  using reward model or rule.
  - 21:     Compute Advantages:  $A_i \leftarrow (r_i - \text{mean}(r))/(\text{std}(r) + \epsilon)$
  - 22:     Compute Ratio  $\rho_i$  and KL divergence terms.
  - 23:      $\mathcal{L}_{\text{task}} = \text{Eq. (13)}$
  - 24:   **end if**
  - 25:   Total Loss:  $\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda \mathcal{L}_{\text{lb}}$
  - 26:   Update  $\theta_r, A_i, B_i \leftarrow \text{AdamW}(\nabla \mathcal{L})$
  - 27: **end for**
  - 28: **return**  $\theta_r, \{A_i, B_i\}$
- 

**Llama-3.1-8B-Instruct** and **Qwen3-8B**. Both models are kept frozen during training, with only the LoRA experts and the demographic-aware router being optimized.

**Demographic Encoder.** To process demographic profiles, we utilize **Qwen3-Embedding-0.6B** as the encoder  $E(\cdot)$ . The encoder takes the linearized demographic string as input with a maximum se-

quence length of 128 tokens. We apply **mean-pooling** over the last hidden states to obtain a fixed-dimensional embedding ( $d_e = 1024$ ). The encoder parameters are frozen throughout all training stages.

**Sparse Cultural Adapters.** Each expert is implemented as a LoRA adapter with rank  $r = 64$  and alpha  $\alpha = 128$ . Adapters are applied to the query ( $W_q$ ) and value ( $W_v$ ) projection matrices in all transformer layers. The router is a 2-layer MLP with a hidden dimension of 256. For each token, the router takes the concatenation of the token’s hidden state and the demographic embedding as input, mapping it to routing logits over  $N = 8$  experts. We select the top  $k = 2$  experts per token.

## D.2 Training Configurations

We perform all experiments on NVIDIA RTX PRO 6000 (96GB) GPUs using the AdamW optimizer with a cosine learning rate schedule. For the Full Fine-Tuning (FFT) baseline, we employ DeepSpeed ZeRO-2 optimization.

**Stage 1: Conditional SFT.** For the initial alignment stage, we train for up to 3 epochs with a learning rate of  $2 \times 10^{-5}$  for Qwen3-8B and  $5 \times 10^{-6}$  for Llama-3.1-8B. The effective batch size is set to 32, and the maximum sequence length is 1024 tokens. We set the load balancing coefficient  $\lambda_{lb} = 0.01$ .

**Stage 2: Conditional Preference Optimization.** For preference alignment (DPO/GRPO), we reduce the learning rate to  $5 \times 10^{-6}$  and train for 1 epoch. For DPO, we set the KL penalty coefficient  $\beta = 0.1$ . For GRPO, we use a group size  $G = 8$  and the same  $\beta$ . The maximum sequence length is increased to 2048 tokens to accommodate longer preference pairs.

**Reward Signal for GRPO.** Following the protocol of Zhang et al. (2025a), we utilize a model-based reward signal derived from GPT-4o. For each generated response  $y_i$  in the group, we compute a pairwise comparison against the base model’s response  $y_{ref}$ . The model is prompted to judge which response better aligns with the user’s demographic profile. We assign a scalar reward  $r_i \in \{1.0, 0.5, 0.0\}$  corresponding to a win, tie, or loss relative to the reference. The specific prompt template used for this judgment is provided in Appendix D.5.

## D.3 Data Construction Protocol

We tailor the data construction strategies for each dataset and training stage as follows.

**WorldValuesBench (WVB).** WVB is exclusively used for the conditional discrimination task. We formulate it as a multiple-choice question answering task.

- **SFT:** The model is presented with the demographic profile, question, and options. We only compute the loss on the token corresponding to the ground-truth option label (e.g., "A", "B"). No preference optimization (DPO/GRPO) is applied to this dataset.

