Cross-Lingual Sentiment Misalignment: Auditing Multilingual Language Models for Inversion Risk, Dialectal Representation, and Affective Stability

Nusrat Jahan Lia
Institute of Information Technology
University of Dhaka, Dhaka, Bangladesh
bsse1306@iit.du.ac.bd

Shubhashis Roy Dipta
University of Maryland, Baltimore County
Baltimore, Maryland, USA
sroydip1@umbc.edu

Abstract

Recent advances in multilingual representation learning aim to bridge the performance gap between high- and low-resource languages, yet their ability to preserve affective meaning across languages remains underexplored, particularly for underrepresented languages like Bengali. This research addresses cross-lingual sentiment misalignment between Bengali and English by introducing a controlled benchmarking framework evaluating four multilingual transformer models on parallel Bengali-English sentence pairs, stratified by dialect, to assess their representational stability. We demonstrate that a compressed model architecture exhibits a 28.7% "Sentiment Inversion Rate," fundamentally misinterpreting positive semantics as negative (or vice versa). Consequently, we identify a cross-lingual sentiment skew that we call "Asymmetric Empathy", where models systematically dampen or artificially amplify the affective weight of Bengali text relative to its exact English counterpart. Finally, we expose a key vulnerability regarding dialectal representation: a "Modern Bias" in the regional model, which exhibits a 57% increase in alignment error when processing the formal Bengali register compared to modern colloquial text. As foundational encoders continue to serve as safety classifiers and reward models for LLM pipelines, cross-lingual reliability becomes a critical concern. We therefore advocate for the integration of "Affective Stability" metrics into future cross-lingual benchmarks to detect and penalize polarity inversions, particularly in low-resource settings.

1 Introduction

Multilingual models have rapidly become the backbone of language-agnostic information access, spanning sentiment analysis, content moderation, retrieval-augmented systems, and downstream knowledge-intensive applications. However, when

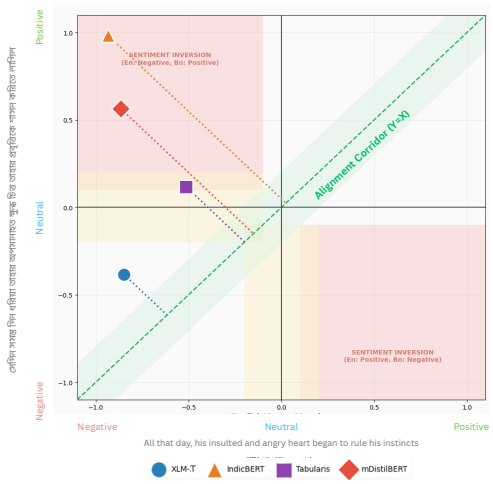

Figure 1: The plot maps the predicted English sentiment vector (X-axis) against the Bengali vector (Y-axis) for a test sentence. The green corridor represents ideal cross-lingual sentiment alignment ($Y \approx X$). While the large-scale XLM-T model successfully preserves polarity, the compressed (mDistilBERT) and regionally specialized (IndicBERT) architectures are projecting a negative English statement into a positive Bengali latent space (Red Zone).

a multilingual model correctly identifies that an English sentence carries negative sentiment but simultaneously classifies its Bengali semantic equivalent as positive (Figure 1), the representational promise of multilingualism collapses in practice.

Multilingual sentiment encoders, including those for Bengali, are deployed as core components in content moderation systems, where biases can propagate to real-world safety failures; for instance, an audit of Bengali sentiment tools revealed identity-based inconsistencies that undermine moderation accuracy (Das et al., 2024). Similarly, Tan et al. (2025) use multilingual classifiers for government-scale moderation in diverse languages with reliance on foundational encoders for LLM safety pipelines. Such instances illustrate the risks of sentiment misalignment in operational safety classifiers.

Proceedings of the 1st Workshop on Multilinguality in the Era of Large Language Models (MeLLM 2026), pages 127–139

To mitigate the *curse of multilinguality*, the phenomenon where adding more languages to a fixed-capacity model degrades per-language performance (Conneau et al., 2020), we investigate the "Sentiment Inversion" crisis and its broader representational implications. We demonstrate that the multilingual curse manifests not only as accuracy degradation but also as affective inversion. We employ a standardized cross-lingual sentiment alignment framework to evaluate the semantic consistency of four multilingual sentiment classifier models on parallel Bengali–English text (see Table 4).

We show that alignment is not uniform across languages or dialects, as summarized in the following key findings:

Capacity Constraints & Alignment Instability: We demonstrate that the capacity-constrained mDistilBERT architecture exhibits a 28.7% "Sentiment Inversion Rate," accompanied by a heavy-tailed distribution of errors (Figure 3), suggesting that model compression may compromise the safety margins required for cross-lingual alignment. We further find that regional specialization alone does not resolve this issue; however, a distilled architecture (Tabularis) that leverages synthetic data can reduce such representational misalignments.

Asymmetric Empathy: We reveal that models exhibit consistent cross-lingual directional bias, systematically dampening or artificially inflating the emotional intensity of non-English text.

The Dialectal Sentiment Gap: We find a "Modern Bias" in which models align well with colloquial text but underperform on formal dialects, which are the backbone of Bengali literature.

2 Related Works

This section reviews prior work in multilingual representation, cross-lingual sentiment analysis, and dialect-aware NLP, with a focus on gaps in affective alignment and evaluation.

2.1 Bengali in the Multilingual NLP Landscape

Bengali remains one of the most underrepresented major languages in NLP despite its global speaker base. Kabir et al. (2024) provide a comprehensive audit of LLM performance on Bengali NLP tasks, identifying key failure modes including language generation errors, verbose mismatching with evaluation metrics, and task-specific weaknesses. Bhowmik et al. (2025) document consistent performance gaps for Bengali relative to English across recent LLMs, tracing these to tokenization inefficiency, where models fragment Bengali script into excessive subword units, degrading semantic coherence. The BnMMLU benchmark further demonstrates that even large-scale frontier models show sublinear returns in Bengali reasoning as model size increases, an empirical fingerprint of the multilingual curse at scale (Joy and Shatabda, 2025; Conneau et al., 2020).

