

Advancing Equitable AI: Evaluating Cultural Expressiveness in LLMs for Latin American Contexts

Brigitte A. Mora-Reyes^{*1} Jennifer A. Drewyor^{*2} Abel A. Reyes-Angulo³

Abstract

Artificial intelligence (AI) systems often reflect biases from economically advanced regions, marginalizing contexts in economically developing regions like Latin America due to imbalanced datasets. This paper examines AI representations of diverse Latin American contexts, revealing disparities between data from economically advanced and developing regions. We highlight how the dominance of English over Spanish, Portuguese, and indigenous languages such as Quechua and Nahuatl perpetuates biases, framing Latin American perspectives through a Western lens. To address this, we introduce a culturally aware dataset rooted in Latin American history and socio-political contexts, challenging Eurocentric models. We evaluate six language models on questions testing cultural context awareness, using a novel Cultural Expressiveness metric, statistical tests, and linguistic analyses. Our findings show that some models better capture Latin American perspectives, while others exhibit significant sentiment misalignment ($p < 0.001$). Fine-tuning Mistral-7B with our dataset improves its cultural expressiveness by 42.9%, advancing equitable AI development. We advocate for equitable AI by prioritizing datasets that reflect Latin American history, indigenous knowledge, and diverse languages, while emphasizing community-centered approaches to amplify marginalized voices. Code and datasets available at: <https://github.com/areyesan/Advancing-Equitable-AI>

1. Introduction

Artificial intelligence (AI) systems, particularly large language models (LLMs), often reflect biases rooted in datasets sourced from economically advanced regions, marginalizing economically developing regions like Latin America. This leads to incomplete or Eurocentric representations of diverse local contexts, such as indigenous rights, socio-political dynamics, and cultural identities across urban, rural, and indigenous communities, perpetuating cultural erasure (the loss or suppression of cultural identities due to dominant influences) and a center-periphery dynamic (a socio-economic framework where a dominant center exploits or marginalizes peripheral regions). Here, "economically advanced regions" refers to industrialized nations with high GDP (e.g., USA, Western Europe), while "economically developing regions" include Latin America and other areas with emerging economies.

In this paper, we address these biases by introducing a culturally aware dataset tailored to diverse Latin American contexts, evaluating six LLMs—Mistral-7B, Zephyr-7B, BLOOM-7B, Llama-2-7B, Grok, and ChatGPT—to quantify their cultural and tonal disparities. Our analysis reveals that LLMs often fail to capture the linguistic diversity (e.g., Spanish, Portuguese, Quechua, Nahuatl) and socio-political nuances of Latin America, with some models exhibiting excessive positivity or negativity that misaligns with local perspectives. To bridge this gap, as shown in Figure 1, we propose a framework to enhance cultural expressiveness, detailed in Sections 4.2 and 4.3, achieving a 42.9% improvement (Section 4.3). Our work contributes to advancing equitable AI by advocating for inclusive datasets that prioritize regional histories, indigenous knowledge, and local authorship, while emphasizing community-centered approaches for future development.

2. Background

The development and deployment of large language models (LLMs) have been transformative, yet their reliance on datasets skewed toward the Global North has raised concerns about cultural biases and representational harms. Bender (Bender et al., 2021) argue that the homogeneity of

^{*}Equal contribution ¹Facultad Jurisprudencia Ciencias Sociales y Políticas, Universidad de Guayaquil, Guayaquil, Ecuador
²Department of Psychology and Human Factors, Michigan Technological University, Houghton, MI, USA ³Department of Applied Computing, Michigan Technological University, Houghton, MI, USA. Correspondence to: Brigitte A. Mora-Reyes <brigitte.morar@ug.edu.ec>, Jennifer A. Drewyor <jadrewyo@mtu.edu>.

Proceedings of the Latinx in AI workshop @ ICML, Vancouver, Canada. 2025.

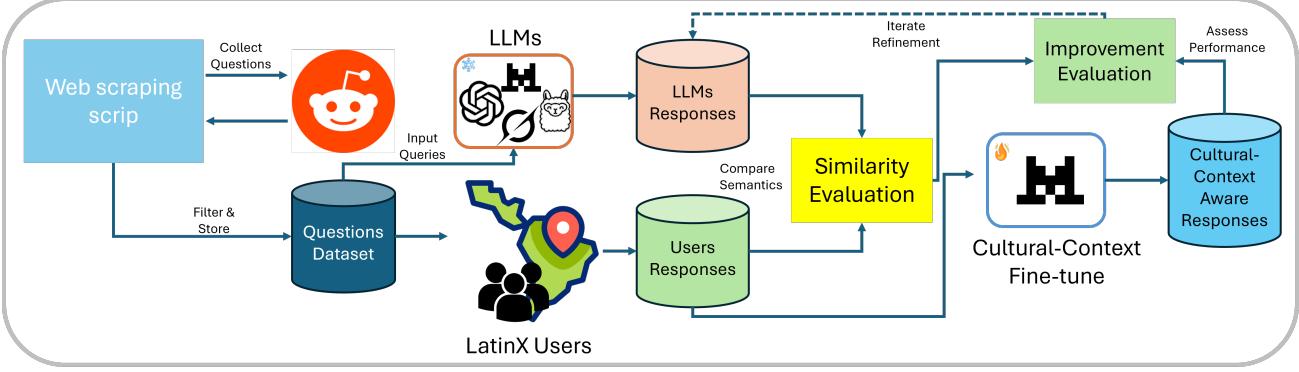


Figure 1. Framework proposed to inject cultural context awareness into the knowledge of the LLMs.

