Alignment but Flattening: Analyzing the Trade-off between Cultural Alignment and Diversity in LLMs

Anonymous ACL submission

Abstract

Cultural alignment fine-tuning has become the de facto paradigm for building culture-aware large language models (LLMs), yet optimizing exclusively for alignment scores provides an incomplete portrait of cultural fidelity. This unidimensional evaluation lens prompts a fundamental question: do models genuinely perceive distinct cultural nuances, or do they merely memorize dominant cultural values? To address this, we propose a synergistic evaluation framework that jointly formalizes cultural alignment and diversity. Through extensive benchmarking of six mainstream LLMs on the World Values Survey, this framework uncovers a systematic and critical trade-off: the pursuit of cultural alignment consistently incurs an acute expense of diversity, leading to severe "cultural flattening." Via a granular investigation into these behavioral responses, we demonstrate that models artificially anchor to dominant majorities, converging onto a monolithic response pattern that systematically wipes out the heterogeneous distributions inherent to human groups. Crucially, our mechanistic analysis reveals that this diversity collapse is not a behavioral anomaly but a structural consequence of the intrinsic low-rank bias in neural network optimization. Therefore, our findings expose the fundamental limitations of current post-training paradigms and call for a shift toward alignment objectives that preserve cross-cultural pluralism.

1 Introduction

Given the rapid progress and the capacity to capture human behavioral patterns (OpenAI et al., 2024; Team et al., 2025; DeepSeek-AI et al., 2025; Yang et al., 2025), large language models (LLMs) are increasingly deployed worldwide as cultural agents, used to simulate the values and opinions of specific sociodemographic groups (Liu et al., 2025; Kolluri et al., 2025; Hu et al., 2026). For such uses, a model's *cultural alignment*, i.e., how faithfully it reproduces the values of the population it represents, becomes essential. To improve it, a growing body of work fine-tunes LLMs on culture-specific data and consistently reports substantial gains in alignment scores (Li et al., 2024a,b; Xu et al., 2024; Jiang et al., 2025), making cultural fine-tuning the de facto recipe for building culture-aware models.

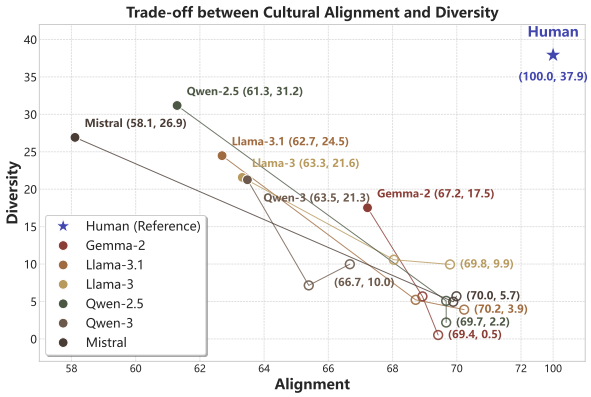

Figure 1: Trade-off between Cultural Alignment and Diversity. Solid dots represent models without fine-tuning, while hollow circles represent fine-tuned ones. For each model, different configurations are connected sequentially in descending order of cultural alignment.

However, a high alignment score alone provides an incomplete picture of cultural fidelity. Since alignment is usually optimized and reported as an aggregate similarity score, a model can improve alignment by tracking dominant response patterns while failing to preserve the diversity of opinions within and across cultures. This raises a central concern that remains underexplored in existing fine-tuning studies: does the pursuit of cultural *alignment* come at the expense of cultural *diversity*? As shown in Figure 1, cultural fine-tuning consistently shifts models toward higher alignment, but also sharply reduces their behavioral diversity, revealing a systematic trade-off between the two.

Unlike prior studies that primarily audit diversity in out-of-the-box models post hoc (Wu et al.,

2024; Murthy et al., 2025), we examine the cultural fine-tuning process itself—the stage at which this trade-off emerges and can therefore be analyzed and potentially mitigated. We first formalize cultural alignment and cultural diversity as two complementary evaluation axes. Alignment measures whether model responses match the corresponding human sociodemographic group, while diversity measures whether responses remain sensitive to differences across cultural personas. This formulation allows us to jointly evaluate whether a model becomes more culturally accurate while still preserving meaningful cross-cultural variation (§3).

Using this framework, we fine-tune six LLMs on culture-specific data derived from the World Values Survey and benchmark their behavioral shifts against real human responses. Across all models, fine-tuning improves cultural alignment, but consistently narrows behavioral diversity. Importantly, this reduction is not merely a uniform decrease in variance. Instead, it takes two distinct forms depending on the question type. For categorical questions, fine-tuned models collapse toward dominant cultural response modes, reducing a broad distribution of human opinions to one or two highly concentrated answers. For ordinal questions, models are susceptible to retreating toward a conservative middle option that covers nearby responses but fails to reproduce the full human distribution. In both cases, personas that should produce culturally differentiated responses are pushed toward more homogeneous outputs, creating the risk that non-dominant cultural perspectives are overwritten (§4).

We further investigate this trade-off from a mechanistic perspective. Building on the low-rank simplicity bias of neural network optimization (Huh et al., 2023; Súkeník et al., 2024; Galanti et al., 2024), we hypothesize that cultural fine-tuning operates under a representational bottleneck: instead of freely expanding the model's capacity to encode every cultural distinction, fine-tuning reallocates cultural representations into a narrow activation subspace. To test this hypothesis, we probe distributed representations over feed-forward network neurons and compare the activated neuron sets before and after fine-tuning. Our analysis shows that fine-tuning uses a largely decoupled but restricted activation space, limiting the model's ability to preserve heterogeneous cultural profiles. This provides evidence that cultural flattening is not only a behavioral artifact, but is connected to how cultural information is reorganized during optimization (§5).

Our main contributions are threefold:

- We introduce a unified evaluation framework that treats cultural alignment and cultural diversity as complementary axes, enabling joint measurement of whether models match human values while preserving cross-cultural variation.
- Applying this framework to six LLMs, we show that alignment gains consistently coincide with reduced behavioral diversity.
- We provide a mechanistic account of cultural flattening grounded in the low-rank simplicity bias, showing that fine-tuning compresses diverse cultural values into a low-rank subspace and thereby marginalizes minority cultures.