For sentiment analysis specifically, BanglaBERT (Bhattacharjee et al., 2022) established a strong monolingual baseline, and ensemble transformer systems (Hoque et al., 2024) have achieved high aggregate accuracy. However, in a landmark audit of Bengali sentiment analysis tools, Das et al. (2024) reveal that aggregate accuracy conceals systematic identity-based biases, with tools exhibiting differential performance across gender, religious, and national identity signals. This colonial impulse in tool design highlights how reductionist representations reanimate historical hierarchies and motivates the external auditing approach we adopt. Our work extends this critical perspective by focusing specifically on cross-lingual affective misalignment and dialectal representational harm.

2.2 The Multilingual Curse and Capacity Constraints

The curse of multilinguality, originally formalized by Conneau et al. (2020), describes the empirical observation that, under fixed model capacity, adding more languages to pretraining initially benefits low-resource languages through positive transfer but eventually degrades per-language performance due to inter-language parameter competition. Recent work has substantially refined this understanding. Blevins et al. (2024) demonstrate that the curse can be partially lifted through Cross-lingual Expert Language Models (X-ELM), which decouple per-language capacity via modular training and outperform jointly trained multilingual models across 16 languages. Foroutan et al. (2025) further argue that the curse arises not from language count per se but from finite model capacity amplifying the impact of noisy, low-quality data in low-resource languages. This has direct implications for Bengali, where pretraining data is both scarce and noisier than for high-resource languages.

2.3 Cross-Lingual Sentiment Analysis and Representational Failures

The adoption of transformer architectures has substantially advanced sentiment analysis in low-resource languages (Bhowmick and Jana, 2021). Cross-lingual transfer from high-resource to low-resource languages has been enabled both through shared representations in pretrained multilingual models (Conneau et al., 2020) and through machine translation strategies (Poncelas et al., 2020). Chen et al. (2025) document that while GPT-4 achieves approximately 84.4% F1 in English sentiment, this drops to around 67% for low-resource languages, and propose adaptive self-alignment strategies with data augmentation to partially close this gap. Recent literature shows that hybrid approaches retaining lexicon features maintain stability advantages over purely neural representations (Mahmud et al., 2024), and ensemble methods can achieve high aggregate accuracy (Hoque et al., 2024).

A growing body of literature documents that high aggregate accuracy routinely masks representational failures (Das et al., 2024). Wasi et al. (2024) show that LLMs acquire social biases through surface linguistic cues, and that Bengali dialectal variation, particularly religious dialect variation, induces systematic performance divergence in large models. Ochieng et al. (2025) extend this critique, demonstrating that reasoning-based LLM sentiment evaluation in low-resource, culturally nuanced contexts reveals failures invisible to label-prediction benchmarks. The CuLEmo benchmark (Belay et al., 2025) further shows that multilingual LLMs systematically fail to capture culturally grounded variations in emotional expression across languages. These cultural layers of failure are distinct from, but related to, the cross-lingual affective misalignment we document.

2.4 Bengali Diglossia and Dialectal NLP

Dialectal variation represents a significant challenge for multilingual NLP, as Wasi et al. (2024) demonstrate through empirical evaluation of LLMs on Bengali religious dialects. Bengali exhibits a well-documented diglossic structure comprising Sadhu Bhasha (formal/literary, Sanskrit-derived vocabulary, archaic conjugation) and Cholito Bhasha (colloquial/standard, simplified morphology, contemporary vocabulary), a stylistic split that presents significant challenges for multilingual NLP due to the frequent blending of these forms in everyday communication (Ayman et al., 2025). The critical insight motivating our dialect stratification is that training data for multilingual models is overwhelmingly drawn from contemporary digital sources such as large-scale web crawls, social media, and news corpora (Conneau et al., 2020; Kakwani et al., 2020), creating a training distribution that is inherently skewed toward the more common, modern Cholito Bengali form (Ayman et al., 2025).

2.5 Benchmarking Gaps and the Need for Affective Stability Metrics

Current multilingual benchmarks, including XTREME, XNLI, and their derivatives, evaluate cross-lingual performance on semantic tasks such as natural language inference, question answering (e.g., MLQA, TyDiQA), and named entity recognition (e.g., WikiAnn), with XNLI focusing on entailment classification (Hu et al., 2020). While effective at measuring semantic transfer, these benchmarks largely overlook affective fidelity. Recent efforts such as MMAFFBen (Liu et al., 2025) begin to address this gap by introducing affective evaluation, but comprehensive measurement of sentiment preservation and cross-lingual affective alignment remains limited. Ochieng et al. (2025) explicitly call for benchmarks that measure LLM sentiment in low-resource, culturally nuanced contexts beyond label accuracy. Miah et al. (2024) note that translation-based cross-lingual sentiment approaches can achieve high aggregate accuracy while failing to capture culturally grounded variations in emotional expression, often introducing translation biases in affective intensity.

We interpret these findings as evidence of directional distortions in how sentiment is mapped across languages, which we formalize as asymmetric empathy. Recent work on multilingual bias evaluation (Wasi et al., 2024; Sadhu et al., 2025) has established that social bias in Bengali LLMs operates across gender and religious lines, but has not examined the cross-lingual affective alignment dimension we investigate. Our proposal for affective stability metrics, which explicitly penalize polarity inversions and dialectal divergence, responds to this benchmarking gap, extending recent work on stability-focused evaluation metrics (Atil et al., 2024) to the multilingual affective domain.

3 Methodology

We employ a controlled experimental framework to quantify cross-lingual sentiment alignment in multilingual transformer architectures. We adopt a *within-model* comparative design in which each transformer processes parallel Bengali-English text pairs independently, enabling direct measurement of semantic divergence without inter-model architectural confounds.

