CulDPD: Dual-Path Decoupling for Cultural Alignment with Cultural Specialization Loss

Anonymous ACL submission

Abstract

Cultural alignment in LLMs remains a challenging task, as current methods often fail to adequately separate cross-cultural consensus from culture-specific patterns, leading to biased or generic outputs. In this paper, we propose CulDPD, a culture dual-path decoupling framework that explicitly separates culture-shared subspace from culture-specific subspace through controlled view perturbation and differentiated representation spaces. CulDPD adopts a dual-path architecture: the culture-shared pathway employs Cultural Sensitivity Masking strategy to suppress explicit cultural identifiers and capture cross-cultural consensus, while the culture-specific pathway leverages multiple experts with dynamic routing to capture culture-specific patterns. Furthermore, we design a dual-level cultural specialization loss, which consists of a Culture Routing Loss that promotes the cultural specialization of individual experts within the culture-specific pathway and a Culture Decoupling Loss that increases the representational distinction between the culture-shared and culture-specific pathways, thereby enhancing the overall decoupling effect. Extensive experiments on multiple cultural benchmark datasets demonstrate that CulDPD improves the alignment of LLMs with diverse cultural contexts.

1 Introduction

Culture is a multi-layered and dynamic construct encompassing multiple dimensions such as language, ethnicity, religion, and social norms (Hofstede, 2011). Significant differences exist between cultures in terms of values, social norms, and behavioral preferences (AlKhamissi et al., 2024; Tao et al., 2024). In Natural Language Processing, the complexity poses substantial technical challenges (Hershcovich et al., 2022). Currently, large language models (LLMs) often exhibit pronounced "cultural bias", as they tend to output values associated with Western, highly educated, industrialized, and democratic societies (i.e., WEIRD values) (Rystrøm et al., 2025; Santurkar et al., 2023), thereby overlooking or marginalizing non-Western cultural perspectives (Sukiennik et al., 2025). To address the issue, cultural alignment has emerged. In simple terms, cultural alignment aims to teach models to "when in Rome, do as the Romans do": the model must not only understand shared norms across cultures but also accurately grasp culture-specific taboos, preferences, and customs. Through such alignment, the model can generate responses that are both professional and contextually appropriate when interacting with users from diverse cultural backgrounds.

Existing cultural alignment methods can generally be divided into two main paradigms: data-centric fine-tuning and prompt-based conditional reasoning (Shen et al., 2026). Data-centric fine-tuning increases cultural data or augments existing data to enable the model to learn culture-related knowledge within a unified parameter space (Li et al., 2024a; Shi et al., 2024; Xu et al., 2025; Yao et al., 2025; Feng et al., 2025). Such methods typically map and optimize signals from different cultural backgrounds into a single parameter space, implicitly blending cross-cultural consensus and culture-specific patterns. Prompt-based conditional reasoning adjusts the model's behavior during the inference phase by introducing external cultural conditions (such as role-playing (Liu et al., 2025b), contextual examples (Choenni and Shutova, 2025), etc.) without requiring additional training. Multi-agent debate and collaboration methods (Ki et al., 2025; Yuan et al., 2024) also fall under this paradigm.

Although these methods have made progress in improving cultural adaptability, they primarily rely on data augmentation or conditional injection to incorporate cultural diversity. Cultural norms comprise both shared and culture-specific

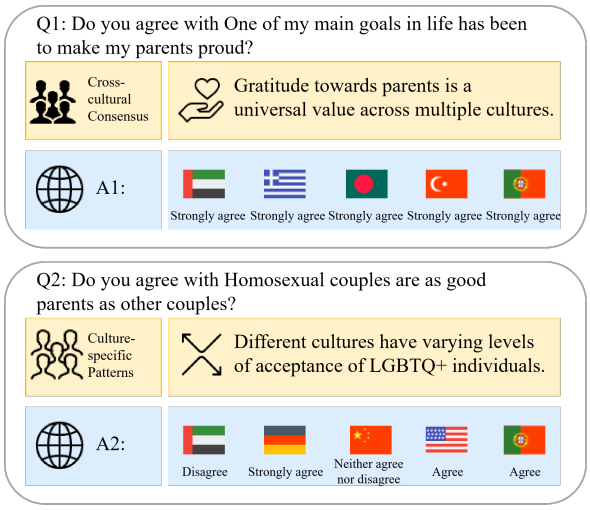


Figure 1: Two Examples: Cross-Cultural Consensus and Culture-Specific Patterns

elements. For example (Figure 1), questions about family pride show cross-cultural consensus, while questions regarding LGBTQ+ parenting highlight culture-specific differences across regions. However, these methods implicitly assume that cross-cultural commonalities and culture-specific norms can be represented within a single shared parameter space. This assumption introduces structural risks. First, mixed encoding may lead to representation collapse, where culture-specific signals are averaged out under cross-cultural training objectives. Second, gradient updates from different cultural distributions may interfere with one another, resulting in cross-group alignment interference. The lack of a structured decoupling mechanism further causes the model, when handling conflicting cultural values, to generate appeasing and moderate responses (Chen et al., 2024; Sun et al., 2026). These issues indicate that mixed modeling is insufficient, and that shared representations and culture-specific representations must be structurally decoupled in an explicit manner.

To this end, we propose a novel **Cul**tural **D**ual-**P**ath **D**ecoupling framework (CulDPD), which consists of a culture-shared pathway and a culture-specific pathway. We first design a Cultural Sensitivity Masking strategy. By performing controlled view perturbation on inputs, it generates masked and unmasked inputs that are semantically related but differ in the strength of cultural cues. The masked and unmasked inputs are respectively fed into the culture-shared pathway and the culture-specific pathway, enabling the culture-shared pathway to focus on extracting stable cross-cultural universal knowledge, while the culture-specific pathway specializes in capturing distinctive patterns across various cultural contexts. The dual-path decoupling design effectively mitigates the representation entanglement caused by the mixture of culture-shared and culture-specific subspaces in existing methods.

In addition, we propose a dual-level Culture Specialization Loss (CSL) to further disentangle the feature representations across cultures. First, the Cultural Routing Loss (CRL) is designed to enhance the cultural specialization of each expert in the culture-specific pathway, enabling each expert to focus on distinct culture-specific patterns. Second, the Culture Decoupling Loss (CDL) is introduced to increase the separation between the culture-shared and culture-specific subspaces, thereby reducing cultural knowledge mixing and interference. Together, two losses facilitate structured modeling of cross-cultural consensus and culture-specific patterns, thereby improving cultural alignment.

In summary, the main contributions are:

- We propose CulDPD, a novel dual-path framework for cultural alignment that explicitly separates culture-shared and culture-specific pathways to disentangle cross-cultural commonalities and culture-specific patterns in the representation space. To further enhance the disentanglement process, a Cultural Sensitivity Masking strategy is introduced to perform controlled input-level view perturbation.
- A dual-level culture specialization loss is introduced to enhance the disentanglement of multicultural knowledge: (1) CRL promotes the cultural specialization of each expert within the culture-specific pathway; (2) CDL enhances the distinguishability between shared and specific representations across pathways.
- Extensive experiments on multiple cultural benchmarks demonstrate that CulDPD consistently achieves strong and stable performance, validating the effectiveness of structured representation decoupling for cultural alignment.

2 Related Work

Existing research on LLM cultural alignment can generally be categorized into two paradigms: data-centric fine-tuning and prompt-based conditional reasoning (Shen et al., 2026).

Data-centric fine-tuning involves constructing or enhancing cultural datasets and fine-tuning the model on this basis, encoding culture-related knowledge into the model's parameters. This approach mainly focuses on the collection, generation, and augmentation of cultural data, such as through manual annotation, multi-agent simulation, or semantic expansion, to improve the model's ability to cover facts, values, and behavioral norms from different cultures (Li et al., 2024a; Shi et al., 2024; Xu et al., 2025; Yao et al., 2025; Feng et al., 2025). However, these methods often collapse diverse cultural signals into a single representation space, leading to representational interference where culture-specific nuances are overwritten by dominant global patterns (Liu et al., 2025a). This lack of structural decoupling results in compromised responses that fail to achieve authentic cultural alignment (Xu et al., 2025).