2 Related Work

Cultural Alignment. With the widespread deployment and application of LLMs globally, imbuing these models with robust cultural awareness has become crucial. To achieve this, prior research has explored various optimization strategies, ranging from lightweight in-context learning (AlKhamissi et al., 2024; Choenni and Shutova, 2025; Ki et al., 2025; Pham et al., 2025; Shetty et al., 2025) and supervised fine-tuning (Huang et al., 2024; Li et al., 2024a,b; Xu et al., 2024; Jiang et al., 2025) to reinforcement learning preference optimization (Feng et al., 2025; Zhao et al., 2026; Yuan et al., 2026). Despite these promising advancements, current evaluation and optimization paradigms remain heavily tethered to alignment-centric metrics (e.g., Jensen-Shannon Divergence or Hofstede's cultural dimensions) that merely measure statistical consistency against dominant human standards, with limited attention to post-alignment behavioral diversity and potentially obscuring underlying representational collapses. In contrast to prior work that largely treats cultural alignment as a linear optimization objective, we simultaneously formalize both alignment and diversity to systematically analyze their interplay, thereby revealing an inherent trade-off where traditional alignment gains come at the cost of diversity preservation. This finding highlights the critical need for a coordinated paradigm that balances cultural alignment with diversity preservation.

Cultural Diversity. Recent studies have begun to note that alignment-oriented optimization may unintentionally suppress behavioral diversity in models (Wu et al., 2024; Murthy et al., 2025), raising concerns about increasing behavioral homogenization after alignment. Separately, several recent works investigate whether deployed models can faithfully reproduce the diversity of human conceptual and behavioral distributions (Santurkar et al., 2023; Hu and Collier, 2024; Schröder et al., 2025; Wang et al., 2025). Their findings suggest that diversity losses are often difficult to recover once models have undergone alignment and deployment. In stark contrast to these existing studies that remain confined to post-hoc, black-box observations of diversity degradation, we pivot the research focus back to the optimization process itself. Specifically, we scrutinize the dynamic trajectories of multi-cultural local fine-tuning to address two fundamental questions: (1) how do superficial gains in localized alignment scores deceptively mask an underlying, latent collapse in representational and behavioral diversity? (2) is this alignment-diversity trade-off merely a superficial behavioral anomaly, or is it an intrinsic, structural consequence of neural network optimization?

3 Measuring the Alignment–Diversity Trade–off

In this work, we propose a framework that measures both the benefits of cultural alignment and the costs of losing cultural diversity.

Tasks. To formalize this, we first delineate how cultural awareness is evaluated: grounded in the sociological premise that culture manifests as structured patterns of behavior (Kroeber and Kluckhohn, 1952), such evaluations conventionally measure the behavioral congruence between prompted model simulations and target human demographics (Joshi et al., 2024). Formally, given a set of cultural contexts C and a questionnaire set Q, we evaluate the model $\mathbb{M}$ through role-play prompting. For each cultural context $c \in C$, the model is conditioned on culture-specific personas $p \sim P_c$ (e.g., "Imagine you are a Chinese female teacher") and asked to respond to survey questions $q \in Q$. This setup enables the model's culturally conditioned behavioral patterns to be elicited and compared against human reference populations.

Alignment. Under this framework, *Cultural Alignment* measures whether the model can faithfully reproduce the behavioral tendencies of the target sociodemographic group it is conditioned to simulate. Formally, we define the alignment score of model $\mathbb{M}$ as:

$$\mathcal{A}(\mathbb{M}) = \mathbb{E}_{q,c,p}\left[\mathcal{S}\big(\mathbb{M}(q|p,c), \mathbb{H}(q|p,c)\big)\right] \quad (1)$$

where $\mathbb{M}(q|p,c)$ denotes the model response under persona (p,c), $\mathbb{H}(q|p,c)$ denotes the corresponding human behavioral reference, and $\mathcal{S}(\cdot,\cdot)$ measures behavioral similarity. A higher $\mathcal{A}(\mathbb{M})$ indicates stronger behavioral consistency with the target cultural group.

Diversity. While alignment evaluates behavioral fidelity within each cultural subgroup, *Cultural Diversity* evaluates whether responses remain behaviorally distinguishable across different cultural subgroups. Specifically, a diverse model should preserve localized behavioral variation instead of collapsing toward a homogeneous response distribution. We therefore define the diversity score as:

$$\mathcal{D}(\mathbb{M}) = \mathbb{E}_{q,p_i,p_j}\left[1 - \mathcal{S}\big(\mathbb{M}(q|p_i,c_i), \mathbb{M}(q|p_j,c_j)\big)\right] \quad (2)$$

where (p_i, c_i) and (p_j, c_j) are personas sampled from different cultural groups ($c_i \neq c_j$). Crucially, distinct from the monotonic optimization of $\mathcal{A}(\mathbb{M})$ where higher scores strictly denote better fidelity, higher diversity alone does not necessarily imply better cultural realism; meaningful diversity should remain grounded in human behavioral distributions. Nevertheless, excessive reductions in this metric indicate cultural flattening, where culturally distinct behaviors collapse into increasingly homogeneous outputs.

Importantly, $\mathcal{A}(\mathbb{M})$ and $\mathcal{D}(\mathbb{M})$ are not independent objectives, but two complementary perspectives derived from the same underlying similarity function $\mathcal{S}(\cdot,\cdot)$. This shared formulation provides a unified basis for comparing behavioral fidelity and cross-cultural differentiation within a common evaluative space. In Section 4.1, we further instantiate $\mathcal{S}(\cdot,\cdot)$ using *Soft Accuracy* and operationalize both objectives under the same measurement framework.

4 Empirical Evidence of the Alignment–Diversity Trade-off

In this section, we empirically assess the alignment–diversity trade-off. We first introduce the evaluation setup (§4.1) and then present the observed

trade-off between cultural alignment and diversity (§4.2). Next, we investigate the behavioral mechanisms through which alignment gains induce unintended reductions in cultural diversity, thereby exposing the inherent marginalization risks these mechanisms impose on under-represented cultural groups (§4.3).

4.1 Experiment Setup

Evaluation Benchmark. Our evaluation is built upon the World Values Survey 7 (WVS-7) (Haerpfer et al., 2022) providing a globally representative sample of social and political values, administered in local languages to ensure cultural fidelity. To capture a broad cultural spectrum, we selected 14 countries across various continents: NGA, MAR, TUN, USA, BOL, URY, CHN, TJK, JOR, RUS, SRB, CYP, AUS, and NZL. For each country, the dataset includes 80 individual respondents, with each respondent covering responses to 31 survey items across representative value topics (e.g., "Basic kinds of attitudes concerning society"), where each item is paraphrased into three distinct variants to obtain multiple responses. Detailed descriptions and statistics of the WVS are presented in Appendix B.