**Community Alignment (CA).** This dataset supports both discrimination and generation tasks.

- **Discrimination Task (SFT):** Similar to WVB, we structure the 4 candidate responses as a multiple-choice problem. The model is trained to predict the label of the response preferred by the demographic group via standard SFT.
- **Generation Task (SFT):** We treat the response selected by the user as the ground truth. The model is conditioned on the profile and context, and trained to generate the selected response text using a standard causal language modeling objective.
- **Generation Task (DPO):** CA provides one chosen response and three rejected responses per sample. We decompose this into three distinct pairwise samples  $(y_w, y_l)$ , pairing the chosen response with each of the three rejected responses.
- **Generation Task (GRPO):** We follow the setting in Zhang et al. (2025a). The model generates a group of responses ( $G = 8$ ), and rewards are calculated using the GPT-4o judge described in Appendix D.

**PRISM.** PRISM focuses on open-ended interaction and naturally contains pairwise preferences.

- **Generation Task (SFT):** We perform SFT on the preferred response in the dataset, conditioning on the interaction history and user profile.

- **Preference Optimization:** Since PRISM data comes as binary preference pairs, DPO uses these pairs directly  $(y_w, y_l)$ . For GRPO, we adopt the same setup as in Community Alignment, sampling multiple outputs for the given context and scoring them using the demographic-aware judge.

#### D.4 Dataset Statistics

We utilize three benchmarks for evaluation: World-ValuesBench (WVB), Community Alignment (CA), and PRISM. For all datasets, we adopt a 10:1 split for training and testing respectively.

**WorldValuesBench (WVB).** Originally containing over 21M samples from 93,278 participants across 65 nations, we perform stratified sampling to obtain 500,000 samples for efficient training and evaluation. Each sample represents a demographic-conditioned value prediction task.

**Community Alignment (CA).** This dataset includes 192,137 pairwise comparisons from users in five nations (US, India, Brazil, France, and Italy). It covers both preference prediction and open-ended generation tasks across five languages.

**PRISM.** PRISM provides 27,111 interaction-level pairwise preferences from 8,016 diverse participants across 75 countries, along with fine-grained individual demographic attributes.

#### D.5 Prompt Templates

We employ specific prompt templates for each dataset to incorporate demographic information. To ensure consistency, we linearize demographic attributes in a fixed order: *Age, Gender, Country, Education, Religion, Ethnicity, Employment*.

**WorldValuesBench (WVB).** For WVB, the demographic profile is prepended to the system prompt to condition the model’s value commitments.

##### WVB System Prompt

You are a person with the following profile: Age: {age}, Gender: {gender}, Country: {country}, Education: {education}, Marital Status: {marital}, Religion: {religion}, Ethnicity: {ethnicity}, Employment: {employment}. You are a helpful assistant that answers survey questions honestly.

##### WVB User Prompt

{Question}? {Options}. You can only choose one option.

**Community Alignment (CA) & PRISM.** For generative tasks, we use a standardized "User Profile" header in the system prompt.

##### Standardized System Prompt (CA/PRISM)

User Profile: Age: {age}, Gender: {gender}, Country: {country}, Education: {education}, Religion: {religion}, Ethnicity: {ethnicity}, Employment: {employment}.

**Expert Verification (GPT-4o Judge).** We employ a GPT-4o judge for evaluating open-ended generation tasks. The judge is provided with 3-shot examples from the training set to ensure calibration. Validation against ground-truth labels confirms high reliability, with the judge achieving an accuracy of **83.3%** on the Community Alignment (CA) dataset and **89.8%** on PRISM.

##### GPT-4o Judge Prompt

**System Prompt:** You are an impartial and culturally aware judge. You will be given a user profile, a conversation context, and two AI responses. Your task is to determine which response is better suited for the specific user described in the profile. Consider the user’s demographics, values, and preferences implied by their profile.

**User Prompt:** Here are some examples of preferences for different users:

Example 1: Profile: {profile\_1} Context: {context\_1} Response A: {response\_a\_1} Response B: {response\_b\_1} Verdict: [[A]] ... (3-shot examples) ...

Now, please evaluate the following case: Profile: {target\_profile} Context: {target\_context} Response A: {target\_response\_a} Response B: {target\_response\_b} Which response is better? Output [[A]], [[B]], or [[Tie]].