3.1 Dataset Specification

We utilize a parallel corpus comprising $n = 7{,}350$ Bengali-English sentence pairs, sourced from the publicly available "BanglaBlend" dataset (Ayman et al., 2025). Formally, the dataset $\mathcal{D}$ is defined as a set of tuples:

$$\mathcal{D} = \{(B_i, E_i, D_i) \mid i \in [1, n]\} \tag{1}$$

where:

- $B_i \in \Sigma^*_{\text{Bengali}}$ represents the i-th Bengali sentence (original text),
- $E_i \in \Sigma^*_{\text{English}}$ represents the corresponding English translation,
- $D_i \in \{\text{Sadhu}, \text{Cholito}\}$ denotes the Bengali dialect classification.

Bengali exhibits diglossia with two primary written forms:

1. **Sadhu Bhasha**: Formal/literary register, characterized by Sanskrit-derived vocabulary and archaic verb conjugations.
2. **Cholito Bhasha**: Colloquial/standard register with relatively simplified morphology and contemporary vocabulary.

3.2 Models Evaluated

We benchmark four multilingual transformer architectures representing distinct design paradigms: XLM-T, a large-scale model fine-tuned on high-volume multilingual social media data; IndicBERT, a regionally specialized encoder for Indian languages; Tabularis, a distilled multilingual model enhanced with synthetic data for broad cross-lingual coverage; and mDistilBERT, a compressed multilingual model designed for efficient zero-shot sentiment transfer. Full repository mappings are provided in Table 4. This selection enables systematic analysis of how scale, regional specialization, synthetic augmentation, and compression interact with cross-lingual affective alignment.

These specific models were selected based on three core inclusion criteria: they are publicly accessible, they are capable of zero-shot Bengali sentiment inference without requiring further task-specific fine-tuning, and together they cover a principled spectrum of architectural capacity (large-scale vs. compressed) and design intent (global vs. regional). Furthermore, because our experimental design isolates cross-lingual representation and affective transfer as the primary variables of interest, we focus exclusively on multilingual architectures, purposefully excluding monolingual models.

3.3 Experimental Design

For each model M and sentence pair (B_i, E_i), we perform independent inference on both languages using the same model weights θ_M:

$$\hat{y}_{B,i} = M_\theta(B_i) \quad \text{[Bengali stream]} \tag{2}$$

$$\hat{y}_{E,i} = M_\theta(E_i) \quad \text{[English stream]} \tag{3}$$

Any divergence between $\hat{y}_{B,i}$ and $\hat{y}_{E,i}$ is attributable to cross-lingual representation, calibration, or decision boundary alignment within the same parameter space.

3.4 Score Normalization and Metric Formulation

3.4.1 Universal Score Normalizer

To enable direct comparison, we define a universal normalization function φ : Predictions $\to [-1, 1]$:

$$\varphi(\hat{y}) = \begin{cases} \hat{y}.\text{score} & \text{if } \hat{y}.\text{label} \in \{\text{positive}\} \\ -\hat{y}.\text{score} & \text{if } \hat{y}.\text{label} \in \{\text{negative}\} \\ 0 & \text{if } \hat{y}.\text{label} \in \{\text{neutral}\} \end{cases} \tag{4}$$

**Note: Standard "Positive" and "Negative" labels map to $\pm s$ for 2-class and 3-class models, while intermediate positive/negative classes are scaled by 0.5 in the 5-class Tabularis model to accommodate the "Very Positive" and "Very Negative" extremes. This maintains a uniform linear spacing across sentiment intensity levels, ensuring that the five classes are equidistant on the sentiment continuum.*

After normalization, we obtain continuous sentiment scores:

$$S_{B,i} = \varphi(\hat{y}_{B,i}) \in [-1, 1] \quad (5)$$

$$S_{E,i} = \varphi(\hat{y}_{E,i}) \in [-1, 1] \quad (6)$$

where $S_{B,i}$ and $S_{E,i}$ denote the Bengali and English sentiment scores, respectively.

3.4.2 Sentence-Level Alignment Metrics

For each sentence pair i, we compute four alignment metrics:

M1. Alignment Divergence

$$D_i = |S_{B,i} - S_{E,i}| \in [0, 2] \quad (7)$$

Interpretation: $D_i = 0.0$ indicates perfect alignment (identical sentiment), $D_i = 0.5$ indicates moderate divergence, $D_i = 2.0$ indicates maximal divergence (opposite extremes).

M2. Directional Bias

$$B_i = S_{E,i} - S_{B,i} \in [-2, 2] \quad (8)$$

Interpretation: $B_i > 0$ indicates that the English text is predicted as more positive (or less negative) than its Bengali counterpart; $B_i < 0$ indicates the reverse; $B_i \approx 0$ indicates minimal cross-lingual divergence for that pair.

M3. Polarity Inversion (Safety Metric)

$$I_i = \mathbb{1}\begin{bmatrix}(S_{B,i} > \tau \wedge S_{E,i} < -\tau) \\ \vee\, (S_{B,i} < -\tau \wedge S_{E,i} > \tau)\end{bmatrix} \quad (9)$$

where $\mathbb{1}[\cdot]$ is the indicator function and $\tau = 0.1$ is a noise threshold to avoid false positives from near-zero scores.

Interpretation: $I_i = 1$ indicates sentiment inversion (e.g., Bengali=Positive, English=Negative); $I_i = 0$ indicates polarity preserved. Inversion is the most severe failure mode, as it indicates misalignment of sentiment direction.

3.5 Population-Level Aggregation and Statistical Computation

To characterize model-level performance and evaluate representational equity across Bengali diglossia, we aggregate the sentence-level metrics into population-level statistics. We compute these across the entire dataset $\mathcal{D}$ as well as its stratified dialectal subsets ($\mathcal{D}_{\text{Sadhu}}$ and $\mathcal{D}_{\text{Cholito}}$).

