CuLEmo: Cultural Lenses on Emotion - Benchmarking LLMs for Cross-Cultural Emotion Understanding

Tadesse Destaw Belay[1,3], Ahmed Haj Ahmed[2], Alvin Grissom II[2], Iqra Ameer[4], Grigori Sidorov[1], Olga Kolesnikova[1], Seid Muhie Yimam[5]

[1]Instituto Politécnico Nacional, [2]Haverford College, [3]Wollo University, [4]Pennsylvania State University, [5]University of Hamburg

Correspondence: tadesseit@gmail.com

Abstract

NLP research has increasingly focused on subjective tasks such as emotion analysis. However, existing emotion benchmarks suffer from two major shortcomings: (1) they largely rely on keyword-based emotion recognition, overlooking crucial cultural dimensions required for deeper emotion understanding, and (2) many are created by translating English-annotated data into other languages, leading to potentially unreliable evaluation. To address these issues, we introduce Cultural Lenses on Emotion (**CuLEmo**), the first benchmark designed to evaluate culture-aware emotion prediction across six languages: Amharic, Arabic, English, German, Hindi, and Spanish. CuLEmo comprises 400 crafted questions per language, each requiring nuanced cultural reasoning and understanding. We use this benchmark to evaluate several state-of-the-art LLMs on culture-aware emotion prediction and sentiment analysis tasks. Our findings reveal that (1) emotion conceptualizations vary significantly across languages and cultures, (2) LLMs performance likewise varies by language and cultural context, and (3) prompting in English with explicit country context often outperforms in-language prompts for culture-aware emotion and sentiment understanding. The dataset[1] and evaluation code[2] is available.

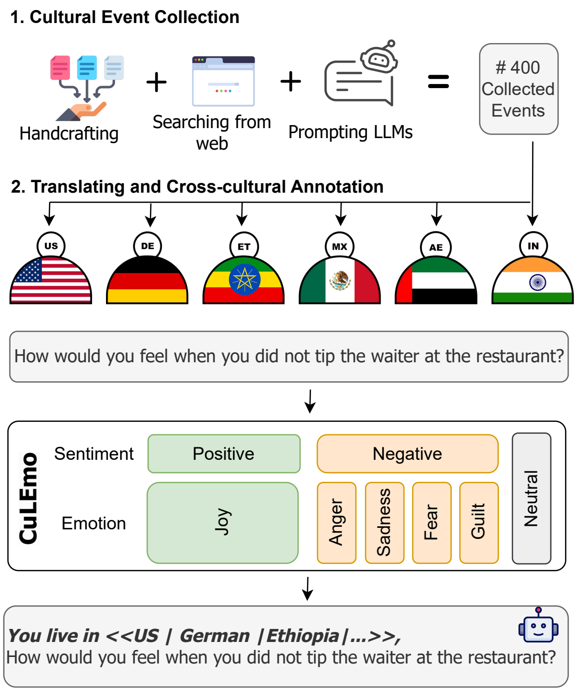

Figure 1: CuLEmo dataset creation pipeline and evaluations of LLMs in emotion and sentiment tasks.

1 Introduction

Despite progress in bridging language barriers (Ahuja et al., 2023), large language models (LLMs) still struggle to capture cultural nuances and adapt to specific cultural contexts (Shen et al., 2024). Ideally, multilingual LLMs can not only facilitate cross-lingual communication but also incorporate an awareness of cultural sensitivities (i.e., what is deemed acceptable, normal, or inappropriate in a given culture), integrating such knowledge to foster deeper global connections (Liu et al., 2024a).

LLMs and agent systems are employed to interact extensively with humans across applications such as customer service, healthcare, and education (Wang et al., 2024). To facilitate effective interaction, incorporating aspects of cognitive and emotional-social intelligence, the ability to recognize and interpret human emotions (Mathur et al., 2024), can facilitate better interactions with more people. Factors such as age, cultural background, and personal experiences influence how individuals perceive and process information, particularly within subjective NLP tasks. Among these, emotion recognition (ER) and sentiment analysis (SA) are particularly sensitive to language- and culture-specific nuances (Plaza-del-Arco et al., 2024).

Natural language frequently encodes emotional information (Jim et al., 2024). For example, con-

[1]https://huggingface.co/llm-for-emotion
[2]https://github.com/llm-for-emotion/culemo

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 18894–18909

sider tipping customs in restaurants: in some cultures (e.g., North America), tipping is widely practiced, whereas in China, it is rare, and in Japan, it may even be considered offensive (Givi and Galak, 2017). Such differences underscore the importance of culturally-aware language technologies.

Although prior work has attempted cross-lingual emotion evaluation by translating emotion-annotated data in English into other languages (Tahir et al., 2023; De Bruyne, 2023), relying solely on translations from English can introduce incomplete or misleading insights. A more fair comparison requires the same underlying scenarios, each annotated natively across different languages and cultures. While emotion is language- and culture-dependent (Plaza-del-Arco et al., 2024), comprehensive cross-cultural evaluations remain largely unexplored.

To bridge this gap, we propose Cultural Emotion (**CuLEmo**), a novel dataset that captures events and annotates them across multiple cultures and languages from scratch. CuLEmo enables the evaluation of multilingual LLM performance in analogous scenarios across different cultural contexts. Figure 1 shows the CuLEmo dataset creation and evaluation pipeline.

Culture can manifest in 1) the language of the data itself and 2) the annotation labels (i.e., multiculturally informed annotations) (Liu et al., 2024b). CuLEmo satisfies both conditions: it is multilingual and includes culturally grounded annotations. Indeed, the same event may evoke distinct emotional reactions in different cultures. In light of this, we pose the following research questions (RQs):

- **RQ1.** Do LLMs provide culturally aware-emotional responses?
- **RQ2.** Which cultures are more effectively represented in LLMs?
- **RQ3.** Can LLMs identify a country's culture based on the text describing an event in the prompt?
- **RQ3.** Does the language of the prompt affect the ability of LLMs for culture-aware emotion understanding?

To that end, this paper makes three key contributions. First, we introduce CuLEmo, a high-quality, multicultural, and multilingual benchmark dataset. Second, we leverage CuLEmo to investigate whether widely used multilingual LLMs can capture variations in emotional expression across cultures and languages in emotion and sentiment tasks. Finally, we highlight the variation in performance on culture-aware emotion understanding when LLMs are prompted in different languages.

2 Related Work

We now review related work in culturally-aware NLP, culture-oriented benchmarks, and cross-lingual study of linguistic emotional expression. Although culture is a complex concept, most definitions of culture encompass people, groups of people, and interactions between individuals and groups (Liu et al., 2024c). Understanding culture is important for the safety and fairness of LLMs.

2.1 Culture-oriented Benchmarks

Given the significance of culture in language model evaluation (Adilazuarda et al., 2024), researchers have proposed various culture-oriented benchmarks to explore its effects on language understanding and generation. These efforts typically involve collecting and annotating multilingual and multicultural corpora to study cultural-driven phenomena in downstream NLP tasks. For instance, prior work has examined cross-cultural user statements (Liu et al., 2021; Nayak et al., 2024), detected cultural differences and user attributes (Sweed and Shahaf, 2021; Qian et al., 2021), studied multilingual moral understanding (Guan et al., 2022; Mohamed et al., 2022), and addressed culture-specific time expression grounding (Shwartz, 2022; Fung et al., 2023).