Prompt-based conditional reasoning applies cultural conditions to the model during the inference stage by introducing context or prompts to simulate specific cultural perspectives. Typical methods include role-playing (Liu et al., 2025b) and contextual examples (Choenni and Shutova, 2025), which guide the model to generate responses aligned with the given cultural background without updating the model's parameters. This paradigm offers high flexibility and controllability, allowing for rapid switching between different cultural conditions. Multi-agent debate and collaboration methods also fall under this paradigm(Ki et al., 2025; Yuan et al., 2024). However, such methods also fail to establish an explicit representation of cross-cultural consensus and culture-specific patterns within the model.

In summary, although the two paradigms mentioned above have made some progress in cultural alignment tasks, they still face a core limitation: the mixed processing of cultural signals makes it difficult for the model to access deep cultural meanings. To address this issue, we propose CulDPD, which adopts an explicit dual-path architecture to disentangle cross-cultural commonalities from culture-specific patterns and introduces a dual-level CSL to further enhance the disentanglement process.

3 CulDPD

The overview of CulDPD is shown in Figure 2. CulDPD is designed with a dual-path framework, leveraging Cultural Sensitivity Masking to decouple culture-specific and culture-shared representations. The model incorporates a dual-level Cultural Specialization Loss, consisting of Culture Routing Loss to encourage expert specialization and Culture Decoupling Loss to ensure the separation of cross-cultural consensus from culture-specific patterns.

3.1 Cultural Sensitivity Masking

Cultural sensitivity information comprises explicit identifiers—including geopolitical entities, religious symbols, and rituals—serving as identity anchors in textual data (Hofmann et al., 2024; Arora et al., 2023; Fung et al., 2024). Table 3 reports detailed examples of culture-sensitive words. While characterizing cultural uniqueness (Hershcovich et al., 2022), their direct exposure can cause over-reliance on culture-specific patterns and neglect of cross-cultural consensus. To this end, we introduce Cultural Sensitivity Masking as a controlled view perturbation mechanism, aiming to attenuate cultural anchors and thus enable the model to distinguish between culture-shared and culture-specific subspaces.

Specifically, we construct a culture-sensitive lexicon covering multiple cultural dimensions through a three-stage process: (i) manually selecting representative seed words; (ii) performing constrained expansion using Gemini-2.5-Flash; and (iii) manually verifying the generated entries. Explicit cultural identifiers in the text are then replaced with a unified placeholder token [MASK]. This operation targets only cultural variables without altering syntactic structures, thereby ensuring semantic equivalence at the sentence level. Detailed procedures and examples are provided in Appendix B.1 and Appendix B.4, respectively.

Through the masking operation $M(\cdot)$, the original input is bifurcated into two views: a complete semantic view containing explicit cultural variables (unmasked data $X_{unmasked}$) and a view with attenuated cultural anchors but retained core semantics (masked data X_{masked}):

$$X_{masked} = M(X_{unmasked}) \tag{1}$$

This masked–unmasked partition constitutes a pair of semantically consistent but culturally perturbed input views $(X_{unmasked}, X_{masked})$, providing an explicit structural prior for the subsequent dual-path decoupling.

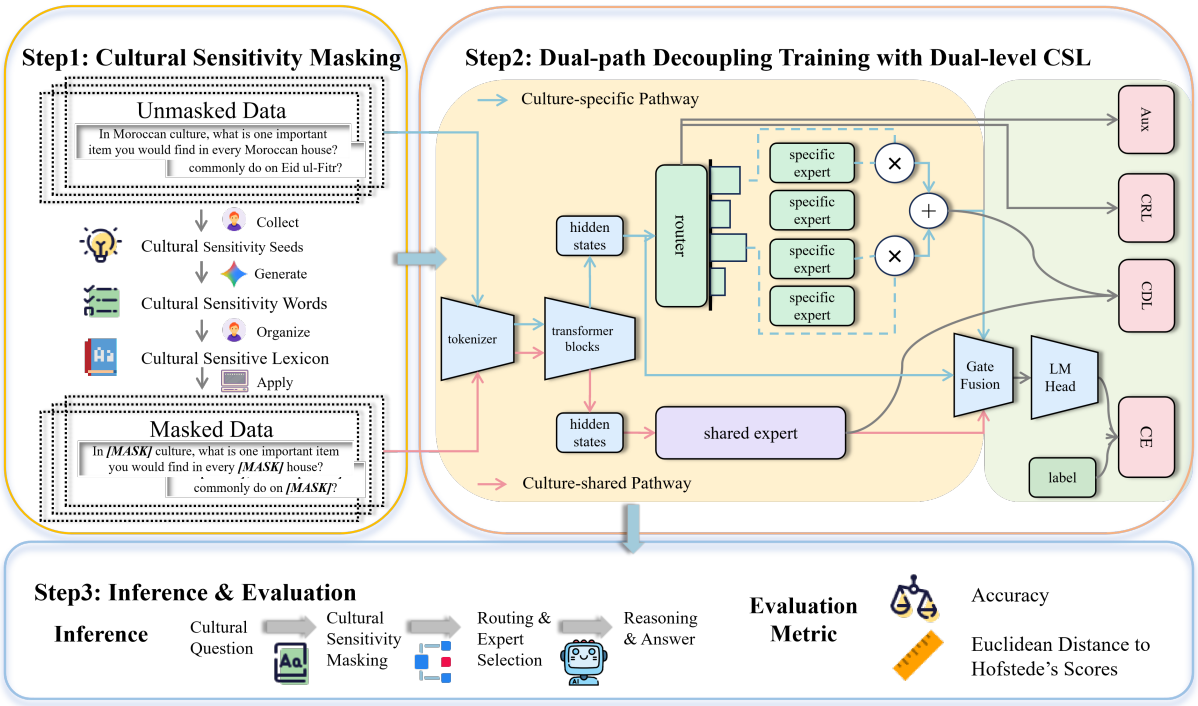


Figure 2: The overview of CulDPD. CulDPD adopts a dual-path architecture that combines Cultural Sensitivity Masking with a dual-level Cultural Specialization Loss to disentangle culture-shared and culture-specific representations, enabling pathway-level cultural decoupling and expert-level cultural specialization.

3.2 Culture Dual-Path Decoupling

Both the masked input X_{masked} and the unmasked input $X_{unmasked}$, obtained through Cultural Sensitivity Masking, are encoded by a shared LLM backbone $f(\cdot)$, thereby yielding representations from two distinct perspectives within the same semantic space. Formally, this is expressed as:

$$h_{shared} = f(X_{masked}) \tag{2}$$

$$h_{specific} = f(X_{unmasked}) \tag{3}$$

Due to the difference in input views, these two representations are structurally guided toward distinct subspaces: h_{shared} emphasizes the culture-shared subspace, while $h_{specific}$ captures information within the culture-specific subspace.

Culture-shared Pathway By suppressing explicit cultural identifiers, the shared pathway focuses on capturing culture-invariant knowledge. This is achieved through a Shared LoRA Expert attached to the LLM backbone to capture universal patterns. For a linear layer with weight W_0, the output e_{shared} is formulated as:

$$e_{shared} = W_0 h_{shared} + BAh_{shared} \tag{4}$$

where A and B are low-rank decomposition matrices of LoRA, designed to learn universal features applicable across diverse cultures.

Culture-specific Pathway The feature representation $h_{specific}$, representing the unmasked view, is processed through a dynamic routing network composed of multiple LoRA experts to capture culture-related norms, values, and behavioral preferences. The gating network computes the routing weights g to activate specific experts:

$$g = \text{Softmax}(W_g h_{specific}) \tag{5}$$

$$e_{specific} = \sum_{i=1}^{n} g_i \cdot E_i(h_{specific}) \tag{6}$$

Each expert E_i specializes in distinct cultural nuances, ensuring the model effectively extracts unique representations across different cultural contexts.

Dynamic Cultural Representation Fusion The final representation is obtained by dynamically fusing culture-shared and culture-specific subspaces via a gating network before prediction. The finalized fused representation h_{cul} is computed as:

$$h_{cul} = \gamma \cdot e_{shared} + (1 - \gamma) \cdot e_{specific} \tag{7}$$

The gating weight γ is dynamically determined by the concatenation of the two feature vectors:

$$\gamma = \sigma\big(W_m[e_{shared}; e_{specific}] + b_m\big) \tag{8}$$

where W_m and b_m are learnable parameters, and $\sigma(\cdot)$ is the sigmoid function that constrains $\gamma \in (0, 1)$, enabling an adaptive balance between the two pathways.