Model Selection. Our evaluated models encompass six mainstream LLMs, spanning four model families with variations in both architectural versions and parameter scales: Llama-3-8B-Instruct, Llama-3.1-8B-Instruct (Touvron et al., 2023), Qwen2.5-7B-Instruct (Bai et al., 2023), Qwen3-4B (Yang et al., 2025), gemma-2-9B-it (Riviere et al., 2024), and Mistral-7b-Instruct-v0.3 (Jiang et al., 2023).

Prompt Design. We instantiate simulation respondents through persona prompts that integrate macro-level national identities with fine-grained demographic attributes, including *country, region, sex, age, social class, education level, and marital status* (AlKhamissi et al., 2024). All prompts are formulated in English as a unified baseline, which helps isolate cultural alignment from the confounding effects of varying multilingual proficiencies across models (Li et al., 2024a; Xu et al., 2024). Detailed prompt templates are provided in Appendix C.

Fine-tuning Methods. Our training corpus is also derived from the WVS-7 to ensure an in-domain evaluation, while maintaining a strict held-out separation between the training and evaluation sets. Specifically, the training data is curated by sampling 20 unique respondents per country and their responses to these aforementioned topics; in this process, we explicitly ensure the diversity of responses to prevent the training set from falling into a single pattern. Based on these data settings, we train two distinct models via supervised fine-tuning, where both models are fine-tuned for a single epoch under identical hyperparameter configurations. We denote the model trained on the full dataset as *SFT*, and the variant trained on the corpus excluding the United States and Australia as *SFT-E*, aiming to evaluate mainstream cultural influences. We provide the training corpus statistics in Appendix B, with detailed fine-tuning implementation described in Appendix D.

Similarity Metric. To operationalize the similarity function $\mathcal{S}(\cdot,\cdot)$ in Eqs. 1 and 2, we instantiate it using *Soft Accuracy (SA)*. Unlike exact-match accuracy, SA assigns partial credit to ordinal-scale responses, thereby better capturing graded behavioral similarity across culturally conditioned outputs.

Formally, for a pair of responses (y_i, y_j) to question q, SA is defined as:

$$SA(y_i, y_j) = \begin{cases} 1 - \dfrac{|y_i - y_j|}{|q| - 1} & \text{if IsOrd}(q), \\ \mathbb{I}(y_i = y_j) & \text{otherwise,} \end{cases} \quad (3)$$

where $\text{IsOrd}(q)$ indicates whether q is an ordinal-scale item, and $|q|$ denotes the number of response options.

Within our framework, the same SA metric serves two complementary roles. In Eq. 1, it measures behavioral consistency between model outputs and human references. In Eq. 2, it measures behavioral similarity across culturally distinct personas, where lower similarity indicates stronger preservation of culturally localized behaviors. However, diversity should remain grounded in realistic human behavioral distributions, as excessively high inter-group similarity may lead to cultural flattening.

4.2 Trade-off Overview

Figure 1 illustrates the joint distribution of cultural alignment and diversity across evaluated models. Along the horizontal axis (x-axis), cultural fine-tuning consistently improves alignment across diverse architectures and training settings, yielding

gains ranging from 2.2 (Gemma-2) to 11.9 (Mistral). This pervasive rightward shift indicates that cultural fine-tuning reliably optimizes models toward higher behavioral consistency with target cultural groups.

However, these alignment gains are accompanied by a substantial reduction in cultural diversity, revealing a systematic trade-off between behavioral fidelity and cultural variation. Along the vertical axis (y-axis), diversity scores for all fine-tuned models collapse below 11.0. This widespread plunge leaves a pronounced gap against the empirical human reference of 37.9 and fails to reach even the lower bound of the vanilla baselines (which uniformly stay above 17.5), statistically implying an over 90% probability that the fine-tuned cohort will generate identical responses even under substantially different cultural contexts. The trade-off is particularly severe in Gemma-2, where a modest alignment gain of 2.2 coincides with a dramatic 17-point reduction in diversity, resulting in a near-zero diversity score of 0.5. Such extreme behavioral convergence suggests that fine-tuning may improve alignment not through nuanced cultural understanding, but by collapsing culturally diverse behaviors toward a dominant response pattern.

Scope of Discussion. To rigorously isolate the origin of this behavioral convergence, we clarify that the observed alignment–diversity trade-off is strictly a *fine-tuning-induced dynamic* within each model's evolutionary trajectory, rather than an intrinsic cross-model trait. As evidenced by our empirical analysis, evaluating different architectures under any uniform, static condition (whether vanilla base W/O, SFT, or SFT-E cohorts) yields no statistically significant correlation between a model's foundational alignment capability and its output diversity ($p > 0.05$).

4.3 Mechanisms of Pseudo-Alignment

We trace the behavioral evolution from base to fine-tuned models against human benchmarks to uncover the underlying mechanics of this optimization trade-off.

To provide an intuitive understanding of these dynamics, we anchor our initial analysis on a granular visualization of the questionnaire item "How often do you attend religious services" (Figure 2). This specific item is strategically selected because its highly heterogeneous human baseline spans the maximum spectrum of seven distinct choice categories, thereby offering a highly sensitive diagnostic baseline for tracking representational erosion.

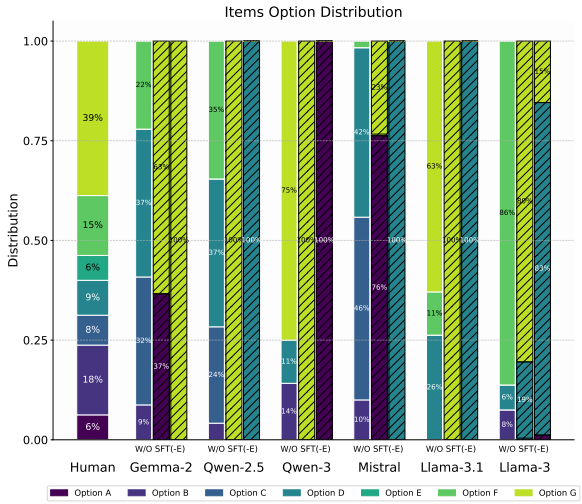

Figure 2: Behavioral response distributions across Human, Base (W/O), and fine-tuned configurations, where hatched bars denote fine-tuned models (Left: SFT; Right: SFT-E).

