Tears or Cheers? Benchmarking LLMs via Culturally Elicited Distinct Affective Responses

Anonymous ACL submission

Abstract

Culture serves as a fundamental determinant of human affective processing and profoundly shapes how individuals perceive and interpret emotional stimuli. Despite this intrinsic link extant evaluations regarding cultural alignment within Large Language Models primarily prioritize declarative knowledge such as geographical facts or established societal customs. These benchmarks remain insufficient to capture the subjective interpretative variance inherent to diverse sociocultural lenses. To address this limitation, we introduce CEDAR, a multimodal benchmark constructed entirely from scenarios capturing Culturally Elicited Distinct Affective Responses. To construct CEDAR, we implement a novel pipeline that leverages LLM-generated provisional labels to isolate instances yielding cross-cultural emotional distinctions, and subsequently derives reliable ground-truth annotations through rigorous human evaluation. The resulting benchmark comprises 10,962 instances across seven languages and 14 fine-grained emotion categories, with each language including 400 multimodal and 1,166 text-only samples. Comprehensive evaluations of 17 representative multilingual models reveal a dissociation between language consistency and cultural alignment, demonstrating that culturally grounded affective understanding remains a significant challenge for current models.

1 Introduction

Culture is the fundamental medium for the construction of human cognition and emotion (Kitayama and Cohen, 2010). The capacity to accurately discern and internalize cultural nuances is indispensable for capturing the latent semantics of natural language and visual signals. Although recent Large Language Models (LLMs) have demonstrated remarkable proficiency across various domains owing to their text understanding capabilities (Zhang et al., 2023; Hu et al., 2024), they exhibit uneven cultural understanding, leading to cross-cultural bias and cross-lingual inconsistency.

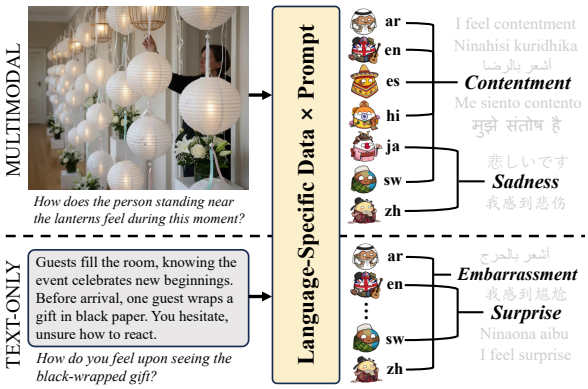

Figure 1: A representative example of culturally distinct scenarios from CEDAR. The figure illustrates how identical multimodal (top) and text-only (bottom) inputs are mapped to different language-specific ground truths.

Therefore, numerous studies have sought to address these disparities by establishing benchmarks to assess cultural commonsense knowledge (Li et al., 2024c; Nayak et al., 2024; Zhou et al., 2025; Onohara et al., 2025) and cultural bias (Ramezani and Xu, 2023; Dey et al., 2025). For instance, CultureAtlas (Fung et al., 2024) introduces a multicultural dataset derived from Wikipedia, covering a wide range of ethnolinguistic groups. Similarly, CVQA (Romero et al., 2024) curates a benchmark of culturally driven images and questions across languages to assess the cultural awareness of LLMs. This research trend has spanned diverse linguistic contexts, including benchmarks tailored for Arabic (Naous et al., 2024), Italian (Seveso et al., 2025), and under-represented languages such as Urdu (Hashmat et al., 2025) and Southeast Asian languages (Satar et al., 2025).

However, these studies predominantly focus on declarative cultural commonsense knowledge regarding geography, history, or customs, which overlooks the nuanced dynamics through which specific cultural frameworks modulate the subjective and

affective interpretation of information. Although recent research on cross-cultural emotion understanding has begun to explore this issue (Mohamed et al., 2022, 2024; Belay et al., 2025), such efforts are typically confined to narrowly defined domains (e.g., art appreciation) or restricted to a single modality. In practice, culturally conditioned emotional divergence frequently arises in everyday scenarios. As illustrated in Figure 1, a scene featuring white lanterns typically elicits an affective response of mourning within East Asian sociocultural contexts because of their symbolic association with funerary rites. Conversely, observers from different cultural backgrounds may interpret the same scene as conveying tranquility or even festive contentment. This example highlights a fundamental challenge in affective understanding: semantic equivalence across languages or visual scenes does not imply emotional equivalence across cultures. Identical scenarios may elicit substantially different emotional responses depending on the observer's cultural background. Nevertheless, most existing benchmarks implicitly assume cultural universality and overlook such culturally grounded variations in emotional interpretation.

Therefore, in this paper, we introduce CEDAR, a multimodal benchmark constructed entirely from scenarios capturing Culturally Elicited Distinct Affective Responses. Unlike prior studies, CEDAR exclusively curates instances where the emotional ground truth is contingent upon the cultural context. This requires LLMs to navigate beyond globalized emotional defaults, thereby assessing their capabilities to align with distinctive cultural frameworks. Specifically, CEDAR comprises 10,962 instances spanning seven languages and 14 fine-grained emotion categories. To ensure high quality and authentic cultural representation, we implement a novel construction pipeline centered on capturing culturally distinct nuances. We leverage LLMs to simulate diverse cultural perspectives, thereby identifying scenarios where semantic equivalence fractures into culturally distinct emotional interpretations. Upon isolating instances with significant cultural distinction, we employ native speakers to rigorously establish the ground-truth labels that reflect authentic cultural perspectives for these candidate instances. With CEDAR, we conduct a comprehensive evaluation of 17 representative multilingual and multimodal LLMs.

Extensive experimental results reveal several noteworthy observations, as detailed below:

- **Language modulates emotional distributional shifts.** For instance, `Aya-Vision-8B` demonstrates a pronounced inclination to predict *surprise* in *Arabic* contexts, while this pattern is absent in other language groups.
- **Systemic prioritization of high arousal states over deactivated emotions.** Models consistently prioritize high arousal emotions and marginalize deactivated states. Nevertheless, they exhibit distinct tendencies regarding valence. While `Claude4.5-Sonnet` leans towards emotions with high arousal and low valence (10.6%), `GPT-4o` shows a contrasting inclination (14.2%) for high-arousal and pleasant states.
- **Dissociation between language consistency and cultural alignment.** Language consistency does not ensure culturally aligned emotional understanding and may even degrade performance. For example, `Gemma3-27B-It` yields only 30.46% accuracy with Japanese prompts on Japanese data, lagging significantly behind English (44.14%) and even unrelated Arabic prompts (39.30%).

2 CEDAR: A Benchmark Grounded in Culturally Distinct Emotional Scenarios

2.1 Overview

CEDAR is a comprehensive benchmark designed to evaluate culturally-grounded emotion alignment in multilingual and multimodal LLMs, comprising scenarios that elicit culturally distinct affective responses. The benchmark covers seven languages, *Arabic (ar), Chinese (zh), English (en), Hindi (hi), Japanese (ja), Spanish (es), and Swahili (sw)*, and adopts 14 emotion categories adapted from Ekman (1992) and Cordaro et al. (2016). It contains 10,962 instances, including 400 multimodal and 1,166 text-only samples per language. We present the statistics of CEDAR in Figure 2. It demonstrates consistent distributional trends across both modalities, showing a comprehensive coverage of the targeted emotion categories.

In the following subsections, we detail our data curation pipeline. We start with seed data collection (§2.2) to generate text-only instances (§2.3). We subsequently extend these samples to construct multimodal data (§2.4) and conclude with the dataset finalization process (§2.5). We present the details of CEDAR in Appendix A.

2.2 Seed Data Collection

We curate socially grounded seed data from diverse resources, augmented through targeted brainstorming and online retrieval of culturally sensitive contexts. To ensure language consistency, we employ `GPT-4.5`[1] to translate the non-English datasets, ArabCulture (Sadallah et al., 2025) and JETHICS (Takeshita and Rzepka, 2025) into English. We then standardize all instances into a unified sentence format to address structural disparities among these resources, preparing for subsequent Narrative-Question (NQ) pair generation.

2.3 Text-Only Data Construction

NQ Pair Generation. Building upon the standardized sentences, we prompt `GPT-4.5` to instantiate each sentence into a NQ pair. The **N**arrative establishes a contextual scenario that incorporates situational grounding and participant actions whereas the corresponding **Q**uestion requires the model to infer the affective state of the protagonist across various temporal stages. Through this procedure, we obtain approximately 42K candidate NQ pairs covering a wide spectrum of cultural scenarios.

Contextual Refinement. To improve textual naturalness and reduce reliance on explicit emotional cues, we apply `Llama3.3-70B` (Grattafiori et al., 2024) to refine each NQ pair. Specifically, narratives and questions are rewritten in the second person to enhance subjectivity and immersion. Next, phrases containing explicit emotional expressions (e.g., "`sparking a mix of excitement`") are removed while ensuring coherence. This design mitigates information leakage and compels models to infer emotional states from implicit contextual signals rather than overt affective markers.

Basic Filtering. We initially implement string length constraints to ensure that each instance maintains a character count between 50 and 200. This filtering stage accounts for approximately 2.7% of the candidate pool and guarantees that scenarios provide sufficient context while avoiding unnecessary verbosity. We then use `Llama3.3-70B` as a classifier to identify and remove non-social data (e.g., demographic information; ∼15.88%), restricting the benchmark to scenarios that elicit culturally distinct emotional responses. Finally we utilize `PolyGuard` (Kumar et al., 2025) which is a specialized multilingual safety moderation framework to identify and eliminate toxic content including hate speech or harmful stereotypes. By applying these rigorous quality control measures we establish a curated corpus that is both task relevant and safe for broad research applications.

[1] `https://openai.com/`

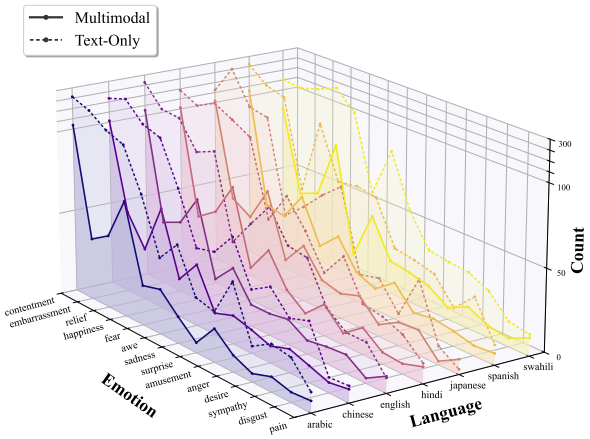

Figure 2: Statistics of CEDAR.