**Prompt Steering (Few-Shot).** The  $k$ -shot baseline retrieves  $k$  demonstrations from the training set matching the user’s country or demographic

cluster to guide the model via in-context learning.

#### Prompt Steering Template

**System:** You are a person from {country}...  
 [Current Target User Profile]  
**User:** {Example 1 Question} **Assistant:**  
 {Example 1 Answer}  
**User:** {Example 2 Question} **Assistant:**  
 {Example 2 Answer}  
 ... ( $k$  examples from matching demographics) ...  
**User:** {Target Question}

## E Analysis Details

### E.1 Visualization of Latent Topology

To visualize the cultural topology learned by the router (Figure 4), we extract the expert activation patterns for users across 65 distinct nations in the WorldValuesBench test set. For a given country  $c$ , we compute the centroid of the routing weights:

$$\bar{g}_c = \frac{1}{|D_c|} \sum_{d \in D_c} \frac{1}{T} \sum_{t=1}^T g(x_t, d) \quad (23)$$

where  $D_c$  is the set of demographic profiles belonging to country  $c$ , and  $g(x_t, d)$  represents the sparse gating probability vector for token  $t$ . We average these vectors across all layers and tokens to obtain a global routing signature  $\bar{g}_c \in \mathbb{R}^N$  for each nation. We then project these high-dimensional signatures into 2D space using t-SNE with a perplexity of 30 and Euclidean distance metric. The resulting clusters reveal that the router learns to group nations based on shared value systems rather than mere geographic proximity.

### E.2 Zero-Shot Generalization Protocol

To rigorously assess zero-shot generalization (Table 2), we adopt a held-out demographic profile protocol. We categorize the 65 nations into 9 distinct cultural clusters (e.g., *English-Speaking*, *Catholic Europe*, *Confucian*) based on the Inglehart-Welzel cultural map. The experiment proceeds as follows:

1. **Exclusion:** Within each cluster  $C_i$ , we randomly select a subset of specific demographic profiles (defined by unique combinations of attributes like age, gender, and education within a country) to hold out from the training set.
2. **Training:** We train CUMA on the remaining dataset, ensuring that the model has seen the

general cultural cluster but not the specific held-out demographic combinations.

3. **Evaluation:** The model is evaluated exclusively on the held-out demographic profiles. This tests the model’s ability to generalize to unseen profiles by leveraging the learned topological structure of the cultural cluster.

## F Impact of Routing Strategy

To validate our hypothesis that conditional capacity separation is strictly required to resolve mean collapse, we compare our standard Top- $k$  (Hard) routing against **Soft Routing**. In the Soft Routing setting, we relax the sparsity constraint ( $k = N$ ), allowing tokens to be processed by a weighted combination of *all* experts:

$$y = \sum_{i=1}^N \text{softmax}(s)_i \cdot E_i(x) \quad (24)$$

This formulation is effectively a dense model with factorized parameters, as every expert contributes to every output.

Table 5 reveals a critical insight: *sparsity is essential for interference mitigation, not just efficiency*. Replacing the discrete Top- $k$  mechanism with Soft Routing (a weighted average of all experts) leads to a marked degradation, with WVB accuracy dropping by 2.56% and EMD rising by 0.0393. While Soft Routing theoretically retains full capacity, it forces distinct cultural gradients to superimpose within a shared linear combination, reintroducing the "mean collapse" pathology of dense models. By enforcing Top- $k$  selection, CUMA creates functionally orthogonal subspaces that shield divergent value systems from mutual interference, ensuring that pluralistic alignment remains distinct rather than diluted.

## G Human Evaluation Audit

To validate the reliability of our GPT-4o-based automated judge, we conducted an expert audit on a 100-sample subset drawn from both the Community Alignment (CA) and PRISM datasets, covering five major cultural clusters: African-Islamic, Confucian, Protestant Europe, Latin America, and English-Speaking.

