What Aggregate Accuracy Hides: Cultural Affective Inequity in Multilingual LLMs

Youngjin Lee [1] **Hayoung Oh** [1]

Abstract

Multilingual LLMs may exhibit similar aggregate accuracy while failing differently across languages and emotions. To analyze cross-cultural disparities hidden by aggregate evaluation, we introduce two complementary disparity metrics: the **Cultural Inequity Score (CIS)**, which quantifies cross-linguistic disparity concentration, and the **Emotion-Stratified Gap (ESG)**, which measures emotion-specific cross-cultural performance gaps. Evaluating four multilingual LLMs on the CEDAR benchmark across six languages, we identify three cases where aggregate evaluation obscures distinct cross-cultural failure patterns: similarly low Swahili accuracy can arise from fundamentally different failure modes across models; a multilingual-oriented model exhibits lower CIS than a larger general-purpose model despite lower aggregate accuracy; and happiness shows the largest cross-cultural ESG across all four models despite relatively high mean accuracy among 14 emotion categories. These results suggest that aggregate accuracy metrics alone may be insufficient for pluralistic affective evaluation, as important distributional patterns of cross-cultural failure can remain obscured without decomposition.

1. Introduction

As multilingual LLMs are deployed across increasingly diverse cultural communities, evaluation must assess not only overall performance, but also how performance disparities are distributed across cultural groups. The dominant evaluation paradigm, however, relies primarily on aggregate accuracy — a scalar summary that can obscure important variation in cross-cultural performance patterns. Under pluralistic evaluation settings, this compression may conceal which language communities experience the largest disparities, which emotion categories contribute disproportionately to those disparities, and whether superficially similar low-performance outcomes arise from different underlying failure behaviors.

We identify three layers of what we term *evaluation compression*. First, aggregate accuracy can conflate qualitatively distinct failure behaviors: instruction-following breakdown and semantically valid but inaccurate affective classification may produce superficially similar low-performance signals despite reflecting different underlying limitations. Second, models with comparable aggregate accuracy can exhibit substantially different distributions of cross-cultural disparity. Third, emotion categories with relatively high mean accuracy can simultaneously exhibit substantial cross-cultural disparity, such that aggregate emotion-level evaluation may obscure uneven performance concentrated within specific language communities.

Existing multilingual evaluation methods typically report per-language accuracy or extreme-group comparisons such as max–min gaps, but can obscure how cross-cultural disparities are distributed across languages and emotion categories. To analyze these hidden disparity patterns, we introduce two complementary evaluation metrics: the **Cultural Inequity Score (CIS)**, which measures how unevenly affective performance is distributed across languages, and the **Emotion-Stratified Gap (ESG)**, which measures emotion-specific cross-cultural performance gaps.

Applying these metrics to four multilingual LLMs evaluated on the CEDAR benchmark (Dai et al., 2026) across six languages, we identify all three forms of evaluation compression. Identical low performance on Swahili corresponds to substantially different failure behaviors across models, ranging from instruction-following breakdown to semantically valid but inaccurate affective classification. Models with similar aggregate accuracy can also exhibit substantially different disparity profiles under CIS. Most notably, happiness — commonly treated as a culturally universal basic emotion under Ekman's universality hypothesis (Ekman & Friesen, 1971) — produces the largest cross-cultural ESG across all four evaluated models despite ranking second in mean accuracy among all 14 emotion categories. Together,

[1]Department of Artificial Intelligence, Sungkyunkwan University, Suwon, South Korea. Correspondence to: Hayoung Oh <hyoh79@g.skku.edu>.

Pluralistic Alignment Workshop @ ICML 2026, Seoul, South Korea.

these findings suggest that aggregate evaluation does not merely simplify cross-cultural disparity — it may obscure its structure in ways that are consequential for pluralistic alignment.

2. Background and Related Work

Evaluation of multilingual LLMs has predominantly relied on aggregate accuracy metrics that summarize performance as a single scalar across languages and tasks. While such metrics efficiently capture overall capability, they may obscure how disparities are distributed across language communities and affective categories under pluralistic evaluation settings. This motivates the need for evaluation methods that characterize cross-cultural affective disparities beyond aggregate accuracy alone.