3.3 Culture Specialization Loss

We propose a dual-level Culture Specialization Loss (CSL), which imposes structural constraints on cultural modeling from the perspective of expert division, aiming to reduce cross-cultural interference and expert redundancy. As shown in Eq. 9, CSL consists of two components: CRL and CDL.

$$L_{CSL} = L_{CRL} + L_{CDL} \tag{9}$$

Culture Routing Loss To address cultural confusion in expert specialization, inspired by (Guo et al., 2025; Dan et al., 2025; Li et al., 2025; Liu et al., 2024), we design CRL to ensure that samples from the same culture exhibit high similarity in expert allocation space, while those from different cultures remain distinct. By guiding experts to specialize in specific cultural patterns, CRL reduces redundancy and improves the model's ability to decouple cultural features as follows:

$$\begin{aligned} L_{CRL} = & \frac{1}{N} \sum_{i=1}^{B} \sum_{j=i+1}^{B} [\delta(cul_i, cul_j)(1 - \text{sim}(i, j)) \\ & + (1 - \delta(cul_i, cul_j))\text{sim}(i, j)] \end{aligned} \tag{10}$$

where B is the batch size, $N = B(B-1)/2$ is the total number of sample pairs, and cul_i denotes the culture of sample i. $\delta(cul_i, cul_j)$ is an indicator function that equals 1 if the cultures are identical and 0 otherwise:

$$\delta(cul_i, cul_j) = \begin{cases} 1, & \text{if } cul_i = cul_j \\ 0, & \text{if } cul_i \neq cul_j \end{cases} \tag{11}$$

$\text{sim}(i, j)$ measures the cosine similarity between the expert assignment weight vectors w of samples i and j:

$$\text{sim}(i, j) = \frac{w_i \cdot w_j}{\|w_i\|\|w_j\|} \tag{12}$$

In short, CRL enforces clear specialization among culture-specific experts by aggregating same-culture allocations while suppressing cross-cultural overlap. This encourages each culture-specific expert to focus more on learning information relevant to a particular culture, thereby strengthening the decoupling of representations across different cultures.

Culture Decoupling Loss We introduce CDL to explicitly increase the distance between culture-shared and culture-specific experts, forcing the model to decouple consensus and specific features. Unlike the batch-wise CRL, CDL operates on each sample by minimizing the cosine similarity between the shared representation e_{shared} and the specific representation $e_{specific}$. This encourages shared experts to focus on universal knowledge while specific experts capture cultural nuances:

$$L_{CDL} = \text{sim}(e_{shared}, e_{specific}) \tag{13}$$

where $\text{sim}(\cdot)$ denotes cosine similarity. This loss directly reshapes the geometric structure of the representation manifold, suppressing the entanglement between shared and specific subspaces.

Final Optimization Objective The model's final optimization objective is composed of three components: basic task loss (CE), load balancing loss (aux), and CSL. Through the joint optimization of these three terms, the model achieves robust task performance while learning structured and highly interpretable cultural representations:

$$L_{total} = L_{CE} + \alpha L_{aux} + \beta L_{CSL} \tag{14}$$

where α and β are hyperparameters that control the load-balancing loss to prevent expert routing collapse and the strength of the culture specialization constraint, respectively.

4 Experimental Setup

4.1 Datasets

Referring to the classification criteria of cultural alignment benchmarks proposed by (Wu et al., 2025), we select four datasets that represent different levels of cultural depth and geographical breadth, in order to comprehensively evaluate the alignment performance of CulDPD in complex scenarios. CulturalBench (Chiu et al., 2024) contains 1,227 multiple-choice questions, covering 45 countries and regions worldwide, and 17 cultural themes. We adopt its CulturalBench-Easy version. CultureLLM (Li et al., 2024a) focuses on value judgments and perspectives on specific issues from different cultural backgrounds, covering 9 languages

	CulturalBench(%)	CultureLLM(%)	NormAd(%)	CultureAtlas(%)
Close-source Models				
GPT-4o	81.34	44.41	68.04	56.98
GPT-4o-mini	81.50	44.50	68.10	57.13
Gemini-2.5	82.07	39.24	**68.21**	54.42
Open-source Models				
Llama-3.1-8B-Instruct as backbone model				
Llama3.1	67.21	17.48	58.15	50.13
Role-Playing	68.03	15.59	58.64	49.54
ICL	70.49	46.55	45.77	23.84
MAD	73.77	51.25	60.46	52.12
DC-Finetune	75.41	88.91	97.35	69.71
Vanilla MoE	78.20	89.91	98.11	73.71
MixLoRA	79.18	89.93	98.48	75.43
CulDPD(Ours)	**82.46**	**91.43**	**99.24**	**78.86**
Qwen2.5-7B-Instruct as backbone model				
Qwen2.5	72.13	27.27	59.78	49.78
Role-Playing	71.31	28.47	59.51	49.94
ICL	72.95	50.95	45.46	36.53
MAD	75.41	57.04	61.60	52.29
DC-Finetune	77.87	89.51	95.45	73.14
Vanilla MoE	78.73	89.93	96.21	76.57
MixLoRA	77.54	89.79	96.59	79.43
CulDPD(Ours)	**83.94**	**92.83**	**98.86**	**81.71**

Table 1: Performance on four benchmark datasets. Note: LlaMA3.1/Qwen2.5, Role-Playing, ICL, and MAD are prompt-based methods that do not require training; DC-Finetune, Vanilla MoE, MixLoRA, and CulDPD are trained methods, with reported results taken from seed = 42. Detailed mean and standard deviation over five random seeds for the trained methods are provided in Appendix D. **Bold** numbers indicate the best performance among methods.

and 7 cultural value themes. NormAd (Rao et al., 2025) and CultureAtlas (Fung et al., 2024) are evaluation benchmarks that focus on assessing the cultural reasoning and knowledge, with 2633 stories and 6238 cultural statements.

4.2 Baselines

We employ several state-of-the-art LLM as baselines: close-source models including GPT-4o, gpt-4o-mini (Hurst et al., 2024) and Gemini-2.5-Flash (Gemini-2.5) and open-source models including Llama3.1-8B-Instruct (Llama3.1) (Grattafiori et al., 2024), Qwen2.5-7B-Instruct (Qwen2.5) (Yang et al., 2025), ICL (Chen et al., 2025; Choenni and Shutova, 2025), Multi-agents debate (MAD) (Ki et al., 2025), Data-centric Fine-tuning (DC-Finetune) (Xu et al., 2025; Li et al., 2024a), Vanilla MoE (Shen et al., 2023) and MixLoRA (Li et al., 2024b). More details can be found in Appendix D.

4.3 Evaluation Metric

We use accuracy as the evaluation metric for the CulturalBench, CultureLLM, NormAd, and CultureAtlas datasets. Accuracy is calculated as the proportion of correctly predicted cultural responses out of the total number of samples, providing a measure of the model's ability to classify responses based on cultural backgrounds.

5 Results

5.1 Overall Performance

Table 1 presents the main results on each benchmark. Our conclusions are as follows.

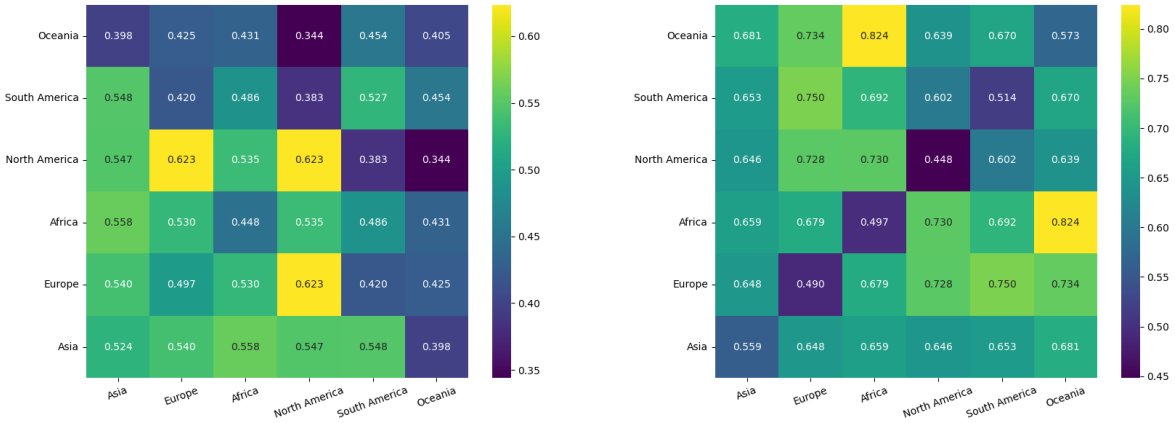
First, CulDPD achieves state-of-the-art performance among open-source models across four benchmark datasets. Second, CulDPD outperforms DC-Finetune especially, indicating that decoupled modeling with explicit expert division is more effective than simple data-driven mixed modeling. Third, CulDPD surpasses Vanilla MoE and MixLoRA, indicating that the performance improvement is not solely attributed to LoRA experts, while Cultural Sensitivity Masking and explicit cultural decoupling also play crucial roles. Finally, CulDPD shows competitive performance against closed-source models like GPT-4o, demonstrating that decoupled modeling effectively mitigates the scale and data disadvantages of open-source models.

5.2 Ablation Study

Table 2 presents the ablation results of CulDPD on four cultural benchmarks. Under both backbone models, the complete model achieves optimal performance, validating the effectiveness of each component. Removing shared experts or cultural

	CulturalBench(%)	CultureLLM(%)	NormAd(%)	CultureAtlas(%)
Llama-3.1-8B-Instruct as backbone model				
CulDPD	**82.46**	**91.43**	**99.24**	**78.86**
w/o shared path	79.53	90.61	98.48	76.57
w/o mask	81.15	91.11	98.86	77.71
w/o specific path	80.49	90.81	98.48	77.14
w/o gate	80.33	91.01	98.86	77.14
w/o CSL	77.87	90.11	98.11	76.00
w/o CRL	80.49	91.01	98.48	77.14
w/o CDL	79.67	90.61	98.11	75.43
Qwen2.5-7B-Instruct as backbone model				
CulDPD	**83.94**	**92.83**	**98.86**	**81.71**
w/o shared path	81.15	91.41	96.97	80.00
w/o mask	81.97	91.91	98.49	80.57
w/o specific path	82.11	91.71	97.35	80.57
w/o gate	80.33	91.81	98.49	81.14
w/o CSL	79.51	90.81	96.59	79.43
w/o CRL	81.30	91.41	97.73	81.14
w/o CDL	80.49	91.01	96.97	80.57

Table 2: Ablation study across four benchmark datasets. **Bold** numbers indicate the best performance among methods.



(a) Without CRL

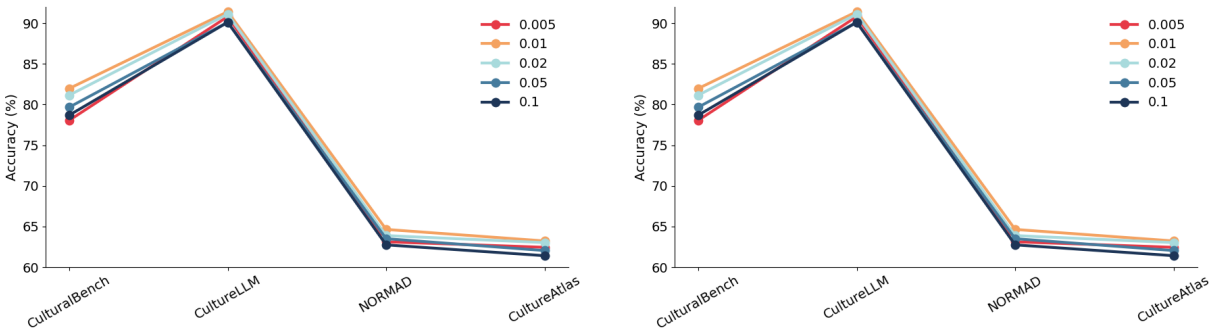
(b) With CRL

Figure 3: Heatmap of Expert Weight Distribution Euclidean Distance Without and With CRL.

information masking results in a consistent decline in performance across all datasets, demonstrating that explicitly modeling cultural commonalities and controlling culturally sensitive cues are crucial for cross-cultural representation learning. In contrast, removing only the masking mechanism or routing experts also leads to performance degradation, indicating that culturally sensitive information control and cultural characteristic experts act on input constraints and difference modeling, respectively, and are complementary. Furthermore, removing the gating mechanism or CSL significantly weakens performance, further validating the key role of dynamic expert weighting and commonality-characteristic joint constraints in cultural alignment.

5.3 Representation Disentanglement Analysis of CRL and CDL

To verify the effect of CRL on culture-specific expert specialization, Figure 3 visualizes the Euclidean distances between expert allocation vectors of samples from different continents on the CulturalBench dataset. For each sample, the expert allocation vector is obtained via the routing network, and pairwise Euclidean distances are computed and aggregated for samples from the same continent versus samples from different continents. Without CRL (Figure 3a), distances between samples from the same versus different continents are similar, indicating "generalization ambiguity." With CRL (Figure 3b), samples from the same continent show higher cohesion while distances across different continents increase significantly, demonstrating



(a) LLaMA-3.1 Backbone (b) Qwen-2.5 Backbone

Figure 4: Effect of the CSL weight β on model performance across four benchmark datasets. Moderate values of β achieve optimal results, while too low or too high values reduce generalization.

that CRL effectively enhances the specialization of cultural experts.

Based on CulturalBench, Figure 5 illustrates the effect of CDL on expert decoupling. For each sample, the Euclidean distance between culture-shared and culture-specific expert output vectors is computed, and the distances are aggregated across all samples to create the boxplot visualization. Before applying CDL, culture-shared and culture-specific expert representations overlap heavily; after applying CDL, the Euclidean distance between them increases notably, demonstrating that CDL effectively promotes cultural decoupling and enhances representation independence.

These results qualitatively support the decoupling objective of CulDPD, indicating that CRL strengthens the specialization of cultural experts, while CDL effectively enhances cultural decoupling.

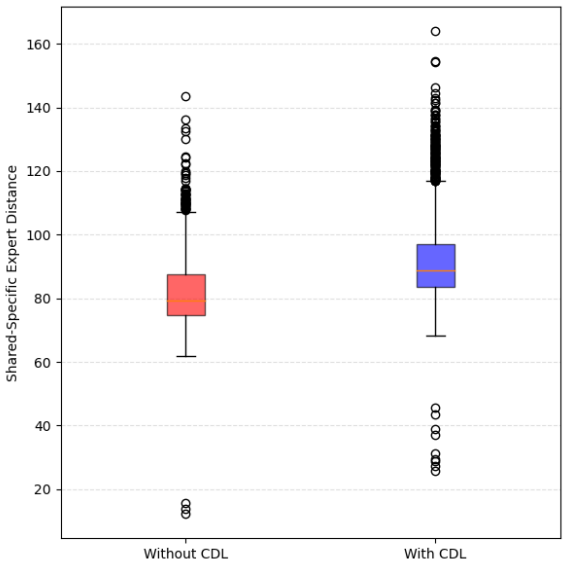


Figure 5: Boxplot of Shared–Specific Expert Euclidean Distance Without and With CDL.

5.4 Parameter Study of the CSL Weight

Figure 4 illustrates the effect of the culture-specific loss weight β on model performance across four benchmark datasets. Moderate values of β achieve optimal performance in all four benchmarks. Lower β values fail to adequately constrain cultural features, limiting the model's ability to capture cross-cultural differences, while excessively high β leads to over-regularization and reduces generalization. This pattern is consistently observed on both LLaMA-3.1 and Qwen-2.5, demonstrating CulDPD's robustness within a certain parameter range.

6 Conclusion

We propose CulDPD, a dual-path decoupling framework for cultural alignment that formulates cultural modeling as a disentangled representation learning problem. By incorporating Cultural Sensitivity Masking, CulDPD constructs input views with different strengths of cultural cues, guiding the model to distinguish between culture-shared and culture-specific subspaces. Through the collaboration of culture-shared experts and culture-specific experts, CulDPD is able to capture both cross-cultural shared knowledge and regional cultural differences. In addition, we design a dual-level CSL to promote the specialization of culture-specific experts and to enhance the separation between culture-shared and culture-specific subspaces. Experiments on multiple cultural benchmarks demonstrate that CulDPD outperforms existing methods in terms of culture alignment, robustness under distribution shifts, and alignment stability.