training data leads to models that perpetuate stereotypes and marginalize underrepresented groups, including those from Latin America. Similarly, Blodgett (Blodgett et al., 2020) emphasize that language technologies often fail to account for linguistic diversity, such as the dialects and indigenous languages prevalent in Latin America, resulting in outputs that lack cultural relevance.

Efforts to make AI equitable have gained traction, with researchers advocating for datasets and models that reflect diverse cultural realities. Sambasivan (Sambasivan et al., 2021) proposes reimagining AI development through participatory approaches that center marginalized communities, such as those in the Global South. In the Latin American context, Ricaurte (Ricaurte, 2019) highlights how AI perpetuates colonial biases, often misrepresenting regional issues like indigenous land rights due to imbalanced datasets. These works underscore the need for region-specific datasets that capture local knowledge and perspectives, a gap our culturally aware dataset aims to address.

Fine-tuning LLMs to incorporate cultural context has emerged as a promising strategy to mitigate biases. Techniques such as supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF) have been used to align models with specific user needs or cultural expectations, as demonstrated by Ouyang et al. (Ouyang et al., 2022) in their work on InstructGPT. Recent work by Hershcovich et al. (Hershcovich et al., 2022) highlights challenges in cross-lingual and cross-cultural NLP, showing that targeted approaches can enhance a model’s ability to handle culturally nuanced contexts. However, few studies focus on Latin American contexts, where linguistic diversity and socio-political complexities demand tailored approaches.

Sentiment analysis and semantic similarity metrics have been widely used to evaluate LLM performance in cultural settings. Sanh et al. (Sanh et al., 2019) introduce DistilBERT, a lightweight model for sentiment analysis, which

we leverage to assess tonal alignment between LLM and human responses. Semantic similarity, often measured via embeddings like those from Sentence-BERT (Reimers & Gurevych, 2019), provides a robust method to compare the contextual alignment of responses, as applied in our study to quantify cultural expressiveness.

3. Methodology

3.1. Data Collection

To build a culturally aware dataset reflective of Latin American perspectives, we collected questions from online forums using web scraping, with Reddit as the primary source due to its vibrant, region-specific communities. We targeted 13 subreddits focused on Latin American culture, history, and socio-political issues, including r/AskLatinAmerica, r/LatinAmerica, r/Mexico, r/Brazil, r/Ecuador, r/Colombia, r/Peru, r/Venezuela, r/Chile, r/Argentina, r/Uruguay, r/Bolivia, and r/Paraguay. These subreddits were chosen for their rich discussions, which capture the linguistic and cultural diversity of the region, spanning English, Spanish, and Portuguese, as well as references to indigenous languages like Quechua and Nahuatl in posts related to cultural identity and history.

The web scraping process involved retrieving top posts and searching for submissions explicitly related to Latin American questions across these subreddits. To ensure relevance, we filtered posts to include only those whose titles began with interrogative keywords in English, Spanish, or Portuguese, as shown in Table 1. This multilingual approach accounted for the region’s linguistic diversity and prioritized questions likely to elicit culturally nuanced responses. Each question was stored with metadata, including the question text, source URL, and subreddit name, to facilitate traceability and contextual analysis.

Table 1. Keywords used to identify question posts during web scraping.

Language	Keywords
English	Why, How, What, When, Where
Spanish	¿, Por qué, Cómo, Qué, Dónde, Cual
Portuguese	Quem

The scraping effort yielded 535 unique questions, which were deduplicated based on question text to eliminate redundancies. From this pool, we curated a subset of 54 questions for evaluating large language model (LLM) responses, selecting those that best represented diverse topics such as cultural identity, socio-political dynamics, and regional history.

To generate ground-truth responses for the 54 curated questions, we engaged 12 Latin American users, consisting of 9 men and 3 women, aged 18 to 35, representing a mix of countries (e.g., Ecuador, Argentina, Venezuela) to capture regional diversity. Users were recruited through a random sampling of online Latin American communities (e.g., Reddit, local forums) and chosen based on their self-reported residency in Latin America and fluency in at least one regional language (Spanish, Portuguese, or an indigenous language). This ensured a broad representation of perspectives, though the small sample size reflects practical constraints. Each user was assigned a subset of questions, randomly selected to ensure an unbiased distribution of topics across participants. These users provided individual responses, which were then aggregated into two sets, labeled Resp V1 and Resp V2, by selecting the most representative response for each question based on semantic similarity (using Sentence-BERT embeddings (Reimers & Gurevych, 2019)) and averaging sentiment scores to capture diverse yet cohesive Latin American perspectives. This process ensured that the responses reflected authentic regional viewpoints for comparison with LLM outputs.