Pluralistic Alignment. Sorensen et al. (2024) formalize the aggregation problem in LLM alignment: standard post-RLHF procedures compress heterogeneous human preferences into a single modal response, with post-aligned models diverging further from real human preference distributions than their pre-aligned counterparts. Subsequent work similarly finds that fine-tuning on aggregated preferences can privilege majority viewpoints while reducing responsiveness to minority perspectives, even when demographic diversity is represented in annotation pools (Bakker et al., 2022; Kharchenko et al., 2024). This literature has primarily examined preference aggregation at the individual level. By comparison, how cross-cultural affective disparity is distributed across language communities has received less systematic analysis.

Multilingual LLM Evaluation. Persistent performance disparities across languages have been documented across both task-specific and generative LLMs, with language resource level often emerging as a major correlate of performance variation (Blasi et al., 2022; Hu et al., 2020). Large-scale multilingual benchmarks further suggest that scaling alone does not eliminate cross-lingual performance gaps, including in frontier models (Ahuja et al., 2023; Han et al., 2025). However, these disparities are typically reported using per-language accuracy or extreme-group comparisons such as max–min gaps, while the broader distributional structure of cross-linguistic disparity remains less systematically characterized.

Affective Computing and Cross-Cultural Emotion Recognition. Cross-cultural variation in emotion recognition is well established in affective science. Ekman & Friesen (1971) proposed the universality of basic emotions, but subsequent work has repeatedly found that recognition accuracy tends to be higher within than across cultural groups, and that cultural distance predicts lower cross-cultural recognition accuracy and larger in-group advantages (Laukka & Elfenbein, 2020). Recent NLP research has likewise begun examining cultural variation in LLM affective behavior, finding that multilingual models may reflect Western affective norms in non-English settings and reproduce nationality-associated emotional stereotypes (Havaldar, 2023; Kamruzzaman et al., 2025). CEDAR (Dai et al., 2026) further demonstrates that language consistency and cultural alignment can diverge substantially in multilingual LLM evaluation.

Collectively, this literature suggests that cross-cultural variation in affective behavior is not represented uniformly in current multilingual LLMs. However, to our knowledge, prior work has not systematically examined how cross-cultural affective disparity is distributed across both languages and emotion categories, or how distinct failure behaviors contribute to those disparities under aggregate evaluation.

3. Methodology

3.1. Benchmark and Cultural Grounding

We evaluate on the CEDAR benchmark (Dai et al., 2026), comprising 1,166 narrative–question pairs per language across seven typologically diverse languages. Importantly, CEDAR is not a translated benchmark: scenarios are independently constructed within each language community to be culturally resonant, rather than adapted from a shared source corpus. This design reduces reliance on translation fidelity and source-language transfer when interpreting cross-language performance differences.

Each instance is labeled with one of 14 culturally contextualized emotion categories: six basic emotions (anger, disgust, fear, happiness, sadness, surprise) and eight social emotions (amusement, awe, contentment, desire, embarrassment, pain, relief, sympathy). Ground-truth labels are elicited from native speaker communities within each language group rather than imposed through externally standardized annotation schemes.

Accordingly, model errors may reflect differences between model predictions and locally grounded affective judgments within each language community. This property is important for our purposes because it allows cross-language accuracy differences to be examined with reduced dependence on translation artifacts or source-language transfer effects, making CEDAR a suitable basis for the distributional disparity analysis conducted in this study.

3.2. Models and Evaluation Protocol

We evaluate four LLMs selected to represent distinct regions of the current model landscape: GPT-4o (OpenAI; parameter count undisclosed) as a high-performance proprietary

model (OpenAI et al., 2024), Mistral-Small-3.2 (Mistral AI; 24B) and LLaMA3.2-11B (Meta; 11B) as general-purpose open-source models (Mistral AI, 2025; Grattafiori et al., 2024), and Aya-Expanse-8B (CohereForAI; 8B) (Dang et al., 2024) as a multilingual-oriented open-source model optimized for cross-lingual generalization. Including Aya-Expanse allows us to examine whether multilingual training objectives are associated with lower cultural affective disparity in this evaluation setting. All models are prompted using the language-specific templates provided by the CEDAR authors, presenting each narrative, question, and closed set of 14 emotion options in the target language. Inference is conducted at temperature $= 0$ and seed $= 42$ for reproducibility.