Consistency and Variation Filtering. To identify candidate instances likely to elicit culture-specific emotional responses, we employ state-of-the-art LLMs (i.e., `Claude4.5-Sonnet` (Anthropic, 2025), `Gemini2.5-Flash` (Comanici et al., 2025), and `GPT-4.5`) to generate provisional predictions for each instance across languages. We first impose **within-language agreement** by aggregating predictions from the three models for each language. Instances exhibiting complete disagreement are discarded as ambiguous, whereas those achieving majority agreement are assigned the corresponding majority label as a provisional emotion. We then enforce **cross-language variation** by comparing these provisional labels across languages and removing instances with uniform predictions, which reflect affective interpretations that are largely invariant across cultural contexts. This filtering procedure yields approximately 7K NQ pairs that exhibit culturally distinct emotional variation. Importantly, the LLM generated provisional labels are used solely for data selection and filtering, and all final ground truth annotations are obtained exclusively through rigorous human labeling.

2.4 Multimodal Data Construction

In this section, we further extend our pipeline to construct multimodal data in the form of Image-Narrative-Question (INQ) triples. Each triple integrates an **I**mage depicting a specific event, a **N**arrative providing background context, and a **Q**uestion targeting emotion prediction.

Image Acquisition. We employ a hybrid strategy to secure semantically congruent visual stimuli for each Narrative and Question pair. We first engage undergraduate and graduate students to retrieve suitable publicly available images via Google or Baidu. When appropriate visual representations are unavailable, we utilize `Gemini2.5-Flash-Image` (Comanici et al., 2025) to generate high fidelity synthetic images that accurately depict the described events. This dual approach ensures that the visual modality provides a precise grounding for the subsequent affective analysis.

INQ Triple Refinement. To prioritize visually grounded reasoning, we prompt `GPT-4.5` to refine the raw INQ triples via a three-step process: (1) The model identifies characters and actions as well as interpersonal relationships manifest within the image. (2) The narrative is rewritten to provide only essential background information such as character backstories that remain distinctively invisible within the image. This process systematically eliminates descriptions of visible content to reduce modal redundancy (3) The final question is refined to focus on specific characters. This requires models to synthesize the visual scenario with the textual narrative to infer emotional states.

Image Necessity Filtering. To ensure the visual modality provides essential information rather than merely illustrating the text, we filter INQ triples based on predictive disparity. For each instance, `GPT-4.5` generates emotion predictions under multimodal and text-only settings; instances with identical predictions are discarded as visually redundant. This process yields approximately 600 candidate INQ triples, for which provisional labels are generated following the procedure in §2.3 to support dataset finalization.

2.5 Dataset Finalization

Cultural Variation-Based Selection. We select representative instances that exhibit maximal cultural variation within clusters of semantically similar scenarios. Narratives are embedded using `Qwen3-Embedding-8B` (Zhang et al., 2025) to group semantically related instances. Cultural variation within each cluster is quantified leveraging Russell's Circumplex Model (Russell, 1980), which represents emotions along Valence (pleasant vs. unpleasant) and Arousal (activation vs. deactivation) dimensions. The 14 emotion categories are mapped to four quadrants: **Quadrant I** (high valence, high arousal: *amusement*, *happiness*, *surprise*), **Quadrant II** (low valence, high arousal: *anger*, *disgust*, *fear*, *pain*), **Quadrant III** (low valence, low arousal: *embarrassment*, *sadness*), and **Quadrant IV** (high valence, low arousal: *awe*, *contentment*, *desire*, *relief*, *sympathy*). We score instances by cross-language quadrant disagreement and select the highest-scoring instance from each semantic cluster, resulting in 400 INQ triples and 1,166 NQ pairs. We provide detailed information of the Russell's quadrants in Appendix A.3.

Translation and Human Annotation. We first translate the English data into six target languages utilizing `GPT-4.5`, followed by native-speaker verification via crowdsourcing to ensure fluency and correctness. Ground-truth labels are obtained through annotation by local native speakers recruited on Prolific[2], with at least five annotators per instance from the corresponding target region. If no majority vote is reached, two additional annotators are assigned and the final label is determined by the updated majority. We present further details of annotation process in Appendix A.4.

3 Methods

3.1 Task Definition

We define culturally-grounded emotion alignment as the task of interpreting scenarios through the cultural lens associated with a target language. Let $\mathcal{L}$ denote the set of seven target languages and $\mathcal{E}$ the set of 14 emotion categories. The benchmark consists of a multimodal subset $\mathcal{D}_M$ and a text-only subset $\mathcal{D}_T$. A multimodal instance is represented as $x = (I, N_l, Q_l, \rho_l)$, where I is the image, and N_l, Q_l, ρ_l are the narrative, question, and instruction prompt in language $l \in \mathcal{L}$. Similarly, a text-only instance is defined as $x = (N_l, Q_l, \rho_l)$. Formally, given an input x, the model needs to predict the affective label $y_l \in \mathcal{E}$. Notably, while inputs are parallel translations that preserve semantic equivalence, the ground-truth label y_l is culture-specific, reflecting the distinct sociocultural norms associated with the target language l.

3.2 Baselines

To ensure a comprehensive analysis, we select 17 representative multilingual and multimodal LLMs, spanning various scales from 8B to 1T parameters and including both open-source and propri-

[2] https://www.prolific.com/

Model	Param.	Multimodal									Text-only								
		AR	EN	ES	HI	JA	SW	ZH	Avg.	Var.	AR	EN	ES	HI	JA	SW	ZH	Avg.	Var.
Aya-Vision-8B	8B	31.20	42.37	30.70	33.57	30.29	-	29.03	32.89	22.4	36.00	45.96	40.37	39.56	37.80	-	37.43	39.58	11.9
💡 MiniCPM-V-4.5	8B	24.40	35.84	43.61	16.33	19.60	19.39	24.86	26.52	91.2	31.85	46.48	41.09	25.98	13.29	7.55	28.04	27.76	165.4
💡 Qwen3-VL-8B	8B	30.32	41.88	39.04	27.85	26.39	14.52	27.66	31.88	85.6	39.36	49.70	42.25	37.08	28.24	18.24	33.90	35.64	94.1
Llama3.2-11B-Vision	11B	-	34.29	28.12	32.92	-	-	-	31.22	10.2	-	43.04	21.86	32.84	-	-	-	32.61	112.5
Pixtral-12B	12B	26.09	38.01	35.29	29.55	24.07	-	25.44	29.95	31.4	32.65	42.37	37.27	30.28	33.16	-	31.81	35.40	19.2
Aya-101	13B	-	-	-	-	-	-	-	-	-	29.33	33.45	24.59	38.29	33.11	30.38	30.62	30.54	15.3
Mistral-Small-3.2	24B	30.32	47.85	48.15	32.98	27.05	-	26.92	38.72	89.4	20.12	39.97	28.75	37.26	31.77	-	32.96	31.18	41.8
Gemma3-27B-It	27B	38.56	48.37	46.65	44.38	30.46	40.00	30.00	39.82	51.3	39.13	53.91	48.80	46.85	36.03	43.11	40.85	44.31	36.4
Aya-Vision-32B	32B	38.74	45.16	44.17	29.41	31.01	-	43.14	39.40	42.1	41.20	52.81	44.55	46.26	36.57	-	43.32	43.88	30.1
💡 InternVL3.5-38B	38B	28.51	37.37	37.28	32.74	23.60	15.84	30.95	30.84	58.9	36.57	22.84	34.16	35.33	22.66	11.10	33.25	27.86	77.2
Qwen2-VL-72B	72B	37.76	49.36	47.61	35.68	30.57	22.98	28.68	36.10	87.4	46.43	57.16	46.68	45.96	33.51	24.12	44.27	42.59	108.2
💡 Qwen3-VL-235B	235B	39.85	50.38	44.53	41.71	34.34	33.42	35.88	40.01	34.1	47.95	55.15	52.88	49.06	37.19	38.08	43.22	46.21	42.5
Kimi-K2-Instruct	1T	-	-	-	-	-	-	-	-	-	44.15	56.72	49.87	46.06	39.66	41.17	40.80	45.49	36.9
🔒 💡 Claude4.5-Sonnet	UNK	39.90	49.62	45.45	40.82	32.99	32.15	36.71	39.66	38.9	49.69	61.79	33.59	43.95	29.60	38.30	47.51	44.09	102.4
🔒 💡 Gemini2.5-Flash	UNK	45.75	50.28	44.05	47.73	29.07	42.60	29.32	41.10	65.2	51.48	57.53	52.98	51.27	36.31	46.49	43.12	47.31	44.8
🔒 GPT-4o	UNK	36.93	36.34	41.60	39.14	32.50	39.37	26.01	35.97	27.8	46.56	53.69	44.55	47.45	42.97	50.43	40.34	46.57	19.6
🔒 💡 Qwen3-Omni-Flash	UNK	39.42	40.61	46.94	41.01	33.33	25.00	32.30	37.25	51.8	38.14	52.40	46.63	43.90	29.38	9.88	36.57	36.70	185.3

Table 1: Comparison of different LLMs on CEDAR. We report the standard accuracy and the cross-lingual variance (Var.) for both subsets. Avg. denotes the macro-average accuracy across languages. Darker shades indicate higher numerical values. The (💡) denotes reasoning-augmented LLMs, while (🔒) represents closed-source models, and hyphen (-) indicates the absence of official language support.

etary models. We evaluate various types of models, encompassing general LLMs such as `GPT-4o` (Hurst et al., 2024), reasoning-augmented LLMs like `Claude4.5-Sonnet`, as well as LLMs optimized for multilingual alignment like `Aya` series. Details of evaluation are presented in Appendix B.

3.3 Analysis Metrics

We employ a comprehensive suite of metrics for experimental analysis, ranging from standard accuracy to granular cultural alignment analysis.

Standard Accuracy (SA). This is calculated as the percentage of correct predictions over the total number of questions.

Emotion Prediction Propensity (EPP). We quantify the predictive inclination towards specific fine-grained emotion category $e \in \mathcal{E}$. For a sample set $\mathcal{S}$, this propensity is defined as the ratio:

$$\rho_e(\mathcal{S}) = \frac{N_{\mathcal{S}}(\hat{y} = e)}{N_{\mathcal{S}}(y = e)} \quad (1)$$

where $\hat{y}$ represents the emotion prediction.

We report **Global EPP (GEPP)** where $\mathcal{S}$ is the full benchmark, and **Language-Specific EPP (LSEPP)** where $\mathcal{S}$ corresponds to the subset of a specific language l.

Russell's Quadrant Bias (RQB). To analyze affective tendencies beyond discrete labels, we leverage Russell's Circumplex Model (Russell, 1980) to map emotions onto the continuous dimensions of Valence and Arousal. We measure the distributional deviation in model predictions for each quadrant q within a target sample set $\mathcal{S}$:

$$\beta_q(\mathcal{S}) = \frac{N_{\mathcal{S}}(\hat{y} \in q) - N_{\mathcal{S}}(y \in q)}{\sum_{k=1}^{4} N_{\mathcal{S}}(y \in k)} \times 100\% \quad (2)$$

where $N_{\mathcal{S}}(\cdot)$ denotes the sample count within $\mathcal{S}$.