Mapping the models' post-fine-tuning response landscape against this rich baseline reveals a clear optimization shortcut: the models heavily anchor to the majority option G, which represents the human plurality choice at a 39% share. Crucially, this optimization yields a rigid rather than adaptive behavior, thereby risking the erroneous misgeneralization of a fixed value paradigm to cultural contexts where it is no longer mainstream. For instance, under SFT-E, Qwen-3 stubbornly over-amplifies Option A into the dominant response even in contexts where it is a minority preference endorsed by only 6% of humans. An analysis of this case allows us to uncover the underlying mechanism behind such alignment gains: **rather than sustaining a nuanced cultural spectrum, fine-tuning induces a deterministic mode collapse that exclusively anchors model outputs to a majority consensus, thereby inflating superficial alignment scores.**

Majority Anchoring Evaluation. To further expose this pseudo-alignment driven by majority anchoring, we track two primary statistical features across the evaluation set: (i) *Majority Choice Rate*, which computes the percentage of responses matching the most frequent human selection to reflect the model's tendency to conform to the plurality consensus; and (ii) *Choice Pool Size*, which records the total number of unique options the model activates to quantify the severity of representational

Model	Majority Choice Rate (%)			Choice Pool Size		
	W/O	SFT	SFT-E	W/O	SFT	SFT-E
Gemma-2-9B-it	40.93	47.59	47.50	**2.48**	**1.07**	**1.09**
Qwen2.5-7B-Instruct	40.59	**55.59**	**52.14**	**2.66**	**1.06**	**1.23**
Qwen3-4B	38.53	**52.29**	45.78	**2.39**	**1.42**	**1.45**
Mistral-7b-Instruct-v0.3	40.16	**54.80**	**51.51**	**2.64**	**1.19**	**1.31**
Llama-3.1-8B-Instruct	41.57	**56.50**	**50.92**	**2.31**	**1.65**	**1.71**
Llama-3-8B-Instruct	40.76	**57.74**	**51.49**	**2.06**	**1.51**	**1.41**
Human Reference	49.98	49.98	49.98	3.61	3.61	3.61

Table 1: The majority choice rate (%) and choice pool size across evaluated models on categorical items. **Bold** values indicate that the majority choice rate is greater than the human reference, or that the choice pool size is less than the human reference.

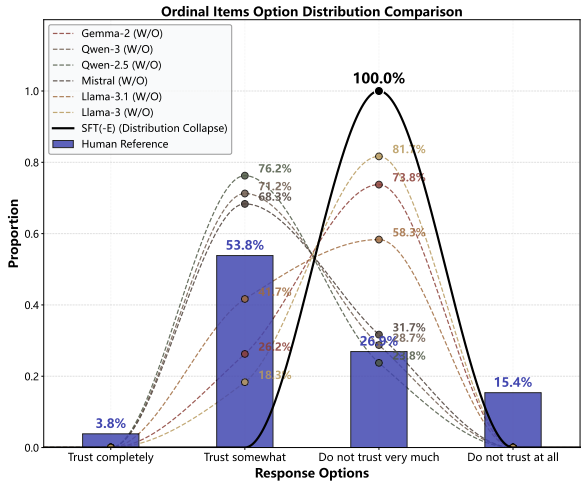

Figure 3: **Ordinal-item option distributions**. Bars indicate the Human Reference distribution. Scatter points show option proportions across settings, where dashed lines and the solid line represent the W/O and SFT(-E) settings respectively, connected via smooth trendlines.

shrinkage and behavioral homogenization.

Table 1 presents the quantitative evaluation of the *Majority Choice Rate* and the *Choice Pool Size* across models before and after fine-tuning, evaluated against the human baseline. The results reveal two clear behavioral shifts: (i) Regarding the *Majority Choice Rate*, models under both SFT and SFT-E exhibit a continuous increase compared to vanilla models (W/O). While this demonstrates that cultural fine-tuning effectively guides models toward the plurality consensus, their scores unexpectedly surpass the human reference line (49.98%), exposing a severe skewness toward dominant options. (ii) Regarding the *Choice Pool Size*, fine-tuned models fall drastically below the W/O baseline (typically > 2.0) and remain far inferior to the human baseline of 3.61. In the worst cases, this metric shrinks close to 1.0, exposing a systemic defect in sustaining cultural diversity. A joint analysis of both features further underscores the dual nature of cultural fine-tuning, revealing the inherent vulnerability behind its superficial alignment efficacy.

Conservative Anchoring Ordinal Items. Intriguingly, when transitioning to ordinal items, the fine-tuning optimization does not always dictate a strict alignment with the absolute human plurality. Instead, we unexpectedly observe a strategic shift where the model retreats toward a moderately conservative stance. As evidenced by Figure 3, while the human majority leans toward "Trust somewhat" (53.8%), the fine-tuned models undergo a complete collapse, concentrating entirely (100.0%) onto "Do not trust very much." This choice, however, covers the broader human opinion spectrum within its neighborhood (53.8%, 26.9%, and 15.4%). Thus, for ordinal items, the model shifts from absolute majority matching to a conservative anchor that maximizes neighborhood coverage.

Marginalization Risks. Crucially, the monolithic response patterns directly expose historically under-represented cohorts to profound marginalization risks. Rather than remaining an idealized theoretical construct, such systemic failures tangibly manifest within our empirical evaluation. For instance, when evaluated under a specific Nigerian context such as a persona portraying a married female from Ogun, Nigeria, fine-tuning consistently disrupts the pre-trained cultural alignment across diverse model families. Characteristically, Qwen2.5-7B experiences an alignment drop from 51.52 to 37.59, which represents a decrease of 27.04%, while Llama-3.1 falls from 50.36 to 38.17, a decline of 24.21%. These localized performance penalties demonstrate that optimization toward general alignment objectives inherently limits a model's capacity to capture nuanced minority perspectives.

To sum up, these combined empirical findings collectively corroborate that **the superficial performance gains from fine-tuning are achieved through a straightforward mechanism: the model captures the most prominent human values at the cost of diversity. Critically, when such single-response models are widely deployed, the underlying risk of cultural flattening demands serious caution.**

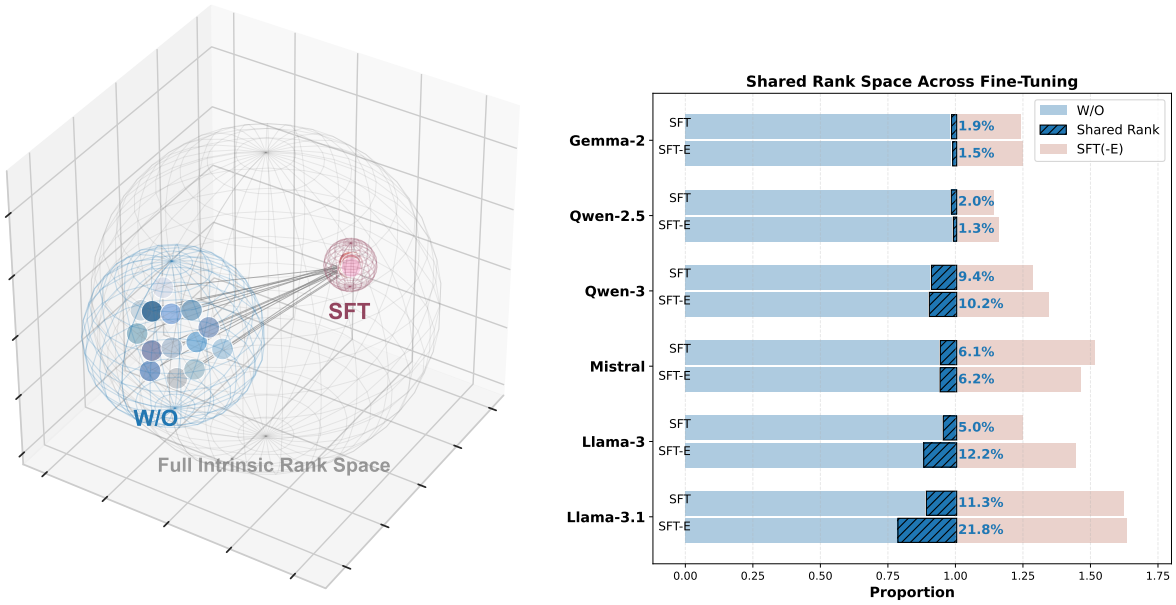

Figure 4: **Rank space distribution across tuning settings.** The left panel shows the projection of activation representation similarity for both W/O and SFT into a shared 3D space via Multidimensional Scaling. The right panel displays the shared rank space across fine-tuning for evaluated models, where the bar charts with diagonal hatching denote the proportion of shared activation ranks across fine-tuning.

5 Mechanistic Analysis

In this section, we provide a mechanistic explanation for the alignment–diversity trade–off. We first propose our theoretical hypothesis alongside a dedicated probing method (§5.1), and then empirically evaluate this underlying mechanism (§5.2).

5.1 Hypothesis and Probing Method

Hypothesis. We hypothesize that this trade-off is driven by the *Low-Rank Simplicity Bias* during neural network optimization, rather than being a mere superficial behavioral anomaly. (Huh et al., 2023; Súkeník et al., 2024; Galanti et al., 2024). Specifically, long-term convergence on massive data during pre-training establishes a stable, low-rank upper bound for the model's representations. Consequently, during fine-tuning, gradient updates are not globally flexible but are strictly confined to a narrow, leftover subspace. This restricted optimization space limits the representational capacity of neurons, ultimately manifests as a decline in behavioral diversity. We provide additional theoretical details and discussions regarding the hypothesis in Appendix F.

Probing Method. To parameterize the compressed rank space, we look into the activation space of *Feed-Forward Network (FFN) neurons*. Existing literature views FFN neurons as key-value memories that store and project conceptual knowledge (Geva et al., 2021; Tang et al., 2024; Ying et al., 2025). Therefore, the variation in the effective rank of the model's optimization space can be directly characterized by the activation patterns of these neurons. Formally, for a decoder-only Transformer at layer l, the FFN processing hidden state h^l is defined as:

$$\mathrm{FFN}^l(h^l) = W^l_{\text{down}} \cdot \theta(W^l_{\text{up}} \cdot h^l) \tag{4}$$

where W^l_{up} and W^l_{down} are weight matrices and $\theta(\cdot)$ is activation function. We denote the i-th element of $\theta(W^l_{\text{up}} \cdot h^l)$ as the i-th neuron $n_{(l,i)}$ in layer l.

During inference, the active dimensions of the rank space are characterized by neuron activations. Specifically, for each question $q \in Q$, the set of activated neurons N_q consists of those whose activations $a_{(l,i)}$ exceed a dynamic threshold $\mu(l)$:

$$N_q = \{n_{(l,i)} \mid a_{(l,i)} > \mu(l)\}, \tag{5}$$

where $\mu(l)$ is defined as the K-th largest activation in layer l (with $K = 100$). By aggregating neurons across all questions Q, we obtain model's total response neuron set $N_{\mathbb{M}}$, corresponding to its effective rank space. A lower cardinality of $N_{\mathbb{M}}$ indicates a more restricted representational space.

5.2 Empirical Analysis

Figure 4 illustrates the rank space distribution of cultural representations across countries, compar-

ing models without (W/O) and with (SFT) cultural fine-tuning. The Euclidean distance within this space measures the representational divergence of countries across different model settings.

Spatial Decoupling. As the left panel of Figure 4 shows, the rank spaces of the vanilla and fine-tuned models exhibit a high degree of geometric decoupling. Specifically, within the full intrinsic rank space of the model, which is represented by the outer grey hypersphere, the cultural representational subspaces for the vanilla and fine-tuned conditions show minimal geometric overlap. This spatial segregation implies that cultural fine-tuning operates under a strict representational bottleneck, forced to optimize within a highly compact and independent subspace. Such structural restriction not only limits the capacity of the model to encode diverse cultural profiles but also introduces the risk of erasing valuable pre-trained cultural knowledge.

To further characterize the precise degree to which the fine-tuned state separates from the vanilla baseline, we evaluate the overlap ratio of activated neurons to quantify the underlying allocation of rank space resources: $|N_{\mathbb{M}}^{(\text{W/O})} \cap N_{\mathbb{M}}^{(\text{SFT(-E)})}|/|N_{\mathbb{M}}^{(\text{W/O})}|$. We plot this metric in the right panel of Figure 4, where the empirical allocations are visually highlighted by the hatched bars. Through this direct comparison, we find that the shared representation space consistently constitutes less than 10% of the baseline space under most evaluation contexts. This numerical insight provides clear evidence that within the cultural optimization space, more than 90% of the activated dimensions emerge as entirely independent of the baseline parameter coordinates. Furthermore, this shifting trend remains largely insensitive to the training corpora, as evidenced by the highly comparable overlap ratios maintained between the standard SFT and enhanced SFT-E settings.