Metric	Tabularis	XLM-T	IndicBERT	mDistilBERT
Mean Div.	0.200	0.276	0.375	0.417
Std Dev.	0.214	0.298	0.607	0.429
Sadhu Div.	0.239	0.286	0.459	0.456
Cholito Div.	0.161	0.266	0.292	0.379
Dialect Gap	0.078	0.020	0.167	0.077
Sadhu Err. Inc. (%)	48.4	7.6	57.1	20.5
Robustness (%)	43.1	42.1	58.3	34.2
Inversions	635	267	1471	2107
Inv. Rate (%)	8.6	3.6	20.0	28.7
Dir. Bias (En-Bn)	0.002	0.057	0.106	-0.066

Table 1: Metric scores per model

Table 2 details the mathematical formulations and interpretations for all population-level evaluation criteria, including overall alignment statistics, safety indicators, and specialized metrics designed to quantify the dialectal gap.

4 Results

Beyond average alignment error, our evaluation identifies three critical failure modes in multilingual model behavior. Table 1 presents the corresponding quantitative results.

4.1 Finding 1: Sentiment Inversion and Alignment Instability

We define a *sentiment inversion* as a case where a Bengali-English translation pair (similar meaning) receives opposite polarity classifications (positive vs. negative). Such inversions represent major alignment failures, as propositional meaning is preserved but affective interpretation is reversed. Across model architectures, inversion rates vary dramatically, revealing a potential relationship between compression, divergence magnitude, and optimization (see Figure 2).

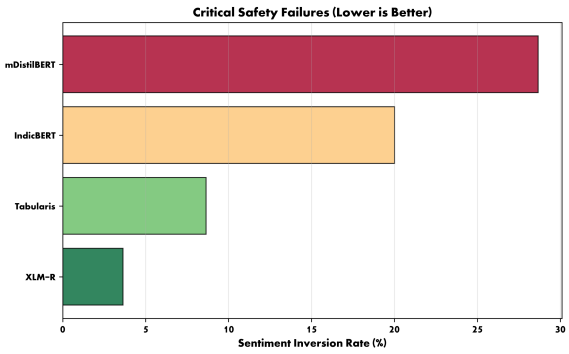

Figure 2: Sentiment Inversion Rate Across Models

- **Compression and Elevated Inversion Risk.** The distilled multilingual architecture (mDistilBERT) exhibits the highest mean alignment divergence and lowest robustness (see Table 1). Nearly one in three sentence pairs

Metric	Formulation	Interpretation
Mean Alignment Error (Divergence)	$\mu_D(M) = \frac{1}{n}\sum_{i=1}^{n} D_i$	Lower μ_D indicates better average cross-lingual consistency.
Std. Deviation (Divergence)	$\sigma_D(M) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(D_i - \mu_D)^2}$	Quantifies the variability in alignment quality across the dataset.
Robustness Index	$\mathcal{R}(M) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[D_i < 0.1] \times 100\%$	% of pairs with negligible divergence; measures the "safe operating zone."
Inversion Rate	$\mathcal{I}_{\text{rate}}(M) = \frac{1}{n}\sum_{i=1}^{n} I_i \times 100\%$	% of sentence pairs exhibiting sentiment polarity flips.
Mean Directional Bias	$\mu_B(M) = \frac{1}{n}\sum_{i=1}^{n} B_i$	> 0: English favored; < 0: Bengali favored; ≈ 0: No systematic language skew.
Formal Penalty (Dialect Gap)	$\Delta_{\text{dialect}}(M) = \mu_D(M, \text{Sadhu}) - \mu_D(M, \text{Cholito})$	> 0 implies a "Modern Bias" where the model struggles with formal registers.
Relative Dialect Error	$\Delta^{\%}_{\text{dialect}}(M) = \frac{\Delta_{\text{dialect}}(M)}{\mu_D(M, \text{Cholito})} \times 100\%$	Normalizes the formal penalty by the baseline colloquial error rate for fair comparison.

Table 2: Summary of population-level alignment and dialectal metrics.

processed by the compressed model receives directly contradictory affective classifications across Bengali and English. This indicates that while compression improves efficiency, it may disproportionately reduce the representational capacity required for reliable affective calibration. As a result, sentiment polarity reversal emerges as a critical failure mode that distorts core cross-lingual meaning.

- **Heavy-Tailed Failure Distribution.** Alignment error density analysis (Figure 3) shows that divergence is not normally distributed. Instead, the compressed architecture exhibits a long right tail, corresponding to extreme polarity flips. While IndicBERT achieves the highest robustness metric, it simultaneously exhibits a massive standard deviation in divergence and a high inversion rate with alignment errors that are not normally distributed. This non-normal distribution has practical implications: mean divergence understates actual risk, and models cannot be reliably characterized by their average behavior for deployment in critical downstream applications.

- **Scale-Driven Inversion Resilience.** The large-scale multilingual model (XLM-T) records the lowest polarity inversion rate (Table 1) across all evaluation pairs. This suggests that massive parameter scale and diverse pre-training may preserve coherent affect mappings more effectively than regional specialization, buffering against semantic instability. Crucially, however, the robust distilled Tabularis model (8.6% inversion rate) implies that compression with data-centric optimization can substantially close the gap between compressed and full-scale architectures. Hence, scale is not the only path to alignment stability. As a DistilBERT-based model fine-tuned with diverse synthetic multilingual data, Tabularis shows that targeted training strategies, rather than scale alone, can drive alignment stability.

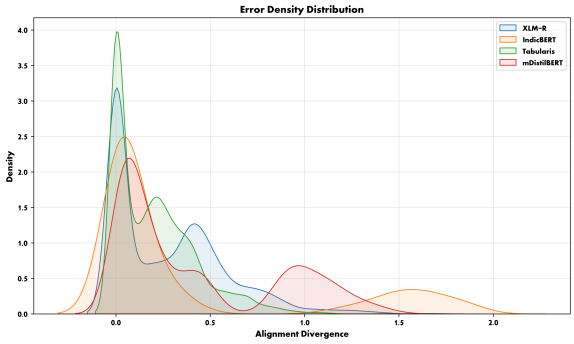

Figure 3: Distribution of Alignment Error Density

4.2 Finding 2: Representational Harm and the Dialectal Gap

To evaluate robustness under Bengali diglossia, we compare alignment divergence across colloquial (Cholito) and formal (Sadhu) variants. A multilingual system should maintain stable cross-lingual

calibration regardless of lexical register. However, we observed a dialectal sensitivity pattern.