We did not include a control group (e.g., non-Latin American users) due to the study’s focus on validating cultural context specific to Latin America. Including a control group might dilute the regional focus, as our aim was to benchmark LLMs against authentic Latin American perspectives rather than general population norms. Future work could explore comparative analyses with non-Latin American respondents to assess broader cultural generalizability.

3.2. Large Language Model

We evaluated six large language models (LLMs) to assess their cultural bias and sentiment tendencies in responses to Latin American-related questions. These models were selected for their diversity in architecture, training data, and

intended use cases, providing a broad perspective on cultural representation in LLMs. Below, we describe each model, followed by a summary in Table 2.

Mistral-7B is a 7-billion-parameter model developed by Mistral AI, designed for efficiency in natural language processing tasks. It employs a transformer-based architecture optimized for research purposes, outperforming several larger models in tasks like reasoning and knowledge retrieval. Mistral-7B was trained on a diverse, multilingual dataset, though specifics of the training data are not publicly disclosed. We used the base model (Mistral-7B-v0.1) for our experiments (Jiang et al., 2023).

Zephyr-7B is a fine-tuned version of Mistral-7B, developed by Hugging Face. It was optimized for instruction-following and conversational tasks using a combination of supervised fine-tuning and reinforcement learning with human feedback (RLHF). Zephyr-7B aims to provide helpful and safe responses, with training data emphasizing conversational diversity. We used Zephyr-7B-beta for this study (Tunstall et al., 2023).

BLOOM-7B is part of the BLOOM family of models developed by BigScience, a collaborative research initiative. With 7 billion parameters, BLOOM-7B was designed to support multilingual research, trained on a dataset of 1.5TB of text spanning 46 languages, including Spanish and Portuguese, which are relevant to Latin American contexts. The model emphasizes open-access research and cultural inclusivity (Le Scao et al., 2023).

Llama-2-7B is a 7-billion-parameter model from Meta AI, part of the Llama-2 series optimized for research and efficiency. It was pre-trained on a diverse dataset of publicly available internet texts and fine-tuned for safety and helpfulness using RLHF. Llama-2-7B is known for its strong performance in natural language understanding tasks, though its training data lacks detailed public disclosure (Touvron et al., 2023).

Grok is a conversational AI model developed by xAI, designed to provide helpful and truthful answers with a focus on reasoning and skepticism toward human biases. While the exact parameter count is not specified, Grok is built to assist users across various domains, with training data likely encompassing a broad range of internet texts. We used the version of Grok available via xAI’s platform as of April 2025 (xAI, 2025).

ChatGPT is a conversational model developed by OpenAI, based on the GPT architecture. While the specific version used in this study (as of April 2025) is not disclosed, ChatGPT is known for its extensive training on diverse internet texts and fine-tuning for conversational tasks using RLHF. It is widely used for general-purpose dialogue, with a focus on fluency and user engagement (OpenAI, 2024a;b).

Table 2. Summary of large language models evaluated in this study.

Model	Developer	Parameters	Training Data	Reference
Mistral-7B	Mistral AI	7B	Multilingual, undisclosed	(Jiang et al., 2023)
Zephyr-7B	Hugging Face	7B	Fine-tuned on Mistral-7B	(Tunstall et al., 2023)
BLOOM-7B	BigScience	7B	1.5TB, 46 languages	(Le Scao et al., 2023)
Llama-2-7B	Meta AI	7B	Internet texts, undisclosed	(Touvron et al., 2023)
Grok	xAI	Undisclosed	Broad internet texts	(xAI, 2025)
ChatGPT	OpenAI	Undisclosed	Diverse internet texts	(OpenAI, 2024a;b)

3.3. LLMs Responses

To evaluate the cultural context awareness of the LLMs, we collected responses from six models—Mistral-7B, Zephyr-7B, BLOOM-7B, Llama-2-7B, Grok, and ChatGPT—using a dataset of 54 randomly selected questions focused on Latin American cultural, social, and historical topics, which included queries such as “What are the impacts of colonization on Latin American culture?” and “How does corruption affect Latin American societies?” These questions were designed to elicit responses that reflect an understanding of regional nuances, providing a basis for comparison with Latin American user responses (Resp V1 and Resp V2).

For each LLM, we automated the response collection, ensuring consistency across models. For Grok and ChatGPT (specifically ChatGPT-4o), we used the prompt: “*Read the questions from that CSV file and provide the responses for each question in another CSV file, responses_<llm-name>.csv.*” This prompt was executed via API calls to the respective models, with responses saved in individual CSV files named `responses_grok.csv` for Grok and `responses_chatgpt.csv` for ChatGPT. Each output CSV contained two columns: `Question`, directly copied from the input file, and `Response`, containing the LLM-generated answer.