Accuracy metric. We adopt total-instance accuracy as the primary evaluation metric, treating invalid responses — outputs that do not match any of the 14 emotion labels — as incorrect. Invalid responses are retained rather than excluded because response-generation failure remains operationally relevant in multilingual evaluation settings. Under this criterion, models are evaluated jointly on affective classification accuracy and the ability to produce valid label outputs within each language context.

LLaMA3.2-11B exhibits an unusually high invalid-response rate for Spanish (45.5%). Re-evaluation at temperature = 0.3 produces a comparable invalid rate (42.5%), suggesting that the behavior reflects a stable instruction-following failure rather than a stochastic decoding artifact. Spanish results for LLaMA3.2-11B should therefore be interpreted with this caveat.

Language exclusion criterion. The primary disparity analysis is conducted on six languages. Swahili is analyzed separately because Aya-Expanse-8B (21.3%) and LLaMA3.2-11B (22.9%) exhibit invalid-response rates exceeding a pre-specified 10% threshold used to identify evaluation settings substantially affected by instruction-following failure. No other language meets this criterion. Swahili results are therefore reported separately in 4.1 as a validity-level decomposition case.

3.3. Cross-Cultural Disparity Metrics

We introduce two complementary metrics for exposing cross-cultural affective disparity patterns that can remain hidden under aggregate evaluation.

Cultural deviation and the sign-cancellation problem. For each model, language $c \in C$, and emotion category $e \in E$, let

$$\delta_{c,e} = \mathrm{acc}_{c,e} - \bar{\mathrm{acc}}_e \tag{1}$$

denote the deviation of per-language emotion accuracy from the cross-language mean for emotion e. Across all model–language–emotion combinations, deviation values are nearly symmetric around zero (50.3% positive, 49.7% negative; range $[-0.504, +0.555]$), indicating that cross-cultural affective variation is not concentrated in a single directional pattern.

We therefore analyze disparity along two complementary dimensions: language-level disparity (CIS) and emotion-level disparity (ESG).

Cultural Inequity Score (CIS). Let acc_c denote a model's total-instance accuracy on language $c \in C$, where $|C| = 6$. CIS is defined as the Gini coefficient of the per-language accuracy distribution:

$$\mathrm{CIS} = \frac{2\sum_{i=1}^{n} i \cdot \tilde{x}_i}{n\sum_{i=1}^{n} \tilde{x}_i} - \frac{n+1}{n} \tag{2}$$

where $\tilde{x}_1 \leq \cdots \leq \tilde{x}_n$ denotes the rank-sorted per-language accuracy sequence and $n = |C|$. CIS $\in [0, 1]$, with lower values indicating more evenly distributed performance across languages and higher values indicating greater concentration of disparity.

We adopt the Gini coefficient because it captures the overall distribution of performance across languages rather than only extreme pairs. Unlike max–min disparity measures, Gini is sensitive to the full distribution of language performance and can therefore capture disparity concentration even when no single language dominates the observed gap. Compared to standard deviation, Gini is less sensitive to absolute accuracy scale, enabling more interpretable comparisons across models with different overall performance levels (Khanuja et al., 2023). We therefore use CIS as a descriptive measure of disparity concentration rather than as a normative measure of multilingual quality.

CIS therefore quantifies the concentration of cross-linguistic performance disparity. Importantly, CIS does not identify the source of that disparity: elevated CIS may reflect instruction-following breakdown, culturally uneven semantic performance, or their combination. Interpretation therefore requires decomposition of the underlying failure behaviors, as demonstrated in §4.1.

Table 1. **Swahili performance decomposition across evaluated models**. Although three models exhibit similarly low aggregate performance on Swahili, invalid-response behavior reveals distinct underlying failure structures. GPT-4o maintains stable performance relative to its six-language average, Mistral-S-3.2 exhibits low accuracy without instruction-following breakdown, and Aya-Expanse-8B and LLaMA3.2-11B exhibit substantial invalid-response failure. Δ denotes the difference between Swahili accuracy and each model's six-language average accuracy.