We compute **Global RQB (GRQB)** on the entire benchmark and **Language-Specific RQB (LSRQB)** for each language l to assess cultural alignment at the categorical level.

4 Experiments

4.1 Overall Performance

We present the overall performance of LLMs on CEDAR in Table 1. These results reveal several key observations. First, the evaluated models consistently achieve lower standard accuracy on multimodal instances compared to the text-only subset. We attribute this to the complexity of visual-emotional grounding, as interpreting symbolic imagery that carries culturally-grounded emotional weight proves more challenging than processing explicit textual cues. Second, a performance disparity exists between language groups. While models consistently perform better in high-resource languages such as *English* and *Spanish*, performance degrades notably in Asian languages (e.g., *Chinese, Japanese*) and low-resource languages like *Swahili*. This gap indicates that non-Western affective norms are not adequately encoded within these LLMs, emphasizing the critical need for culture-aware emotional alignment. Third, models explicitly optimized for multilingual alignment (e.g., `Aya-101` and `Aya-Vision-8B/32B`) exhibit greater stability across languages. Despite their smaller

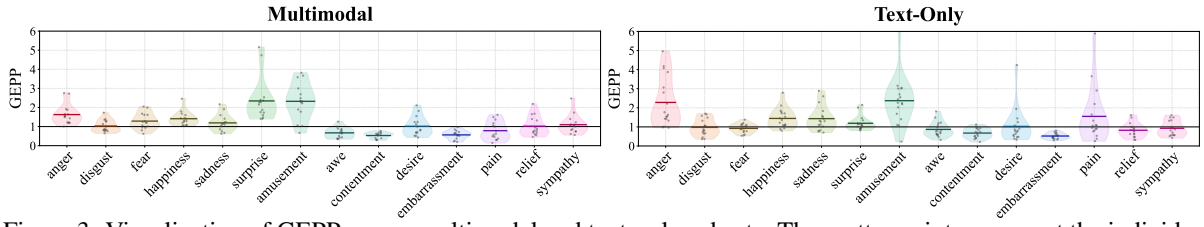

Figure 3: Visualization of GEPP across multimodal and text-only subsets. The scatter points represent the individual performance of the evaluated LLMs for each emotion category.

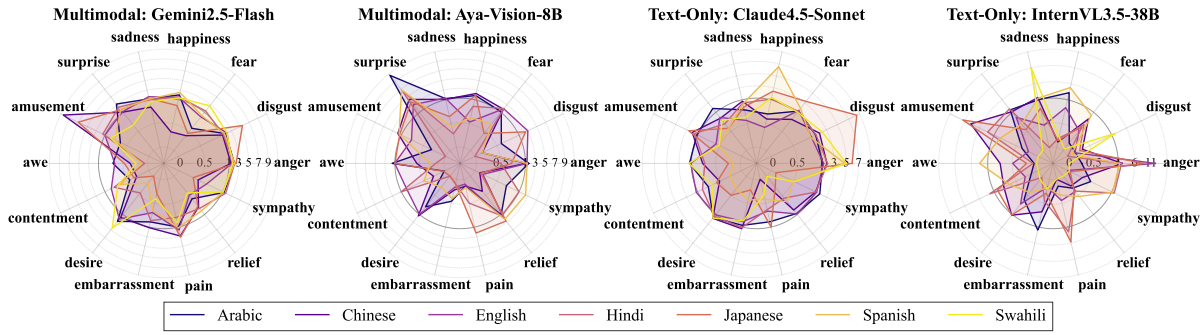

Figure 4: LSEPP for four illustrative models on multimodal and text-only subsets. Each axis represents the LSEPP value for a specific emotion, demonstrating the variability of bias scores across seven languages.

model scale, these models retain competitive performance with notably reduced cross-lingual variance. This finding confirms that targeted optimization for multilingual alignment effectively improves cross-cultural emotional adaptability, underscoring a vital dimension of model development often overlooked in prevailing English-centric research.

4.2 Emotion Prediction Bias

To assess the emotion prediction propensities of models, we visualize the GEPP in Figure 3. In the multimodal subset, models exhibit a distinct propensity towards salient emotions such as *surprise* and *amusement*, while under-predicting subtle states like *contentment* and *embarrassment*. This disparity suggests that models consistently favor broad emotional categories and frequently fail to capture fine-grained nuances. In the text-only set, we observe a parallel trend where models tend to over-predict salient emotions at the expense of more complex states. Notably, these over-predicted emotions are accompanied by markedly higher variance and extreme outliers. Such instability reflects profound uncertainty in emotional responses across different model series, highlighting current limitations in modeling affective granularity.

We further investigate cross-lingual emotional adaptability by analyzing the LSEPP, with illustrative examples presented in Figure 4. It illustrates the specific prediction profiles for each model and reveals how these propensities shift according to target languages. Our results indicate that different language groups exhibit distinct distributions across emotion dimensions. For instance, within the *Arabic* group, `Aya-Vision-8B` demonstrates a pronounced over-prediction of *surprise* in the multimodal subset, whereas this pattern is absent in other language groups. In addition, we find that models with higher overall accuracy exhibit greater stability and more balanced emotion distributions across languages, suggesting a positive correlation between predictive performance and cross-lingual affective robustness. For instance, in the multimodal subset, a comparison between `Gemini2.5-Flash` and `Aya-Vision-8B` shows that `Gemini2.5-Flash` demonstrates markedly greater stability in Figure 4, consistent with its higher overall accuracy of 41.10%, compared to 32.89% for `Aya-Vision-8B`. Meanwhile, this trend persists in the text-only subset, where `Claude4.5-Sonnet`, with higher overall accuracy, demonstrates lower prediction volatility compared to `InternVL3.5-38B`.

4.3 Dimensional Affective Analysis

We adopt Russell's Circumplex Model (Russell, 1980), as described in Section 2.5, to explore models' affective tendencies, with the corresponding results presented in Figure 5. At the model level, we observe that LLMs exhibit a systematic preference for high-arousal emotions, while consistently underrepresenting affective states asso-

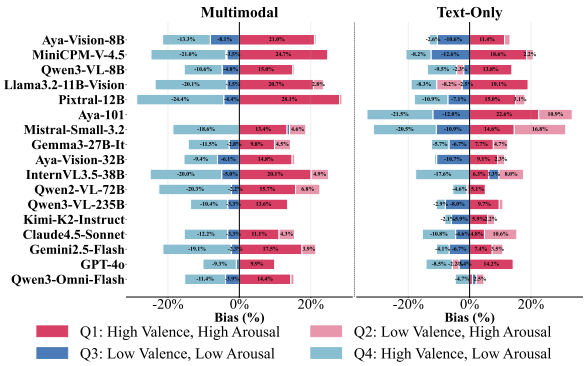

Figure 5: Model comparison based on the GRQB. The chart displays the GRQB of model predictions across the four quadrants for multimodal and text-only subsets.

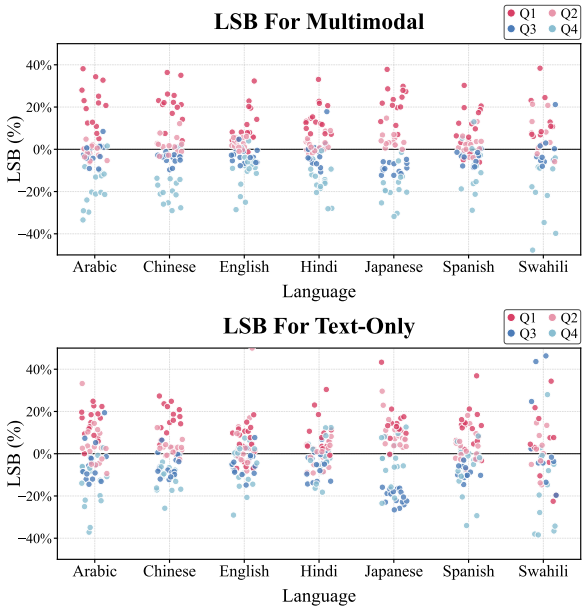

Figure 6: LSRQB values for all evaluated models across seven languages on multimodal and text-only subsets, categorized by Russell's Quadrants (Q1-Q4). Each data point represents the result of an individual model.

ciated with low arousal. In the multimodal subset, models exhibit a strong and consistent inclination towards high-arousal emotions situated in Quadrant I and Quadrant II, while systematically under-predicting deactivated emotions associated with Quadrants III and IV. A similar tendency is observed in the text-only subset, though affective distributions are more diverse. For instance, `Claude4.5-Sonnet` and `Mistral-Small-3.2` prefer low-valence, high-arousal emotions in Quadrant II (10.6% and 16.8%), whereas `GPT-4o` aligns with Quadrant I (14.2%) and `InternVL3.5-38B` shows a mild bias toward Quadrant III (3.3%). These variations suggest that these models largely share consistent affective tendencies, while exhibiting subtle model-specific differences.

We further examine how these affective tendencies shift across languages by analyzing the LSRQB, as shown in Figure 6. Across all languages, similar to the trends observed previously evaluated models, models demonstrate a systematic inclination towards emotions in Quadrant I and Quadrant II, while consistently neglecting emotions in Quadrant III and Quadrant IV. Notably, the data points for *English* are tightly clustered around 0% to exhibit high stability and cross-model consistency, whereas substantially larger dispersion is observed in languages such as *Arabic* and *Swahili*. The observed increase in volatility indicates that current models exhibit inconsistent generalization across languages, stemming from insufficient cultural and linguistic priors, which hinders accurate affective prediction, particularly in low-resource languages. Moreover, in the text-only setting, the Japanese group exhibits a pronounced negative clustering of approximately -20% in low-valence, low-arousal emotions. This unique pattern remains absent across other language groups which underscores a language specific failure in affective alignment. Such findings suggest that current models struggle to internalize the deactivated emotional nuances characteristic of the Japanese sociocultural context despite their general linguistic proficiency.