Limitations

This work still has limitations. First, Utilizing "country" as a primary proxy for culture is a necessary simplification in current work, yet it risks obscuring the intersectional nature of identity, where sub-regional, ethnic, and class-based variances create significant internal heterogeneity(Taras et al., 2014). Second, CulDPD training relies on explicit cultural annotations or prior cultural information. If such information is missing, noisy, or the cultural cues are implicitly strong, it could impact the model's alignment performance, especially in cases where data quality and availability are low. In addition, due to the limited availability of open-ended cultural datasets and the inherent constraints of current benchmark evaluation protocols, the present study mainly evaluates performance on structured multiple-choice cultural reasoning benchmarks. Therefore, broader aspects of cultural understanding, such as stereotype avoidance, open-ended cultural reasoning, culturally adaptive generation, and deeper modeling of cultural diversity, remain insufficiently explored. These limitations provide an opportunity for future exploration in more fine-grained, dynamic, and weakly-supervised cultural modeling.

References

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.

Arnav Arora, Lucie-aimée Kaffee, and Isabelle Augenstein. 2023. Probing pre-trained language models for cross-cultural differences in values. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 114–130, Dubrovnik, Croatia. Association for Computational Linguistics.

Quan Ze Chen, Kevin Feng, Chan Young Park, and Amy X Zhang. 2025. Spica: Retrieving scenarios for pluralistic in-context alignment. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 748–765.

Wei Chen, Zhen Huang, Liang Xie, Binbin Lin, Houqiang Li, Le Lu, Xinmei Tian, Deng Cai, Yonggang Zhang, Wenxiao Wan, Xu Shen, and Jieping Ye. 2024. From yes-men to truth-tellers: addressing sycophancy in large language models with pinpoint tuning. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and 1 others. 2024. Culturalbench: a robust, diverse and challenging benchmark on measuring (the lack of) cultural knowledge of llms.

Rochelle Choenni and Ekaterina Shutova. 2025. Self-alignment: Improving alignment of cultural values in llms via in-context learning. *Preprint*, arXiv:2408.16482.

Yuhao Dan, Jie Zhou, Qin Chen, Junfeng Tian, and Liang He. 2025. P-react: Synthesizing topic-adaptive reactions of personality traits via mixture of specialized lora experts. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6342–6362.

Ruixiang Feng, Shen Gao, Xiuying Chen, Lisi Chen, and Shuo Shang. 2025. Culfit: A fine-grained cultural-aware llm training paradigm via multilingual critique data synthesis. *arXiv preprint arXiv:2505.19484*.

Yi Fung, Ruining Zhao, Jae Doo, Chenkai Sun, and Heng Ji. 2024. Massively multi-cultural knowledge acquisition & lm benchmarking. *arXiv preprint arXiv:2402.09369*.

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.

Hongcan Guo, Haolang Lu, Guoshun Nan, Bolun Chu, Jialin Zhuang, Yuan Yang, Wenhao Che, Xinye Cao, Sicong Leng, Qimei Cui, and Xudong Jiang. 2025. Advancing expert specialization for better moe. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

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

Valentin Hofmann, Goran Glavaš, Nikola Ljubešić, Janet B Pierrehumbert, and Hinrich Schütze. 2024. Geographic adaptation of pretrained language models. *Transactions of the Association for Computational Linguistics*, 12:411–431.

Geert Hofstede. 2011. Dimensionalizing cultures: The hofstede model in context. *Online readings in psychology and culture*, 2(1):8.

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.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, and 399 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Dayeon Ki, Rachel Rudinger, Tianyi Zhou, and Marine Carpuat. 2025. Multiple llm agents debate for equitable cultural alignment. *arXiv preprint arXiv:2505.24671*.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024a. Culturellm: Incorporating cultural differences into large language models. *Advances in Neural Information Processing Systems*, 37:84799–84838.

Dengchun Li, Yingzi Ma, Naizheng Wang, Zhengmao Ye, Zhiyuan Cheng, Yinghao Tang, Yan Zhang, Lei Duan, Jie Zuo, Cal Yang, and 1 others. 2024b. Mixlora: Enhancing large language models fine-tuning with lora-based mixture of experts. *arXiv preprint arXiv:2404.15159*.

Junzhuo Li, Bo Wang, Xiuze Zhou, and Xuming Hu. 2025. Dynamic expert specialization: Towards catastrophic forgetting-free multi-domain moe adaptation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 18489–18504.

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

Dongxu Liu, Bing Xu, Yinzhuo Chen, Bufan Xu, Wenpeng Lu, Muyun Yang, and Tiejun Zhao. 2024. Pmol: Parameter efficient moe for preference mixing of llm alignment. *Preprint*, arXiv:2411.01245.

Haijiang Liu, Yong Cao, Xun Wu, Chen Qiu, Jinguang Gu, Maofu Liu, and Daniel Hershcovich. 2025b. Towards realistic evaluation of cultural value alignment in large language models: Diversity enhancement for survey response simulation. *Inf. Process. Manage.*, 62(4).

Vlad-Andrei Negru, Camelia Lemnaru, Mihai Surdeanu, and Rodica Potolea. 2026. Understanding cultural alignment in multilingual llms via natural debate statements. *Preprint*, arXiv:2602.12878.

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

Jonathan Hvithamar Rystrøm, Hannah Rose Kirk, and Scott Hale. 2025. Multilingual != multicultural: Evaluating gaps between multilingual capabilities and cultural alignment in LLMs. In *Proceedings of Interdisciplinary Workshop on Observations of Misunderstood, Misguided and Malicious Use of Language Models*, pages 74–85, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? ICML'23. JMLR.org.

Lingzhi Shen, Xiaohao Cai, Yunfei Long, Imran Razzak, Guanming Chen, and Shoaib Jameel. 2026. Calm: Culturally self-aware language models. *Preprint*, arXiv:2601.03483.

Sheng Shen, Le Hou, Yanqi Zhou, Nan Du, Shayne Longpre, Jason Wei, Hyung Won Chung, Barret Zoph, William Fedus, Xinyun Chen, and 1 others. 2023. Mixture-of-experts meets instruction tuning: A winning combination for large language models. *arXiv preprint arXiv:2305.14705*.

Weiyan 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 *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4996–5025.

Nicholas Sukiennik, Chen Gao, Fengli Xu, and Yong Li. 2025. An evaluation of cultural value alignment in llm. *ArXiv*, abs/2504.08863.

Ao Sun, Xiaoyu Wang, Zhe Tan, Yu Li, Jiachen Zhu, Shu Su, and Yuheng Jia. 2026. Cuma: Aligning llms with sparse cultural values via demographic-aware mixture of adapters. *Preprint*, arXiv:2601.04885.

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).

Vas Taras, Piers Steel, and Bradley L Kirkman. 2014. Does country equal culture? beyond geography in the search for cultural entities. In *Academy of Management Proceedings*, volume 2014, page 17239. Academy of Management Briarcliff Manor, NY 10510.

Meng-Chen Wu, Si-Chi Chin, Tess Wood, Ayush Goyal, and Narayanan Sadagopan. 2025. Incorporating diverse perspectives in cultural alignment: Survey of evaluation benchmarks through a three-dimensional framework. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17037–17072.

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.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Jing Yao, Xiaoyuan Yi, Jindong Wang, Zhicheng Dou, and Xing Xie. 2025. Caredio: Cultural alignment of llm via representativeness and distinctiveness guided data optimization. *arXiv preprint arXiv:2504.08820*.

Jiahao Yuan, Zixiang Di, Shangzixin Zhao, Zhiqing Cui, Hanqing Wang, Guisong Yang, and Usman Naseem. 2024. Cultural palette: Pluralising culture alignment via multi-agent palette. *arXiv preprint arXiv:2412.11167*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

A Prompts

A.1 Zero-shot Prompt

We formulate the prompt for zero-shot method using the following format:

```
### Question: Give me the answer from 1 to n:
{Question} {Options}
This question was raised in response to
     ↪ {Country}.
You can only choose one option.
### Answer:
```

A.2 Role-Playing Prompt

We formulate the prompt for role-playing method using the following format:

```
Role: You are a model specialized in
     ↪  understanding of {Country}'s culture.

### Question: Give me the answer from 1 to n:
{Question} {Options}
This question was raised in response to
     ↪ {Country}.
You can only choose one option.

### Answer:
```

A.3 ICL Prompt

We formulate the prompt for ICL(Choenni and Shutova, 2025) method using the following format:

```
Context:
{Question_1} {Options_1} {Ans_1}
{Question_2} {Options_2} {Ans_2}
{Question_3} {Options_3} {Ans_3}
{Question_4} {Options_4} {Ans_4}
{Question_5} {Options_5} {Ans_5}

### Question: Give me the answer from 1 to n:
{Question} {Options}
This question was raised in response to
     ↪ {Country}.
You can only choose one option.

### Answer:
```

A.4 Multi-agents Debate (MAD) Prompt

We define the prompt structure for the MAD method across four distinct stages: initial persona setting, iterative debate rounds, and final synthesis.