Model	SW acc	Invalid rate	avg-6 acc	Δ
GPT-4o	0.487	0.26%	0.464	+0.023
Mistral-S-3.2	0.247	0.00%	0.386	-0.139
LLaMA3.2-11B	0.110	22.90%	0.236	-0.126
Aya-Expanse-8B	0.036	21.27%	0.291	-0.255

Emotion-Stratified Gap (ESG). For each emotion category $e \in E$, ESG measures the worst-case cross-cultural performance gap:

$$\mathrm{ESG}_e = \max_c(\mathrm{acc}_{c,e}) - \min_c(\mathrm{acc}_{c,e}) \tag{3}$$

Whereas CIS characterizes disparity across languages at the aggregate level, ESG captures how strongly disparity concentrates within individual emotion categories. ESG intentionally adopts a worst-case formulation because its purpose is diagnostic rather than distributive: we aim to identify emotion categories exhibiting the largest cross-cultural divergence rather than characterize the full disparity distribution. A model with uniformly moderate cross-cultural gaps and one with an extreme gap concentrated in a single emotion can therefore exhibit similar CIS values but substantially different ESG structure.

We report model-averaged ESG_e across all 14 emotion categories, along with per-model $\mathrm{ESG}_{\mathrm{mean}}$ and $\mathrm{ESG}_{\mathrm{max}}$ as summary statistics.

Relationship between CIS and ESG. CIS and ESG capture different dimensions of cross-cultural affective disparity. CIS characterizes how unevenly overall performance is distributed across language communities, whereas ESG identifies which emotion categories exhibit the largest cross-cultural gaps. Together, the two metrics make visible disparity structures that aggregate accuracy alone does not capture.

4. Results

We organize the results around three forms of evaluation compression: (1) validity-level compression (Swahili), (2) distributional compression (CIS), and (3) emotion-level compression (ESG).

4.1. Validity-Level Compression: The Swahili Case

We begin with Swahili, which illustrates the first layer of *evaluation compression*: aggregate evaluation collapses qualitatively distinct failure modes into a similar low-performance signal.

Under aggregate evaluation, three of four models appear uniformly weak on Swahili (Mistral-S-3.2: $\Delta = -0.139$; LLaMA3.2-11B: $\Delta = -0.126$; Aya-Expanse-8B: $\Delta = -0.255$ relative to each model's six-language average). GPT-4o is the sole exception, achieving slightly higher accuracy on Swahili than its six-language average (0.487 vs. 0.464, $\Delta = +0.023$). However, decomposition by invalid-response behavior suggests that these models do not fail in the same way (Table 1).

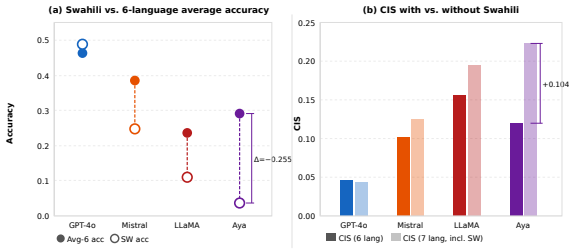

Figure 1. **Validity-level compression in the Swahili evaluation setting.** (a) Swahili accuracy relative to each model's six-language average accuracy. Although Mistral-S-3.2, LLaMA3.2-11B, and Aya-Expanse-8B all appear to underperform on Swahili under aggregate evaluation, the magnitude and structure of the deviation differ substantially across models. GPT-4o is the sole model whose Swahili accuracy exceeds its six-language average. (b) CIS computed with and without Swahili. Including Swahili substantially inflates CIS for Aya-Expanse-8B and LLaMA3.2-11B due to high invalid-response rates, while GPT-4o remains comparatively stable. The figure illustrates how aggregate inequity metrics can conflate instruction-following breakdown with cross-cultural performance disparity.

Instruction-following failure vs. semantically valid degradation. LLaMA3.2-11B and Aya-Expanse-8B exhibit high invalid-response rates on Swahili (22.9% and 21.3%, respectively), indicating substantial breakdown in closed-set instruction following. This pattern is particularly pronounced for Aya-Expanse: its highest invalid-response rate among the remaining languages is only 1.54% (Japanese), making the Swahili invalid rate approximately $13.8\times$ larger. These results suggest a localized response-generation failure in the Swahili evaluation setting rather than uniformly unstable multilingual behavior.

Mistral-S-3.2 exhibits a different pattern. Its Swahili invalid-response rate is 0.0% across all 1,166 instances, indicating fully valid closed-set responses throughout evaluation. Nevertheless, Swahili accuracy remains substantially below the model's six-language average (0.247 vs. 0.386, $\Delta = -0.139$). Unlike Aya and LLaMA, Mistral main-

tains response validity while exhibiting degraded affective classification performance.