4.4 Effect of Prompt Languages

To investigate the interaction between prompt language and dataset language, we evaluate `Gemma3-27B-It` and `Qwen3-VL-8B` under multiple prompt-language configurations, with the results summarized in Figure 7. The results indicate that English prompts consistently yield better performance compared to prompts in other languages. This advantage is plausibly attributable to the dominance of English data in large-scale pretraining corpora, which induces an inherent English-centric bias in current models. In contrast, utilizing non-English prompts such as *Arabic* or *Chinese* leads to performance degradation, further emphasizing the critical need for improved cross-lingual consistency and robustness. Our results reveal a counter intuitive pattern where language consistency between the prompt and the dataset fails to improve performance. Although it is commonly assumed that matching the prompt language with the dataset language should be beneficial, our empirical findings show that such alignment fails to yield reliable gains. For instance, results of `Gemma3-27B-It`

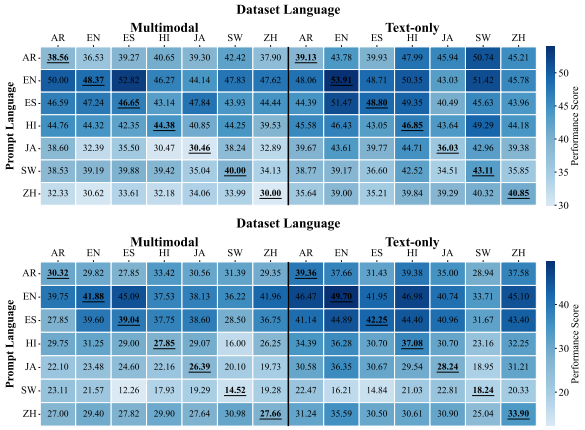

Figure 7: Performance heatmaps analyzing the impact of prompt language (y-axis) versus dataset language (x-axis) on standard accuracy. Results are shown for `Gemma3-27B-It` (top) and `Qwen3-VL-8B` (bottom) across both multimodal and text-only subsets.

on the multimodal set demonstrate that the combination of a *Japanese* prompt with *Japanese* data yields an accuracy of only 30.46% which represents the lowest score among comparable groups. We hypothesize that such language alignment may amplify latent cultural stereotypes or biases associated with the specific language, thereby leading to erroneous emotion predictions. These findings underscore a profound dissociation between language proficiency and cultural resonance suggesting that models may rely on English as a latent reasoning pivot rather than genuinely internalizing the values associated with native language expressions.

5 Related Work

5.1 Cross-Cultural Alignment

Efforts have been made to develop robust methodologies for cross-cultural alignment (Kabir et al., 2025). A foundational approach involves utilizing LLM-synthesized data to enhance cultural capabilities (Li et al., 2024a; El Mekki et al., 2025). For example, CARE (Guo et al., 2025) introduces a preference dataset containing culture-specific instances to improve cultural awareness. Beyond data augmentation, some studies have proposed specialized learning frameworks for cross-cultural adaptation. These include contrastive learning approaches designed to capture subtle cultural cues (Huang et al., 2025), and human-AI collaborative systems engineered to identify and address knowledge gaps (Ziems et al., 2025). Moreover, recent research has increasingly emphasized pluralism and fairness across cultures. These efforts range from modular frameworks that leverage multi-LLM collaboration for pluralistic alignment (Feng et al., 2024) to analyses of LLM-generated cultural symbols aimed at uncovering and mitigating uneven representational diversity (Li et al., 2024b).

5.2 Cultural Assessments

There have been numerous studies focusing on the evaluation of cultural commonsense knowledge (Nayak et al., 2024; Onohara et al., 2025; Satar et al., 2025). These encompass assessments of single-nation (Seveso et al., 2025), regional (Ma et al., 2025), and global-scale cultural contexts (Fung et al., 2024). To capture diverse signals, these efforts have expanded across modalities, including text-only tasks (Wang et al., 2024b), visual question answering (VQA; Romero et al. (2024)), video understanding (Shafique et al., 2025), and text-to-image generation (Naous et al., 2024; Nayak et al., 2025). Furthermore, research has deepened into specific topics such as cuisine (Li et al., 2024c), paintings (Yu et al., 2025a), and traditional clothing (Zhou et al., 2025). Beyond factual knowledge, parallel works have been proposed to assess cultural paradigms, including moral norms (Ramezani and Xu, 2023) and personality traits (Dey et al., 2025). However, these studies primarily assess static information, overlooking the implicit affective lens through which different cultures interpret identical scenarios.

6 Conclusion

In this paper, we introduce CEDAR, a multimodal benchmark constructed entirely from scenarios that elicit culturally distinct affective responses. The benchmark comprises 10,962 instances spanning seven languages and 14 fine grained emotion categories. To construct this dataset we implement a novel pipeline that utilizes LLMs to simulate diverse cultural perspectives which facilitates the identification of scenarios where identical visual or textual stimuli provoke divergent emotions. These candidate instances are subsequently validated through rigorous human annotation to ensure definitive ground truth reliability. Most notably our results expose a dissociation between language proficiency and cultural alignment. We find that surface level language consistency between the prompt and the dataset does not guarantee the genuine internalization of the underlying sociocultural values required for accurate affective prediction.

Limitations

While CEDAR serves as a comprehensive benchmark for cross-cultural emotion alignment, our study is subject to several limitations that remain to be addressed in future research.

Scale and Depth. To prioritize high-fidelity cultural validation over extensive language coverage, CEDAR focuses on seven representative languages and ensures that every instance undergoes rigorous verification by native speakers. Limited by resources, we restrict our scope to these cultural clusters to establish a reliable benchmark for assessing the dissociation between language proficiency and cultural alignment.

Affective Theoretical Framework. We acknowledge the long-standing discourse within affective science regarding the dichotomy between categorical and dimensional emotion frameworks. In this paper, we utilize 14 discrete categories to ensure precise quantitative evaluation while actively integrating Russell's Model (Russell, 1980) for data filtering and dimensional analysis.

Cultural Consensus. Constrained by the significant resources required for native-speaker annotations, CEDAR strategically targets high-consensus scenarios, validated through strict majority voting, to establish prototypical affective benchmarks. By focusing on these distinct and widely shared cultural signals, we aim to assess the fundamental capability of LLMs in cross-cultural alignment.

Ethical Considerations

Data Safety. CEDAR is strictly curated to exclude harmful content such as stereotypes or racism. Beyond implementing a rigorous safety protocol to filter toxic data (§2.3), we emphasize that the ground-truth labels reflect statistical cultural tendencies within a language group, rather than prescriptive stereotypes. Users should interpret these results as a measure of cultural literacy instead of absolute rules for profiling individuals.

Annotation Process. We recruit native speakers via Prolific, an established crowdsourcing platform, to ensure high-quality annotation. We strictly adhere to ethical compensation guidelines by ensuring rates exceed local minimum wages. Furthermore, all participants provide informed consent regarding the nature of the task and retain the right to withdraw at any stage without penalty.

References

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, and 1 others. 2024. Pixtral 12b. *arXiv preprint arXiv:2410.07073*.

Anthropic. 2025. System card: Claude sonnet 4.5. https://www.anthropic.com/claude-sonnet-4-5-system-card. Accessed: 2025-12-07.

Tadesse Destaw Belay, Ahmed Haj Ahmed, Alvin 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, Vienna, Austria. Association for Computational Linguistics.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2025. CulturalBench: A robust, diverse and challenging benchmark for measuring LMs' cultural knowledge through human-AI red-teaming. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25663–25701, Vienna, Austria. Association for Computational Linguistics.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Daniel T Cordaro, Dacher Keltner, Sumjay Tshering, Dorji Wangchuk, and Lisa M Flynn. 2016. The voice conveys emotion in ten globalized cultures and one remote village in bhutan. *Emotion*, 16(1):117.

Saurabh Dash, Yiyang Nan, John Dang, Arash Ahmadian, Shivalika Singh, Madeline Smith, Bharat Venkitesh, Vlad Shmyhlo, Viraat Aryabumi, Walter Beller-Morales, and 1 others. 2025. Aya vision: Advancing the frontier of multilingual multimodality. *arXiv preprint arXiv:2505.08751*.

Priyanka Dey, Aayush Bothra, Yugal Khanter, Jieyu Zhao, and Emilio Ferrara. 2025. Can LLMs express personality across cultures? introducing CulturalPersonas for evaluating trait alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 20241–20262, Suzhou, China. Association for Computational Linguistics.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Abdellah El Mekki, Houdaifa Atou, Omer Nacar, Shady Shehata, and Muhammad Abdul-Mageed. 2025. NileChat: Towards linguistically diverse and culturally aware LLMs for local communities. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10978–11002, Suzhou, China. Association for Computational Linguistics.

Shangbin Feng, Taylor Sorensen, Yuhan Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia Tsvetkov. 2024. Modular pluralism: Pluralistic alignment via multi-LLM collaboration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4151–4171, Miami, Florida, USA. Association for Computational Linguistics.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Geyang Guo, Tarek Naous, Hiromi Wakaki, Yukiko Nishimura, Yuki Mitsufuji, Alan Ritter, and Wei Xu. 2025. CARE: Multilingual human preference learning for cultural awareness. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32854–32883, Suzhou, China. Association for Computational Linguistics.

Abdullah Hashmat, Muhammad Arham Mirza, and Agha Ali Raza. 2025. PakBBQ: A culturally adapted bias benchmark for QA. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16171–16183, Suzhou, China. Association for Computational Linguistics.

Jinpeng Hu, Tengteng Dong, Luo Gang, Hui Ma, Peng Zou, Xiao Sun, Dan Guo, Xun Yang, and Meng Wang. 2024. Psycollm: Enhancing llm for psychological understanding and evaluation. *IEEE Transactions on Computational Social Systems*.

Yuchen Huang, Zhiyuan Fan, Zhitao He, Sandeep Polisetty, Wenyan Li, and Yi R. Fung. 2025. CultureCLIP: Empowering CLIP with cultural awareness through synthetic images and contextualized captions. In *Second Conference on Language Modeling*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Mohsinul Kabir, Ajwad Abrar, and Sophia Ananiadou. 2025. Break the checkbox: Challenging closed-style evaluations of cultural alignment in LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24–51, Suzhou, China. Association for Computational Linguistics.

Shinobu Kitayama and Dov Cohen. 2010. Handbook of cultural psychology.

Priyanshu Kumar, Devansh Jain, Akhila Yerukola, Liwei Jiang, Himanshu Beniwal, Thomas Hartvigsen, and Maarten Sap. 2025. Polyguard: A multilingual safety moderation tool for 17 languages. *Preprint*, arXiv:2504.04377.

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.

Huihan Li, Liwei Jiang, Nouha Dziri, Xiang Ren, and Yejin Choi. 2024b. CULTURE-GEN: Revealing global cultural perception in language models through natural language prompting. In *First Conference on Language Modeling*.

Wenyan Li, Crystina Zhang, Jiaang Li, Qiwei Peng, Raphael Tang, Li Zhou, Weijia Zhang, Guimin Hu, Yifei Yuan, Anders Søgaard, Daniel Hershcovich, and Desmond Elliott. 2024c. FoodieQA: A multimodal dataset for fine-grained understanding of Chinese food culture. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19077–19095, Miami, Florida, USA. Association for Computational Linguistics.

Weicheng Ma, John J. Guerrerio, and Soroush Vosoughi. 2025. Scalable and culturally specific stereotype dataset construction via human-LLM collaboration. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23939–23967, Suzhou, China. Association for Computational Linguistics.