- **Modern-Register Overfitting.** Both IndicBERT and Tabularis exhibit sharp increases in divergence when processing formal Sadhu text (see Table 1). This indicates that alignment quality is concentrated in modern and high-frequency lexical distributions, while archaic or formal constructions fall outside the model's calibrated semantic manifold. We term this phenomenon *Modern Bias*: strong alignment in contemporary usage, but underperformance in formal registers. The *Modern Bias* finding has direct consequences for linguistic equity: users who communicate in formal registers including academic, literary, and administrative Bengali, receive poorer cross-lingual sentiment alignment than users of colloquial varieties. As a result, current multilingual systems risk institutionalizing a structural inequity in which access to reliable inference is contingent on conforming to simplified or non-native linguistic norms.

  Conversely, XLM-T demonstrates strong dialectal resilience, likely due to its large-scale multilingual training over diverse text distributions, which provides broader lexical and syntactic coverage and supports stable affective mappings across regional variation.

4.3 Finding 3: Asymmetric Empathy and Directional Bias

In a multilingual architecture, the system must preserve not just the polarity but the intensity of user intent. We evaluate this using *Directional Bias* (English score − Bengali score). An ideally aligned system should yield a distribution centered at zero with low variance. Instead, we observed two distinct alignment regimes (as seen in Table 1, Figure 4 and Figure 5)

- **Compression-Induced Bengali Positivity Skew.** The distilled architecture (mDistilBERT) exhibits a negative directional bias, indicating that it systematically scores Bengali text as more positive (or less negative) than its exact English translation. For a human user in a safety-critical context: the model may artificially dampen the severity of a negative Bengali sentiment, underweighting the severity of negative Bengali content relative to equivalent English content. As demonstrated in the case study (Figure 5), mDistilBERT exhibits a safety failure by correctly scoring an English statement as deeply negative (-0.981) while assigning a positive score (+0.533) for its exact Bengali equivalent.

- **Regional English Optimism Bias.** Conversely, IndicBERT demonstrates a positive mean directional bias, assigning higher positivity (or lower negativity) to English inputs relative to their Bengali counterparts. This imposes an equity penalty on Bengali users, as their neutral or moderately negative statements are penalized with harsher negative classifications compared to English speakers.

These findings indicate that cross-lingual affective misalignment is consistently and directionally structured rather than stochastic, which can lead to downstream distortions in applications relying on Bengali sentiment signals.

4.4 Proposed Metric: Affective Stability Index

Our empirical evaluation reveals that aggregate alignment error (μ_D) alone is insufficient to capture the critical safety failures inherent in cross-lingual representation. Specifically, models with relatively moderate average divergence may still exhibit high rates of sentiment inversion ($\mathcal{I}_{rate}$) (Table 1). To address this benchmarking gap and formally quantify cross-lingual reliability, we introduce the Affective Stability Index ($\mathcal{AS}$). We define Affective Stability as a composite metric that rewards tight semantic alignment while strictly penalizing polarity inversions. Utilizing the population-level metrics defined in Section 3.5, it is computed as:

$$\mathcal{AS}(M) = \left(1 - \frac{\mu_D(M)}{2}\right) \times (1 - \mathcal{I}_{rate}(M)) \tag{10}$$

The first term normalizes the Mean Alignment Divergence (μ_D) into a similarity score bounded by $[0, 1]$, as the maximum theoretical divergence in our normalized space is 2.0. The second term acts as a strict penalty mask based on the Inversion Rate ($\mathcal{I}_{rate}$), expressed as a probability. An $\mathcal{AS}$ score of 1.0 indicates perfect cross-lingual affective fidelity, whereas lower scores reflect compounding representational misalignments.

Table 3 presents the Affective Stability scores for all evaluated architectures. The results validate our observations regarding scale and compression. While the large-scale XLM-T maintains high affective stability, mDistilBERT suffers a degradation in overall reliability. Notably, the distilled Tabularis model achieves an $\mathcal{AS}$ score highly competitive with XLM-T. This empirically demonstrates that targeted training strategies and synthetic data utilization can preserve affective calibration even under capacity constraints. A calibration study is provided in the Appendix Section B.

Model	Mean Div. (μ_D)	Inv. Rate ($\mathcal{I}_{rate}$)	Affective Stability ($\mathcal{AS}$)
XLM-T	0.276	0.036	**0.831**
Tabularis	0.200	0.086	0.823
IndicBERT	0.375	0.200	0.650
mDistilBERT	0.417	0.287	0.564

Table 3: Affective Stability ($\mathcal{AS}$) of evaluated models.

5 Discussion: Alignment Under Linguistic Pluralism

Our findings instantiate the multilingual curse at an affective, rather than purely accuracy-based, level. The inversion gap we observe is consistent with Blevins et al. (2024)'s theoretical framing: fixed model capacity induces inter-language parameter competition that degrades per-language representation. We hypothesize that compression amplifies this competition in ways that may disrupt the fine-grained representational signals required for stable sentiment polarity mapping.

Our results suggest two distinct intervention points for multilingual model development. First, at the pretraining data level: the *Modern Bias* finding points to a gap in training corpus composition for Bengali, even when utilizing synthetic data. Formal Sadhu text is underrepresented in web-crawled multilingual corpora (dominated by social media and news in contemporary registers), and correcting this imbalance would directly address dialectal representational harm. Second, at the post-distillation fine-tuning level: our finding that Tabularis substantially outperforms mDistilBERT despite both being distilled architectures demonstrates that targeted fine-tuning with synthetic multilingual corpora can preserve affective calibration through compression. Addressing both stages is particularly vital in low-resource settings like Bengali, where computationally efficient, compressed multilingual models must still deliver equitable and reliable affective understanding across all linguistic communities.