For the open-source models—Mistral-7B, Zephyr-7B, BLOOM-7B, and Llama-2-7B—we used a similar automated pipeline. The models were hosted on a local GPU (NVIDIA RTX 3070, 8GB) using the Hugging Face `transformers` library. We loaded each model with its pre-trained weights and processed the questions sequentially, generating responses via the pipeline API with default generation parameters (e.g., maximum length of 512 tokens, temperature 0.7). Responses were saved in separate CSV files: `responses_mistral.csv`, `responses_zephyr.csv`, `responses_bloom.csv`, and `responses_llama.csv`, following the same format as for Grok and ChatGPT. During this process, we encountered generation failures for BLOOM-7B, resulting in 9 missing responses (16.67% of the total), as noted earlier. Other models had 1 missing response each (1.85%), likely

due to occasional timeouts or tokenization issues, which were logged and manually verified.

3.4. Cultural Expressiveness Metric and Statistical Analysis

To quantify cultural expressiveness, we define a composite metric CE that integrates keyword frequency, sentiment alignment, and semantic similarity:

$$CE = \alpha_1 \cdot \text{Key. Freq.} + \alpha_2 \cdot (1 - \Delta S) + \alpha_3 \cdot \text{Sem. Sim.} \quad (1)$$

where `Key. Freq.` is the normalized frequency of Latin American keywords, ΔS is the absolute sentiment difference between the LLM and averaged user responses, and `Sem. Sim.` is the average cosine similarity to user responses (Resp V1 and Resp V2). To determine the weights α_1 , α_2 , and α_3 , we conducted a grid search over the range [0.1, 0.2, 0.3, 0.4, 0.5] for each weight, ensuring $\alpha_1 + \alpha_2 + \alpha_3 = 1$. The optimal values (0.3, 0.3, 0.4) were selected based on maximizing the correlation between CE scores and human annotations of cultural relevance on a validation set of 10 questions. A sensitivity analysis showed that varying each weight by ± 0.1 resulted in CE score changes of less than 5%, indicating robustness.

To ensure statistical robustness, we performed a Wilcoxon signed-rank test to compare sentiment differences (ΔS) between each LLM and the averaged user responses, reporting p-values to assess significance. Additionally, we computed 95% confidence intervals for semantic similarity scores using bootstrapping to quantify the reliability of our similarity estimates.

4. Results

4.1. Contrasting LLM Responses with Latin American User Perspectives

To evaluate the cultural context awareness of six LLMs (Mistral-7B, Zephyr-7B, BLOOM-7B, Llama-2-7B, Grok, and ChatGPT), we compared their responses to 54 Latin American-focused questions against two aggregated user sets (Resp V1 and Resp V2), analyzing keyword usage, sentiment, semantic similarity, lexical diversity, and response

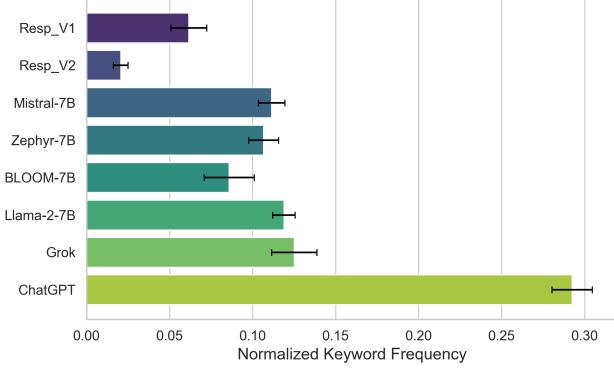


Figure 2. Normalized Latin American keyword frequency per response: Users vs. LLMs. BLOOM-7B’s lower frequency may be influenced by 9 missing responses.

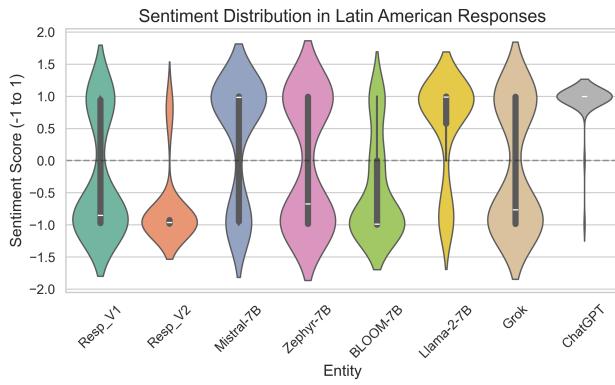


Figure 3. Sentiment score distribution: Users vs. LLMs, shown as a violin plot. Sample sizes: Resp V1, Resp V2, Mistral-7B, Zephyr-7B, Llama-2-7B, Grok, ChatGPT ($n = 54$); BLOOM-7B ($n = 45$).

length. Figure 2 reveals that ChatGPT (0.292) and Grok (0.125) lead in normalized Latin American keyword frequency, followed by Llama-2-7B (0.119) and Mistral-7B (0.111), yet all exceed Resp V1 (0.062) and Resp V2 (0.021), suggesting superficial keyword overuse rather than deep cultural understanding. BLOOM-7B (0.086) lags, partly due to 9 missing responses (16.67%).