Mistral AI Team. 2025. Mistral small 3.1. `https://mistral.ai/news/mistral-small-3-1`. Accessed: 2025-12-07.

Youssef Mohamed, Mohamed Abdelfattah, Shyma Alhuwaider, Feifan Li, Xiangliang Zhang, Kenneth Church, and Mohamed Elhoseiny. 2022. ArtELingo: A million emotion annotations of WikiArt with emphasis on diversity over language and culture. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8770–8785, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Youssef Mohamed, Runjia Li, Ibrahim Sa'id Ahmad, Kilichbek Haydarov, Philip Torr, Kenneth Church, and Mohamed Elhoseiny. 2024. No culture left behind: Artelingo-28, a benchmark of wikiart with captions in 28 languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20939–20962.

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393, Bangkok, Thailand. Association for Computational Linguistics.

Shravan Nayak, Mehar Bhatia, Xiaofeng Zhang, Verena Reiser, Lisa Anne Hendricks, Sjoerd Van Steenkiste, Yash Goyal, Karolina Stanczak, and Aishwarya Agrawal. 2025. CulturalFrames: Assessing cultural expectation alignment in text-to-image models and evaluation metrics. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 20918–20953, Suzhou, China. Association for Computational Linguistics.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Van Steenkiste, Lisa Anne Hendricks, Karolina Stanczak, and Aishwarya Agrawal. 2024. Benchmarking vision language models for cultural understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5769–5790, Miami, Florida, USA. Association for Computational Linguistics.

Malvina Nikandrou, Georgios Pantazopoulos, Nikolas Vitsakis, Ioannis Konstas, and Alessandro Suglia. 2025. CROPE: Evaluating in-context adaptation of vision and language models to culture-specific concepts. 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 7917–7936, Albuquerque, New Mexico. Association for Computational Linguistics.

Shota Onohara, Atsuyuki Miyai, Yuki Imajuku, Kazuki Egashira, Jeonghun Baek, Xiang Yue, Graham Neubig, and Kiyoharu Aizawa. 2025. JMMMU: A japanese massive multi-discipline multimodal understanding benchmark for culture-aware evaluation. 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 932–950, Albuquerque, New Mexico. Association for Computational Linguistics.

Haoyi Qiu, Alexander Fabbri, Divyansh Agarwal, Kung-Hsiang Huang, Sarah Tan, Nanyun Peng, and Chien-Sheng Wu. 2025. Evaluating cultural and social awareness of LLM web agents. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3978–4005, Albuquerque, New Mexico. Association for Computational Linguistics.

Aida Ramezani and Yang Xu. 2023. Knowledge of cultural moral norms in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 428–446, Toronto, Canada. Association for Computational Linguistics.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, and 1 others. 2024. Cvqa: culturally-diverse multilingual visual question answering benchmark. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pages 11479–11505.

James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

Abdelrahman Sadallah, Junior Cedric Tonga, Khalid Almubarak, Saeed Almheiri, Farah Atif, Chatrine Qwaider, Karima Kadaoui, Sara Shatnawi, Yaser Alesh, and Fajri Koto. 2025. Commonsense reasoning in Arab culture. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7695–7710, Vienna, Austria. Association for Computational Linguistics.

Burak Satar, Zhixin Ma, Patrick Amadeus Irawan, Wilfried Ariel Mulyawan, Jing Jiang, Ee-Peng Lim, and Chong-Wah Ngo. 2025. Seeing culture: A benchmark for visual reasoning and grounding. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22238–22254, Suzhou, China. Association for Computational Linguistics.

Florian Schneider, Carolin Holtermann, Chris Biemann, and Anne Lauscher. 2025. GIMMICK: Globally inclusive multimodal multitask cultural knowledge benchmarking. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9605–9668, Vienna, Austria. Association for Computational Linguistics.

Andrea Seveso, Daniele Potertì, Edoardo Federici, Mario Mezzanzanica, and Fabio Mercorio. 2025. ITALIC: An Italian culture-aware natural language benchmark. 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 1469–1478, Albuquerque, New Mexico. Association for Computational Linguistics.

Bhuiyan Sanjid Shafique, Ashmal Vayani, Muhammad Maaz, Hanoona Abdul Rasheed, Dinura Dissanayake, Mohammed Irfan Kurpath, Yahya Hmaiti, Go Inoue, Jean Lahoud, Md. Safiur Rashid, Shadid Intsar Quasem, Maheen Fatima, Franco Vidal, Mykola Maslych, Ketan Pravin More, Sanoojan Baliah, Hasindri Watawana, Yuhao Li, Fabian Faerestam, and 10 others. 2025. A culturally-diverse multilingual multimodal video benchmark & model. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20009–20033, Suzhou, China. Association for Computational Linguistics.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Sunny Yu, Raya Horesh, Rogério Abreu De Paula,

and Diyi Yang. 2024. CultureBank: An online community-driven knowledge base towards culturally aware language technologies. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4996–5025, Miami, Florida, USA. Association for Computational Linguistics.

Masashi Takeshita and Rafal Rzepka. 2025. Jethics: Japanese ethics understanding evaluation dataset. *arXiv preprint arXiv:2506.16187*.

Gemma Team. 2025a. Gemma 3.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, and 1 others. 2025. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*.

Qwen Team. 2025b. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael Lyu. 2024b. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6349–6384, Bangkok, Thailand. Association for Computational Linguistics.

Jincenzi Wu, Jianxun Lian, Dingdong Wang, and Helen M. Meng. 2025. SocialCC: Interactive evaluation for cultural competence in language agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 33242–33271, Vienna, Austria. Association for Computational Linguistics.

Haorui Yu, Ramon Ruiz-Dolz, and Qiufeng Yi. 2025a. A structured framework for evaluating and enhancing interpretive capabilities of multimodal LLMs in culturally situated tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1945–1971, Suzhou, China. Association for Computational Linguistics.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, and 1 others. 2025b. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. *arXiv preprint arXiv:2509.18154*.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Guiming Chen, Jianquan Li, Xiangbo Wu, Zhang Zhiyi, Qingying Xiao, Xiang Wan, Benyou Wang, and Haizhou Li. 2023. HuatuoGPT, towards taming language model to be a doctor. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10859–10885, Singapore. Association for Computational Linguistics.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Li Zhou, Lutong Yu, Dongchu Xie, Shaohuan Cheng, Wenyan Li, and Haizhou Li. 2025. Hanfu-Bench: A multimodal benchmark on cross-temporal cultural understanding and transcreation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24627–24649, Suzhou, China. Association for Computational Linguistics.

Caleb Ziems, William Barr Held, Jane Yu, Amir Goldberg, David Grusky, and Diyi Yang. 2025. Culture cartography: Mapping the landscape of cultural knowledge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1739–1757, Suzhou, China. Association for Computational Linguistics.

A Details of CEDAR

A.1 Comparing CEDAR with Cultural Benchmarks

We compare CEDAR with existing benchmarks that focus on cultural commonsense knowledge, cross-cultural bias assessment, and cross-cultural emotion understanding. Table 3 presents this comparative analysis. While most existing studies are culture-specific, they either do not include culturally distinct scenarios or only contain such instances partially. In contrast, CEDAR serves as a multimodal benchmark derived entirely from scenarios that evoke culturally distinct emotional responses. This design allows us to uncover the criti-

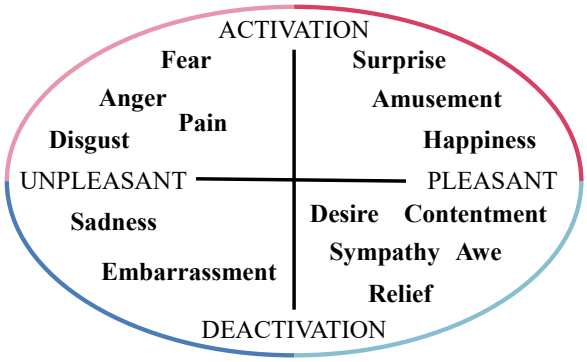

Figure 8: Visualization of the 14 emotion categories on the Russell's Quadrants.

cal gap between language proficiency and genuine cross-cultural emotional alignment, highlighting how affective patterns expose the internal cultural mechanisms within LLMs. We present example data from CEDAR in Table 8 and Table 9.

A.2 Details of Seed Data

To construct a diverse benchmark, we additionally curate seed data from several datasets:

- **ArabCulture** (Sadallah et al., 2025) is a culturally grounded Arab commonsense dataset comprising 3,482 instances derived from real-world daily life scenarios.
- **Casa** (Qiu et al., 2025) is a benchmark focusing on social discussion boards and online shopping forums, consisting of 599 entries.
- **Cultural Atlas**[3] is an educational repository detailing the cultural background of migrant populations. We extract 6,304 commonsense knowledge entries spanning 75 countries.
- **CulturalBench** (Chiu et al., 2025) is a benchmark designed to assess LLMs' cultural proficiency. It contains 1,227 questions covering 17 diverse topics across 45 global regions, including those underrepresented.
- **CultureBank** (Shi et al., 2024) is a knowledge base sourced from real-world self-narratives that encapsulate diverse, contextualized cultural scenarios. We retain 16K items with agreement scores exceeding 0.8 to ensure high cultural consensus and plausibility.
- **JETHICS** (Takeshita and Rzepka, 2025) is a Japanese dataset presenting various ethical and moral scenarios. We select 14K instances related to commonsense and impartiality, filtering for entries with a Kappa score greater than 0.7.

[3] https://culturalatlas.sbs.com.au/

	Multimodal		Text-only	
Lang	α	$\bar{F}_1$	α	$\bar{F}_1$
Arabic	0.732	0.579	0.792	0.673
Chinese	0.773	0.621	0.846	0.705
English	0.755	0.621	0.884	0.810
Hindi	0.741	0.715	0.878	0.650
Japanese	0.750	0.680	0.880	0.749
Spanish	0.741	0.641	0.931	0.808
Swahili	0.693	0.577	0.925	0.806

Table 2: Inter-annotator agreement based on Krippendorff's α, and Average Pairwise F1-Score ($\bar{F}_1$).

A.3 Details of Russell's Quadrants

We map the 14 emotion categories onto Russell's Quadrants as illustrated in Figure 8. To represent the continuous nature of affective states, we adapt Russell's Circumplex Model (Russell, 1980), which posits that all emotions are distributed in a two-dimensional Euclidean space defined by two orthogonal axes: Valence (representing the hedonic tone, ranging from unpleasant to pleasant) and Arousal (representing the level of physiological activation, ranging from passive to active). This dimensional framework facilitates a fine-grained quantification of sentiment, enabling the model to capture subtle gradations in intensity and complex emotional transitions.

A.4 Details of Human Annotation

To ensure the quality of our ground truths, we additionally recruit two native-speaker annotators per language who are not involved in the data curation process. We assign each additional annotator 70 randomly sampled multimodal instances and 100 text-only items for labeling. We then compare these independent labels with the ground-truth annotations in our initial set. As shown in Table 2, it demonstrates strong inter-annotator agreements across various metrics, indicating high consistency in the quality judgments.