The multilingual NLP evaluation landscape currently lacks standardized metrics for cross-lingual affective consistency. Existing benchmarks (Hu et al., 2020; Ruder et al., 2021; Han et al., 2025; Goldman et al., 2025) primarily measure semantic and syntactic transfer but do not penalize polarity inversions or account for affective fidelity. We propose that multilingual benchmarks incorporate the metric suite introduced in this work (Inversion Rate, Affective Stability) as standard evaluation dimensions. This is especially critical for downstream applications that depend on affective signals: content moderation, mental health monitoring, customer feedback analysis, and social listening systems operating across languages all face systematic failure risks due to sentiment inversion.

We further argue that the choice is not binary between scale and alignment; targeted distillation strategies, data augmentation, and curated metrics that explicitly prioritize affective calibration can simultaneously address efficiency constraints and ensure representational equity, particularly in low-resource deployment settings.

6 Conclusion

We present a cross-lingual sentiment alignment audit comparing four transformer architectures on Bengali-English parallel data stratified by dialect. Our findings reveal that current multilingual models exhibit structured affective representational failures: sentiment inversion under compression, dialectal bias against formal registers, and directional asymmetry in emotional intensity calibration. These failures are distinct from accuracy degradation measured by standard benchmarks and constitute specific threats to equitable language technology access for Bengali users.

While this study establishes a quantitative framework for evaluating cross-dialectal robustness, the observed affective failures among smaller models should not be interpreted as strictly causal effects of model scale alone, since the evaluated open-weights models also differ in architecture, tokenizer design, and pre-training distributions. Future studies should consider exploring this avenue. The proposed Affective Stability index adopts a conservative $\tau = 0.1$ threshold to reduce low-magnitude scoring fluctuations across models, rather than relying on a task-specific optimized parameter. Alternative calibration strategies may further be explored

to refine sensitivity. Finally, dialectal boundaries and sentiment interpretation inherently contain a degree of linguistic subjectivity, particularly in colloquial or pragmatically ambiguous cases. Hence, a small portion of observed inversions may reflect annotation uncertainty or dialect-dependent interpretation rather than purely systematic representational misalignment.

We argue that building better multilingual representations requires evaluative interventions that make affective alignment failures visible. Future multilingual benchmarks should incorporate Inversion Rate and Affective Stability as standard dimensions. Future training and alignment research should address the formal register gap in Bengali pretraining data and explore dialect-stratified post-training alignment as a path to equitable compression. We believe these directions are generalizable beyond Bengali to the broader landscape of low-resource and dialectally diverse languages underrepresented in current multilingual NLP domain.

References

Berk Atil, Alexa Chittams, Liseng Fu, Ferhan Ture, Lixinyu Xu, and Breck Baldwin. 2024. Llm stability: A detailed analysis with some surprises. *arXiv preprint arXiv:2408.04667*, 1.

Umme Ayman, Chayti Saha, Azmain Mahtab Rahat, and Sharun Akter Khushbu. 2025. Banglablend: A large-scale nobel dataset of bangla sentences categorized by saint and common form of bangla language. *Data in Brief*, 58:111240.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond. In *Proceedings of the thirteenth language resources and evaluation conference*, pages 258–266.

Tadesse Destaw Belay, Ahmed Haj Ahmed, Alvin C Grissom Ii, Iqra Ameer, Grigori Sidorov, Olga Kolesnikova, and Seid Muhie Yimam. 2025. Culemo: Cultural lenses on emotion-benchmarking llms for cross-cultural emotion understanding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18894–18909.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, Kazi Samin Mubasshir, Md Saiful Islam, Anindya Iqbal, M Sohel Rahman, and Rifat Shahriyar. 2022. Banglabert: Language model pretraining and benchmarks for low-resource language understanding evaluation in bangla. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1318–1327.

Anirban Bhowmick and Abhik Jana. 2021. Sentiment analysis for bengali using transformer based models. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 481–486.

Shimanto Bhowmik, Tawsif Tashwar Dipto, Md Sazzad Islam, Sheryl Hsu, and Tahsin Reasat. 2025. Evaluating llms' multilingual capabilities for bengali: Benchmark creation and performance analysis. *arXiv preprint arXiv:2507.23248*.

Terra Blevins, Tomasz Limisiewicz, Suchin Gururangan, Margaret Li, Hila Gonen, Noah A Smith, and Luke Zettlemoyer. 2024. Breaking the curse of multilinguality with cross-lingual expert language models. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 10822–10837.

Vadim Borisov, Samuel Gyamfi, and Richard H. Schreiber. 2025. Multilingual sentiment analysis. Revision 69afb83.

Li Chen, Shifeng Shang, and Yawen Wang. 2025. Bridging resource gaps in cross-lingual sentiment analysis: adaptive self-alignment with data augmentation and transfer learning. *PeerJ Computer Science*, 11:e2851.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8440–8451.

Dipto Das, Shion Guha, Jed R Brubaker, and Bryan Semaan. 2024. The"colonial impulse" of natural language processing: An audit of bengali sentiment analysis tools and their identity-based biases. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–18.

Negar Foroutan, Paul Teiletche, Ayush Kumar Tarun, and Antoine Bosselut. 2025. Revisiting multilingual data mixtures in language model pretraining. *arXiv preprint arXiv:2510.25947*.

Omer Goldman, Uri Shaham, Dan Malkin, Sivan Eiger, Avinatan Hassidim, Yossi Matias, Joshua Maynez, Adi Mayrav Gilady, Jason Riesa, Shruti Rijhwani, and 1 others. 2025. Eclektic: a novel challenge set for evaluation of cross-lingual knowledge transfer. *arXiv preprint arXiv:2502.21228*.