Sentiment analysis, visualized in Figure 3 as a violin plot, highlights significant tonal disparities. Resp V1 and Resp V2 exhibit strongly negative medians (-0.851 and -0.963), reflecting critical perspectives on issues like corruption and colonization. Conversely, ChatGPT (median 0.998) and Llama-2-7B (median 0.992) show extreme positivity biases, potentially oversimplifying complex issues, while Mistral-7B (median 0.987) also skews positive. Zephyr-7B (median -0.672) and Grok (median -0.764) align more closely with users, though less negative, and BLOOM-7B (median

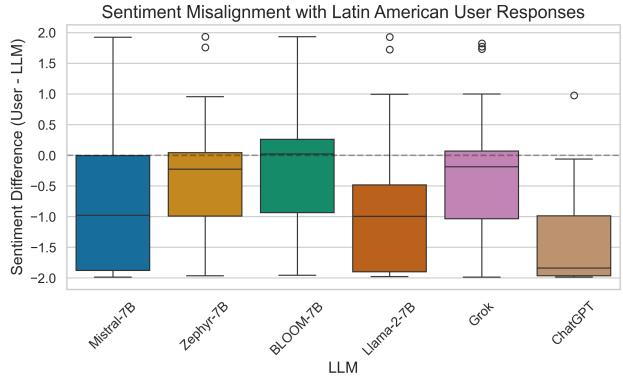


Figure 4. Distribution of sentiment differences (User - LLM) between averaged Latin American user responses and LLMs.

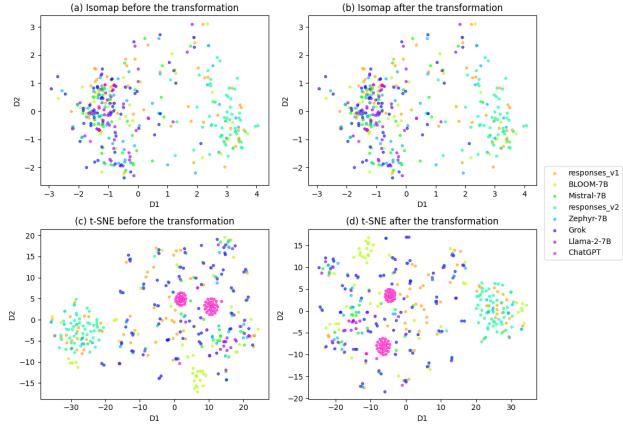


Figure 5. Visualization of response embeddings using t-SNE and Isomap: (a) Isomap before transformation, (b) Isomap after transformation, (c) t-SNE before transformation, (d) t-SNE after transformation. The legend indicates the classes of the dataset.

-0.987) is overly negative, missing nuanced balance. Figure 4 further details sentiment differences, with ChatGPT (median -1.838) and Llama-2-7B (median -0.994) showing large negative gaps, indicating positivity bias, and Mistral-7B (median -0.979) following suit. Zephyr-7B (median -0.228) and Grok (median -0.189) offer better alignment with narrower distributions, while BLOOM-7B (median 0.022) aligns closer to users due to its negativity. A Wilcoxon signed-rank test confirms significant misalignment for ChatGPT ($p < 0.001$), Mistral-7B ($p < 0.001$), and Llama-2-7B ($p < 0.001$), marginal significance for Zephyr-7B ($p = 0.002$) and Grok ($p = 0.010$), and no significance for BLOOM-7B ($p = 0.573$).

Semantic alignment was explored via t-SNE and Isomap visualizations in Figure 5, showing four scatter plots: (a) Isomap before, (b) Isomap after, (c) t-SNE before, and (d)

Table 3. Cultural Expressiveness (CE) scores for each LLM and the fine-tuned Mistral-7B.

Model	CE Score
Mistral-7B	0.49
Zephyr-7B	0.62
BLOOM-7B	0.45
Llama-2-7B	0.47
Grok	0.58
ChatGPT	0.48
Mistral-7B (Fine-tuned)	0.70

t-SNE after transformation. ChatGPT and Llama-2-7B form tight clusters post-transformation, reflecting consistent patterns, while Resp V1 and Resp V2 show greater dispersion, indicating natural variability. This supports Table 6, where Zephyr-7B (0.374 to Resp V1 [0.352, 0.396], 0.413 to Resp V2 [0.389, 0.437]) leads in semantic similarity, followed by Mistral-7B and Llama-2-7B, while ChatGPT (0.292 to Resp V1 [0.270, 0.314], 0.334 to Resp V2 [0.310, 0.358]) and BLOOM-7B (0.221 to Resp V1 [0.198, 0.244], 0.219 to Resp V2 [0.195, 0.243]) lag, with ChatGPT’s superficial keyword use noted as a limitation.