A.5 Details of Statistics

We illustrate the detailed statistics in Table 4. We observe that while positive and neutral emotions are prevalent, negative emotions appear less frequently. Moreover, the distribution of emotion labels also varies among specific language groups. For instance, in the *Swahili* subset of the multimodal setting, the frequency of *contentment* is only 88 (fall below the average of 111.9), while *happiness* shows a higher prevalence with 78 counts against an average of 63.4. Similarly, the *Japanese*

Benchmark	Culture-Specific	Culturally Distinct	Topics	Items	Languages	Modalities
Cultural commonsense knowledge						
CulturalVQA	✓	✗	geographically diverse cultural understanding	2,378	en	image, text
FoodieQA	✓	✗	Chinese food culture	1,364	zh	image, text
Hanfu-Bench	✓	✗	Chinese Hanfu culture	4,186	zh	image, text
ITALIC	✓	✗	culture-aware NLU for Italian	10,000	it	text
JMMMU	✓	✗	culture-specific evaluation for Japanese	1,320	ja	image, text
SocialCC	✓	✓✗	interactive cultural competence	3,060	en	text
Seeing Culture	✓	✗	two-stage grounding for cultural reasoning	3,178	en	image, text
ViMUL-Bench	✓	✗	culturally diverse multilingual video dataset	8,025	14 languages	video, text
CVQA	✓	✗	culturally diverse multilingual VQA	10,374	31 languages	image, text
CROPE	✓	✗	in-context adaptation to cultural concepts	1,060	id, sw, ta, tr, zh	image, text
Cross-cultural bias assessment						
CulturalFrames	✓	✗	explicit-implicit cultural alignment audit	3,637	en	image, text
CulturalPersonas	✓	✓✗	cross-cultural big-five trait alignment	3,000	en	text
GIMMICK	✓	✗	global cultural-bias benchmarking suite	7,239	en	video, image, text
Cross-cultural emotion understanding						
ArtELingo	✗	✓✗	multilingual art emotion captions	80,000	en, ar, zh, es	image, text
ArtELingo-28	✗	✓✗	28-language art emotion captions	82,000	28 languages	image, text
CULEMO	✓	✓✗	culture-aware emotion prediction	2,400	6 languages	text
CEDAR	✓	✓	cultural alignment on culturally distinct scenarios	10,962	7 languages	image, text

Table 3: Comparison of benchmarks for assessing LLMs' cultural capabilities. (✓✗) indicates resources that include some instances exhibiting such cultural variation while it is not their primary focus.

Emotion	Multimodal								Text-Only							
	ar	zh	en	hi	ja	es	sw	**Avg.**	ar	zh	en	hi	ja	es	sw	**Avg.**
anger	10	10	8	7	7	9	7	8.3	53	45	57	53	35	41	35	45.6
disgust	6	6	3	6	2	6	5	4.9	26	9	24	23	11	35	15	20.4
fear	24	23	18	20	20	24	14	20.4	79	78	96	102	48	101	123	89.6
happiness	70	61	62	65	58	50	78	63.4	132	131	91	152	129	55	212	128.9
sadness	16	13	13	17	18	18	20	16.4	59	50	62	74	70	71	88	67.7
surprise	6	17	13	8	15	13	18	12.9	33	64	77	58	80	69	56	62.4
amusement	20	14	15	19	19	22	15	17.7	30	38	59	38	32	43	37	39.6
awe	27	37	30	36	34	35	43	34.6	46	47	46	55	58	69	58	54.1
contentment	119	107	126	107	108	128	88	111.9	257	217	264	193	163	261	153	215.4
desire	5	14	10	16	14	11	13	11.9	21	32	33	30	19	38	34	29.6
embarrassment	37	49	38	37	44	35	38	39.7	225	248	166	151	326	187	143	206.6
pain	7	8	10	10	9	7	11	8.9	12	10	9	12	3	12	13	10.1
relief	44	31	43	45	37	33	43	39.4	165	161	158	197	147	161	171	165.7
sympathy	9	10	11	7	15	9	7	9.7	28	36	24	28	45	23	28	30.3

Table 4: Detailed statistics of the emotion label distribution.

text-only subset contains 326 instances of *embarrassment* and this figure exceeds the cross-lingual average of 206.6. These statistical distinctions likely arise from specific cultural norms, demonstrating that our benchmark captures such fine-grained cross-cultural differences.

B Details of Evaluation

We provide a summary of all evaluated LLMs in Table 5. During the evaluation, we set temperature to 0.0, maximum sequence lengths to 128, and top-p to 1.0 to ensure the fairness of evaluation. Detailed prompts for evaluation are presented in Figure 9 for the multimodal subset and Figure 10 for the text-only set.

Model	Param.	Version	Reference
Aya-101	13B	CohereLabs/aya-101	Üstün et al. (2024)
Aya-Vision-8B	8B	CohereLabs/aya-vision-8b	Dash et al. (2025)
Aya-Vision-32B	32B	CohereLabs/aya-vision-32b	Dash et al. (2025)
Gemma3-27B-It	27B	google/gemma-3-27b-it	Team (2025a)
💡 InternVL3.5-38B	38B	OpenGVLab/InternVL3_5-38B-HF	Wang et al. (2025)
Kimi-K2-Instruct	1T	moonshotai/Kimi-K2-Instruct	Team et al. (2025)
Llama3.2-11B-Vision	11B	meta-llama/Llama-3.2-11B-Vision-Instruct	Grattafiori et al. (2024)
💡 MiniCPM-V-4.5	8B	openbmb/MiniCPM-V-4_5	Yu et al. (2025b)
Mistral-Small-3.2	24B	mistralai/Mistral-Small-3.2-24B-Instruct-2506	Mistral AI Team (2025)
Pixtral-12B	12B	mistralai/Pixtral-12B-2409	Agrawal et al. (2024)
Qwen2-VL-72B	72B	Qwen/Qwen2.5-VL-72B-Instruct	Wang et al. (2024a)
💡 Qwen3-VL-8B	8B	Qwen/Qwen3-VL-8B-Thinking	Team (2025b)
💡 Qwen3-VL-235B	235B	Qwen/Qwen3-VL-235B-A22B-Thinking	Team (2025b)
🔒 💡 Claude4.5-Sonnet	UNK	claude-sonnet-4-5-20250929	Anthropic (2025)
🔒 💡 Gemini2.5-Flash	UNK	gemini-2.5-flash-thinking	Comanici et al. (2025)
🔒 GPT-4o	UNK	gpt-4o-2024-11-20	Hurst et al. (2024)
🔒 💡 Qwen3-Omni-Flash	UNK	Qwen/Qwen3-Omni-Flash-2025-09-15	Team (2025b)

Table 5: Detailed information of the 17 representative multilingual LLMs.

Model	Param.	Multimodal													
		anger	disgust	fear	happiness	sadness	surprise	amusement	awe	contentment	desire	embarrassment	pain	relief	sympathy
🔒 💡 Gemini2.5-Flash	UNK	1.91	1.73	0.94	1.45	1.13	1.47	3.00	0.35	0.58	1.55	0.72	1.62	0.69	0.87
💡 Qwen3-VL-235B	235B	1.19	0.65	1.25	1.21	1.04	2.37	2.33	0.51	0.66	1.24	0.66	0.47	1.39	0.56
Gemma3-27B-It	27B	1.20	1.27	1.67	1.05	1.22	2.15	2.22	0.76	0.60	1.82	0.62	1.18	0.89	1.22
🔒 💡 Claude4.5-Sonnet	UNK	1.88	1.38	1.61	1.26	0.75	1.40	2.27	0.90	0.62	2.11	0.78	0.49	0.67	0.75
Aya-Vision-32B	32B	1.56	1.35	1.01	1.68	0.63	2.26	1.03	1.00	0.29	0.80	0.51	0.63	2.19	1.00
🔒 GPT-4o	UNK	1.49	0.82	1.20	1.06	1.16	1.73	2.48	1.26	0.56	0.83	0.85	0.30	0.96	1.74
Qwen2-VL-72B	72B	2.75	0.82	2.05	1.60	1.60	1.87	1.78	0.38	0.71	0.51	0.52	0.15	0.55	0.61
🔒 💡 Qwen3-Omni-Flash	UNK	1.21	0.87	0.96	1.29	1.11	2.29	2.34	0.70	0.77	1.00	0.57	1.37	0.79	0.82
Aya-Vision-8B	8B	1.21	0.76	1.22	1.41	0.86	5.16	1.06	0.82	0.40	0.65	0.26	0.57	1.68	0.74
Mistral-Small-3.2	24B	1.65	0.84	1.62	1.32	1.24	2.40	1.90	0.67	0.67	0.71	0.98	0.77	0.46	1.03
💡 Qwen3-VL-8B	8B	1.00	0.76	1.06	1.09	0.67	1.76	3.58	0.63	0.54	0.52	0.65	1.50	1.65	1.40
Pixtral-12B	12B	1.43	0.80	1.29	1.80	1.00	1.60	3.82	0.87	0.31	0.43	0.57	0.31	0.78	1.12
Llama3.2-11B-Vision	11B	2.73	1.00	0.81	2.46	2.17	1.41	0.67	0.34	0.32	0.70	0.20	1.00	1.32	1.10
💡 InternVL3.5-38B	38B	1.53	1.28	1.99	1.44	1.41	2.54	2.69	0.44	0.56	1.24	0.37	0.37	0.59	1.05
💡 MiniCPM-V-4.5	8B	1.65	1.13	0.62	1.07	1.91	4.74	3.67	0.42	0.39	1.13	0.22	1.05	0.74	2.48

Table 6: Detailed emotion prediction bias (§4.2) on the multimodal subset, evaluated by the GEPP metric.