Spatial Restriction. Even though optimization occurs in a highly decoupled space, we observe that the model's total effective rank space does not undergo a corresponding expansion. Instead, this process is confined to an exceptionally restricted subspace. Indeed, the newly occupied rank space consistently fails to compensate for the overall dimensionality; for most models, the incremental space accounts for less than 50% of the original dimensions, dropping to a striking 15% in Qwen-2.5.

In summary, these results strongly align with our hypothesis that cultural fine-tuning is confined to a strict rank bottleneck that limits the capacity of the model to encode nuanced cultural profiles. This structural constraint provides a geometric perspective to explain why the reduction in cultural diversity emerges as a prevailing tendency throughout the alignment process.

To clarify, this paper focuses on characterizing the trade-off between alignment and diversity; the mechanistic analysis is leveraged to substantiate this observed behavior, rather than redirecting our primary scope toward theoretical neural network derivation.

6 Conclusion

In this work, we first mathematically formalize cultural alignment and diversity into a synergistic evaluation framework, through which we uncover a systematic and critical trade-off across mainstream LLMs: the pursuit of cultural alignment consistently incurs an acute expense of diversity, forcing models to anchor exclusively to dominant majorities in a monolithic, single-point response mode, ultimately leading to severe cultural flattening. Crucially, going beyond post-hoc behavioral auditing, we provide a definitive mechanistic account of this phenomenon grounded in the low-rank simplicity bias lemma of neural network optimization, empirically validating that the cultural fine-tuning process segregates and compresses diverse cultural representations into an exceptionally restricted, decoupled activation subspace. This mathematically explains why alignment inherently reduces diversity, thereby offering a fundamentally new lens for guiding future pluralistic alignment research.

Limitations

We now discuss the limitations of our work:

- On the one hand, our empirical analysis is limited to supervised fine-tuning (SFT) methods. While SFT is the most common and standard way to align LLMs; more complex optimization algorithms might show completely different behavioral dynamics regarding cultural entropy, which we have not yet systematically explored.
- On the other hand, our diagnostic evaluation relies entirely on static benchmarks, which fail to capture the highly context-dependent and fluid nature of cultural values. Moving forward, a critical research avenue is to bridge the gap between

static datasets and complex, open-ended environments, particularly concerning how interactive deployment affects long-term bias entrenchment within model representations.

References

Badr AlKhamissi, Muhammad N. ElNokrashy, Mai AlKhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. *ArXiv*, abs/2402.13231.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenhang Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, K. Lu, and 31 others. 2023. Qwen technical report. *ArXiv*, abs/2309.16609.

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

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

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. *Preprint*, arXiv:2505.19484.

Tomer Galanti, Zachary S. Siegel, Aparna Gupte, and Tomaso Poggio. 2024. Sgd and weight decay secretly minimize the rank of your neural network. *Preprint*, arXiv:2206.05794.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. *Preprint*, arXiv:2012.14913.

Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, and Bi Puranen. 2022. World values survey wave 7 (2017-2020) cross-national data-set.

Tiancheng Hu, Joachim Baumann, Lorenzo Lupo, Nigel Collier, Dirk Hovy, and Paul Röttger. 2026. Simbench: Benchmarking the ability of large language models to simulate human behaviors. *Preprint*, arXiv:2510.17516.

Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in LLM simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10289–10307, Bangkok, Thailand. Association for Computational Linguistics.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Song Dingjie, Zhihong Chen, Mosen Alharthi, Bang An, Juncai He, and 1 others. 2024. Acegpt, localizing large language models in arabic. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8139–8163.

Minyoung Huh, Hossein Mobahi, Richard Zhang, Brian Cheung, Pulkit Agrawal, and Phillip Isola. 2023. The low-rank simplicity bias in deep networks. *Preprint*, arXiv:2103.10427.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *ArXiv*, abs/2310.06825.

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

Nitish Joshi, Javier Rando, Abulhair Saparov, Najoung Kim, and He He. 2024. Personas as a way to model truthfulness in language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6346–6359.

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

Akaash Kolluri, Shengguang Wu, Joon Sung Park, and Michael S. Bernstein. 2025. Finetuning LLMs for human behavior prediction in social science experiments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30096–30111, Suzhou, China. Association for Computational Linguistics.

Alfred Louis Kroeber and Clyde Kluckhohn. 1952. Culture: A critical review of concepts and definitions. *Papers. Peabody Museum of Archaeology & Ethnology, Harvard University*.

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.

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. 2024b. Culturepark: Boosting cross-cultural understanding in large language models. *Advances in Neural Information Processing Systems*, 37:65183–65216.

Chen Cecilia Liu, Anna Korhonen, and Iryna Gurevych. 2025. Cultural learning-based culture adaptation of language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3114–3134, Vienna, Austria. Association for Computational Linguistics.

Sonia Krishna Murthy, Tomer Ullman, and Jennifer Hu. 2025. One fish, two fish, but not the whole sea: Alignment reduces language models' conceptual diversity. 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 11241–11258, Albuquerque, New Mexico. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Viet Thanh Pham, Zhuang Li, Lizhen Qu, and Gholamreza Haffari. 2025. CultureInstruct: Curating multi-cultural instructions at scale. 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 9207–9228, Albuquerque, New Mexico. Association for Computational Linguistics.

Gemma Team Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, L'eonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ram'e, Johan Ferret, Peter Liu, Pouya Dehghani Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, and 176 others. 2024. Gemma 2: Improving open language models at a practical size. *ArXiv*, abs/2408.00118.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? *Preprint*, arXiv:2303.17548.

Sarah Schröder, Thekla Morgenroth, Ulrike Kuhl, Valerie Vaquet, and Benjamin Paaßen. 2025. Large language models do not simulate human psychology. *ArXiv*, abs/2508.06950.

Anudeex Shetty, Amin Beheshti, Mark Dras, and Usman Naseem. 2025. Vital: A new dataset for benchmarking pluralistic alignment in healthcare. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22954–22974.

Peter Súkeník, Marco Mondelli, and Christoph Lampert. 2024. Neural collapse versus low-rank bias: Is deep neural collapse really optimal? *Preprint*, arXiv:2405.14468.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. *ArXiv*, abs/2402.16438.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1332 others. 2025. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

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

Angelina Wang, Jamie Morgenstern, and John P. Dickerson. 2025. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Preprint*, arXiv:2402.01908.

Fan Wu, Emily Black, and Varun Chandrasekaran. 2024. Generative monoculture in large language models. *Preprint*, arXiv:2407.02209.

Shaoyang Xu, Yongqi Leng, Linhao Yu, and Deyi Xiong. 2024. Self-pluralising culture alignment for large language models. In *North American Chapter of the Association for Computational Linguistics*.

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.

Jiahao Ying, Wei Tang, Yiran Zhao, Yixin Cao, Yu Rong, and Wenxuan Zhang. 2025. Disentangling language and culture for evaluating multilingual large language models. In *Annual Meeting of the Association for Computational Linguistics*.

Jiahao Yuan, Zixiang Di, Shangzixin Zhao, Zhiqing Cui, Hanqing Wang, Guisong Yang, and Usman Naseem. 2026. Cultural palette: Pluralising culture alignment via multi-agent palette. *Preprint*, arXiv:2412.11167.

Weixiang Zhao, Haozhen Li, Yanyan Zhao, Haixiao Liu, Biye Li, Ting Liu, and Bing Qin. 2026. Culturerl: Internalizing cultural principles in large language models via norm-driven reinforcement learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(44):38120–38128.

A The Usage of AI

In this work, the application of AI is strictly limited to aiding and polishing academic writing, e.g., the description of evaluation framework, with no involvement in core research processes.

B Detailed Description and Statistical Analysis of the WVS

B.1 Detailed Description of WVS

The World Values Survey (WVS) collects responses to a range of questions broadly categorized into social, cultural, material, governmental, ethical, and economic domains, with data drawn from demographically controlled population samples worldwide (Haerpfer et al., 2022). Its seventh wave (WVS-7), conducted between 2017 and 2021, incorporates region-specific modules alongside globally standardized categories. Comprising 259 questions, WVS-7 is designed to include indicators aligned with multiple United Nations Sustainable Development Goals. Administered as a questionnaire, the survey targets selected samples from the general population, with all questions localized to the native or dominant regional languages.

The value topics used in this work are as follows: (1) Friends, (2) People of a different race, (3) Immigrants/foreign workers, (4) Basic kinds of attitudes concerning society, (5) People of another religion, (6) People of another nationality, (7) Major Companies, (8) Private banks, (9) The United Nations (UN), (10) International Monetary Found (IMF), (11) The World Bank (WB), (12) The World Health Organization (WHO), (13) Increases the crime rate, (14) Increases the risks of terrorism, (15) Helps poor people establish new lives, (16) Losing my job or not finding a job, (17) Not being able to give one's children a good education, (18) Freedom and Equality - Which more important, (19) Freedom and security - Which more important, (20) How often do you attend religious services, (21) To make sense of life after death vs To make sense of life in this world, (22) Interest in politics, (23) Signing a petition, (24) Joining in boycotts, (25) local level, (26) Votes are counted fairly, (27) Election officials are fair, (28) How much would you say the political system in your country allows people like you to have a say in what the government does?, (29) Having a strong leader who does not have to bother with parliament and elections, (30) Having experts, not government, make decisions according to what they think is best for the country, (31) Having a system governed by religious law in which there are no political parties or elections.

Although the WVS-7 data was collected between 2017 and 2021 and may have certain temporal limitations, it remains one of the most recent and comprehensive datasets available for this research; furthermore, our sampling within each culture covers diverse demographic variables—including gender, age, and social class—to mitigate representation bias. The concerns regarding the values timeliness, contextual dynamics and sample representativeness reflect systemic challenges within the field, rather than being specific to this work. Constructing such corpora involves significant resource requirements and privacy considerations, which transcend the scope of this study.

B.2 Statistical Analysis of the Fine-tuning Corpus

Country	#Number
United States	1842
Bolivia	1746
Uruguay	1743
China	1410
Tajikistan	1695
Jordan	1662
Russia	1773
Serbia	1830
Cyprus	1674
Nigeria	1788
Morocco	1860
Tunisia	1779
Australia	1839
New Zealand	1683
SFT	24324
SFT-E	20643

Table 2: Distribution of the dataset across different countries and training configurations.

Table 2 provides an overview of the dataset composition across the fourteen target countries, as well as the total volume of samples used for the SFT and SFT-E training configurations. To ensure the quality and reliability of the training corpus, a rigorous data cleaning process was conducted. Specifically, entries containing null values, empty strings, or ambiguous labels such as "nonknown" were identified and filtered out prior to training.

This preprocessing step ensures that the models are aligned using high-quality, meaningful cultural data, thereby enhancing the robustness of the resulting cultural language models.

C Prompt Design

Imaging you are a {marital_status} {sex} from {region}, {country}. You are {age} years of age and completed {education} education level. You consider yourself part of the {social_class}.
Answer the following question from this perspective. Others will read what you choose; your goal is to convince them it was chosen from the perspective of the persona described above. Select exactly one option. Do not include any extra commentary. Answer by typing the number corresponding to your chosen answer.

Figure 5: Persona Prompting used in this work.

As described in Figure 5, the instruction first initiates contextual persona inception by dynamically injecting comprehensive demographic profiles, including marital status, sex, geographic region, country, age, education level, and social class, which shifts the model's attention weights away from globally averaged pre-training patterns toward targeted, localized cultural priors.

Concurrently, the prompt applies strict Behavioral Conditioning by embedding a socially conditioned game-theoretic objective ("your goal is to convince them it was chosen from the perspective of the persona described above"), which successfully strips away the over-smoothed safety neutrality typical of conversational assistants.

Finally, through rigid Output Structure Formalization, the model is restricted to a single categorical choice token and explicitly barred from generating any extra commentary ("Select exactly one option. Do not include any extra commentary"), ensuring a clean and deterministic extraction of the model's underlying associative probability distributions while exposing the structural space collapse and marginalization risks induced by the fine-tuning process.

D Detailed SFT Implementation

Training Protocol and Hyperparameters To understand the mechanistic alignment-diversity trade-off, we fine-tune the backbone models using the `trl` library's `SFTTrainer` on downstream cultural instruction datasets. All models are optimized for 1 epoch with a linear learning rate scheduler.

Evaluation and Hardware Configuration Crucially, to eliminate downstream generation stochasticity and guarantee strict empirical reproducibility, the text generation during the evaluation phase is executed using greedy decoding by anchoring the temperature parameter precisely to $\tau = 0$. Both the pipeline training and downstream diagnostic evaluations are implemented on a distributed hardware configuration consisting of four NVIDIA Tesla P100 GPUs, utilizing PyTorch and the Hugging Face `Transformers` library ecosystem.