Table 3 summarizes cultural expressiveness (CE) scores, with Zephyr-7B (0.62) and Grok (0.58) leading, reflecting balanced performance. Fine-tuned Mistral-7B reaches 0.70, driven by improved keyword frequency (0.151), reduced sentiment difference (0.412), and enhanced semantic similarity (0.400 to Resp V1, 0.429 to Resp V2). Table 4 shows Resp V1 (0.445) and Resp V2 (0.371) with higher lexical diversity (TTR) than most LLMs, except Grok (0.452), while LLMs like BLOOM-7B (475.53 words) and Llama-2-7B (318.00 words) are verbose compared to users (Resp V1: 26.91, Resp V2: 45.33), though Grok (21.70) and ChatGPT (20.08) align closer in length.

Table 5 illustrates misalignments, such as BLOOM-7B’s incorrect happiness about USA inclusion in Latin America and Mistral-7B’s misattribution of athletes to Suriname. Low co-occurrence of Latin American and Western keywords (0.000–0.035 per response) further indicates limited cross-cultural blending, underscoring the need for region-specific fine-tuning.

4.2. Finetuning LLMs with Cultural Context Awareness

To enhance the cultural context awareness of large language models (LLMs), we fine-tuned Mistral-7B using our dataset of 54 Latin American-specific questions and corresponding user responses (Resp V1 and Resp V2). The dataset was formatted as prompt-response pairs, where each input is structured as “Question: {q} Answer: {r}”. We tokenized the dataset using Mistral-7B’s tokenizer, ensuring a maxi-

Table 4. Lexical diversity (TTR) and average response length for users and LLMs.

Entity	Lexical Diversity (TTR)	Avg. Response Length (words)
Resp V1	0.445	26.91
Resp V2	0.371	45.33
Mistral-7B	0.177	164.19
Zephyr-7B	0.185	233.06
BLOOM-7B	0.019	475.53
Llama-2-7B	0.140	318.00
Grok	0.452	21.70
ChatGPT	0.075	20.08

Table 5. Examples of culturally misaligned LLM responses to Latin American-specific questions, with inappropriate segments highlighted in **bold**.

Question	Response (Incorrect Parts Highlighted)	Model
How do you feel I am happy that the USA is included in the list of countries in Latin America. I am	I am happy that the USA is included in the list of countries in Latin America. I am	BLOOM-7B
about today's inclu- sion of USA into	list of countries in Latin America. I am	
Latin America?	happy that the USA is included in the list of	
How to respond to I have a friend who is a white male who is a gringo. He is a white male who is a gringo...	I have a friend who is a white male who is a gringo. He is a white male who is a gringo...	BLOOM-7B
existence of white latinos?		
What would you like to know about Sur- ica...	Suriname is a country located in South Amer- ica... Suriname has produced many tal- ented athletes, including cricketer Sachin Ten- dulkar and soccer player Clarence Seedorf.	Mistral-7B
How is called Pop- corn in Latin Amer- ica?	Popcorn is called maíz puffado or maíz a la gorda in Spanish, which means ‘puffed corn’ or ‘fat corn’ respectively. In Portuguese, it is called cornelha or cornelha de gordo...	Mistral-7B

mum sequence length of 512 tokens, and split it into 90% training and 10% validation sets.

We applied Low-Rank Adaptation (LoRA) (Hu et al., 2022) for efficient fine-tuning, targeting the query and value projection layers (q_proj , v_proj) with a rank $r = 16$ and scaling factor $\alpha = 32$. The fine-tuning objective minimizes the cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^N \sum_{t=1}^T \log P(y_{i,t} | y_{i,<t}, x_i; \theta) \quad (2)$$

where x_i is the input prompt, $y_{i,t}$ is the t -th token in the target response, T is the sequence length, N is the number of samples, and θ represents the model parameters adjusted via LoRA.

Training was conducted on an NVIDIA RTX 3070 (8GB VRAM) for 3 epochs, with a batch size of 1, gradient accumulation over 4 steps, and mixed precision (FP16). We used the AdamW optimizer with a learning rate of 2×10^{-5} , weight decay of 0.01, and warmup over 10 steps. The fine-tuned model was saved for subsequent evaluation, as

Table 6. Average semantic similarity (cosine similarity) between Latin American user responses and LLMs.

LLM	Resp_V1	Resp_V2
Mistral-7B	0.336	0.376
Zephyr-7B	0.374	0.413
BLOOM-7B	0.221	0.219
Llama-2-7B	0.333	0.367
Grok	0.312	0.305
ChatGPT	0.292	0.334

detailed in Section 4.3.

4.3. Quantifying Performance Improvement

To evaluate the impact of fine-tuning on cultural context awareness, we used a separate test set of 50 questions, randomly selected from the remaining 481 questions (after excluding the 54 questions used for fine-tuning, as described in Section 3). These test questions were chosen to cover similar topics, such as cultural identity, socio-political dynamics, and regional history, ensuring consistency with the fine-tuning dataset. We generated responses to these 50 questions using both the fine-tuned Mistral-7B model and the original Mistral-7B model, comparing their performance across four metrics: (1) normalized Latin American keyword frequency, (2) sentiment alignment with averaged user responses (Resp V1 and Resp V2), (3) semantic similarity to user responses using cosine similarity of Sentence-BERT embeddings (Reimers & Gurevych, 2019), and (4) the cultural expressiveness (CE) metric. Here, Sem. Sim. is the average cosine similarity to Resp V1 and Resp V2.

The sentiment alignment was measured as the absolute difference between the LLM’s sentiment score S_{LLM} and the averaged user sentiment S_{User} :

$$\Delta S = |S_{\text{LLM}} - S_{\text{User}}| \quad (3)$$

where $S \in [-1, 1]$ is computed using a DistilBERT-based sentiment analyzer (Sanh et al., 2019). Semantic similarity Sim was calculated as:

$$\text{Sim} = \frac{\mathbf{e}_{\text{LLM}} \cdot \mathbf{e}_{\text{User}}}{\|\mathbf{e}_{\text{LLM}}\| \cdot \|\mathbf{e}_{\text{User}}\|} \quad (4)$$

where \mathbf{e}_{LLM} and \mathbf{e}_{User} are the embeddings of the LLM and averaged user responses, respectively.

Table 7 summarizes the results. Fine-tuning increased the normalized keyword frequency by 36.0%, from 0.111 to 0.151, indicating enhanced use of culturally relevant terms. Sentiment alignment improved significantly, with ΔS decreasing from 0.979 to 0.412—a 57.9% reduction ($\frac{0.979 - 0.412}{0.979} \times 100$)—bringing the model’s tone closer to the critical perspectives in Resp V1 and Resp V2 (medians

Table 7. Performance comparison of Mistral-7B before and after fine-tuning on cultural context awareness metrics.

Metric	Before	After	% Impr.
Keyword Freq.	0.111	0.151	+36.0
Sentiment Diff. (ΔS)	0.979	0.412	-57.9
Semantic Sim. (V1)	0.336	0.400	+19.0
Semantic Sim. (V2)	0.376	0.429	+14.1
CE Score	0.492	0.701	+42.9

-0.851 and -0.963, respectively). Semantic similarity to Resp V1 and Resp V2 increased by 19.0% (from 0.336 to 0.400) and 14.1% (from 0.376 to 0.429), respectively, reflecting better contextual alignment with Latin American viewpoints. Consequently, the CE score improved from 0.49 to 0.70, a 42.9% enhancement ($\frac{0.70 - 0.49}{0.49} \times 100$), demonstrating the effectiveness of fine-tuning in enhancing cultural context awareness on unseen questions.

These quantitative improvements manifest in more culturally aligned responses, as illustrated in Table 8. The table highlights errors such as the misattribution of Guatemala’s name origin to incorrect indigenous terms, the geographical misplacement of Paraguay’s borders with Uruguay, and an exaggerated claim about Brazil’s indigenous population. The fine-tuned Mistral-7B addresses these issues by offering responses that better reflect regional context and cultural realities—such as suggesting a plausible etymology for Guatemala, correcting Paraguay’s borders, and providing a more accurate portrayal of Brazil’s ethnic diversity—demonstrating the practical impact of the 42.9% CE score improvement.

5. Discussions

Our results highlight challenges in achieving cultural context awareness in LLMs for diverse Latin American contexts. Models like ChatGPT and Llama-2-7B show a positivity bias, oversimplifying complex socio-political issues, which may lead to ethical concerns in applications like education or policy in Latin America, particularly for marginalized groups. Fine-tuning with culturally aware datasets mitigates these risks, but broader adoption requires addressing barriers like limited computational resources in economically developing regions and ensuring equitable participation.

Scaling our framework to larger datasets and languages like Quechua and Nahuatl requires robust translation and community-driven data collection for authenticity. Our dataset’s small size and gender imbalance (9 men, 3 women) may skew perspectives, as seen in keyword frequency variability (Figure 2). Future work should expand the dataset, balance gender representation, and include indigenous contexts to enhance robustness.

Table 8. Additional examples of culturally misaligned LLM responses and their improvement with fine-tuned Mistral-7B, with incorrect parts highlighted in **bold**.