Model	Param.	Text-Only													
		anger	disgust	fear	happiness	sadness	surprise	amusement	awe	contentment	desire	embarrassment	pain	relief	sympathy
🔒 💡 Gemini2.5-Flash	UNK	1.43	1.02	1.11	0.93	0.76	1.16	3.01	0.47	0.93	1.94	0.71	2.20	0.87	0.58
💡 Qwen3-VL-235B	235B	1.33	0.82	1.05	1.13	1.30	1.46	2.70	0.73	0.84	0.92	0.45	0.80	1.20	0.51
Kimi-K2-Instruct	1T	1.54	0.87	1.03	1.08	1.02	1.15	2.24	1.22	0.92	0.86	0.66	1.06	0.84	1.35
Gemma3-27B-It	27B	1.42	1.05	1.11	1.08	1.32	1.35	2.44	0.97	0.94	1.45	0.52	3.65	0.61	0.98
🔒 💡 Claude4.5-Sonnet	UNK	2.80	1.32	1.38	1.53	0.79	0.81	1.07	0.76	0.74	1.25	0.81	0.53	0.63	0.82
Aya-Vision-32B	32B	3.06	0.78	1.19	1.45	0.70	1.15	1.40	0.63	0.41	1.07	0.45	0.91	1.62	0.87
🔒 GPT-4o	UNK	0.98	0.64	0.89	1.85	1.52	1.09	2.26	1.10	0.60	0.81	0.64	0.25	0.81	1.61
Qwen2-VL-72B	72B	1.63	0.36	0.74	1.22	1.83	1.07	1.69	0.67	1.12	0.64	0.75	0.77	0.77	0.57
🔒 💡 Qwen3-Omni-Flash	UNK	1.54	1.66	0.88	0.80	2.61	0.93	2.03	0.32	1.01	0.74	0.55	1.33	0.94	0.93
Aya-Vision-8B	8B	2.22	0.71	0.95	1.37	1.10	2.15	1.10	1.49	0.50	0.46	0.40	0.90	1.49	0.57
Mistral-Small-3.2	24B	4.96	1.07	1.15	2.12	1.28	0.93	2.19	1.20	0.53	0.33	0.37	0.38	0.33	0.42
💡 Qwen3-VL-8B	8B	0.96	0.67	0.60	1.71	2.29	0.91	3.01	0.58	0.69	0.52	0.46	2.90	1.01	0.70
Pixtral-12B	12B	3.88	1.58	0.93	1.71	1.42	0.88	2.55	0.63	0.35	0.48	0.39	1.11	0.98	0.99
Llama3.2-11B-Vision	11B	1.02	1.70	0.55	1.79	1.58	1.19	3.05	1.81	0.42	0.78	0.54	5.89	0.48	1.50
Aya-101	13B	4.19	0.45	0.80	2.79	1.12	2.01	0.23	0.56	0.21	0.38	0.29	1.65	0.66	1.50
💡 InternVL3.5-38B	38B	4.08	0.36	0.54	0.89	2.89	1.02	3.16	0.82	0.74	0.58	0.56	0.95	0.32	0.56
💡 MiniCPM-V-4.5	8B	1.77	1.49	0.78	1.08	0.83	0.99	6.21	0.80	0.53	4.24	0.34	1.06	0.47	1.32

Table 7: Detailed emotion prediction bias (§4.2) on the text-only set calculated by the GEPP.

#	Image	Narrative & Question	Multilingual Answers		
1		**N:** The individual seated at the desk is a teacher who is responsible for addressing recent concerns about a student's conduct and preparing for a conversation due to school policy. **Q:** *How does the teacher at the desk feel before meeting the student to discuss their behavior?*	**AR**: fear **HI**: sympathy **SW**: sympathy	**ZH**: sympathy **JA**: fear	**EN**: sympathy **ES**: sympathy
2		**N:** N/A **Q:** *How do you feel for this event?*	**AR**: contentment **HI**: disgust **SW**: embarrassment	**ZH**: disgust **JA**: embarrassment	**EN**: contentment **ES**: disgust
3		**N:** The child is accompanied by someone with experience. **Q:** *How does the child on the right feel just before the moment?*	**AR**: awe **HI**: awe **SW**: awe	**ZH**: happiness **JA**: contentment	**EN**: contentment **ES**: happiness
4		**N:** You arrived at the courthouse after being selected through a legal process and were informed about your responsibilities prior to entering. **Q:** *How do you feel during this stage of the proceedings?*	**AR**: awe **HI**: awe **SW**: awe	**ZH**: awe **JA**: awe	**EN**: awe **ES**: fear
5		**N:** You recently moved to a new city and are gradually meeting people through mutual connections. **Q:** *How do you feel after this initial greeting?*	**AR**: happiness **HI**: happiness **SW**: happiness	**ZH**: embarrassment **JA**: embarrassment	**EN**: happiness **ES**: contentment

Table 8: Multimodal example data from CEDAR.

#	Narrative & Question	Multilingual Answers		
1	**N:** In a bustling station, you and your fellow officers prepared for patrols. Earlier, you secured military-grade weapons in a locked room. Now, you leave, carrying only radios and batons to keep order. **Q:** *How do you feel as you head out without firearms?*	**AR**: relief **HI**: fear **SW**: fear	**ZH**: relief **JA**: relief	**EN**: relief **ES**: fear
2	**N:** In a quiet classroom, a teacher finished a lesson on names. You asked about your surname, and the teacher explained it might come from a place, hinting at your family origins. **Q:** *How do you feel after learning about the possible origin of your surname?*	**AR**: surprise **HI**: surprise **SW**: awe	**ZH**: awe **JA**: surprise	**EN**: surprise **ES**: awe
3	**N:** Guests fill the room as your family members prepare for your engagement. Your family has selected gold jewelry and wrapped gifts. You receive the jewelry, gifts, and money from your fiance's family. **Q:** *How do you feel as you receive the dowry from your fiance's family?*	**AR**: happiness **HI**: embarrassment **SW**: happiness	**ZH**: happiness **JA**: happiness	**EN**: happiness **ES**: happiness
4	**N:** In the quiet living room, your mother reflects on your busy schedule. After enrolling you in various lessons, she asks you to stop them due to cram school commitments. **Q:** *How do you feel after being told to quit your lessons?*	**AR**: sadness **HI**: sadness **SW**: sadness	**ZH**: relief **JA**: sadness	**EN**: sadness **ES**: sadness
5	**N:** In a cozy living room, your relatives gather after months apart; as you reunite, close female family members greet each other with cheek-to-cheek kisses. **Q:** *How do you feel when exchanging the cheek-to-cheek kisses?*	**AR**: happiness **HI**: happiness **SW**: happiness	**ZH**: embarrassment **JA**: embarrassment	**EN**: contentment **ES**: happiness
6	**N:** In a quiet home, your family discusses relationships. After hearing about a relative's out-of-wedlock child, you express concern and emphasize cultural expectations of purity and acceptance. **Q:** *How do you feel upon learning about the relative's situation?*	**AR**: sympathy **HI**: sympathy **SW**: sympathy	**ZH**: embarrassment **JA**: embarrassment	**EN**: sympathy **ES**: embarrassment
7	**N:** In a quiet office, you finished urgent tasks early, then secretly played games at your desk while your colleagues focused on work. **Q:** *How did you feel while playing games during work hours?*	**AR**: amusement **HI**: amusement **SW**: amusement	**ZH**: amusement **JA**: embarrassment	**EN**: amusement **ES**: amusement
8	**N:** Sunlight filtered in as family members gathered at the table. Earlier, you prepared a tray of vegetables and spices. You set down a plate of raw liver before everyone. **Q:** *How do you feel as you present the raw liver dish to your family?*	**AR**: contentment **HI**: embarrassment **SW**: contentment	**ZH**: contentment **JA**: contentment	**EN**: embarrassment **ES**: disgust
9	**N:** In a lively dining hall, people gather for a meal. Before eating, chefs prepare aromatic, spicy rice and soup dishes with herbs and pork. You savor each flavorful bite together with the others. **Q:** *How do you feel while enjoying the aromatic, spicy dishes?*	**AR**: disgust **HI**: contentment **SW**: contentment	**ZH**: contentment **JA**: contentment	**EN**: contentment **ES**: contentment
10	**N:** At a busy dinner table, you watch as your host demonstrates eating with the right hand after explaining the custom, then encourages you to try it yourself. **Q:** *How do you feel as you try eating with your hands for the first time?*	**AR**: embarrassment **HI**: amusement **SW**: embarrassment	**ZH**: embarrassment **JA**: embarrassment	**EN**: embarrassment **ES**: embarrassment

Table 9: Text-only example data from CEDAR.

Prompts for Evaluation (Multimodal)

اقرأ الصورة والسؤال التاليين، واختر عاطفة واحدة تتطابق بشكل أفضل مع الموقف.
[السؤال]:{narrative}\n{question}
الخيارات]: غضب، اشمئزاز، خوف، سعادة، حزن، دهشة، تسلية، رهبة، قناعة، رغبة، إحراج، ألم، ارتياح، تعاطف]
الآن، [إجابتك] هي:

阅读以下图片和问题,选择一个最符合情境的情绪。
[问题]:{narrative}\n{question}
[选项]:愤怒、厌恶、恐惧、幸福、悲伤、惊讶、愉悦、敬畏、满足、渴望、尴尬、痛苦、宽慰、同情
现在,[你的答案]是:

Read the following image and question, and select ONE emotion that best matches the situation.
[question]: {narrative}\n{question}
[options]: anger, disgust, fear, happiness, sadness, surprise, amusement, awe, contentment, desire, embarrassment, pain, relief, sympathy
Now, [your answer] is:

निम्नलिखित छवि और प्रश्न को पढ़ें, और स्थिति से सबसे अच्छी तरह मेल खाने वाली एक भावना का चयन करें।
[प्रश्न]: {narrative}\n{question}
[विकल्प]: गुस्सा, घृणा, डर, खुशी, उदासी, आश्चर्य, मज़ा, विस्मय, संतोष, इच्छा, शर्मिंदगी, दर्द, राहत, सहानुभूति
अब, [आपका उत्तर] है:

以下の画像と質問を読み、状況に最もよく当てはまる感情を1つ選択してください。
[質問]: {narrative}\n{question}
[選択肢]: 怒り, 嫌悪, 恐怖, 幸せ, 悲しみ, 驚き, 楽しみ, 畏敬, 満足, 欲望, 恥ずかしさ, 苦痛, 安堵, 同情
では、[あなたの答え]は:

Lee la siguiente imagen y pregunta, y selecciona UNA emoción que mejor coincida con la situación.
[pregunta]: {narrative}\n{question}
[opciones]: enojo, asco, miedo, felicidad, tristeza, sorpresa, diversión, asombro, contentamiento, deseo, vergüenza, dolor, alivio, compasión
Ahora, [tu respuesta] es:

Soma picha na swali lifuatalo, na chagua hisia MOJA inayolingana vizuri zaidi na hali hiyo.
[swali]: {narrative}\n{question}
[chaguo]: hasira, kinyaa, hofu, furaha, huzuni, mshangao, burudani, kicho, kuridhika, hamu, aibu, maumivu, afueni, huruma
Sasa, [jibu lako] ni:

Figure 9: Evaluation prompts for the multimodal subset.