Figure 1 visualizes both the relative Swahili performance deviation and the resulting inflation of CIS under the seven-language setting.

GPT-4o as a contrast case. GPT-4o provides an important contrast. Its Swahili accuracy slightly exceeds its six-language average ($+0.023$), suggesting that the degradation observed in other models is unlikely to arise uniformly from Swahili itself. Instead, performance degradation appears to depend substantially on model-specific behavior under the Swahili evaluation setting.

Compression effects on disparity measurement. These distinctions also affect measurement of disparity. When Swahili is included in the seven-language analysis, Aya-Expanse-8B's CIS increases from 0.120 to 0.223 (+86.5%), driven largely by the collapse of Swahili accuracy to 0.036. However, CIS describes the unevenness of performance disparities across languages rather than the source of those disparities. In this case, the increase may partly reflect concentrated instruction-following failure rather than culturally grounded affective disparity alone.

By contrast, GPT-4o's CIS slightly decreases when Swahili is included ($0.0454 \rightarrow 0.0427$, -6.0%), consistent with its comparatively stable Swahili performance. These results suggest that aggregate disparity metrics alone may obscure distinctions between response-generation failure and semantically valid but inaccurate prediction behavior unless validity-level behavior is examined separately. Validity-level decomposition therefore provides an additional interpretive layer for distinguishing response-generation breakdown from affective classification degradation in the presence of valid outputs.

4.2. Distributional Compression: CIS and the Equity-Accuracy Gap

Having established that aggregate evaluation conflates distinct failure modes at the validity level, we turn to a second layer of compression: the distributional structure of cross-cultural disparity within the six-language primary analysis.

Table 2 and Figure 2 summarize aggregate accuracy and cross-linguistic disparity structure across all four models. CIS broadly follows average accuracy ordering within this model set. However, rank consistency is not the primary contribution of CIS. Bootstrap confidence intervals overlap across all model pairs (Table 2), CIS is interpreted here as a descriptive characterization of disparity structure rather than a confirmatory ranking metric. Its value lies in making visible the cross-linguistic disparity patterns that aggregate accuracy compresses into a single scalar summary.

The Aya finding. Aggregate accuracy identifies Aya-Expanse-8B as the second-weakest model (avg $= 0.291$ vs. LLaMA3.2-11B's 0.236). CIS reveals a different structure: Aya exhibits lower cross-linguistic disparity than LLaMA (0.120 vs. 0.156), despite its relatively low overall accuracy. This pattern is corroborated by English premium, defined as the gap between English accuracy and the six-language mean. Aya's English premium (+0.052) is comparable to GPT-4o's (+0.063) and substantially smaller than LLaMA's (+0.156), indicating that Aya's failures are distributed more evenly across languages.

Mistral-S-3.2 exhibits the strongest English concentration among the four models (+0.134), combining high English performance with the largest cross-linguistic accuracy range. Aggregate accuracy therefore obscures an important distinction: Aya underperforms globally, but its failures are distributed more evenly across cultural communities than those of LLaMA or Mistral.

These patterns are consistent with the interpretation that multilingual optimization may reduce concentration of cultural performance gaps even when overall accuracy remains comparatively low. We note, however, that Aya-Expanse-8B and LLaMA3.2-11B differ in model size, preventing causal attribution to training design alone.

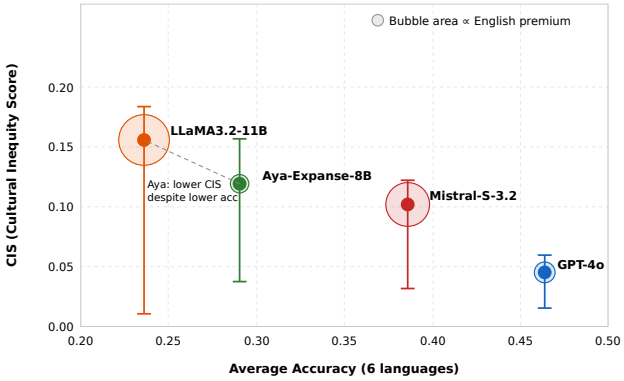

Figure 2. **Aggregate Accuracy, CIS, and English Performance Concentration** Relationship between aggregate accuracy and Cultural Inequity Score (CIS) across evaluated models. Each point represents a model in the six-language primary analysis; error bars denote bootstrap 95% confidence intervals, and bubble area is proportional to English performance premium. Models with similar aggregate accuracy can exhibit substantially different cross-linguistic disparity structures.