Wenhan Han, Yifan Zhang, Zhixun Chen, Binbin Liu, Haobin Lin, Bingni Zhang, Taifeng Wang, Mykola Pechenizkiy, Meng Fang, and Yin Zheng. 2025. Mubench: Assessment of multilingual capabilities of large language models across 61 languages. *arXiv preprint arXiv:2506.19468*.

Md Nesarul Hoque, Umme Salma, Md Jamal Uddin, Md Martuza Ahamad, and Sakifa Aktar. 2024. Exploring transformer models in the sentiment analysis task for the under-resource bengali language. *Natural Language Processing Journal*, 8:100091.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International conference on machine learning*, pages 4411–4421. PMLR.

Saman Sarker Joy and Swakkhar Shatabda. 2025. Bnmmlu: Measuring massive multitask language understanding in bengali. *arXiv preprint arXiv:2505.18951*.

Mohsinul Kabir, Mohammed Saidul Islam, Md Tahmid Rahman Laskar, Mir Tafseer Nayeem, M Saiful Bari, and Enamul Hoque. 2024. Benllm-eval: A comprehensive evaluation into the potentials and pitfalls of large language models on bengali nlp. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2238–2252.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul NC, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the association for computational linguistics: EMNLP 2020*, pages 4948–4961.

Zhiwei Liu, Lingfei Qian, Qianqian Xie, Jimin Huang, Kailai Yang, and Sophia Ananiadou. 2025. Mmaffben: a multilingual and multimodal affective analysis benchmark for evaluating llms and vlms. *arXiv preprint arXiv:2505.24423*.

Hemal Mahmud, Hasan Mahmud, and Mohammad Rifat Ahmmad Rashid. 2024. Enhancing sentiment analysis in bengali texts: A hybrid approach using lexicon-based algorithm and pretrained language model bangla-bert. *arXiv preprint arXiv:2411.19584*.

Md Saef Ullah Miah, Md Mohsin Kabir, Talha Bin Sarwar, Mejdl Safran, Sultan Alfarhood, and Md F Mridha. 2024. A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and llm. *Scientific Reports*, 14(1):9603.

Millicent Ochieng, Anja Thieme, Ignatius Ezeani, Risa Ueno, Samuel Maina, Keshet Ronen, Javier Gonzalez, and Jacki O'Neill. 2025. Reasoning beyond labels: Measuring llm sentiment in low-resource, culturally nuanced contexts. *arXiv preprint arXiv:2508.04199*.

Alberto Poncelas, Pintu Lohar, James Hadley, and Andy Way. 2020. The impact of indirect machine translation on sentiment classification. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 78–88.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and 1 others. 2021. Xtreme-r: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245.

Jayanta Sadhu, Maneesha Rani Saha, and Rifat Shahriyar. 2025. Social bias in large language models for bangla: An empirical study on gender and religious bias. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 204–218.

Leanne Tan, Gabriel Chua, Ziyu Ge, and Roy Ka-Wei Lee. 2025. Lionguard 2: Building lightweight, data-efficient & localised multilingual content moderators. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 264–285.

Azmine Toushik Wasi, Raima Islam, Mst Rafia Islam, Taki Hasan Rafi, and Dong-Kyu Chae. 2024. Exploring bengali religious dialect biases in large language models with evaluation perspectives. *arXiv preprint arXiv:2407.18376*.

A Supplementary Figures and Tables

This appendix contains the visualizations and table referenced in the main findings of the paper.

Model Name	Repository (HuggingFace)
XLM-T	`cardiffnlp/XLM-Roberta-sentiment` (Barbieri et al., 2022)
IndicBERT	`ai4bharat/IndicBERTv2-sentiment`
Tabularis	`tabularisai/multilingual-sentiment`
mDistilBERT	`lxyuan/distilbert-multilingual`

Table 4: Model Repository Mapping

Note on Tabularis: This model is a fine-tuned version of the model `distilbert/distilbert-base-multilingual-cased` for multilingual sentiment analysis. It utilizes synthetic data from multiple sources to achieve robust performance across different languages and cultural contexts (Borisov et al., 2025).