Cultural Agent Initialization (Round 0):

```
Role: You are a cultural expert specializing in
     ↪ {Country}'s societal norms, values,
     ↪  and perspectives. Your objective is to
     ↪  provide a culture-specific response to
     ↪  the following query.

### Question: Give me the answer from 1 to n:
{Question} {Options}
This question was raised in response to
     ↪ {Country}.
You can only choose one option.

### Answer:
```

Multi-cultural Debate (Round 1):

```
Task: You are participating in a multicultural
     ↪  expert panel. Review your previous
     ↪  stance alongside insights from other
     ↪  cultural representatives.

### Current State:
- Your Previous Answer (Round 0): {your_answer}
- Perspectives from Other Cultures:
     ↪ {other_perspectives}

### Refinement Principles:
1. Cultural Integrity: Maintain your cultural
     ↪  lens; do not compromise your
     ↪  perspective solely for consensus.
2. Analytical Rigor: Only update your answer if
     ↪  other perspectives reveal a factual
     ↪  oversight or a critical cultural
     ↪  dimension previously missed.
3. Pluralism: Acknowledge that cultural
     ↪  divergence is expected and valuable.
     ↪  If your original reasoning remains
     ↪  culturally sound, stand by it.

### Updated Answer:
```

Final Consensus Building (Round 2):

```
Task: This is the concluding round of the
     ↪  multicultural deliberation. Finalize
     ↪  your cultural expert testimony.

### Discussion History:
- Round 0 (Initial): {your_r0_answer}
- Round 1 (Updated): {your_r1_answer} (Change
     ↪  Status: {changed_r1})
- Peer Perspectives (Round 1):
     ↪  {other_perspectives_r1}

### Final Instructions:
- Deliver your final culture-centric judgment.
- Prioritize cultural authenticity over
     ↪  majority agreement.
- Ensure your selection represents the most
     ↪  nuanced understanding of {Country}'s
     ↪  values.

### Final Answer:
```

Multicultural Synthesizer:

```
Role: You are a Multicultural Meta-Synthesizer.
     ↪  Your task is to aggregate diverse
     ↪  expert insights and determine the
     ↪  final optimal response.

### Expert Inputs (Round 2):
- Asia: {asia_answer}
- West: {western_answer}
- South America: {south_america_answer}
- Oceania: {oceania_answer}
- Africa: {africa_answer}

### Decision Logic:
1. Merit-based Weighting: Evaluate the depth
     ↪  and relevance of each culture's
     ↪  reasoning rather than applying a
     ↪  simple majority vote.
2. Contextual Relevance: Identify which
     ↪  specific cultural insights are most
     ↪  pivotal to the question's context.
3. Expert Credibility: Consider if one region's
     ↪  expertise carries higher localized
     ↪  authority for this specific query.
4. Synthesis: Select the single option that best
     ↪  integrates the collective intelligence
     ↪  of all represented cultures.

### Final Synthesized Answer:
```

B Cultural Sensitivity Masking

B.1 Details of the Cultural Sensitivity Masking Operation

We construct the Cultural Sensitivity Lexicon in three stages: (i) manually selecting representative seed terms across multiple cultural dimensions, (ii) performing LLM-based constrained expansion using Gemini-2.5-Flash, and (iii) manually verifying entries with annotators. The finalized lexicon comprises 1,091 entries, including 729 single-word terms (66.82%) and 362 multi-word expressions (33.18%). Using this lexicon, identified cultural identifiers in the text are replaced with a unified placeholder [MASK], applied only to explicit cultural markers while preserving task-relevant content and syntactic structure. This ensures that the underlying semantics remain consistent between masked and unmasked views.

Specifically, we first manually select a set of highly representative cultural sensitivity seed terms based on social norms and core values from diverse cultural backgrounds. These terms cover multiple key dimensions, including geopolitical entities (e.g., country names, disputed regions), religious symbols (e.g., clerical titles, sacred sites), and core rituals (e.g., festival names, specific ceremonies), forming the foundational anchors for cultural characteristics.

Subsequently, we leverage the robust semantic association capabilities of LLM to perform multi-dimensional expansion guided by the initial seeds. Using carefully designed prompts, LLM identifies and complements synonymous expressions, specific appellations, and metaphors within various cultural contexts. This process constructs a comprehensive and fine-grained collection of culture-sensitive terms, effectively addressing the limitations of manual coverage.

To ensure professional rigor and accuracy, we implement a manual verification stage. Human reviewers evaluate the LLM-generated candidate terms one by one, eliminating items that might cause ambiguity or lose their cultural sensitivity in specific contexts. This guarantees that every entry in the lexicon accurately reflects authentic cultural nuances.

Based on the finalized lexicon, we employ a matching mechanism to replace culture-sensitive identifiers in the text with a unified placeholder [MASK]. This operation is strictly confined to explicit cultural markers, ensuring that no task-relevant core semantic content is removed. Meanwhile, the original syntactic structure remains unaltered, thereby physically attenuating the cultural signal intensity while logically preserving the underlying semantic consistency between the masked and unmasked views.

Importantly, the masking mechanism is independent of any specific cultural taxonomy. The constructed lexicon merely instantiates the cultural variable under investigation. As long as the target cultural attribute can be operationalized

through identifiable lexical or contextual markers, the framework can be extended to alternative cultural dimensions (e.g., religion, language, or social identity) without modifying the dual-path structure. Therefore, the masking remains modular and extensible.

B.2 LLM-Based Cultural Sensitivity Lexicon Expansion Prompt

We define cultural vocabulary expansion as a constrained generation task. An initial seed set obtained from manual selection is input into a LLM under specific multidimensional constraints. As shown in Eq. (15), the LLM culturally sensitive vocabulary expansion process is represented as a formal function f_{LLM}.

$$\mathcal{L}_{expanded} = f_{LLM}(\mathcal{S}_{seed}, \mathcal{C}_{dim}, \mathcal{R}_{const}) \quad (15)$$

Where $\mathcal{S}_{seed}$ denotes the initial manually-curated seed lexicon, $\mathcal{C}_{dim}$ represents the predefined cultural dimension space (e.g., country, religion, and customs), and $\mathcal{R}_{const}$ signifies the set of constraints, such as the exclusion of generic terms and the inclusion of multi-word phrases. The resulting $\mathcal{L}_{expanded}$ constitutes the enhanced culturally sensitive lexicon.

The prompt aims to guide LLM to cover multiple cultural dimensions while rigorously filtering non-specific generalized concepts. This systematic approach ensures that the generated cultural indicators are reproducible and interpretable.

```
Input: S_seed ⊂ C_dim

Prompt:
### Instruction:
Based on the given seeds S_seed,
Expand lexicon within the scope of C_dim.

### Constraints:
1. Focus on explicit cultural identifiers.
2. Exclude generic terms.
3. Include multi-word expressions.

Output: L_expanded
```

B.3 Statistics of the Cultural Sensitivity Lexicon

To systematically characterize the coverage of cultural sensitivity information, we conduct category- and scale-level statistics on the final cultural sensitivity dictionary and organize the results by cultural dimensions. Table 3 summarizes the distribution and composition of different categories of cultural sensitivity terms, covering both single-word entries and multi-word phrases. These statistics indicate that the constructed dictionary achieves broad coverage across multiple cultural dimensions, thereby avoiding a masking process that focuses only on a single type of explicit cultural cue.

B.4 Examples of Cultural Sensitivity Masking

Table 4 illustrates the concrete application of the Cultural Sensitivity Masking strategy on real-world samples through several examples. By comparing the input texts before and after masking, one can clearly observe that cultural indicators are replaced with a unified placeholder, while the syntactic structure and task semantics remain unchanged.

B.5 Robustness to Cultural Sensitive Lexicon Quality

To evaluate the impact of the culturally sensitive dictionary on model performance, we conducted controlled degradation experiments, randomly removing 10%, 20%, and 30% of dictionary entries. The experiments were performed on the Llama-3.1-8B and Qwen2.5-7B backbone networks, and the results are shown in Table 5. On all four benchmark datasets, even removing up to 30% of dictionary entries resulted in negligible performance degradation. These results demonstrate that CulDPD's two-path architecture is robust to moderate incompleteness in the masked dictionary and does not heavily rely on a perfect dictionary.

C Case study

Figure 6 presents representative model behaviors on two types of tasks: culture-specific pattern understanding and cross-cultural consensus judgment, comparing the performance of Base (Llama3.1), DC-Finetune, Vanilla MoE, MixLoRA, and CulDPD. In addition, we further analyze the outputs generated by the shared-only and specific-only pathways of CulDPD to qualitatively examine the functional disentanglement between culture-shared reasoning and culture-specific reasoning.

Case 1 (subfigure 6a) provides two examples related to culture-specific patterns. The first question examines the practical function of shutters in Spanish culture, where the correct answer is that they block out almost all light to adapt to the hot climate. Base incorrectly interprets them as decorative designs, while DC-

Category	Description	Single-word Terms	Multi-word Terms	#Percentage(%)
Country / Nationality	Explicit references to countries, regions, or national identities that anchor an utterance to a specific national context.	China, Japan, German, American, etc.	United States, South Korea, Middle East, etc.	23.85
Ethnicity	Terms referring to ethnic groups or ethnic identities with culturally specific historical and social meanings.	Han, Arab, Latino, Slavic, etc.	Native American, African American, etc.	6.15
Religion	References to religions or belief systems encoding culturally grounded value systems and norms.	Islam, Christian, Buddhist, Hindu, etc.	Roman Catholic, Sunni Islam, etc.	6.82
Religious Practice / Symbol	Culture-specific religious rituals, symbols, or practices conveying normative expectations.	mosque, church, Bible, Quran, etc.	Friday prayer, Holy Communion, etc.	8.17
Holiday / Festival	Culturally specific public or religious holidays signaling temporal and cultural context.	Ramadan, Easter, Diwali, etc.	Spring Festival, Christmas Eve, etc.	4.42
Custom / Tradition	Social customs or traditions reflecting culturally shared behavioral norms.	tipping, bowing, handshake, etc.	family reunion dinner, tea ceremony, etc.	14.33
Social Role / Title	Culture-dependent social roles or titles encoding hierarchical or relational norms.	elder, priest, monk, rabbi, etc.	community leader, village chief, etc.	5.48
Institution / System	Culture-specific political, legal, or social institutions shaped by historical context.	parliament, monarchy, caste, etc.	civil service exam, welfare state, etc.	7.12
Language / Script Marker	Explicit mentions of languages, scripts, or writing systems associated with specific cultures.	Arabic, Latin, Kanji, etc.	simplified Chinese, Arabic script, etc.	9.13
Cuisine / Food Culture	Food-related terms strongly tied to cultural identity and everyday practices.	sushi, curry, pasta, etc.	halal food, traditional cuisine, etc.	14.33

Table 3: Category-wise Statistics of the Cultural Sensitivity Lexicon.

Finetune, Vanilla MoE, and MixLoRA only recognize their role in filtering strong sunlight. In contrast, both CulDPD(Specific-only) and CulDPD correctly select the light-blocking option, whereas CulDPD(Shared-only) fails on this example.

The second question concerns the key factor for attaining higher social status in Vietnamese culture, where the correct answer is obtaining a university degree. Base and DC-Finetune favor family inheritance or athletic achievement, respectively, reflecting a misinterpretation of the "education-oriented" value emphasized in Vietnamese culture. Similarly, CulDPD(Specific-only) and CulDPD correctly predict the answer, while CulDPD(Shared-only) fails to identify this culturally dependent educational preference.

Case 2 (subfigure 6b) presents a cross-cultural consensus question—"Is work a duty towards society?"—evaluated under multiple national contexts. The results show that Base produces unstable and biased responses across countries, while DC-Finetune, Vanilla MoE, and MixLoRA provide correct predictions only under a subset of country conditions. In contrast, CulDPD maintains consistent and correct predictions across all national contexts. More importantly, CulDPD(Shared-only) also produces stable and correct predictions across all country settings, whereas CulDPD(Specific-only) exhibits noticeable fluctuations under several country conditions. Overall, these case studies provide qualitative evidence that the two pathways in CulDPD exhibit distinct functional roles in modeling cross-cultural shared consensus and culture-specific reasoning patterns.

D Details in experiments

We conduct our experiments using the LLaMA-Factory [1] (Zheng et al., 2024) framework, which provides a unified and efficient pipeline for fine-tuning large language models. All models are trained on a node with 4 NVIDIA V100 GPUs (32GB VRAM). To handle the computational constraints, we leverage the Accelerate backend and mixed-precision training to balance training speed and numerical stability. For the training configu-

[1] https://github.com/hiyouga/LlamaFactory

Unmasked: In Moroccan culture, what is one important item you would find in every Moroccan house? 1. A baseball bat 2. A traditional Moroccan teapot 3. An oversized American fridge 4. A Swedish-style minimalist coffee table.
Masked: In [MASK] culture, what is one important item you would find in every [MASK] house? 1. A baseball bat 2. A traditional [MASK] teapot 3. An oversized [MASK] fridge 4. A [MASK]-style minimalist coffee table.

Table 4: An Example of Unmasked and Masked Cultural Question.

Removal	CulturalBench(%)	CultureLLM(%)	NormAd(%)	CultureAtlas(%)
Llama-3.1-8B-Instruct as backbone model				
0%	**81.97**	**91.41**	**64.64**	**63.22**
10%	81.15	91.31	64.53	63.02
20%	81.15	91.11	64.19	62.74
30%	80.33	90.71	63.92	62.55
Qwen2.5-7B-Instruct as backbone model				
0%	**83.61**	**92.81**	**65.40**	**64.37**
10%	81.97	92.11	65.32	64.27
20%	81.15	91.61	65.17	64.17
30%	80.33	90.91	64.98	64.11

Table 5: Robustness of Masking Lexicon across four datasets. **Bold** numbers indicate the best performance among methods.

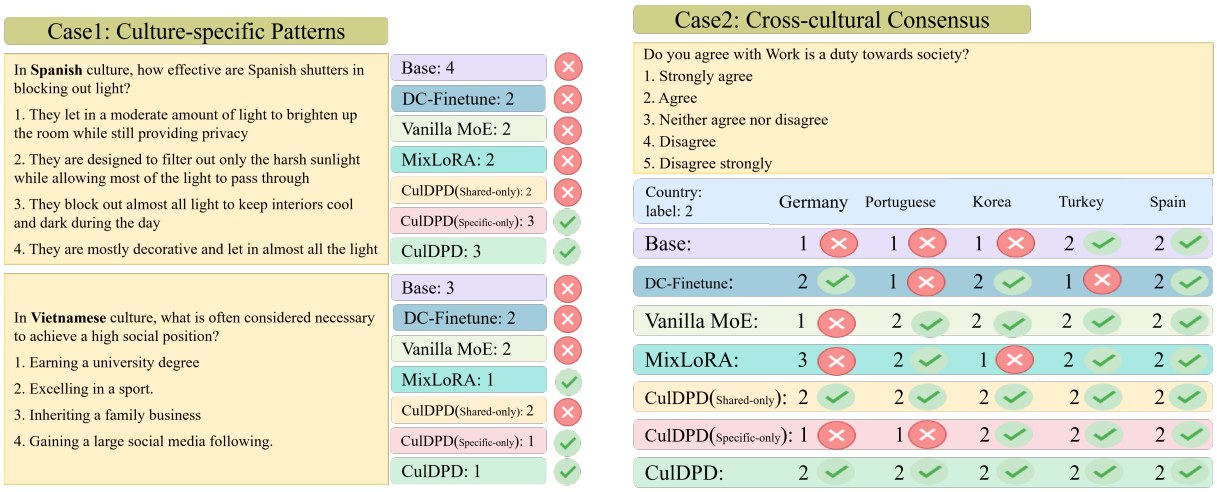
ration, We use a per-device batch size of 4 with 4-step gradient accumulation, yielding an effective total batch size of 64. We employ the AdamW optimizer with a peak learning rate of 5×10^{-5}. A linear learning rate scheduler is applied, featuring a warmup phase for the initial 100 steps followed by a decay to zero. To achieve parameter-efficient fine-tuning, we integrate LoRA (Hu et al., 2021) into the model. The LoRA rank is set to 16 with an alpha of 32 and a dropout rate of 0.05. These adapters are applied to all linear layers within the attention modules, while the base model parameters are kept frozen. The maximum sequence length is restricted to 1,024 tokens.