Resource-Level Disparity Structure. A second structure emerges when CIS is computed separately for high-resource languages (English, Chinese, Spanish) and low-resource languages (Arabic, Hindi, Japanese). Contrary to the expectation that disparity should be greatest among low-

Table 2. **Aggregate Performance and Cross-Cultural Disparity Structure Across Models**. Mean accuracy alone does not fully characterize cross-cultural performance structure: models with similar aggregate performance can exhibit substantially different disparity concentration (CIS), English performance concentration (EN Premium), and emotion-level disparity patterns (ESG). Happiness produces the largest emotion-stratified gap across all four models despite relatively high mean accuracy.

Model	Mean Acc	CIS	95% CI	EN Premium	ESG_mean	Happiness ESG
GPT-4o	0.464	0.045	[0.015, 0.059]	+0.063	0.376	0.692
Mistral-S-3.2	0.386	0.102	[0.032, 0.122]	+0.134	0.579	0.817
Aya-Expanse-8B	0.291	0.120	[0.038, 0.157]	+0.052	0.443	0.808
LLaMA3.2-11B	0.236	0.156	[0.011, 0.184]	+0.156	0.523	0.912

resource languages, CIS is consistently higher within the high-resource subset across all four models (e.g., GPT-4o: 0.057 vs. 0.021; LLaMA3.2-11B: 0.182 vs. 0.091).

This pattern suggests that cross-linguistic disparity is not explained by resource availability alone. Low-resource languages cluster at similarly reduced performance levels, whereas high-resource languages diverge more substantially from one another. Chinese is the worst-performing language for three of four models despite its high-resource status (Appendix Table A2), indicating that high training-resource availability does not necessarily correspond to more evenly distributed affective performance.

What aggregate accuracy compresses. CIS reveals how cross-linguistic performance disparities are distributed across cultural communities rather than reducing them to a single aggregate ranking. However, CIS itself aggregates across emotion categories: a model with uniformly moderate cross-cultural gaps and one with a severe disparity concentrated in a single emotion category can produce similar CIS values. Section 4.3 examines this third layer of compression.

4.3. Emotion-Level Compression: ESG and the Happiness Paradox

CIS characterizes disparity across languages but aggregates over emotion categories by design. ESG instead decomposes disparity along the emotion axis, revealing that cross-cultural affective disparity is unevenly distributed across emotion categories.

ESG rank ordering is largely model-dependent — with one consistent exception. Emotion-level ESG rankings vary substantially across models (Kendall's $\tau = -0.055$ to 0.231, all $p > .27$), suggesting that cross-cultural disparity does not follow a stable universal hierarchy of emotion difficulty. Against this backdrop of inconsistency, one pattern remains invariant across all evaluated models: happiness produces the maximum ESG in every case.

The happiness outlier. Happiness exhibits the highest model-averaged ESG (0.807), exceeding the second-ranked emotion, amusement (0.557), by 0.250 (Figure 3; Appendix Table A1). It also produces the maximum ESG for all four evaluated models individually: GPT-4o (0.692), Mistral-S-3.2 (0.817), Aya-Expanse-8B (0.808), and LLaMA3.2-11B (0.912).

This result contrasts with assumptions commonly associated with Ekman's universality hypothesis (Ekman & Friesen, 1971), in which happiness is treated as a culturally universal basic emotion. In our evaluation, however, happiness exhibits the largest cross-cultural disparity among all 14 emotion categories despite ranking second in mean accuracy (0.564).

The specific languages advantaged or disadvantaged vary across models, suggesting that the observed pattern is unlikely to arise from a single language-specific annotation effect alone. One possible interpretation is that multilingual LLMs may encode culturally uneven expectations for happiness recognition that remain difficult to detect under aggregate evaluation. However, the present analysis does not identify the mechanism underlying this disparity pattern, and alternative explanations — including benchmark-specific effects, response-format interactions, or uneven multilingual supervision — remain possible.