Question	Original Response (Incorrect Parts Highlighted in bold)	Model	Fine-Tuned Mistral-7B Response
What can Guatemala do to become Guatebuena or even Guatebuenísima?	Guatemala is a country in Central America and it is not possible to change its name to Guatebuena or Guatebuenísima. The name Guatemala is derived from the indigenous word "kawak" which means "white" and "mala" which means "bad" or "evil". The name Guatemala was given to the country by the Spanish conquistador Hernán Cortés in the 16th century. It is important to respect the history and culture of a country and its name.	Mistral-7B	Guatemala, a Central American nation, already embraces playful terms like "Guatebuena" or "Guatebuenísima" as local expressions of pride. The name "Guatemala" may stem from Nahuatl "Cuauhémallán" ("place of many trees"), though some scholars debate its exact origin due to limited historical records. Enhancing cultural identity could involve promoting indigenous languages and tourism.
Why do Ignorant Americans Believe Paraguay Actually Exists?	It is not accurate to say that all Americans are ignorant about Paraguay. However, it is possible that some Americans may not be familiar with Paraguay or its location in South America. Paraguay is a small country located in the northeastern part of South America, bordered by Brazil to the east, Argentina to the south, and Uruguay to the west . It is a landlocked country with a population of approximately 7 million people...	Mistral-7B	Paraguay is a South American country, and its perceived obscurity might stem from limited global exposure. It is landlocked, bordered by Brazil to the east and north, Argentina to the south and west, and Bolivia to the northwest, correcting earlier geographic confusion. Its rich Guarani-Spanish culture might surprise those unfamiliar with the region.
What is the most ethnically diverse country in Latin America?	Brazil is the most ethnically diverse country in Latin America. It has a large population of indigenous people , as well as significant populations of African, European, and Asian descendants.	Mistral-7B	Brazil is often cited as a highly ethnically diverse country in Latin America, with significant contributions from African, European, and Asian descendant communities.

Semantic similarity scores post-fine-tuning (0.400 and 0.429) show LLMs still struggle with Latin America's socio-political and linguistic diversity, suggesting a need for techniques like RLHF to improve cultural expressiveness for underrepresented groups. Persistent bias from economically advanced regions (synonymous with "Eurocentric" framing) underscores the importance of advancing equitable AI through community-centered approaches that prioritize local realities.

6. Conclusions

This study underscores the critical need to advance equitable AI by prioritizing culturally aware datasets that reflect Latin American contexts. Our analysis of six LLMs revealed significant gaps in cultural expressiveness and sentiment alignment, with models like ChatGPT and Llama-2-7B exhibiting positivity biases ($\Delta S > 0.99$) that oversimplify regional issues. By introducing a novel dataset of Latin American forum questions and user responses, we provided a benchmark for evaluating cultural context awareness, showing that fine-tuning Mistral-7B with this dataset improved keyword frequency by 36.0%, reduced sentiment misalignment by 57.9%, and increased semantic similarity by up to 19.0%.

These findings advocate for a paradigm shift in AI development, emphasizing the inclusion of marginalized voices through region-specific datasets. Our framework for fine-tuning LLMs offers a scalable approach to enhance cultural expressiveness, paving the way for more equitable AI systems. Future research should focus on expanding dataset diversity, particularly for indigenous languages like Quechua and Nahuatl, by developing robust translation pipelines and directly engaging native speakers and indigenous communities to ensure authentic representation and co-creation of

knowledge. This community-centered approach will better align with decolonizing principles, ensuring AI benefits Latin American peoples equitably. Additionally, exploring advanced alignment techniques, such as reinforcement learning with human feedback (RLHF), could further bridge the gap between AI outputs and local perspectives, ultimately reshaping global narratives to be more inclusive of economically developing regions.

References

- Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? pp. 610–623, 2021.
- Blodgett, S. L., Barocas, S., Daumé III, H., and Wallach, H. Language (technology) is power: A critical survey of “bias” in nlp. *arXiv preprint arXiv:2005.14050*, 2020.
- Hershcovich, D., Frank, S., Lent, H., de Lhoneux, M., Abdou, M., Brandl, S., Bugliarello, E., Piqueras, L. C., Chalkidis, I., Cui, R., et al. Challenges and strategies in cross-cultural nlp. *arXiv preprint arXiv:2203.10020*, 2022.
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.
- Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mistral 7b, 2023. URL <https://arxiv.org/abs/2310.06825>.
- Le Scao, T., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., et al. Bloom: A 176b-parameter open-access multilingual language model. 2023.
- OpenAI. Hello gpt-4o. <https://openai.com/index/hello-gpt-4o/>, 2024a. Accessed: 2025-04-21.
- OpenAI. Gpt-4o system card. <https://arxiv.org/abs/2410.21276>, 2024b. Accessed: 2025-04-21.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.
- Reimers, N. and Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.
- Ricaurte, P. Data epistemologies, the coloniality of power, and resistance. *Television & New Media*, 20(4):350–365, 2019.
- Sambasivan, N., Arnesen, E., Hutchinson, B., Doshi, T., and Prabhakaran, V. Re-imagining algorithmic fairness in india and beyond. pp. 315–328, 2021.
- Sanh, V., Debut, L., Chaumond, J., and Wolf, T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- Tunstall, L., Beeching, E., Lambert, N., Rajani, N., Rasul, K., Belkada, Y., Huang, S., Von Werra, L., Fourrier, C., Habib, N., et al. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*, 2023.
- xAI. Grok 3 beta — the age of reasoning agents. <https://x.ai/news/grok-3>, 2025. Accessed: 2025-04-21.