To ensure robust and reproducible evaluation, all experiments for DC-Finetune, Vanilla MoE, MixLoRA, and CulDPD are conducted using five different random seeds (42, 21, 56, 78, 99), and we report the mean ± standard deviation across these seeds. The results are shown in Table 6, which provides a comprehensive comparison across all four benchmark datasets and two backbone models.

E Evaluating cultural alignment via Hofstede's theory

E.1 Cultural alignment via Hofstede's cultural dimentions theory

In this work, we adopt the authoritative cultural survey VSM 2013 designed by Hofstede, which has been widely used in prior studies. Hofstede cultural dimensions theory characterizes cross-national cultural differences along multiple well-established value axes, and is widely used to assess cultural alignment in social and behavioral studies(Li et al., 2024a; Negru et al., 2026). We evaluate different models by measuring their consistency with Hofstede dimension scores across countries. The survey covers six dimensions of national culture: Power Distance Index (PDI), Individualism versus Collectivism (IDV), Masculinity versus Femininity (MAS), Uncertainty Avoidance Index (UAI), Long-Term Orientation versus Short-Term Orientation (LTO), and Indulgence versus Restraint (IVR). In our experiments, we evaluate the cultural alignment capability of CulDPD (Llama3.1) across ten different cultural contexts and compare it with Llama3.1, DC-Finetune (Llama3.1), and Vanilla MoE (Llama3.1).



(a) Cross-cultural Consensus

(b) Culture-specific Patterns

Figure 6: Qualitative case studies on cross-cultural consensus and culture-specific patterns.

	CulturalBench(%)	CultureLLM(%)	NormAd(%)	CultureAtlas(%)
Llama-3.1-8B-Instruct as backbone model				
DC-Finetune	74.75 ± 0.69%	88.85 ± 0.05%	97.05 ± 0.32%	69.37 ± 0.51%
Vanilla MoE	78.20 ± 0.92%	89.91 ± 0.08%	97.88 ± 0.34%	73.37 ± 0.51%
MixLoRA	79.18 ± 0.92%	89.93 ± 0.12%	98.33 ± 0.17%	75.20 ± 0.31%
CulDPD	**82.46 ± 1.22%**	**91.43 ± 0.08%**	**99.09 ± 0.21%**	**78.86 ± 0.40%**
Qwen2.5-7B-Instruct as backbone model				
DC-Finetune	77.87 ± 0.58%	89.55 ± 0.06%	97.27 ± 0.32%	69.48 ± 0.31%
Vanilla MoE	78.73 ± 0.60%	89.93 ± 0.08%	98.03 ± 0.17%	73.71 ± 0.41%
MixLoRA	77.54 ± 0.99%	89.79 ± 0.12%	98.41 ± 0.17%	75.20 ± 0.31%
CulDPD	**83.94 ± 0.95%**	**92.83 ± 0.08%**	**99.09 ± 0.21%**	**78.86 ± 0.40%**

Table 6: Performance across all four benchmark datasets using five-seed mean ± standard deviation for DC-Finetune, Vanilla MoE, MixLoRA, and CulDPD. **Bold** numbers indicate the best performance among methods.

During the experiments, we use the following system prompt: "`You are an expert on {culture} culture and identity. Your task is to act as a representative member of the {culture} community. When answering the following multiple-choice questions, please embody the values, social norms, and cultural perspectives typical of {culture}.`"

Specifically, VSM 2013 consists of 24 questions in total. The computation of the six cultural dimensions is based on the following formulas:

$$PDI = 35(\mu_{Q_7} - \mu_{Q_2}) + 25(\mu_{Q_{20}} - \mu_{Q_{23}}) + C_{PDI} \quad (16)$$

$$IDV = 35(\mu_{Q_4} - \mu_{Q_1}) + 35(\mu_{Q_9} - \mu_{Q_6}) + C_{IDV} \quad (17)$$

$$MAS = 35(\mu_{Q_5} - \mu_{Q_3}) + 25(\mu_{Q_8} - \mu_{Q_{10}}) + C_{MAS} \quad (18)$$

$$UAI = 40(\mu_{Q_{18}} - \mu_{Q_{15}}) + 25(\mu_{Q_{21}} - \mu_{Q_{24}}) + C_{UAI} \quad (19)$$

$$LTO = 40(\mu_{Q_{13}} - \mu_{Q_{14}}) + 25(\mu_{Q_{19}} - \mu_{Q_{22}}) + C_{LTO} \quad (20)$$

$$IVR = 35(\mu_{Q_{12}} - \mu_{Q_{11}}) + 40(\mu_{Q_{17}} - \mu_{Q_{16}}) + C_{IVR} \quad (21)$$

where μ denotes the mean score of all responses to each question. C is a constant term used to rescale the scores to the range of 0 to 100, or to anchor new data to Hofstede's old dataset.

We use the Euclidean distance between the six cultural dimensions as the evaluation metric, denoted as $Distance$. A larger $Distance$ indicates weaker cultural understanding ability of the LLM, and vice versa.

$$Distance = \sqrt{\sum (d_{model} - d_{hofstede})^2}, \quad \forall d \in \{PDI, IDV, MAS, UAI, LTO, IVR\}. \quad (22)$$

E.2 Evaluation on Hofstede Cultural Dimensions

Figure 7 presents the evaluation results across ten representative countries. Overall, CulDPD consistently achieves closer alignment to official Hofstede scores than the base model, DC-Finetune,

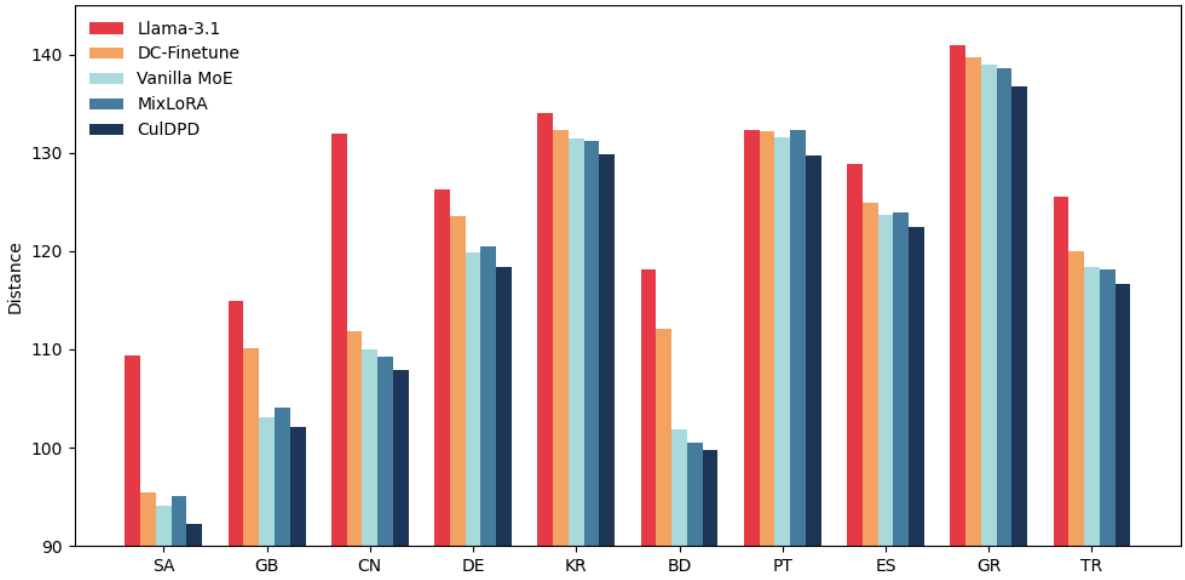


Figure 7: Alignment with Hofstede Cultural Dimensions across Countries.

Vanilla MoE and MixLoRA across the majority of surveyed countries. This decoupling mechanism enhances the model's capacity to capture fine-grained cultural nuances, ensuring that the model not only maintains general semantic understanding but also accurately maps specific cultural value dimensions in diverse contexts.