E Fine-tuning Results

Country	W/O	SFT	SFT-E
United States	71.69	71.59	70.69
Bolivia	71.76	74.01	72.91
Uruguay	67.60	73.33	72.06
China	71.17	73.51	71.74
Tajikistan	67.66	72.83	73.12
Jordan	68.97	70.23	68.78
Russia	70.23	71.41	69.53
Serbia	69.58	71.62	70.66
Cyprus	66.90	74.76	68.11
Nigeria	71.14	71.00	68.43
Morocco	71.18	74.40	71.37
Tunisia	71.70	74.51	72.30
Australia	71.86	76.49	72.11
New Zealand	70.62	72.98	68.99

Table 3: Llama model metric performance. where text in red indicates that fine-tuning impairs the model's cultural alignment.

Country	W/O	SFT	SFT-E
United States	71.19	70.01	72.18
Bolivia	70.26	73.14	75.84
Uruguay	66.52	73.41	75.26
China	71.39	71.14	74.16
Tajikistan	67.42	75.39	74.53
Jordan	72.05	70.53	74.17
Russia	72.49	70.59	73.58
Serbia	71.52	71.34	73.74
Cyprus	69.55	70.58	71.40
Nigeria	70.01	71.04	74.48
Morocco	69.74	74.05	75.21
Tunisia	71.14	73.07	75.66
Australia	74.57	75.63	73.40
New Zealand	72.91	71.12	69.18

Table 4: Qwen model metric performance. where text in red indicates that fine-tuning impairs the model's cultural alignment.

The experimental results demonstrate that fine-tuning is a highly effective method for enhancing cultural alignment. Across all evaluated models, both the SFT and SFT-E strategies consistently produce superior performance metrics compared to the

Country	W/O	SFT	SFT-E
United States	68.45	72.34	75.23
Bolivia	69.13	74.72	74.39
Uruguay	68.60	74.19	72.83
China	63.77	75.61	78.64
Tajikistan	68.74	74.48	70.89
Jordan	68.09	73.31	72.97
Russia	63.15	73.22	71.99
Serbia	67.97	73.11	73.50
Cyprus	64.33	71.16	69.68
Nigeria	69.76	72.36	73.73
Morocco	70.23	76.11	73.21
Tunisia	69.23	75.41	75.63
Australia	69.25	75.77	72.44
New Zealand	68.18	71.86	70.12

Table 5: Mistral model metric performance. where text in red indicates that fine-tuning impairs the model's cultural alignment.

Country	W/O	SFT	SFT-E
United States	74.10	73.51	75.57
Bolivia	73.29	74.04	74.68
Uruguay	70.96	72.76	72.63
China	74.06	76.72	76.27
Tajikistan	69.37	73.07	71.06
Jordan	70.25	72.79	72.93
Russia	72.52	73.94	72.38
Serbia	73.38	73.76	73.00
Cyprus	71.11	74.24	69.80
Nigeria	70.62	71.26	74.30
Morocco	72.91	74.47	73.38
Tunisia	72.61	75.77	76.03
Australia	77.43	76.70	72.42
New Zealand	77.04	74.38	69.38

Table 6: gemma model metric performance. where text in red indicates that fine-tuning impairs the model's cultural alignment.

baseline (W/O), confirming that fine-tuning successfully integrates target cultural preferences into the models.

The SFT strategy exhibits superior stability, consistently delivering reliable performance gains across the majority of countries and models, which makes it a robust approach for general cultural alignment. In contrast, the SFT-E strategy demonstrates higher volatility; while it achieves substantial performance breakthroughs in specific regions, it remains prone to occasional regressions. This suggests that the SFT-E strategy, despite its potential for deeper alignment, is more sensitive to distribution shifts within the training data.

Extreme cases illustrate the nuances of these strategies. For instance, the significant gain observed in Mistral-China (from 63.77 to 78.64) underscores how targeted fine-tuning can effectively

Country	W/O	SFT	SFT-E
United States	71.77	70.73	70.00
Bolivia	71.01	72.25	72.69
Uruguay	66.44	73.14	72.27
China	73.24	74.65	72.51
Tajikistan	67.67	73.84	72.99
Jordan	72.13	72.21	69.41
Russia	72.86	71.69	69.66
Serbia	71.33	72.08	70.74
Cyprus	69.11	69.27	69.14
Nigeria	69.86	72.04	70.60
Morocco	69.19	74.19	72.88
Tunisia	71.22	74.59	73.80
Australia	73.68	78.06	74.48
New Zealand	72.43	74.56	70.93

Table 7: Llama3.1 model metric performance. where text in red indicates that fine-tuning impairs the model's cultural alignment.

bridge initial gaps in cultural knowledge. Conversely, regressions such as those seen in Qwen-New Zealand highlight the risks of overfitting or catastrophic forgetting, where aggressive training may disrupt pre-existing, well-aligned knowledge. Finally, cases showing marginal gains, such as Llama3.1-Cyprus, suggest that some models already possess a high degree of cultural awareness, leading to diminishing returns from further fine-tuning.

F Low-Rank Simplicity Bias in Cultural Alignment

The observed decline in cultural diversity during fine-tuning is rooted in the intrinsic low-rank simplicity bias of neural network optimization. During Supervised Fine-Tuning (SFT), gradient updates are not globally flexible; instead, they are strictly confined to a narrow, low-rank manifold established during the pre-training phase. This restriction creates a rank bottleneck that forces the model to prioritize a limited set of representational dimensions.

Polysemantic Competition and Interference. Due to this restricted rank, neurons are forced to become polysemantic, encoding multiple cultural features within the same limited subspace. In our task, when the model optimizes for a specific cultural value (alignment), polysemantic collisions occur. Dominant alignment signals, which provide the steepest gradient for loss reduction, effectively interfere with and suppress more nuanced, marginalized cultural signals. This "interference" prevents the model from simultaneously representing the

full spectrum of cultural heterogeneity

Rank Starvation and Geometric Collapse. This phenomenon leads to a state of rank starvation, where the available degrees of freedom in the representational manifold are insufficient to accommodate high-dimensional cultural diversity. Optimization dynamics under weight decay and SGD favor solutions that collapse features toward their class means, a process known as Neural Collapse. Consequently, while the model successfully "aligns" to target cultural averages, it achieves this by physically "squeezing out" the variability of thought

To sum up, In the context of cultural SFT, the "Flattening" effect is not merely a data frequency issue but a physical consequence of resource allocation under rank constraints. The model essentially trades off its latent representational capacity for optimization efficiency, leading to a singular, low-rank representation that lacks the diverse "rank budget" necessary to simulate heterogeneous human populations.