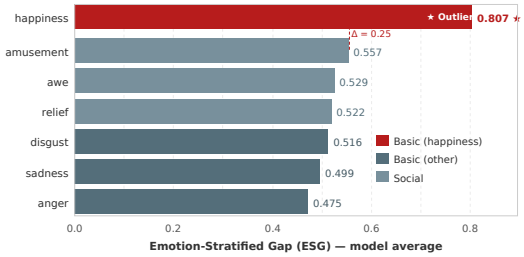

Figure 3. **Model-averaged ESG for the seven highest-gap emotion categories**. happiness (ESG = 0.807) exceeds the second-ranked emotion, amusement (0.557), by a margin of 0.25, and is the highest-ESG emotion across all four models individually. The gap between happiness and the remaining emotions is model-invariant (Kendall τ range: -0.055 to 0.231; all $p > .27$). Full ESG values for all 14 emotion categories are reported in Appendix Table A1.

Social vs. basic emotions. Excluding happiness as a pre-specified outlier, social and basic emotions exhibit broadly comparable ESG levels. Two of four models show higher ESG for social emotions, but one-sided Mann–Whitney U tests remain non-significant for all models ($p > .30$). We therefore interpret this pattern as suggestive rather than confirmatory.

Taken together, these findings suggest that aggregate emotion-level accuracy alone may obscure substantial cross-cultural disparity concentrated within specific emotion categories, even when overall emotion recognition performance appears comparatively strong.

5. Discussion and Limitations

Aggregate metrics remain useful summaries of overall performance, but our results suggest they can obscure how cross-cultural disparity is distributed across languages and emotion categories. Superficially similar aggregate outcomes often correspond to qualitatively different underlying behaviors — including instruction-following breakdown, semantically valid but inaccurate affective classification, and emotion-category-level disparity concentration — which may require different forms of analysis and intervention. CIS and ESG are therefore best interpreted as complementary decomposition tools rather than replacements for aggregate evaluation.

While CEDAR reports language-specific accuracy variation and prediction-bias analyses, our goal is different. Rather than identifying which languages perform better or worse, CIS characterizes how disparity is distributed across the entire language set, while ESG identifies emotion categories where disparity concentrates. The validity-level decomposition further distinguishes instruction-following failure from semantically valid affective degradation, a distinction not directly visible under aggregate benchmark reporting.

The happiness outlier illustrates a case where aggregate evaluation is particularly misleading: happiness ranks second in mean accuracy yet consistently produces the largest cross-cultural ESG across all four models. One possible interpretation is that multilingual LLMs encode culturally uneven expectations for happiness recognition in ways that aggregate evaluation cannot detect. The present analysis does not identify the underlying mechanism directly, nor does it address human affective universality.

We do not interpret elevated CIS or ESG as evidence of model bias alone. Because CEDAR labels reflect community-specific affective judgments, observed disparities may arise from a combination of cultural variation and model-specific representation differences.

Several limitations apply. All analyses are based on four models and six languages on a single narrative benchmark, limiting generalizability. CIS quantifies disparity concentration but not its underlying causes, making validity-level decomposition necessary for interpretation. The resource-level CIS pattern is reported as an empirical observation rather than a causal claim.

6. Conclusion

This paper examined how aggregate evaluation can obscure the structure of cross-cultural affective performance in multilingual LLMs. Across three layers of analysis, models with similar aggregate outcomes exhibited substantially different disparity structures — in failure concentration, emotion-specific gaps, and response validity behavior — suggesting that aggregate metrics alone may provide an incomplete characterization of cross-cultural alignment. The most consistent finding was the happiness outlier: happiness produced the largest cross-cultural ESG across all four models despite ranking second in mean accuracy, raising the possibility that multilingual LLMs do not process happiness-related affective judgments uniformly across language communities. Future work should examine whether these compression phenomena generalize across larger model sets, additional languages, and interaction settings beyond narrative emotion elicitation.

Impact Statement

This paper examines cross-cultural disparity patterns in multilingual affective evaluation. Our findings may help improve evaluation practices for multilingual LLMs by identifying failure patterns that can remain hidden under aggregate metrics. However, CIS and ESG quantify distributional patterns within the evaluated benchmark setting and should not be interpreted as comprehensive measures of a model's cultural appropriateness or as definitive rankings of cultural equity across models. Observed disparities may reflect a combination of cultural variation, benchmark characteristics, and model-specific representation differences rather than model bias alone. We therefore emphasize that the proposed metrics are intended as complementary analytical tools rather than normative measures of cultural fairness.