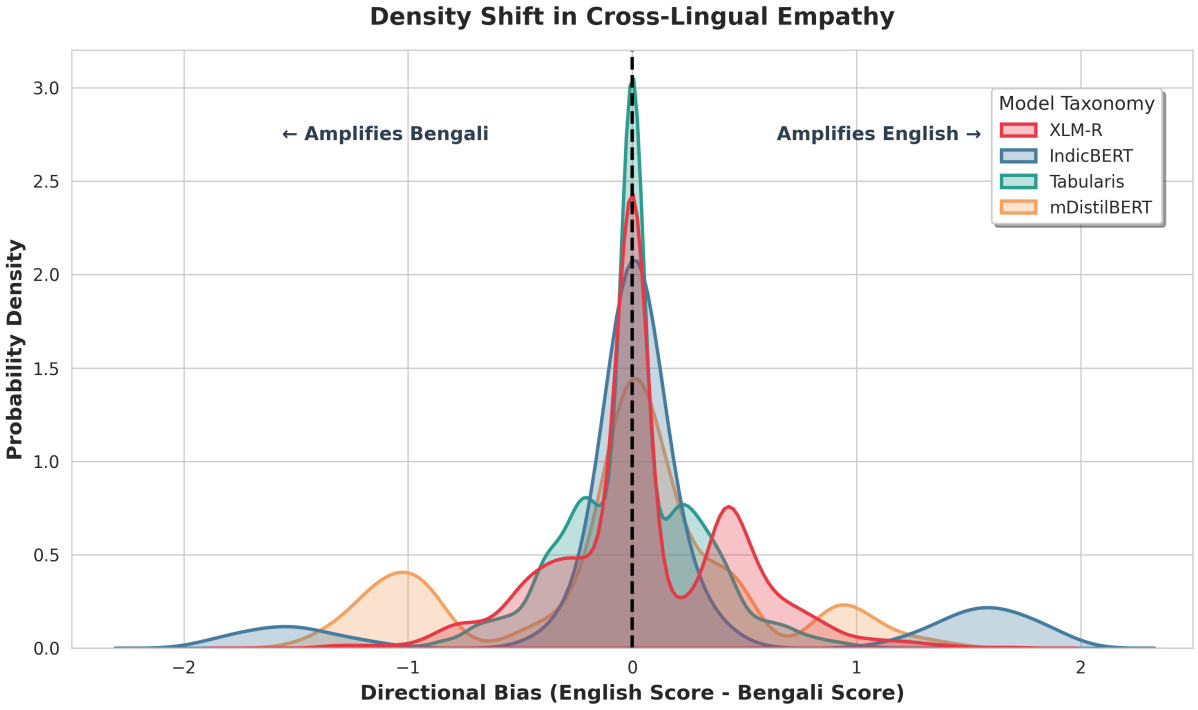

Figure 4: Directional Bias in Sentiment Scores (English − Bengali)

Model	English Input (E_i)	Bengali Input (B_i)	S_E	S_B	Bias (B_i)
IndicBERT	In football this is called the killer ball or the final ball	ফুটবলে এটিকে বলা হয় কিলার বল বা ফাইনাল বল	+0.665	-0.806	+1.471
	Interpretation: *The model successfully maps the English sports metaphor to a positive sentiment. However, it fails to contextually ground the Bengali transliteration of "killer"*				
mDistilBERT	This is a complete disaster and I hate it.	এটি একটি সম্পূর্ণ বিপর্যয় এবং আমি এটি ঘৃণা করি।	-0.981	+0.533	-1.514
	Interpretation: *The model exhibits a critical safety failure (Sentiment Inversion). While the English negative sentiment is correctly identified (-0.981), the Bengali translation is hallucinated as Positive (+0.533)*				

Figure 5: Illustrative case study validating Asymmetric Empathy in cross-lingual sentiment alignment, demonstrating severe instance-level directional bias.

Table 5: Calibration study comparing alternative formulations for affective stability measurement.

Formulation	Tabularis	XLM-T	IndicBERT	mDistilBERT	Limitation
Arithmetic Mean: $\frac{a+b}{2}$	0.907	0.913	0.806	0.752	Allows compensatory behavior where one strong component masks severe degradation in the other.
Harmonic Mean: $\frac{2ab}{a+b}$	0.907	0.910	0.806	0.750	Still partially compensatory and insufficiently suppresses asymmetric failures.
Geometric Mean: $\sqrt{ab}$	0.907	0.912	0.806	0.751	Produces weak penalties under severe degradation in one component.
Euclidean Aggregation: $1 - \sqrt{\frac{(1-a)^2+(1-b)^2}{2}}$	0.907	0.900	0.806	0.748	Smooths large failures too aggressively and weakens joint sensitivity.
Proposed AS: ab	**0.823**	**0.831**	**0.650**	**0.564**	**Jointly penalizes semantic divergence and polarity inversion while remaining bounded and non-compensatory.**

B Calibration Study of the Affective Stability Index

While Mean Alignment Divergence (μ_D) provides a useful aggregate estimate of cross-lingual representational shift, it does not fully capture critical affective failures. In particular, models with relatively moderate divergence may still exhibit substantial rates of sentiment polarity inversion. To address this limitation, we introduced the *Affective Stability Index* ($\mathcal{AS}$), a composite metric designed to jointly measure semantic alignment fidelity and affective consistency.

Recall that the proposed metric is defined as:

$$AS(M) = \left(1 - \frac{\mu_D(M)}{2}\right) \times \left(1 - I_{\text{rate}}(M)\right), \tag{11}$$

where:

- $\mu_D(M)$ denotes the Mean Alignment Divergence,
- $I_{\text{rate}}(M)$ denotes the normalized sentiment inversion rate.

For convenience, we define:

$$a(M) = 1 - \frac{\mu_D(M)}{2}, \qquad b(M) = 1 - I_{\text{rate}}(M). \tag{12}$$

Here, $a(M)$ denotes normalized semantic alignment fidelity, obtained by converting divergence into a bounded similarity score, while $b(M)$ denotes affective directional consistency, capturing the probability of preserving sentiment polarity. We intentionally restrict the metric to these two variables. Several additional statistics reported in Table 1 were excluded because they are either redundant, dataset-dependent, or diagnostic rather than foundational. For example:

- standard deviation of divergence strongly correlates with mean divergence,
- dialect-specific divergences are already reflected in $\mu_D(M)$,
- raw inversion counts depend on dataset size,
- directional bias is signed and may artificially cancel instability effects.

Thus, $\mathcal{AS}$ follows a minimal sufficient design principle, combining one magnitude-sensitive statistic and one structure-sensitive statistic without introducing redundant penalization.

B.1 Calibration Against Alternative Formulations

To evaluate the suitability of the proposed formulation, we compare $\mathcal{AS}$ against several alternative aggregation functions constructed from the same underlying variables. Results are shown in Table 5.

B.2 Rationale for Multiplicative Aggregation

Additive formulations such as arithmetic means or weighted sums violate the non-compensatory requirement. A model may achieve a relatively high aggregate score despite instability in one dimension. For example, if $a = 1.0$ but $b = 0.4$, the arithmetic mean still yields 0.7. This substantially overstates practical measures.

In contrast, the proposed multiplicative formulation yields:

$$AS = ab = 1.0 \times 0.4 = 0.4, \tag{13}$$

which more faithfully reflects representational failure under polarity inversion.

The multiplicative form introduces interaction sensitivity:

$$\frac{\partial AS}{\partial a} = b, \qquad \frac{\partial AS}{\partial b} = a. \tag{14}$$

Thus, sensitivity to one component depends directly on the quality of the other component. This prevents inflation under asymmetric failures and ensures that high Affective Stability values are only achievable when both affective alignment and directional consistency are simultaneously preserved.