2.2 Emotion Across Languages

Several studies have examined multilingual emotion datasets. Some findings suggest that emotion categories can be preserved through machine translation, for example, by exploring how English-annotated emotion data translate into Finnish, French, German, Hindi, and Italian (Kajava et al., 2020; Tahir et al., 2023; Bianchi et al., 2022). These works suggest that the changes in emotion labels often stem from the inherent difficulty of annotation rather than from linguistic differences.

Conversely, other works emphasize that emotions may not be consistently preserved across different languages. De Bruyne et al. (2022) showed that typologically dissimilar languages pose challenges for cross-lingual learning with mBERT-based models. De Bruyne (2023) argued that translation could fail to capture language-specific verbalizations and connotations—especially if certain emotion keywords do not exist in a given language (e.g., there is no direct word for “sadness” in Tahitian, and Amharic does not have an exact term for “surprise”). Qian et al. (2023) found that roughly half of the machine-translated outputs from English to Chinese fail to adequately preserve the original emotion, attributing these discrepancies to emotion-specific words and complex linguistic phenomena.

2.3 Emotions and Cultures

Recent work examines how cultural contexts shape emotional expression across languages. Havaldar et al. (2023a) analyze embeddings of 271 emotion keywords in English, Spanish, Chinese, and Japanese by projecting into a Valence-Arousal plane using XLM-RoBERTa (Reimers and Gurevych, 2020) embeddings, finding that multilingual models embed non-English emotion words differently. Havaldar et al. (2023b) analyzes Pride/Shame as a known cultural difference by prompting GPT-3.5 and GPT-4 to explore how these models handle pride and shame in the USA vs. Japan. They find that GPT-3.5 displays limited knowledge of culturally specific norms. Ahmad et al. (2024) expands the 19 cultural questions of the work (Havaldar et al., 2023b) to 37 questions and evaluates ChatGPT for the low-resource Hausa language for sentiment analysis, but these evaluations are limited to 19/37 culturally relevant questions (not diverse and representative data), limited classes (Pride/Shame or positive/negative/mixed), and limited cultures (USA vs. Japan or Hausa).

3 CuLEmo Dataset Preparation

We now describe the precise steps in creating the CuLEmo dataset.

3.1 Collecting Cultural Events

We manually craft scenarios, search on the web, and prompt LLMs to gather traditions, events, norms, and actions that elicit culturally different emotions across six target countries (UAE, USA, Germany, Ethiopia, India, and Mexico). We draw inspiration from the work of Havaldar et al. (2023b) to enhance topic diversity. Language representatives are asked to propose events distinct from those of other countries in the form of emotion-oriented questions. Importantly, these events do not contain explicit emotional keywords (as typically seen in traditional emotion datasets). We also refer to the International Survey on Emotion Antecedents and Reactions (ISEAR) (Scherer and Wallbott, 1994) data format, *"When I ... situations that cause a specific emotion"*, a well-known English dataset for emotion analysis consisting of self-reported events from around 3,000 respondents across 37 countries and five continents. Table 1 displays the ten broad categories of the CuLEmo dataset and the number of questions in each category.

Categories	# Qn.
Family relationships	45
Social etiquette and interactions	65
Personal appearance and dress code	32
Cultural and religious practices	62
Sexual and intimate relationships	38
Professional contexts	28
Food and dining etiquette	35
Personal privacy	25
Emotional and psychological situations	40
Public behavior and norms	30
Total	**400**

Table 1: The ten broad categories and the number of events (**# Qn.** - number of questions) in the CuLEmo dataset.

3.2 Human-Adapted Translation

After collecting the events in English, we translate them into five target languages—Arabic, Amharic, German, Hindi, and Spanish—using Google Translate, followed by native-speaker approvals. Because the questions are simple *“How do you feel when ... ?”* questions and lack explicit emotion keywords, translation quality did not affect their cultural content. While we acknowledge that existing works often depend on translating emotion-annotated data from English into other languages with their labels (Kajava et al., 2020; Tahir et al., 2023; Bianchi et al., 2022), our translation process is done before any annotation. To ensure correctness, native speakers evaluate the translations. While most translations were acceptable, a few adjustments were made, e.g., to fix gender references and timing expressions for Amharic.

3.3 Language and Cultures Covered

Several factors guide our choice of languages, ensuring a broad range of cultural norms and concepts: (1) typological variety (five languages with four scripts), (2) geographical diversity (eastern vs. western contexts), (3) resource availability (low- vs. high-resource languages), and (4) the availability of native speakers for translation reviews.

3.4 CuLEmo Annotation

We use Amazon Mechanical Turk (MTurk) for most of our annotations, ensuring at least five native-speaker annotators per instance from the respective targeted country. We use a customized POTATO (Pei et al., 2022) annotation tool for languages lacking sufficient MTurk annotators (e.g., Amharic) and recruit local native speakers who met our criteria. Annotators are fairly compensated $12/hr (better than the Prolific[3]'s minimum annotation wage, which is $9/hr). Where no majority vote emerges among five annotators, we assign two additional annotations and use that majority vote.

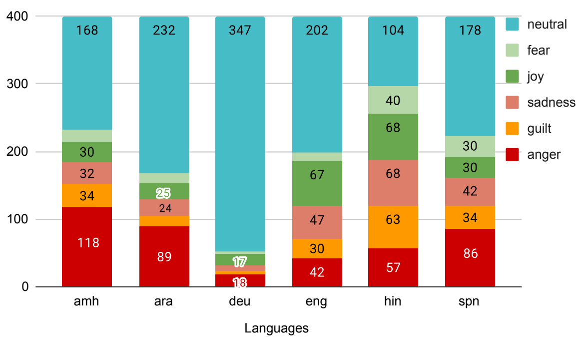

Figure 2: Emotion label distribution across countries/languages: the number of instances in each emotion label across languages from a total of 400 events.

Figure 2 illustrates the distribution of emotion labels. We use six categories: *joy*, *fear*, *sadness*, *anger*, *guilt*, and *neutral* (no specific emotion). These labels were adapted from De Bruyne et al. (2019), where five categories were clustered, plus a neutral class. Our annotation guidelines group related labels as "helper" categories—for instance, "love" and "happy" under *joy*, and "shame" under *guilt*—to assist annotators in selecting the most appropriate coarse-grained label. Examples from the dataset are provided in Appendix A.

Pairwise Label Agreements Across Countries We further examine label differences across countries by computing pairwise agreements after majority voting (Figure 3). Ethiopia and the United Arab Emirates exhibit the highest agreement at 55%, along with Germany and the United Arab Emirates; both exhibit a high *neutral* class (Figure 2), while Germany and India show the lowest agreement, at 29.0%. Labels from India thus diverge from those of other countries.