References

Ahuja, S. et al. MEGAVERSE: Benchmarking large language models across languages, modalities, models and tasks, 2023.

Bakker, M. A. et al. Fine-tuning language models to find agreement among humans with diverse preferences, 2022.

Blasi, D., Anastasopoulos, A., and Neubig, G. Systematic inequalities in language technology performance across

the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 5486–5505, 2022. doi: 10.18653/v1/2022.acl-long.376.

Dai, C. et al. Tears or Cheers? benchmarking llms via culturally elicited distinct affective responses, 2026.

Dang, J. et al. Aya Expanse: Combining research breakthroughs for a new multilingual frontier, 2024.

Ekman, P. and Friesen, W. V. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2):124–129, 1971. doi: 10.1037/h0030377.

Grattafiori, A. et al. The Llama 3 herd of models, 2024.

Han, W. et al. MuBench: Assessment of multilingual capabilities of large language models across 61 languages, 2025.

Havaldar, S. Multilingual language models are not multicultural: A case study in emotion, 2023.

Hu, J. et al. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization, 2020.

Kamruzzaman, M. et al. From anger to joy: How nationality personas shape emotion attribution in large language models, 2025.

Khanuja, S., Ruder, S., and Talukdar, P. Evaluating the diversity, equity, and inclusion of NLP technology: A case study for indian languages. In *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 1763–1777, 2023. doi: 10.18653/v1/2023.findings-eacl.131.

Kharchenko, J. et al. How well do llms represent values across cultures? empirical analysis of llm responses based on hofstede cultural dimensions, 2024.

Laukka, P. and Elfenbein, H. A. Cross-cultural emotion recognition and in-group advantage in vocal expression: A meta-analysis. *Emotion Review*, 13(1):3–11, 2020. doi: 10.1177/1754073919897295.

Mistral AI. Mistral small 3.2, 2025. Model card.

OpenAI et al. GPT-4o system card, 2024.

Sorensen, T. et al. A roadmap to pluralistic alignment, 2024.

A. Additional Results

Table A1. **Model-Averaged ESG Across Emotion Categories** Model-averaged Emotion-Stratified Gap (ESG) across all 14 emotion categories. Happiness exhibits the largest cross-cultural disparity despite relatively high aggregate accuracy, substantially exceeding the second-ranked emotion category (amusement). "Worst Language (modal)" denotes the language most frequently associated with the lowest performance across evaluated models.

Emotion	Type	ESG	Worst Language (modal)
Happiness	Basic	0.8095	Chinese
Amusement	Social	0.5569	Swahili
Awe	Social	0.5215	Swahili
Disgust	Basic	0.5161	Hindi
Relief	Social	0.5157	Swahili
Anger	Basic	0.4950	Arabic
Sadness	Basic	0.4534	Swahili
Surprise	Basic	0.4431	Swahili
Sympathy	Social	0.4411	Chinese
Pain	Social	0.4292	Japanese
Desire	Social	0.3927	Arabic
Fear	Basic	0.3752	Arabic
Contentment	Social	0.3640	Arabic
Embarrassment	Social	0.3056	Swahili

Table A2. **Per-Language Total-Instance Accuracy Across Evaluated Models** Per-language total-instance accuracy across evaluated models. Chinese is the lowest-performing language for three of four models despite its high-resource status. English consistently exhibits the highest or near-highest accuracy across models. Swahili results are reported separately in the primary analysis due to elevated invalid-response rates for Aya-Expanse-8B and LLaMA3.2-11B.

Model	Arabic	Chinese	English	Hindi	Japanese	Spanish	Swahili
GPT-4o	0.486	0.407	0.527	0.449	0.443	0.473	0.487
Mistral-S-3.2	0.383	0.334	0.520	0.337	0.307	0.434	0.247
LLaMA3.2-11B	0.273	0.182	0.392	0.190	0.184	0.196	0.110
Aya-Expanse-8B	0.352	0.182	0.343	0.353	0.246	0.268	0.036