**Audit Protocol.** Five researchers with academic backgrounds in Sociology and Computer Science

Strategy	WorldValuesBench			Community Alignment (CA)			PRISM
	Acc $\uparrow$	Macro-F1 $\uparrow$	EMD $\downarrow$	Acc $\uparrow$	Macro-F1 $\uparrow$	Win%	Win%
CUMA (Top- $k$ )	<b>50.64</b>	<b>31.50</b>	<b>0.1876</b>	<b>52.45</b>	<b>50.10</b>	<b>78.2</b>	<b>76.8</b>
Soft Routing	48.08	28.73	0.2269	-	-	73.0	71.0

Table 5: **Top- $k$  vs. Soft Routing on Qwen3-8B.** Top- $k$  routing significantly outperforms Soft routing.

served as expert auditors. To prevent subjective bias, we adopted a reference-based protocol: auditors did not provide personal preferences but instead verified whether a response accurately reflected the specific value dimensions of the target culture as defined by the World Values Survey (WVS). Evaluators used objective cultural profiles, such as the Survival vs. Self-Expression and Traditional vs. Secular-Rational indices, as external ground truth. All evaluators were blind to model identities and the prior scores assigned by GPT-4o.

**Results.** Table 6 presents the results for the Qwen3-8B (GRPO) setting. We quantify agreement using Cohen’s Kappa ( $\kappa$ ), defined as:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad (25)$$

where  $p_o$  is the observed agreement and  $p_e$  is the expected agreement by chance. The average  $\kappa = 0.84$  indicates strong consistency between the expert annotations and the automated judge, confirming the reliability of our evaluation protocol.

Method	Expert Win%	GPT-4o Win%	$\kappa$
P-Tuning v2	55%	60%	0.81
LoRA	59%	64%	0.83
DoRA	61%	66%	0.82
MixLoRA	68%	71%	0.85
HydraLoRA	68%	72%	0.82
CUMA (ours)	<b>74%</b>	<b>78%</b>	0.84

Table 6: **Human Expert Audit** on a 100-sample subset (Qwen3-8B, GRPO). Expert Win-Rate denotes the percentage of cases where human auditors preferred the method’s output over the base model, compared against the GPT-4o judge’s verdict.

## H Inference Efficiency Analysis

Table 7 reports the practical efficiency of CUMA and baselines, measured on a single NVIDIA RTX PRO 6000 (96GB) GPU with batch size 1 for the Qwen3-8B backbone. Results are averaged over 3 independent runs.

Method	Peak VRAM (GB)	First Token Latency (ms)	Throughput (tok/s)	Trainable Params (%)
LoRA	15.4	45.5	44.2	0.37%
DoRA	15.5	46.8	43.1	0.38%
P-Tuning v2	15.8	48.2	42.5	0.94%
MixLoRA	16.5	105.4	36.5	3.01%
HydraLoRA	16.4	112.2	35.8	2.31%
CUMA ( $r=8$ )	16.2	124.5	34.8	1.53%

Table 7: **Inference Efficiency Comparison** on Qwen3-8B with a single RTX PRO 6000 GPU. Sparsely activated methods (MixLoRA, HydraLoRA, CUMA) incur modest overhead from loading multiple expert modules, while maintaining practical throughput.

Sparsely activated methods exhibit slightly higher peak VRAM usage compared to dense baselines, primarily due to the simultaneous loading of multiple expert modules. The additional first-token latency arises from the demographic encoding and routing computation. Given the significant improvements in cultural alignment (Table 1), this efficiency trade-off represents a modest cost for robust pluralistic modeling.

## I Routing Overlap Analysis

To examine whether CUMA adaptively shares or separates parameters based on the nature of the task, we measure *routing overlap*, defined as the percentage of shared experts activated by different demographic groups for the same query.

**Protocol.** We evaluate on two contrasting datasets: (1) GSM8K (Cobbe et al., 2021), a mathematical reasoning benchmark representing consensus knowledge where correct answers are culturally invariant; and (2) WorldValuesBench, representing *non-consensus* cultural values where preferences diverge across demographics.

**Results.** On GSM8K, the average routing overlap reaches 76.4%, indicating that CUMA routes most demographic groups to the same experts for consensus tasks. In contrast, on WorldValuesBench, the average overlap drops significantly to 18.2%. This confirms that CUMA learns to share parameters for consensus tasks while isolating conflicting cultural updates into specialized subspaces.