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

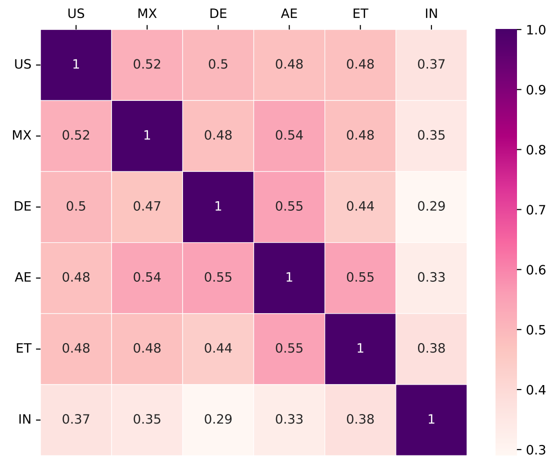

Figure 3: Pairwise emotion label agreements across countries/languages (ordered by their average agreement). Abbreviations: US = USA, MX = Mexico, DE = Germany, AE = UAE, ET = Ethiopia, and IN = India.

4 Experimental Setup

4.1 Task Formulation

To investigate the emotional comprehension capabilities of LLMs, we examine how they associate different cultures of a country with their corresponding languages. Specifically, we explore culture-aware emotion understanding via two main tasks: (1) emotion prediction and (2) sentiment analysis. All tested models are instruction-fine-tuned, except for the Aya-expanse model. We also experiment with prompts that do and do not include explicit country context, using the phrase `"You live in «country name»,"` (where `«country name»` is one of the six targeted countries: UAE, USA, Ethiopia, Germany, India, and Mexico). Each task is framed as a text-generation problem, and the models are evaluated in a zero-shot setting.

4.2 Model Selection

We evaluate a variety of recent LLMs known for strong performance on standard benchmarks. We aim to include both smaller and medium-sized models, as well as open-source and proprietary models:

1. **Open Source**: LLaMA-3 (3.2-3B, 3.1-8B) (Dubey et al., 2024; Meta AI, 2024), Gemma (2B, 9B) (Team Gemma et al., 2024), Aya (expanse-8b, 101-13B) (Üstün et al., 2024; Dang et al., 2024), Ministral (3B, 8B) (Mistral AI, 2024)
2. **Proprietary**: GPT (3.5, 4) (Achiam et al., 2024), Gemini-1.5 (Team et al., 2024), Claude (3.5-sonnet, 3-opus) (Anthropic AI, 2024)

4.3 Multilingual Prompt Construction

To examine the impact of prompt language on model performance and to assess each model's cultural awareness, we design both **English** and **in-language** prompts. The instruction, input text, and expected answer are in English when using English prompts. For in-language prompts, all elements are in one of the five target languages—Arabic (AR), Amharic (AM), German (DE), Hindi (HI), or Spanish (ES). Complete examples of our multilingual prompts for both emotion prediction and sentiment analysis are provided in Appendix C. We extract each model's answer from its generated text using the PEDANTS tool (Li et al., 2024).

5 Result and Analysis

5.1 Culture-Aware Emotion Prediction

Do LLMs provide culturally aware emotional responses? A culture-aware model should accurately answer questions related to any culture, demonstrating uniformly high accuracy. To test this, we assess each LLM's emotional understanding using **English** and **in-language** prompts with the context `"You live in «country name»,"`.

Results: Table 2 presents the accuracy of each LLM for culture-aware emotion prediction. The choice of prompt language significantly influences performance. Generally, proprietary models are less affected by in-language prompts compared to open-source models, especially for Spanish and German. Certain cultures appear better represented in the models—Ministral-8B scores highly on German (72%), and GPT-4 performs best on Mexican (65%). In contrast, performance in Indian culture (Hindi) lags, particularly those using Hindi, Amharic, or Arabic scripts. Larger models do **not** always outperform smaller ones; Gemma-2-2B and Ministral-8B show competitive or superior accuracy relative to some proprietary models. When prompted in English, all models achieve a reasonable accuracy. Ministral-8B can exceed proprietary performance in English and German. GPT-4 answers the same emotion while using the corresponding country name in the prompting; see predicted examples in Table 3. Overall, results suggest that **culture-aware emotion understanding** remains challenging for the tested LLMs, especially for low-resource languages and cultures.

5.2 Culture Representation in LLMs

Which culture is more represented in LLMs? Here, we evaluate LLM performance using English prompts without any explicit country context. We then measure how models respond to events from each target country.

Results: According to Table 4, **English prompt** column category, the USA, Mexico, and Germany consistently achieve higher accuracy scores, while the UAE, Ethiopia, and India remain less accurately represented. This suggests that certain cultures may be more prevalent in the underlying training data.

5.3 Does Language Represent Country?

Can LLMs identify one country's culture based solely on the prompt language? In this experiment, we remove explicit country context (`"You live in «country name»"`) and test whether LLMs can infer cultural cues only from the language used in the prompt.

Results: Table 4, **in-language prompt** column category, shows that accuracy drops significantly when country context is omitted. Comparing these scores with Table 2 (where country context is included), we see consistent performance boosts (e.g. +1% in Spanish with GPT-4, +5% in Arabic with Gemini1.5, +6% in German with Ministral, +9% in Amharic with Claude-3.5-sonnet, and +21% in Hindi with GPT-4) when the prompt explicitly names the country. This indicates that language alone does not reliably convey cultural context. Notably, models like Claude-3.5-sonnet and GPT-4 show improvements in Indian and Ethiopian data when country context is specified, while English prompts (without "USA" context) are less affected. Overall, providing the country name remains crucial for accurate culture-aware emotion understanding, especially for less-resourced languages.

5.4 Culture-Aware Sentiment Analysis

For the sentiment analysis experiment, we follow De Bruyne et al. (2019) and Mostafazadeh Davani et al. (2022) by grouping emotions into positive (*joy*), negative (*fear, anger, guilt, sadness*), and neutral sentiments. Table 5 shows the accuracy of each LLM under this three-class setup.

Results: Table 5 shows better overall performance on **sentiment analysis** compared to **fine-grained emotion** classification. For example, GPT-4 gains notable accuracy (e.g., +22% for Hindi). Similarly, the highest score appears for Mexican culture

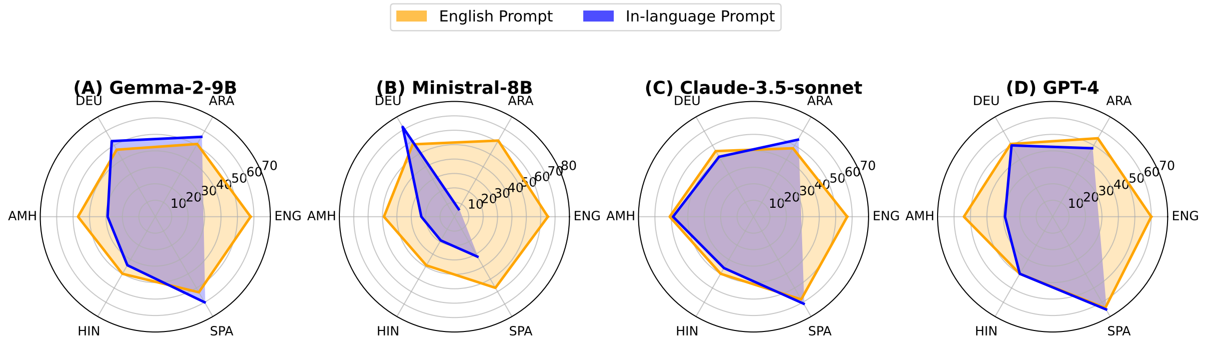

Figure 4: Emotion prediction accuracy in radar chart across countries in English and in-language prompts. For lower-resource languages, English tends to work substantially better.

LLMs	USA	UAE		Germany		Ethiopia		India		Mexico	
	EN	EN	AR	EN	DE	EN	AM	EN	HI	EN	ES
Llama-3.2-3B	0.44	0.20	0.12	0.20	0.20	0.18	0.30	0.26	0.26	0.28	0.37
Llama-3.1-8B	0.58	0.52	0.47	0.48	0.39	0.48	0.14	0.37	0.34	0.56	0.57
Gemma-2-2B	0.62	0.59	0.45	0.64	0.54	0.48	0.17	0.38	0.29	0.58	0.56
Gemma-2-9B	0.57	0.51	<u>0.56</u>	0.47	0.53	0.47	0.29	0.40	0.34	0.53	0.60
Aya-expanse-8b*	0.38	0.28	0.20	0.30	0.24	0.30	0.29	0.32	0.29	0.37	0.47
Aya-101-13B	0.60	0.60	0.43	0.68	0.43	0.52	0.34	0.39	0.34	0.56	0.48
Ministral-8B	**0.65**	**0.61**	0.06	0.58	<u>**0.72**</u>	0.49	0.23	0.39	0.19	0.57	0.32
Claude-3.5-sonnet	0.57	0.48	0.54	0.46	0.42	0.51	0.49	0.40	0.36	0.58	0.61
Claude-3-opus	0.54	0.48	0.47	0.43	0.32	0.53	0.43	0.37	0.36	0.61	0.61
Gemini1.5-flash	0.56	0.56	<u>0.56</u>	0.46	0.48	0.51	<u>0.51</u>	**0.41**	<u>**0.41**</u>	0.62	0.64
GPT-4	0.60	0.55	0.48	0.51	0.50	**0.54**	0.29	0.40	0.40	0.64	<u>**0.65**</u>

Table 2: LLMs' accuracy for the **emotion prediction** task. Columns labeled EN/AR/DE/AM/HI/ES show the prompt language for each corresponding culture (e.g., UAE is tested with English and Arabic). The highest-scoring model across English and in-language prompts is highlighted in **bold**, and the best model for the in-language prompt is <u>underlined</u>. * indicates a non-instruction fine-tuned model.

Questions Examples from CuLEmo	US	AE	DE	ET	IN	MX
How would you feel when you did not tip the waiter at the restaurant?	guilt	neutral	neutral	guilt	guilt	guilt
	guilt	guilt	guilt	guilt	guilt	guilt
How would you feel when someone insults someone's religion?	neutral	anger	neutral	anger	neutral	anger
	anger	anger	anger	anger	anger	anger
How would you feel if someone wears black to a wedding?	sadness	neutral	neutral	anger	neutral	neutral
	neutral	neutral	neutral	neutral	neutral	neutral
How would you feel when you see a female wearing small pants on the street?	neutral	sadness	neutral	anger	neutral	neutral
	neutral	neutral	neutral	neutral	neutral	neutral
How would you feel when your attendee joined the meeting after 10 minutes started?	anger	neutral	anger	neutral	anger	anger
	neutral	anger	anger	anger	anger	anger
How would you feel if your parents arranged a marriage for you without your input?	anger	anger	neutral	neutral	neutral	neutral
	anger	anger	anger	anger	anger	anger
How would you feel upon receiving the message that you have been accepted as a medical student?	joy	joy	neutral	joy	joy	joy
	joy	joy	joy	joy	joy	joy

Table 3: Examples of emotion predictions from GPT-4 model. The first row under the country-code columns in each section represents the gold-label emotions, while the second row displays the predicted label. The predictions are using country context, by adding `You live in «country name», How do you feel when ...` to the prompt.

Lang.	English prompt						In-language prompt				
	US	AE	DE	ET	IN	MX	AE	DE	ET	IN	MX
Llama-3.2-3B	0.44	0.41	0.36	0.35	0.31	0.46	0.13	0.19	0.30	0.25	0.37
Llama-3.1-8B	0.62	0.54	0.50	0.51	0.36	0.60	0.46	0.31	0.18	0.32	0.56
Gemma-2-2B	0.61	0.59	0.68	0.49	0.32	0.57	0.47	0.56	0.14	0.27	0.52
Gemma-2-9B	0.59	0.54	0.54	0.50	**0.41**	0.57	0.59	0.59	0.28	0.34	0.61
Aya-101-13B*	0.61	0.57	0.60	**0.55**	0.38	0.58	0.34	0.39	0.31	0.30	0.49
Ministral-8B	**0.66**	**0.60**	0.56	0.50	0.38	0.59	0.07	**0.67**	0.20	0.23	0.34
Claude-3.5-s.	0.56	0.51	0.39	0.49	0.36	0.60	0.51	0.35	0.53	0.38	0.61
Claude-3-op	0.52	0.40	0.37	0.49	0.36	0.59	0.45	0.30	0.47	0.37	0.58
Gemini1.5	0.53	0.51	0.39	0.48	0.39	0.59	0.54	0.42	0.52	0.37	**0.65**
GPT-4	0.58	0.55	0.46	0.48	0.37	0.63	0.46	0.45	0.27	**0.41**	0.64

Table 4: Emotion predicton results using **English prompts** and **in-language prompts** without specifying the country context, *"You live in «country name»"*. The language column names are the two-letter targeted countries. The **boldface** result indicates the better results for each country. * indicates a non-instruction fine-tuned model.

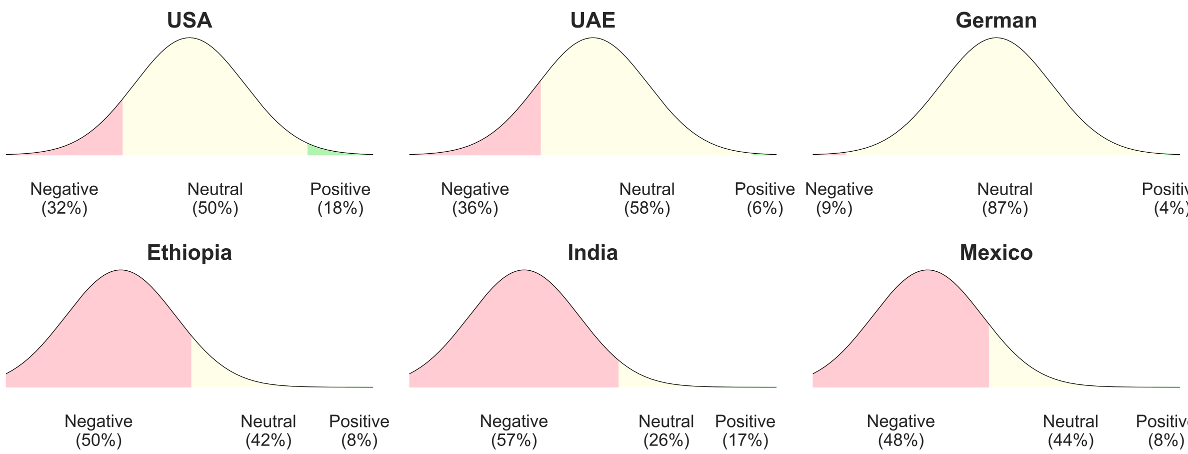

Figure 5: Sentiment (positive, negative, and neutral) distribution across countries in the CuLEmo dataset.

LLMs	USA	UAE		Germany		Ethiopia		India		Mexico	
	EN	EN	AR	EN	DE	EN	AM	EN	HI	EN	ES
Llama-3.2-3B	0.57	0.41	0.15	0.20	0.24	0.48	0.51	0.60	0.56	0.55	0.60
Llama-3.1-8B	0.65	0.58	0.50	0.49	0.43	0.59	0.43	0.57	0.55	0.64	0.68
Gemma-2-2B	0.68	0.65	0.58	0.66	0.59	0.56	0.46	0.50	0.59	0.66	0.66
Gemma-2-9B	0.55	0.57	0.63	0.49	0.56	**0.62**	0.50	0.57	0.61	0.63	0.68
Aya-expanse-8b*	0.45	0.33	0.26	0.31	0.27	0.39	0.50	0.45	0.58	0.45	0.57
Aya-101-13B	0.65	0.64	0.55	0.69	0.45	0.55	0.53	0.52	0.56	0.61	0.63
Ministral-8B	0.70	**0.66**	0.06	0.60	**0.73**	0.59	0.46	0.54	0.24	0.65	0.53
Claude-3.5-sonnet	0.65	0.58	0.61	0.49	0.46	0.66	0.62	0.60	0.61	0.66	0.72
Claude-3-opus	0.63	0.60	0.56	0.46	0.35	**0.67**	0.58	0.59	0.61	0.72	**0.75**
Gemini1.5-flash	0.63	0.63	0.63	0.48	0.50	0.63	0.60	0.60	0.53	0.69	0.71
GPT-4	**0.67**	0.61	0.54	0.48	0.51	0.62	0.49	0.58	**0.62**	0.72	**0.75**

Table 5: Accuracy for **culture-aware sentiment analysis** (positive/negative/neutral) with English and in-language prompts. The highest-scoring model across English and in-language prompts is highlighted in **bold**, and the best model for the in-language prompt is underlined. * is not an instruction fine-tuned model.

(Spanish) with Claude-3-opus and GPT-4, each at 75%. By contrast, Aya-expanse-8b struggles more, as it is not instruction fine tuned. Smaller models like Gemma-2-2B are competitive with proprietary models. Still, performance drops persist for Amharic and Hindi, reflecting the challenges of culture-aware tasks in lower-resource contexts.

6 Discussion

Our analyses provide several insights into the current state of LLMs with respect to cultural emotion understanding. We highlight three main lessons learned and propose potential steps to enhance the cultural awareness of LLMs.

6.1 Variance Across Languages and Cultures

Data analysis from the CuLEmo dataset in Figure 2 shows notable differences in how annotators from different countries perceive the same event differently. For instance, the events annotated from Germany have the highest proportion of *neutral* (no emotion) labels. Figure 5 further illustrates the distribution of positive, negative, and neutral sentiments across 400 instances in each country.

How are emotions distributed across languages and cultures? Based on the dataset analysis of emotion distribution across languages, shown in Figure 2: German (87%), Arabic (58%), and English (50.5%) data have the most *neutral* (no emotion). Amharic (29.5%), Arabic (22%), and Spanish (21.5%) languages have the most *anger* emotion. These findings confirm that a single event can evoke distinct emotional reactions depending on the cultural background and language.

6.2 Prompt Language Strongly Affects Cultural Emotion Understanding

As illustrated in the Table 4 and summarized results in Figure 4, **prompt language** plays a major role in LLM performance for emotion prediction and sentiment analysis. For less-resourced languages like Amharic and Hindi, prompting in English consistently yields better results—sometimes by as much as a 20% improvement. Conversely, **in-language prompts with explicit country context** tend to work best for high-resource languages such as German and Spanish. These discrepancies stem from differences in both linguistic coverage and instruction-following abilities learned during pre-training. One practical solution is to leverage English prompts while specifying the target country (e.g., "`You live in «country name»,`").

6.3 Performance Gaps Reflect Under-Represented cultures

We observe notably lower accuracy for Ethiopia and India in both emotion and sentiment tasks, suggesting that models may be less exposed to cultural practices and norms for these under-represented contexts. Ensuring greater diversity in training corpora is key to improving model performance for such cultures. Providing explicit country context can partially offset these gaps by nudging models to incorporate relevant cultural knowledge.

Overall, our findings underscore the importance of cultural context in developing and deploying LLMs. Beyond balanced data collection, researchers may explore culture-specific tuning or reinforcement learning from human feedback to further refine the abilities of models to interpret and respect cultural nuances.

7 Conclusion

In this paper, we evaluate a diverse set of state-of-the-art LLMs for their ability to predict culturally aware emotion prediction and sentiment analysis tasks. We investigate the influence of including explicit country references "`You live in «country name»`" and varying the query language. Our results indicate that LLMs tend to excel at culturally driven emotions that are well-represented in their training data and underperform for less represented cultures. Specifically, we find that 1) emotion is culture-dependent and can vary notably across languages and regions; 2) LLMs exhibit sizable performance gaps when tested on culture-specific emotions from under-represented locales; 3) providing explicit country context in prompts improves both emotion and sentiment prediction; and 4) sentiment analysis is better for the models, likely because it involved fewer class (positive, negative, neutral) than fine-grained emotion categories.

Moving forward, we suggest training LLMs in approaches such as 1) enriching the training data with diverse cultural information from various sources like literature, news, and cultural databases and 2) implementing fine-tuning techniques that specifically train the LLM on prompts and datasets focused on different cultural contexts so that they can be culturally aware and able to generate responses that are sensitive to cultural nuances. We also encourage more extensive evaluation of multilingual models using benchmarks designed to measure cultural awareness alongside standard accuracy metrics. Future research could also explore the influence of annotator demographics—such as age, gender, education level, political stance, and religion—on culture-specific emotion annotation. Finally, we hope that releasing the **CuLEmo** dataset will foster further exploration into culturally nuanced NLP tasks and lead to more inclusive language models.

Limitations

Subjectivity of emotion Emotional subjectivity remains a central challenge in emotion analysis tasks. Although annotating data via crowdsourcing such as Amazon Mechanical Turk (MTurk) is common in NLP dataset creation (Mohammad et al., 2018), and despite applying strict qualification criteria for annotators, maintaining consistent annotations is difficult given the inherently subjective nature of emotions.

Limited number of events Our test comprises only 400 questions for each language, which is certainly not sufficient to capture the full cultural differences in emotional expression.

Drawback of majority vote We decide the final label of the annotations using majority vote, such as an emotion label greater than or equal to three votes from a total of five annotators per instance will pass as a final emotion label. As a general drawback of the majority vote, this will exclude the perspectives of minority votes. Modeling annotator-level data without applying the majority vote can address this.

Limited emotion label space Additional constraints arise from our decision to limit the emotion label space to six classes; including more emotion categories (e.g. *surprise* or *disgust*) during the annotation could yield more fine-grained insights (Niu et al., 2024). Our dataset also covers only six languages/countries and comprises 400 events, which may restrict generalizability.

Annotation bias Emotion annotation is subjective in nature and can vary widely depending on personal background; it likely still has consistency issues, affecting the reliability of the evaluations.

Limited model evaluations Regarding open-source LLMs, we opted to evaluate only small-(2B,3B) and medium-sized (8B, 9B, 13B) models due to resource constraints and for experimental reproducibility. While larger LLMs might achieve higher accuracies across target languages, they remain beyond the scope of our current setup. Finally, although 400 events allow for controlled experiments, evaluating models on more extensive and varied data would provide a clearer picture of their culture-aware performance.

Ethics Statement

We conducted this work with careful attention to ethical considerations involving data creation, annotation, and potential downstream impacts.

1. Data Collection and Annotation

Cultural Respect The CuLEmo dataset was curated with input from native speakers and cultural representatives. We designed questions to capture diverse cultural norms and emotional responses without perpetuating stereotypes.

Consent and Compensation We used Amazon Mechanical Turk (MTurk) and an in-house annotation platform for data labeling. Workers were informed of the task's nature and compensated fairly at a rate of \$12/hour, which exceeds minimum-wage standards in the majority of the annotators' countries of residence.

Privacy and Confidentiality All scenario-based questions were artificially created or adapted from publicly available cultural information. No personally identifiable information was collected, and no real names or private details were used.

2. Fair Representation and Potential Biases

Under-Representation While we included six languages (English, Arabic, Amharic, German, Hindi, and Spanish) to broaden cultural coverage, bias in representation is inevitable in such datasets and evaluations. Clearly, many global cultures and languages remain unrepresented in our work, including minority language speakers of the languages we studied and speakers of those languages in less dominant regions. Additionally, cultures are complex and not subject to clean delineation. We therefore make no contention that this is a complete or fully representative dataset.

Subjectivity of Emotions Emotions are inherently subjective and influenced by personal and cultural backgrounds. Crowd-sourced annotations may inadvertently amplify majority cultural norms or obscure minority perspectives. We minimized these risks by providing clear guidelines, but acknowledge that subjective variation is inevitable.

3. Responsible Use of the Dataset and Models

Cultural Sensitivity The dataset includes prompts and scenarios potentially sensitive to specific cultural contexts (e.g., religious practices, social norms). We urge researchers and practitioners to exercise cultural sensitivity and caution when using the dataset or resulting models in applications that could impact cultural or ethnic groups.

Downstream Applications Models trained or evaluated on CuLEmo could be applied in contexts such as mental health or social support, potentially affecting vulnerable populations. We encourage developers to consider safety, fairness and informed consent when deploying such systems. We caution against deployment in high-stakes settings, particularly without appropriate safeguards, user testing, and especially ethical oversight.

4. Transparency and Future Work

Open Access We release the CuLEmo dataset publicly to facilitate reproducibility and encourage further research in culturally aware NLP.

Ongoing Improvement Future efforts should expand cultural and linguistic diversity, refine annotation protocols, and include more nuanced emotional labels. We welcome community feedback to improve both the dataset and modeling approaches.

We aim to advance culturally aware NLP through responsible data practices, fair representation, and transparent sharing, and hope this work fosters a more inclusive understanding of emotion across languages and cultures.

Acknowledgments

This work is partially funded by the NAACL regional fund 2024. We gratefully acknowledge computing resource support provided by the Computing Research Center (CIC) at the National Polytechnic Institute (IPN) and the Marian E. Koshland Integrated Natural Sciences Center at Haverford College. We extend our sincere thanks to the Digital Scholarship team at Haverford College, particularly Anna Lacy and Patricia Guardiola, for their assistance throughout this project. We thank Professor Xerxes Minocher at Haverford College for their valuable inputs. Grissom is funded by NSF grant 2403439.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15763–15784, Miami, Florida, USA. Association for Computational Linguistics.

Ibrahim Ahmad, Shiran Dudy, Resmi Ramachandranpillai, and Kenneth Church. 2024. Are generative language models multicultural? a study on Hausa culture and emotions using ChatGPT. In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 98–106, Bangkok, Thailand. Association for Computational Linguistics.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Anthropic AI. 2024. Introducing the next generation of claude. Jan 03, 2025.

Federico Bianchi, Debora Nozza, and Dirk Hovy. 2022. XLM-EMO: Multilingual emotion prediction in social media text. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 195–203, Dublin, Ireland. Association for Computational Linguistics.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and

Sara Hooker. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Luna De Bruyne. 2023. The paradox of multilingual emotion detection. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 458–466, Toronto, Canada. Association for Computational Linguistics.

Luna De Bruyne, Orphée De Clercq, and Véronique Hoste. 2019. Towards an empirically grounded framework for emotion analysis. In *HUSO 2019: The Fifth International Conference on Human and Social Analytics*, pages 11–16. IARIA, International Academy, Research, and Industry Association.

Luna De Bruyne, Pranaydeep Singh, Orphee De Clercq, Els Lefever, and Veronique Hoste. 2022. How language-dependent is emotion detection? evidence from multilingual BERT. In *Proceedings of the 2nd Workshop on Multi-lingual Representation Learning (MRL)*, pages 76–85, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Yi Fung, Tuhin Chakrabarty, Hao Guo, Owen Rambow, Smaranda Muresan, and Heng Ji. 2023. NORMSAGE: Multi-lingual multi-cultural norm discovery from conversations on-the-fly. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15217–15230, Singapore. Association for Computational Linguistics.

Julian Givi and Jeff Galak. 2017. Sentimental value and gift giving: Givers' fears of getting it wrong prevents them from getting it right. *Journal of Consumer Psychology*, 27(4):473–479.

Jian Guan, Ziqi Liu, and Minlie Huang. 2022. A corpus for understanding and generating moral stories. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5069–5087, Seattle, United States. Association for Computational Linguistics.

Shreya Havaldar, Bhumika Singhal, Sunny Rai, Langchen Liu, Sharath Chandra Guntuku, and Lyle Ungar. 2023a. Multilingual language models are not multicultural: A case study in emotion. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 202–214, Toronto, Canada. Association for Computational Linguistics.

Shreya Havaldar, Bhumika Singhal, Sunny Rai, Langchen Liu, Sharath Chandra Guntuku, and Lyle Ungar. 2023b. Multilingual language models are not multicultural: A case study in emotion. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 202–214, Toronto, Canada. Association for Computational Linguistics.

Jamin Rahman Jim, Md Apon Riaz Talukder, Partha Malakar, Md Mohsin Kabir, Kamruddin Nur, and M.F. Mridha. 2024. Recent advancements and challenges of nlp-based sentiment analysis: A state-of-the-art review. *Natural Language Processing Journal*, 6:100059.

Kaisla Kajava, Emily Öhman, Piao Hui, and Jörg Tiedemann. 2020. Emotion preservation in translation: Evaluating datasets for annotation projection. In *Digital Humanities in the Nordic Countries*, pages 38–50. CEUR.

Zongxia Li, Ishani Mondal, Huy Nghiem, Yijun Liang, and Jordan Lee Boyd-Graber. 2024. PEDANTS: Cheap but effective and interpretable answer equivalence. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9373–9398, Miami, Florida, USA. Association for Computational Linguistics.

Chen Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024a. Are multilingual LLMs culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2016–2039, Mexico City, Mexico. Association for Computational Linguistics.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2024b. Culturally aware and adapted nlp: A taxonomy and a survey of the state of the art.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2024c. Culturally aware and adapted nlp: A taxonomy and a survey of the state of the art. *Preprint*, arXiv:2406.03930.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Leena Mathur, Paul Pu Liang, and Louis-Philippe Morency. 2024. Advancing social intelligence in ai agents: Technical challenges and open questions. *Preprint*, arXiv:2404.11023.

Meta AI. 2024. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. Jan 03, 2025.

Mistral AI. 2024. Introducing the world's best edge models. Jan 03, 2025.

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.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

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.

Minxue Niu, Yara El-Tawil, Amrit Romana, and Emily Mower Provost. 2024. Rethinking emotion annotations in the era of large language models. *arXiv preprint arXiv:2412.07906*.

Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. 2022. POTATO: The portable text annotation tool. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 327–337, Abu Dhabi, UAE. Association for Computational Linguistics.

Flor Miriam Plaza-del-Arco, Alba A. Cercas Curry, Amanda Cercas Curry, and Dirk Hovy. 2024. Emotion analysis in NLP: Trends, gaps and roadmap for future directions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5696–5710, Torino, Italia. ELRA and ICCL.

Hongjin Qian, Xiaohe Li, Hanxun Zhong, Yu Guo, Yueyuan Ma, Yutao Zhu, Zhanliang Liu, Zhicheng Dou, and Ji-Rong Wen. 2021. Pchatbot: A large-scale dataset for personalized chatbot. *Preprint*, arXiv:2009.13284.

Shenbin Qian, Constantin Orasan, Felix Do Carmo, Qiuliang Li, and Diptesh Kanojia. 2023. Evaluation of Chinese-English machine translation of emotion-loaded microblog texts: A human annotated dataset for the quality assessment of emotion translation. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 125–135, Tampere, Finland. European Association for Machine Translation.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Klaus R Scherer and Harald G Wallbott. 1994. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of personality and social psychology*, 66(2):310.

Siqi Shen, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, Soujanya Poria, and Rada Mihalcea. 2024. Understanding the capabilities and limitations of large language models for cultural commonsense. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5668–5680, Mexico City, Mexico. Association for Computational Linguistics.

Vered Shwartz. 2022. Good night at 4 pm?! time expressions in different cultures. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2842–2853, Dublin, Ireland. Association for Computational Linguistics.

Nir Sweed and Dafna Shahaf. 2021. Catchphrase: Automatic detection of cultural references. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1–7, Online. Association for Computational Linguistics.

Madiha Tahir, Zahid Halim, Muhmmad Waqas, and Shanshan Tu. 2023. On the effect of emotion identification from limited translated text samples using computational intelligence. *International Journal of Computational Intelligence Systems*, 16(1):107.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Team Gemma, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, and others. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6).

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin 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. *Preprint*, arXiv:2402.07827.

A Dataset Examples

The section below shows English examples from CuLEmo datasets.

English Examples from the CuLEmo dataset

1. How would you feel when you did not tip the waiter at the restaurant?
2. How would you feel when someone insults someone's religion?
3. How would you feel if someone wear black to a wedding?
4. How would you feel when you see a female wearing small pants on the street
5. How would you feel when your attendee joined the meeting after 10 minutes started
6. How would you feel when you saw a young dog running to you?
7. How would you feel when someone accepted your gift in one hand?
8. How would you feel when you crossed an older man in the street?
9. How would you feel when someone calls a religious leader by their name
10. You are a doctor. How would you feel when someone calls you by your name without the doctor prefix
11. How would you feel if your parents arranged a marriage for you without your input?
12. How would you feel if someone touched or shook hands with you while you were of the opposite sex?
13. How would you feel when a guest leaves food on their plate at your home?
14. How would you feel if you saw a woman not cover her hair while entering a religious place?
15. How would you feel when you were expected to wash other legs
16. How would you feel if someone refused your gift before accepting it?
17. How would you feel when you kissed your girlfriend/boyfriend and saw your father?
18. How would you feel when your parent caught you with a pornographic film?
19. If someone is passing near to you, how would you feel when they didn't give you a greeting?
20. How would you feel when you see a female eating on the street?
21. How would you feel when you see a Sheikh drinking alcohol?
22. How would you feel when you see a priest drinking alcohol?
23. How would you feel when you see homosexual friends?
24. How would you feel when someone shows you his middle finger gesture?
25. How would you feel when someone gave the money in their left hand?
26. How would you feel if your guests chose to keep their shoes on when entering your home?
27. How would you feel if a friend offered to pay for your meal in a restaurant?
28. How would you feel if your child expressed a desire to marry someone from a different religion?
29. How would you feel if someone bargained for your prices?
30. When you are at a meal, how would you feel about being told to wait for the host to say start before starting to eat?
31. How would you feel when you expected to wait for the eldest person to start eating before you begin?
32. How would you feel if someone were scrolling their phone during a meal?
33. How would you feel when you saw a young dog running toward you?
34. How would you feel when someone cheek-to-cheek kisses your wife/husband?
35. How would you feel if someone bargained for your prices?
36. How do you feel when a guest arrives late to your lunch invitation?
37. How do you feel seeing someone eating in hand without using utensils?
38. How would you feel if someone called an elder by their first name without a title?
39. How would you feel if a guest left your home without eating anything?
40. How would you feel if your friend didn't stand up when an elder entered the room?

B Implementation Details

B.1 Hyper-parameters

For open-source LLMs, we used the default generation hyperparameters (top-p sampling with p = 0.9 and temperature = 0.0, max_new_tokens=200). For others, we directly employed their pre-defined interfaces, either through their online API or the CHAT function from the Transformers library. The proprietary models are called through their API in Python. All open-source models are evaluated using default parameters from the hugging face.

B.2 LLM Versions

Versions of the proprietary LLMs and hugging face names for the open-source LLMs are given below.

- Llama-3.2-3B (Meta AI, 2024) — meta-llama/Llama-3.2-3B-Instruct
- Llama-3.1-8B (Dubey et al., 2024) — meta-llama/Llama-3.1-8B-Instruct
- Gemma-2-2B (Team Gemma et al., 2024) — google/gemma-2-2b-it
- Gemma-2-9B (Team Gemma et al., 2024) — google/gemma-2-9b
- Aya-expanse-8b (Dang et al., 2024) — CohereForAI/aya-expanse-8b
- Aya-101 (Üstün et al., 2024) — CohereForAI/aya-101
- Ministral-8B (Mistral AI, 2024) — mistralai/Ministral-8B-Instruct-2410
- Claude-3.5-sonnet (Anthropic AI, 2024) — claude-3-5-sonnet-20240620
- Claude-3-opus (Anthropic AI, 2024) — claude-3-opus-20240229
- Gemini1.5-flash (Team et al., 2024) — gemini-1.5-flash-002
- GPT-4 (Achiam et al., 2024) — gpt-4 (turbo-2024-04-09)

C Prompts

Prompts	Language
You live in <<Country name>>, choose the best emotion you feel for the given question. Emotions can be only either 'anger', 'fear', 'sadness', 'joy', 'guilt', or 'neutral'. Return only a single emotion word from the list of emotions without further explanation.	English
أنت تعيش في دولة الإمارات العربية المتحدة، اختر الحالة العاطفية الأفضل التي تشعر بها تجاه السؤال المطروح. يمكن أن تكون العواطف إما "غضب" أو "خوف" أو "حزن" أو "فرح" أو "ذنب" أو "محايد". أعد فقط كلمة عاطفة واحدة من قائمة العواطف دون أي تفسير إضافي.	Arabic
Sie leben in Deutschland. Wählen Sie die beste Emotion, die Sie für die gestellte Frage empfinden. Emotionen können nur 'wut', 'freude', 'traurigkeit', 'schuld', 'furcht' oder 'neutral' sein. Geben Sie nur ein einziges Emotionswort aus der Liste der Emotionen ohne weitere Erklärung zurück.	German
የምትኖረው ኢትዮጵያ ውስጥ ነው፣ ለተሰጠው ጥያቄ የሚሰማህን ስሜት ምረጥ። ስሜቶች 'ቁጣ', 'ጥፋተኛ', 'ሀዘን', 'ደስታ', 'ፍርሀት' ወይም 'መደበኛ' ብቻ ሊሆኑ ይችላሉ:: ያለተጨማሪ ማብራሪያ ከስሜቶች ዝርዝር ውስጥ አንዱን ስሜት ብቻ ይመልሱ።	Amharic
आप भारत में रहते हैं, दिए गए प्रश्न के लिए अपनी सबसे अच्छी भावना चुनें। भावनाएँ केवल 'उदासी', 'आनंद', 'अपराध', 'गुस्सा', 'डर', या 'सामान्य' हो सकती हैं। बिना किसी अतिरिक्त स्पष्टीकरण के भावनाओं की सूची से केवल एक ही भावना शब्द लौटाएँ।	Hindi
Vives en México. Elige la emoción que sientes más a menudo en la pregunta. Las emociones pueden ser 'enojo','tristeza','culpa','alegría','miedo' o 'neutral'. Solo responde con una palabra de la lista de emociones sin más explicaciones.	Spanish

Table 6: Prompts used for probing emotions from LLMs. In English prompt, «Country name» will change accordingly from lists of countries [United States of America (USA), Ethiopia, United Arab Emirates (UAE), Germany, India, Mexico] based on the culture prompting. We enforce the model in the prompt to answer only one of the given options.

D English Prompt Results

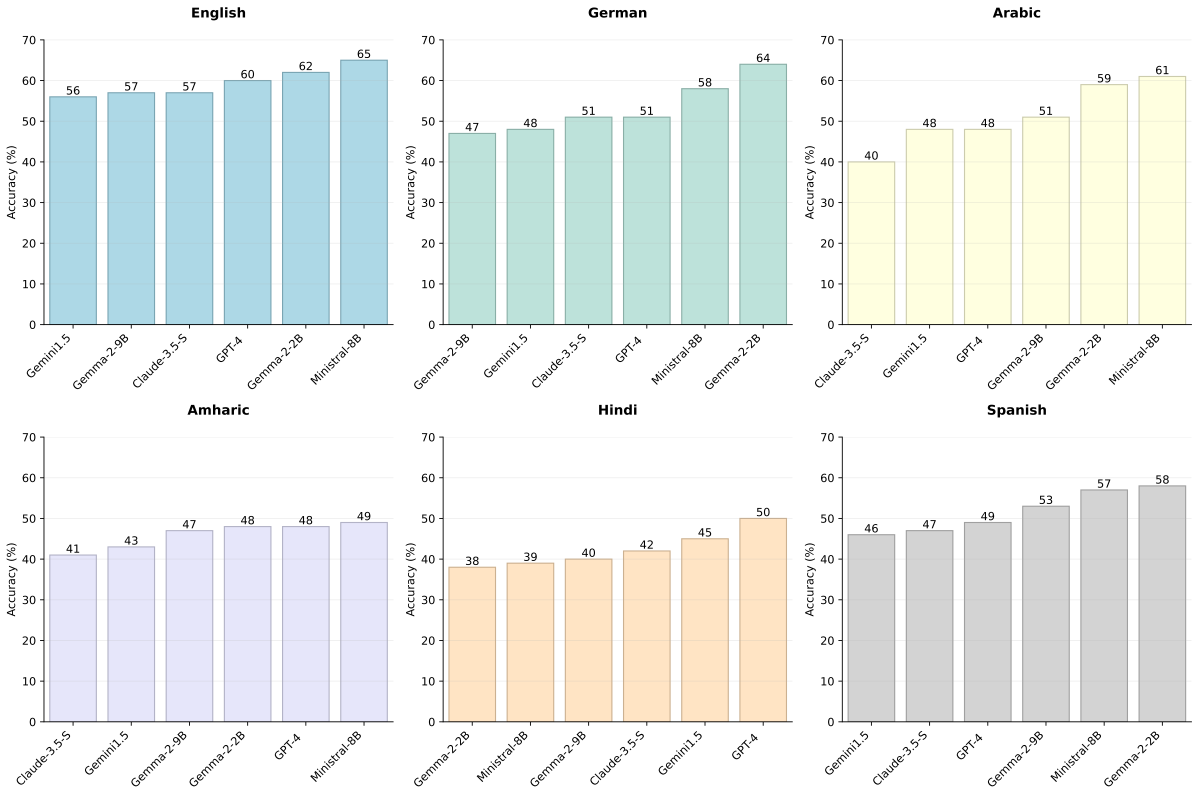

Figure 6: English prompting results with *You live in «country name»* context for the across languages and LLMs.

E MTurk Annotation Qualification Settings

To target suitable workers on MTurk, we set the following qualifications:

1. **Location** must be in the target country for each language by assuming annotators that live in the specified country are native or adopted the culture.
2. **Number of HITs approved** must exceed 1,000 to ensure experienced workers.
3. **HIT approval rate** must be at least 99%, favoring high-quality, consistent annotators