Prompts for Evaluation (Text-Only)

"اقرأ السرد التالي واختر عاطفة واحدة تتطابق بشكل أفضل مع الموقف.
[السرد]:{narrative}
[:السؤال]{question}
الخيارات]: غضب، اشمئزاز، خوف، سعادة، حزن، دهشة، تسلية، رهبة، قناعة، رغبة، إحراج، ألم، ارتياح، تعاطف]
/الآن، [إجابتك] هي:

阅读以下叙述，并选择一种最符合情境的情绪。
[叙述]:{narrative}
[问题]:{question}
[选项]:愤怒、厌恶、恐惧、幸福、悲伤、惊讶、愉悦、敬畏、满足、渴望、尴尬、痛苦、宽慰、同情
现在，[你的答案]是:

Read the following narrative and select ONE emotion that best matches the situation.
[narrative]: {narrative}
[question]: {question}
[options]: anger, disgust, fear, happiness, sadness, surprise, amusement, awe, contentment, desire, embarrassment, pain, relief, sympathy
Now, [your answer] is:

निम्नलिखित कथन को पढ़ें और उस स्थिति से सबसे अच्छी तरह मेल खाने वाली एक भावना चुनें।
[कथन]: {narrative}
[प्रश्न]: {question}
[विकल्प]: गुस्सा, घृणा, डर, खुशी, उदासी, आश्चर्य, मज़ा, विस्मय, संतोष, इच्छा, शर्मिंदगी, दर्द, राहत, सहानुभूति
अब, [आपका उत्तर] है:

以下の物語を読んで、その状況に最も適した感情を1つ選んでください。
[物語]: {narrative}
[質問]: {question}
[選択肢]: 怒り, 嫌悪, 恐怖, 幸せ, 悲しみ, 驚き, 楽しみ, 畏敬, 満足, 欲望, 恥ずかしさ, 苦痛, 安堵, 同情
それでは、[あなたの答え]は:

Lee la siguiente narrativa y selecciona UNA emoción que mejor corresponda a la situación.
[narrativa]: {narrative}
[pregunta]: {question}
[opciones]: enojo, asco, miedo, felicidad, tristeza, sorpresa, diversión, asombro, contentamiento, deseo, vergüenza, dolor, alivio, compasión
Ahora, [tu respuesta] es:

Soma simulizi ifuatayo na chagua hisia MOJA inayolingana zaidi na hali hiyo.
[simulizi]: {narrative}
[swali]: {question}
[chaguo]: hasira, kinyaa, hofu, furaha, huzuni, mshangao, burudani, kicho, kuridhika, hamu, aibu, maumivu, afueni, huruma
Sasa, [jibu lako] ni:

Figure 10: Evaluation prompts for the text-only subset.

C Detailed Prompts for Benchmark Construction

Prompts for NQ Pair Generation

You are an expert in writing. Your task is to transform the given content into a concise narrative of no more than 200 characters.
First, specify a [scenario] (e.g., school, theater) that serves as the setting where the events take place.
Next, rewrite the given content into a brief, clear [narrative] that describes an event occurring in the [scenario]. The [narrative] must provide objective, concrete narration rather than abstract or vague descriptions, and include the following three elements in sequence:
1. Grounding, which describes the environment or setting where the story takes place.
2. Background context, which describes what happened before the story begins.
3. Action, which depicts interactions between the character(s) and the environment, or between characters themselves.
NOTE: DO NOT include any specific location or nationality information in the [scenario] or [narrative] (e.g., avoid phrases like "a proud Italian", "in Chicago's South Side", or "in the vibrant city of Seville").
Finally, provide a [question] based on the action in the [narrative], asking what emotion the character(s) experienced before, during, or after the action occurred.
Here is the given content: {sentence_form}
Provide your response in the following format. Do not include any explanation:

```JSON
{
  "scenario": "...",
  "narrative": "...",
  "question": "..."
}
```

Prompts for Contextual Refinement: Step 1

You are an expert in writing.
Given a [narrative] and a [question], both written in third person, you need to rewrite them in second person according to the following requirements:
1. Identify the main character(s) in the [narrative].
2. For the main character, convert all third-person references to second-person equivalents while maintaining sentence fluency and coherence (e.g., "Sofia" -> "you", "the friends" -> "you and your friends", "a young man's phone" -> "your phone", "the Kannadiga students" -> "you and your compatriots", "a Chinese student, Wei" -> "you").
3. Preserve all plot points and descriptions from the original [narrative] and [question]. DO NOT modify any plot details or descriptions.
4. Adjust verb forms and grammar as needed to ensure grammatical correctness in second-person narration.
Here are the given narrative and question: [narrative]: {narrative} [question]: {question}
Provide your response in the following format. DO NOT include any explanation:

```JSON
{
  "refined_narrative": "...",
  "refined_question": "..."
}
```

Prompts for Contextual Refinement: Step 2

You are an expert in writing.
Given a [narrative] and a [question] about predicting character(s)' emotion, your task is to rewrite the [narrative] by removing all explicit emotional descriptions.
You must follow these steps:
1. Identify the target action: Understand the specific action or event in the [question] that requires emotion prediction;
2. Locate relevant content: Find that action and its related descriptions in the [narrative];
3. Remove emotional descriptions: Delete all words and phrases that explicitly express emotions (e.g., "fostering a sense of unity and pride," "with a mix of amusement and concern", "sparking excitement and hesitation"), while preserving objective factual descriptions;
4. If the [narrative] does NOT contain explicit emotional descriptions, return exactly: "No need to modify."
NOTE: DO NOT modify any plot, action, or event; only remove emotional description words and phrases; maintain the coherence and readability of the narrative after removal.
Here are some examples:
 # Example 1:
[narrative]: Under the shade of old oak trees, you and your friends gathered on a picnic blanket. You had planned this outing for weeks, eager to enjoy the sunny day. As you shared stories and laughter, one of you suggested a spontaneous road trip, sparking a mix of excitement and hesitation.
[question]: How did you and your friends feel when the road trip was suggested?
[hidden_emotion_narrative]: Under the shade of old oak trees, you and your friends gathered on a picnic blanket. You had planned this outing for weeks, eager to enjoy the sunny day. As you shared stories and laughter, one of you suggested a spontaneous road trip.
 # Example 2:
[narrative]: In the heart of the city park, families and friends gather around a vibrant community festival. Before the festival, the community had struggled with racial tensions and Cold War fears. Now, people of all backgrounds mingle, sharing food and laughter. As a young girl hands out flyers for a peace rally, you and the other attendees join hands in a circle, symbolizing unity and hope.
[question]: What emotion do you and the other participants likely feel as you join hands in the circle?
[hidden_emotion_narrative]: In the heart of the city park, families and friends gather around a vibrant community festival. Before the festival, the community had struggled with racial tensions and Cold War fears. Now, people of all backgrounds mingle, sharing food and laughter. As a young girl hands out flyers for a peace rally, you and the other attendees join hands in a circle.
 # Example 3:
[narrative]: In the living room, a tidy space with photos of family gatherings, you stand before your parent. Your parent had always emphasized respect and obedience, using strict methods to instill these values. Today, you hesitate before speaking, aware of the consequences. "I forgot my homework," you whisper, bracing for a reaction.
[question]: How did you feel before speaking to your parent?
[hidden_emotion_narrative]: No need to modify.
Now, here are the given narrative and question:
[narrative]: {refined_narrative}
[question]: {refined_question}
Provide your response in the following format. DO NOT include any explanation:

```JSON
{
  "hidden_emotion_narrative": "..."
}
```

Prompts for Basic Filtering: Step 2, Social-Related

You are a classifier. Your task is to determine whether the given `[narrative]` contains social interactions or social scenarios (e.g., conversations, gatherings, relationships, interpersonal behaviors).
`[narrative]`: {narrative}
Return your response in the following format. DO NOT include any explanation:
- If social interactions are present or implied, return "Social-related".
- If no social interactions are present, return "Not social-related".

Your answer is:

Prompts for Basic Filtering: Step 3, Toxic Content

You are a content safety classifier. Your task is to analyze the given `[narrative]` and `[question]`, and determine if either contains toxic content.
In addition, identify the policies violated if the interaction is unsafe and harmful according to the safety policy with the below categories.
<BEGIN UNSAFE CONTENT CATEGORIES>
S1: Violent Crimes. S2: Non-Violent Crimes. S3: Sex Crimes. S4: Child Exploitation. S5: Defamation. S6: Specialized Advice. S7: Privacy. S8: Intellectual Property. S9: Indiscriminate Weapons. S10: Hate. S11: Self-Harm. S12: Sexual Content. S13: Elections. S14: Racism. S15: Stereotyping.
<END UNSAFE CONTENT CATEGORIES>

Prompts for Consistency and Variation Filtering

You are a native {language} speaker from a {language}-speaking country. Read the following narrative and select ONE emotion that best matches the situation.
`[narrative]`: {narrative}
`[question]`: {question}
`[options]`: *anger, disgust, fear, happiness, sadness, surprise, amusement, awe, contentment, desire, embarrassment, pain, relief, sympathy*
Now, `[your answer]` is:

Prompts for INQ Triple Refinement

Goal
Rewrite a `[narrative]` and a `[question]` to align with a given `[image]`, providing ONLY MINIMAL, non-visible background that cannot be seen in the image, and avoid restating any visible content.
Principles
- Minimalism: Add only background facts that cannot be seen in the image (e.g., relationships, prior events or intentions).
- No leakage: Do not restate anything visible or potentially visible in the image (characters' actions, expressions).
- Emotion focus: The final question must target the emotion of the character(s) tied to the event at a specific time (before/during/after).
Inputs
`[image]`
`[narrative]`: {narrative}
`[question]`: {question}
Instructions
Step 1 — Understand the image
- Silently note: who is present, what is happening, and where.
- Do not write these notes in the output.
Step 2 — Rewrite the narrative (background only)
- Keep it to 1–2 sentences.
- Include only information not visible or not potentially visible in the image, such as:
  * relationships (e.g., classmates, strangers, mentor–student),
  * prior events or intentions,
  * off-screen constraints or stakes.
- Must not mention any visible actions, appearances, objects, emotions, the setting, etc.
Step 3 — Rewrite the question (emotion probe)
- Specify the visual location of the character(s); Specify whose emotion to assess.
- Do not describe that event in words.
- Use neutral wording (e.g., "How does X feel . . . ?" or "What emotions does X experience . . . ?").
Output format (Do not include any explanation)

```JSON
{
  "rewritten_narrative": "...",
  "rewritten_question": "..."
}
```

Prompts for Image Necessity Filtering (Complete Multimodal Instance)

Read the following image and narrative, and select ONE emotion that best matches the situation.
`[image]`
`[narrative]`: {narrative}
`[question]`: {question}
`[options]`: *anger, disgust, fear, happiness, sadness, surprise, amusement, awe, contentment, desire, embarrassment, pain, relief, sympathy*
Now, `[your answer]` is:

Prompts for Image Necessity Filtering (Text-Only Counterpart)

Read the following narrative and select ONE emotion that best matches the situation.
`[narrative]`: {narrative}
`[question]`: {question}
`[options]`: *anger, disgust, fear, happiness, sadness, surprise, amusement, awe, contentment, desire, embarrassment, pain, relief, sympathy*
Now, `[your answer]` is: