Beyond Metrics: Evaluating LLMs' Effectiveness in Culturally Nuanced, Low-Resource Real-World Scenarios

Millicent Ochieng[†] **Varun Gumma**[‡] **Sunayana Sitaram**[‡] **Jindong Wang**[§*]
Vishrav Chaudhary[¶*] **Keshet Ronen**[◊] **Kalika Bali**[‡] **Jacki O'Neill**[†]
[†]**Microsoft Research Africa** [‡]**Microsoft Research India** [§]**William & Mary**
[¶]**Meta** [◊]**University of Washington**
Contact: {mochieng, jacki.oneill}@microsoft.com

Abstract

The deployment of Large Language Models (LLMs) in real-world applications presents both opportunities and challenges, particularly in multilingual and code-mixed communication settings. This research evaluates the performance of seven leading LLMs in sentiment analysis on a dataset derived from multilingual and code-mixed WhatsApp chats, including Swahili, English and Sheng. Our evaluation includes both quantitative analysis using metrics like F1 score and qualitative assessment of LLMs' explanations for their predictions. We find that, while Mistral-7b and Mixtral-8x7b achieved high F1 scores, they and other LLMs such as GPT-3.5-Turbo, Llama-2-70b, and Gemma-7b struggled with understanding linguistic and contextual nuances, as well as lack of transparency in their decision-making process as observed from their explanations. In contrast, GPT-4 and GPT-4-Turbo excelled in grasping diverse linguistic inputs and managing various contextual information, demonstrating high consistency with human alignment and transparency in their decision-making process. The LLMs however, encountered difficulties in incorporating cultural nuance especially in non-English settings with GPT-4s doing so inconsistently. The findings emphasize the necessity of continuous improvement of LLMs to effectively tackle the challenges of culturally nuanced, low-resource real-world settings and the need for developing evaluation benchmarks for capturing these issues.

1 Introduction

Large Language Models (LLMs) have ushered in major advancements in language processing, demonstrating exceptional ability to process everyday language commands and handle textual tasks such as Question Answering, Sentiment Analysis, Summarization, among others (OpenAI, 2023a; Brown et al., 2020; Chowdhery et al., 2022; Anil et al., 2023; Touvron et al., 2023).

Despite LLMs advancements, their effectiveness is predominantly observed in Latin Script languages with abundant training data, such as English, which constitutes a significant proportion of their training corpus (Raffel et al., 2020; Common Crawl, 2023; Together Computer, 2023; Longpre et al., 2023). Although English is not the mother tongue of the majority of the world's population, 93% of GPT-3's training data consists of English content (Brown et al., 2020). Studies reveal that languages with medium to low amounts of training data like Swahili still present challenges for these models, highlighting they are far from achieving parity with English (Ahuja et al., 2023a,b; Robinson et al., 2023). The picture is further complicated given that 60% of the world population speaks two or more languages[1]. In such settings, code-mixing[2] is a prevalent aspect of natural language use. Consequently, the performance of these models in real-world settings, especially in low-resource code-mixed and culturally diverse environments, remains an area of significant interest.

This study investigates the effectiveness of seven prominent LLMs on a sentiment analysis task on a dataset derived from WhatsApp chats. The dataset exhibits extensive code-mixing, encompassing multilingual conversations in English, Swahili, and Sheng[3] in 'chat speak' e.g. using emojis, abbreviations, colloquial chat message spellings and misspellings. With LLMs' ability to process and produce human-like text, this task aims to evaluate

[*] Work done while at Microsoft.

[1] https://ilanguages.org/bilingual.php

[2] the practice of alternating between two or more languages or dialects in a conversational turn

[3] a dynamic urban slang from Nairobi, Kenya, blending Swahili, English, and local languages, evolving continually among the youth.

their understanding of the nuances present in the dataset. We supplement the quantitative analysis of the LLMs performance with a systematic qualitative analysis of the explanations the models provide for their predictions. While studies such as (Narang et al., 2020; Wiegreffe et al., 2021; Majumder et al., 2021; Wiegreffe et al., 2022) have demonstrated the capability of LLMs to generate natural language explanations alongside predictions, enhancing explainability and improving the faithfulness of AI systems, it remains uncertain whether these explanations directly influence the decision-making process. However, we expect, and indeed do see, a correlation between models' predictions and their explanations. We used the explanations as a method of interrogating, to some extent, the models ability to process the cultural and linguistic nuances of the messages. By looking beyond the numbers, this method enables us to get some sense of how well the different LLMs handle the complex interactional features present in a real-world multilingual dataset. We demonstrate the value of using qualitative HCI methods alongside traditional performance metrics. Our contributions are as follows: (1) we evaluate and compare the performance of seven advanced LLMs including GPT-4, GPT-4-Turbo, GPT-3.5-Turbo, Llama-2-70b, Mistral-7b, Mixtral-8x7b and Gemma-7b on a sentiment analysis task using a novel WhatsApp chat dataset; (2) we identify differences in the interpretation strategies employed by different LLMs, highlighting the diversity in their approach to processing complex linguistic data; (3) we highlight the value of real-world, multilingual, and code-mixed datasets in assessing the performance of LLMs; (4) we show how qualitative HCI methods can be used in NLP to get a deeper understanding of model performance. Our findings reveal that, while LLMs like Mistral-7b and Mixtral-8x7b achieved high F1 scores in sentiment analysis in the dataset, they and other LLMs such as GPT-3.5-Turbo, Llama-2-70b, and Gemma-7b seem to be less robust at handling linguistic, cultural, and contextual nuances. Further, there was a lack of transparency in their generated explanations. In contrast, LLMs like GPT-4 and GPT-4-Turbo deployed diverse linguistic and contextual information in their explanations, demonstrating high consistency with human judgement. All the LLMs however, struggled to incorporate the more complex cultural nuances in the WhatsApp dataset especially in non-English settings - even GPT-4 and GPT-4-Turbo did so inconsistently.

2 Evaluation Dataset and Task

2.1 Dataset

The WhatsApp Chat Dataset: Our study employed a distinctive dataset originally collected by Karusala et al. (2021) further annotated by Mondal et al. (2021), with all ethical considerations and privacy measures observed as described below. It features multilingual exchanges among young people living with HIV in informal settlements in Nairobi, Kenya, captured within two health-focused WhatsApp chat groups moderated by a medical facilitator. The total number of messages are 6,556 and the conversations are predominantly in English, enriched with a considerable use of Swahili, Sheng, and code-mixing. The data annotation included sentiment and word-level language identification for each message. As Karusala et al. (2021) describe, recruited participants signed a consent form outlining study procedures, data anonymization, and security measures. All messages were anonymized and translated into English by a native speaker. Each chat message in the dataset included an anonymized speaker ID, timestamp, original message, and English translation. Due to the sensitive nature of the content, the dataset is not publicly available, but researchers can request access by contacting the authors. We specifically selected this dataset because it consists of real WhatsApp interactions between participants and a medical facilitator occurring as part of a Global Health research intervention. Additionally, its authentic representation of real-world, code-mixed communication aligns with our core research focus.

Pre-Processing: Considering that the data originates from WhatsApp conversations, it exhibits a casual, conversational style, often with short interactions. We retained only turns with three or more words providing more valuable data for sentiment analysis. Contrary to typical processing methods, we do not perform punctuation or emoji normalization on the data, as these elements are integral to the communication. The resulting dataset consisted of 3,719 messages with an average of eleven words per message.

2.2 Evaluation Task

Sentiment Analysis: The core of our evaluation focuses on a sentiment detection task because of

Language	# of Neutral Messages	# of Positive Messages	# of Negative Messages	# of Messages Per Language	Average # of Tokens Per Message	Total # of Tokens Per Message
Monolingual						
En	1303	54	24	1381	12	16902
Sw	270	-	32	302	4	1264
Sh	2	-	-	2	3	6
Multilingual						
En-Sw	631	5	19	655	9	5582
Sw-Sh	143	-	12	155	5	705
En-Sh	51	2	1	54	6	301
En-Sw-Sh	190	3	16	209	10	2100
En-CM	10	-	1	11	9	94
Sw-CM	29	-	2	31	4	124
En-Sw-CM	60	-	6	66	12	812
En-Sh-CM	2	-	-	2	12	24
Sw-Sh-CM	20	1	1	22	5	106
En-Sw-Sh-CM	36	-	3	39	14	542
Other	96	1	-	98	20	795
En-Other	359	10	4	373	20	7622
Sw-Other	73	-	5	78	4	342
Sh-Other	1		-	1	4	4
En-Sw-Other	119	2	2	123	11	1396
En-Sh-Other	16	-	2	18	10	188
Sw-Sh-Other	19	-	1	20	6	110
En-Other-CM	4	-	-	4	6	24
Sw-Other-CM	4	-	3	7	6	41
En-Sw-Sh-Other	37	2	1	40	12	481
En-Sw-Other-CM	10	-	2	12	13	161
En-Sh-Other-CM	3	-	-	3	4	12
Sw-Sh-Other-CM	4	-	-	4	8	33
En-Sw-Sh-Other-CM	9	-	-	9	17	151
Total	**3501**	**80**	**137**	**3719**	**246**	**39922**

Table 1: Message Distribution by Language. This table displays the count of neutral, positive, and negative messages, total messages per language, average tokens per message, and total tokens per message for each language studied.

its real-world application for such chat groups. We wished to support the facilitator by for example flagging negative messages. Table 1 illustrates the sentiment distribution according to human annotators within our pre-processed dataset, heavily skewed towards the Neutral class. This imbalance highlights the evaluation challenge of accurately identifying the less frequent Negative and Positive sentiments, testing the LLMs' ability to detect sentiment cues in a predominantly Neutral context.

Languages in the WhatsApp Dataset: Table 1 describes the statistics of languages within the dataset defined as: *En (English), Sw (Swahili), Sh (Sheng) and CM (Code-Mixed).* These include messages in single language (*Monolingual*) and messages in more than one language (*Multilingual*). The dataset includes an 'Other' category used for words that do not fit the primary categories due to uncertainty, named entities, or other unique factors.

3 Experimental Setup

3.1 Models

We evaluated three OpenAI models: `GPT-4-Turbo`, `GPT-4-32k` (OpenAI, 2023b), and `GPT-3.5-Turbo` (Ouyang et al., 2022), with `GPT-4-32k` being the latest iteration and known for its enhanced performance on text processing and generation. `GPT-4-Turbo` and `GPT-3.5-Turbo` are optimized versions designed for more efficient processing without significantly compromising performance. From the open-source collection, we select Meta's `Llama-2-70b-chat` (Touvron et al., 2023), an LLM known for its efficiency and chat functionality. Additionally from Mistral AI, we include `Mistral-7B-Instruct-v0.2` (Jiang et al., 2023) and `Mixtral-8x7B-Instruct-v0.1` (Jiang et al., 2024), the former being popular for its exceptional ability to follow instructions and the latter for its innovative architecture which makes it excel in mathematics, code generation, and multilingual tasks. Lastly, we include Google's `Gemma-7b-it` (Mesnard et al., 2024), a state-of-the-art language model that excels in language understanding, reasoning, and safety, outperforming comparable models in numerous academic benchmarks. Throughout this paper, we refer to the mentioned models as: GPT-4, GPT-4-Turbo, GPT-3.5-Turbo, Llama-2-70b, Mistral-7b, Mixtral-8x7b and Gemma-7b.

3.2 Model Evaluation

Different prompting approaches (Brown et al., 2020; Chen et al., 2023) have been shown to effectively guide LLMs contextually, towards desired outputs. Investigations reveal that the quality of prompts provided, have a profound influence on the performance of LLMs (Liu et al., 2023; Hada et al., 2023). Leveraging this technique, we craft a detailed prompt to guide the LLMs to function as specialized NLP assistants for sentiment analysis. The prompt directs models to identify sentiments as Positive, Negative, or Neutral. Figure 3 illustrates the standardized prompt that we used for evaluating all seven LLMs, facilitating a fair and consistent comparison of their performance. We employ the same sentiment definitions given to human annotators during the dataset's sentiment annotation phase. Furthermore, we direct the models to justify their sentiment classifications in 200 words or less, focusing on the text spans that influenced their decisions. We conduct the evaluation on the entire pre-processed dataset and employ the weighted F1-score[4] metric instead of accuracy due to the skew in our dataset.

4 Qualitative Analysis

We supplemented our quantitative evaluation with in-depth qualitative analyses. Considering the skew towards the neutral class in our dataset, as illustrated in Table 1, and to ensure a balanced and rigorous analysis, we selected a sample comprising a total of 261 messages at random including both monolingual and multilingual messages for in-depth examination. For monolingual, the sample included 150 messages divided equally among the three sentiment categories: Negative, Positive, and Neutral. For the multilingual case, the sample entailed 50 Neutral, 50 Negative and 11 Positive messages whereby the positive messages represented all the positive labels in this category. The first author being proficient with the languages in the WhatsApp dataset, analyzed all the 261 sampled messages, identifying patterns in the data. We borrowed five criteria as outlined by (Chang et al., 2023) on *'how to evaluate'* to guide our human evaluation. These criteria, described in Table 4, were linguistic accuracy, contextual and cultural relevance, fluency in maintaining consistency, alignment with human expectations and transparency in the LLMs' decision-making process. While these criteria have been designed for quantitative evaluation, we used them to provide a structure for analysing the set of justifications for each message across models. We supplemented this structured analysis with a more in-depth ethnomethodologically-informed approach (Crabtree et al., 2000, 2006; O'Neill and Martin, 2003; Martin et al., 2014) where the first author and the last author (who is skilled in ethnomethodologically-informed analysis) together analysed each turn (message) in detail, to understand how the justifications produced by the different models related to the original and the human-translated message, how they related to one another and how they related to the sentiment prediction. In these sessions, the two authors looked in detail at the messages and model justifications and identified emergent patterns, interrogating and refining them. This analysis was deeply qualitative, aiming to derive insights into differences between models in their justifications. As an additional sanity check, we invited three other native speakers to review a set of 15 messages, selected from the 261 messages the first author had analysed. The reviewers conducted their assessments independently and reconvened to discuss their findings, along with the first author and were all confident about the consistency of the findings. In this paper, for reasons of space, we use a small number of examples to illustrate the patterns that we found in the data.

[4] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

5 Results and Discussion

In this section, we explore the results of our study by discussing both quantitative and qualitative findings of models performance on the WhatsApp dataset, beginning with the quantitative results measured by the F1 score. Following this, we will delve into the qualitative findings, discussing insights from the models' justifications and their implications for language processing strategies.

5.1 F1 Score-Based Models Comparison

As evidenced by the F1 Score comparison of the seven models in Figure 1, the Mistral-7b model demonstrates a higher performance in sentiment analysis on the WhatsApp dataset, closely followed by GPT-4. Conversely, the Llama-2-70b model exhibits the weakest performance.

As Table 1 illustrates, the majority of positive and neutral sentiments were expressed in English, whereas most negative sentiments were conveyed in Swahili. This supports the findings of (Rudra et al., 2016), which suggest that people are more likely to express negative opinions in non-English languages. Further analysis of F1 scores by sentiment and language, as shown in Figures 2 and 4,

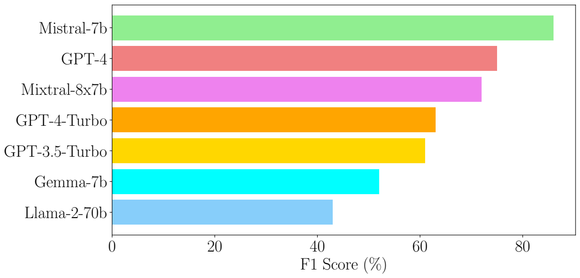

Figure 1: Overall F1 score comparison of the models.

highlights the distinct capabilities of various models. Specifically, Mistral-7b excels in identifying neutral sentiments, predominantly in English followed closely by GPT-4s, Mixtral-8x7b, and GPT-3.5-Turbo, with Gemma-7b and Llama-2-70b trailing. Conversely, GPT-4 and GPT-4-Turbo demonstrate superior performance in accurately classifying the rare negative sentiments, predominantly in Swahili and code-mixed. These findings are consistent with those from standard NLP benchmarks such as those reported in (Ahuja et al., 2023b), particularly in non-English contexts, specifically low-resource languages like Swahili and code-mixed languages. In these settings, larger models such as OpenAI's GPT-4s frequently outperform other LLMs.

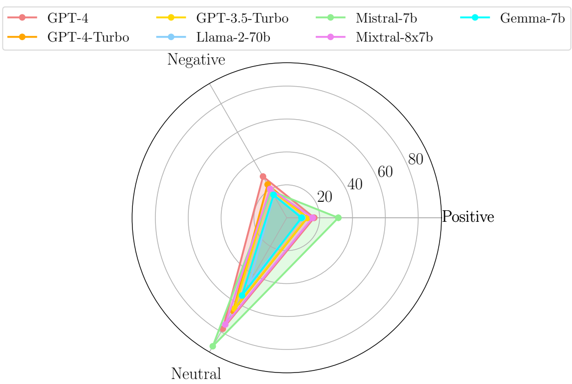

Figure 2: Comparison of F1 scores for the models across Positive, Negative, and Neutral sentiments.

5.2 Insights from the Justifications

As instructed by our detailed prompt in Figure 3, all the seven LLMs produced their predictions along with justifications. Across the board, even where sentiment predictions were the same for all the models, we noticed distinct differences in the justifications provided. Some LLMs consistently incorporated words or spans of text as part of their justifications (as requested in the prompt) others did so less frequently. Similarly some regularly translated non-English terms into English, others did not. See the examples presented in this Section's Tables 2 and 3 and in Appendix §A.4. In Table 2, GPT-3.5-Turbo and Mistral-7b do not provide any spans of message text in their justifications.

Which models perform best at accurately interpreting linguistic nuances and textual inaccuracies such as spelling errors, local abbreviations and grammatical inaccuracies? As illustrated by the examples in Tables 3 and 5 and Appendix §A.4, GPT-4 and GPT-4-Turbo stand out with superior performance in languages like Swahili and Sheng, and in code-mixed scenarios where they effectively handle the linguistic nuances between English and local languages. In particular, they were more consistently correct in their **translation** of the non-English word spans used in their justifications. They maintained strong performance even in rare sentiment classes, outperforming models like GPT-3.5-Turbo, which, though proficient, does not reach the high level of linguistic performance exhibited by the GPT-4s. Taking the example in Table 6 to illustrate - a code-mixed English-Swahili-Sheng message, GPT-4 and GPT4-Turbo provide correct interpretations of the Swahili and Sheng in their justifications, with GPT-4 Turbo even identifying *'kuniboo'* (Translation: *'bore me'*) as Sheng. Gemma-7b also identifies Sheng, but wrongly identifies the whole sentence as Sheng, and mistranslates it. The other models all provide mistranslations in their justifications. Models such as Llama-2-7b, Mistral-7b, Mixtral-8x7b and Gemma-7b face difficulties with Swahili and Sheng, as evidenced by their often incorrect translations of Swahili words and phrases. They often **prioritized English** in mixed-language settings resulting in either incorrect predictions or justifications when key sentiment indicators lie in the non-English segments which was mostly the case for Negative sentiments. Similarly with regards to LLMs' robustness to textual inaccuracies, the GPT-4s accurately interpreted messages with irregularities. However, the remaining models were less reliable, struggling with noisy data especially non-English texts; an example is shown in Table 5.

Nonetheless, overall all the seven LLMs have demonstrated proficiency in English messages in the WhatsApp dataset. However, even in English the models can fail to predict the correct sentiment, and their justifications reflect the sentiment that they predicted. Let's take the example in Appendix Table 7, for a human reviewer, this message is a clear example of a social media chain message - typically 'copy and paste' messages requiring the reader to either like, respond, or forward. The linguistic indicators of this are the instruction at the start *"Send to everyone you love..."* and the conclusion *"You are lovable if you get FIVE sent back to you"*. This can be read as an instruction or perhaps a playful activity. None of the models predicted the correct sentiment (Neutral), all predicting Positive. In their justifications, none of them identified this as a chain message, instruction, activity or similar - even where they highlighted the phrase *"You are lovable if you get FIVE"*.

Do the models utilize the surrounding textual context and cultural subtleties to determine sentiment? From our analysis and as illustrated in Table 2, GPT-4 and GPT-4-Turbo effectively utilized context in their justifications for their sentiment predictions. This is evidenced by their use of relevant word spans and their correct explanations of the meanings of phrases, leading to accurate predictions and coherent interpretation. GPT-3.5-Turbo lagged slightly due to occasional oversights in contextual (phrase) information. The remaining models including Llama-2-7b, Mistral-7b, Mixtral-8x7b and Gemma-7b often used word-level rather than phrase level justifications, especially in multilingual and code-mixed texts, leading to misinterpretation of meaning and incorrect predictions or justification. With regards to cultural relevance, the LLMs generally struggled to incorporate cultural nuances in the dataset, see example in Table 3. However, in specific scenarios, models like OpenAI's GPT-4s, which excel in grasping linguistic subtleties and leveraging contextual information, demonstrated proficiency in incorporating cultural aspects into their interpretations.

Are the models fluent in maintaining consistency in their interpretation across similar sentiment scenarios within the WhatsApp dataset? Our findings show that models like GPT-4 and GPT-4-Turbo demonstrate high consistency, reliably applying their analytical capabilities in sentiment predictions across both English and non-English language settings. In contrast, other models performed better on English and Neutral sentiments but lagged behind in non-English cases exhibiting less consistency, often varying in their justifications and output even under similar conditions. This inconsistency can lead to unpredictability in performance in complex multilingual environments, highlighting the need for models to have stable and reliable interpretation mechanisms when deployed in varied real-world applications.

Do the models' predictions and interpretation conform to typical human expectation? We observed that LLMs such as OpenAI's GPT-4s, which demonstrated robust linguistic and contextual comprehension in sentiment detection for the dataset, often aligned with human expectations, consistently producing accurate predictions along with correct justifications. However, the rest of the LLMs, while mostly aligned in English contexts, did not perform as well in non-English settings.

Were the models transparent in their decision-making process in the WhatsApp dataset? Our analysis revealed that Mistral and Mixtral models, despite achieving high F1 scores, often lacked correct and transparent interpretations in their justifications. This was particularly evident in non-English contexts, reflecting issues we observed with LLMs such as GPT-3.5-Turbo, Llama-2-7b, and Gemma-7b. Frequently, these LLMs incorrectly translated the text or failed to disclose the specific text spans that informed their conclusions. This opacity makes it challenging to trust their judgments. That they often reached the correct prediction could be because the sentiment could be determined despite the incorrect translation (assuming some correlation between the prediction and the justification). The implications of this are that these models might well be useful for tasks like sentiment prediction, but less useful for tasks which require a more accurate interpretation of language and semantics - including translation tasks. In contrast, OpenAI's GPT-4s showcased a robust multilingual and contextual performance. These models consistently highlighted relevant phrases or text spans that justified their conclusions, generally leading to correct and transparent outcomes. However, when cultural factors influenced their decision-making process, the models' performance were mixed, displaying transparent interpretation in some instances but falling short in others. This inconsistency raises concerns about the models' reliability in culturally nuanced scenarios.

6 Related Works

Multilingualism in LLMs: LLMs' effectiveness in multilingual contexts continually attracts atten-

Message	GPT-4	GPT-4-Turbo	GPT-3.5-Turbo	Llama-2-70b	Mistral-7b	Mixtral-8x7b	Gemma-7b
Original: *"Hi guys meeting yetu imekuewaje."* **Translation:** *"Hey guys how was the meeting."* **Actual Sentiment:** Neutral	The message 'Hi guys meeting yetu imekuewaje' is asking about the status of a meeting, which is a neutral inquiry and does not express any positive or negative sentiment. **Prediction:** Neutral	The message 'Hi guys meeting yetu imekuewaje' is a neutral sentiment as it is a general inquiry about the status or outcome of a meeting without any explicit positive or negative emotion or opinion. **Prediction:** Neutral	The sentiment is Neutral. The message is a simple greeting and does not express a positive or negative sentiment. **Prediction:** Neutral	The word 'meeting' suggests a gathering or event, which is typically a positive or neutral activity. The phrase 'imekuewaje' is in Swahili and roughly translates to 'it's okay' or 'it's all good' in English, which implies a positive or neutral sentiment. Therefore, the overall sentiment of the message is likely Neutral. **Prediction:** Neutral	The message is a Swahili greeting, which is a normal aspect and does not express any particular sentiment. **Prediction:** Neutral	The given message 'Hi guys meeting yetu imekuewaje' is a greeting and an invitation to a meeting. The word 'guys' is a friendly way to refer to a group of people, 'hi' is a common greeting, and 'yetu' means 'our' in Swahili, indicating inclusivity. The word 'imekuewaje' is in Sheng, a slang language in Kenya, and it means 'has been organized' or 'has been set'. Therefore, the sentiment expressed in this message is Neutral, as it is a factual statement and a simple greeting without expressing any positive or negative emotions. **Prediction:** Neutral	The phrase 'Hi guys meeting yetu imekuewaje' is neutral. The speaker is simply greeting the recipients and does not express any positive or negative sentiment. **Prediction:** Neutral

Table 2: In this example, only GPT-4 and GPT-4-Turbo showed excellence in both linguistic and contextual understanding while remaining models misinterpreted the message as simply a greeting, due to a misunderstanding of the Swahili words. Interestingly all LLMs made the correct prediction, and their explanations correlated with the prediction they made, but only the explanations by GPT-4 and GPT-4 Turbo matches human expectations.

Message	GPT-4	GPT-4-Turbo	GPT-3.5-Turbo	Llama-2-70b	Mistral-7b	Mixtral-8x7b	Gemma-7b
Original: *"Kama hauko shule shindaapo"* **Translation:** *"If you are not in school stay there"* **Actual Sentiment:** Negative	The message 'Kama hauko shule shindaapo' translates to 'If you are not in school, stay there'. This is a neutral statement as it does not express any positive or negative sentiment. It is simply a factual advice or suggestion. **Prediction:** Neutral	The message 'Kama hauko shule shindaapo' translates to 'If you are not at school, stay where you are' in English. This message seems to be giving neutral advice or instruction without any explicit positive or negative emotion. **Prediction:** Neutral	The sentiment is Negative. The phrase 'shindaapo' expresses a negative feeling or view about not being in school. **Prediction:** Negative	The sentence 'Kama hauko shule shindaapo' expresses a negative sentiment. The word 'kama' means 'like' in Swahili, and 'hauko' means 'there isn't' or 'there is no'. The phrase 'shule shindaapo' means 'school is closed'. Therefore, the speaker is expressing a negative sentiment towards the fact that school is closed. The tone of the sentence is also melancholic, indicating a negative emotion. **Prediction:** Negative	The given query in Swahili translates to 'If school had not closed'. It is a statement expressing a factual situation and does not contain any positive or negative sentiment. **Prediction:** Neutral	The query 'Kama hauko shule shindaapo' is in Swahili and translates to 'If you hadn't gone to school, you would have passed'. The sentence does not express any positive or negative sentiment. It is a hypothetical situation and there is no sentiment involved. Therefore, the sentiment is Neutral. **Prediction:** Neutral	The phrase 'Kama hauko shule shindaapo' is a Swahili phrase that translates to 'You are not going to school today'. The speaker is expressing a negative sentiment as they are saying that the recipient will not be able to attend school. The words 'shule' and 'shindaapo' are the key words used to convey this sentiment. **Prediction:** Negative

Table 3: The Swahili phrase carries a negative sentiment as it is used to dismissively suggest that someone who is not in school should refrain from participating in discussions, reflecting a cultural emphasis on educational status as a basis for contributing to intellectual conversations. All LLMs did not understand the cultural context here.

tion. Studies by (Lai et al., 2023; Ahuja et al., 2023a; Hendy et al., 2023; Ahuja et al., 2023b), have thoroughly evaluated LLMs across various languages using standard NLP benchmarks with (Ojo et al., 2023; Robinson et al., 2023) focusing on African languages. However, these studies typically depend on NLP benchmarks that may not closely reflect the complexity of real-world contexts, often missing subtleties and scenarios that occur in everyday use. Moreover, the LLMs might have already encountered some benchmark content online, a concern highlighted by (Sainz et al., 2023; Ahuja et al., 2023b).

Models' Challenges in Code-Mixed Text Processing: Code-mixing - the blending of two or more languages within a single utterance of a conversation (Poplack, 2001), is common in multilingual communities, including Kenya (Orao, 2012). The phenomenon of code-mixing presents unique challenges in the field of NLP. (Zhang et al., 2023; Doğruöz et al., 2023; Kaji and Shah, 2023) emphasizes the lack of training data as one of the main challenges, attributing to the complexity of processing code-mixed language.

Importance of Real-World Data in LLMs Evaluation: (Wibowo et al., 2023) introduces COPAL-ID, a culturally rich Indonesian dataset that challenges even advanced models like GPT-4, highlighting the need for nuanced datasets in LLMs evaluation. (Chiu et al., 2024) present Cultural-Teaming, an AI-assisted interactive red-teaming approach that enhances the creation of multicultural evaluation datasets, revealing significant gaps in LLMs' understanding of diverse cultural contexts through the development of the challenging CULTURALBENCH-V0.1 dataset. (Zheng et al., 2023) curates LMSYS-Chat-1M, a dataset of one million real-world conversations with 25 LLMs, designed to enhance understanding and development of LLMs capabilities in diverse interaction scenarios. Our work extends the efforts on LLMs evaluation using real-world datasets by employing a code-mixed WhatsApp dataset, reflecting a linguistic phenomena absent in curated datasets. Our evaluation combines quantitative and qualitative analysis of LLMs' performance and decision-making processes.

7 Conclusion

Our study utilized a multilingual and code-mixed WhatsApp dataset to assess the effectiveness of seven LLMs on a sentiment analysis task. Our evaluation includes both quantitative analysis using metrics like F1 score and qualitative assessment of LLMs' explanations for their predictions. Our comparative analysis revealed that, while Mistral-7b and Mixtral-8x7b achieved high F1 scores, they and other LLMs such as GPT-3.5-Turbo, Llama-2-70b, and Gemma-7b struggled with understanding linguistic and contextual nuances, as well as lack of transparency in their decision-making process as observed from their explanations. In contrast, GPT-4 and GPT-4-Turbo excelled in grasping diverse linguistic inputs and managing various contextual information, demonstrating high consistency with human alignment and transparency in their decision-making process. The LLMs however, encountered difficulties in incorporating cultural nuance especially in non-English settings with the GPT-4s doing so inconsistently. Our evaluation, which leverages real-world data, substantiates the robustness observed in NLP benchmarks, particularly highlighting the superior performance of larger models like OpenAI's GPT-4s in handling low-resource and code-mixed languages. The study highlights the importance of using real-world data for LLMs evaluation. In addition, it advocates for combining qualitative methods from Human-Computer Interaction (HCI) with NLP to gain deeper insights into model performance.

Future research should explore the integration of linguistic diversity and cultural intelligence into model training and evaluation frameworks. Additionally, further work is needed to bridge the gap between quantitative performance metrics and qualitative understandings of model behavior. Lastly, future research should also focus on investigating the relationship between explanations and AI decision-making, for example by quantifying correlations across different NLP tasks. This will ensure that future models are not only effective but also interpretable and aligned with human expectations.

Limitations

This study, though comprehensive, has several limitations. It primarily examines texts in Swahili, English, Sheng, and their code-mixed variants, overlooking the vast array of global languages and dialects. Additionally, the focus on seven specific LLMs provides insights but excludes other emerging LLMs. Lastly, while combining quantitative metrics and qualitative analysis, the balance may constrain the depth of qualitative insights due to the dataset's scale and the subjective nature of qualitative evaluation.

Ethics Statement

Given the use of real-world WhatsApp chat data, ethical considerations were paramount. All data was anonymized, screened for sensitive content, and used in accordance with ethical guidelines for research involving human subjects.

Acknowledgements

The authors thank Muchai Mercy, Kemunto Ochwang'i, and Faith Ngetich for their invaluable help with analyzing the LLMs' interpretative capabilities.

References

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. 2023a. 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.

Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Maxamed Axmed, Kalika Bali, and Sunayana Sitaram. 2023b. Megaverse: Benchmarking large language models across languages, modalities, models and tasks. *ArXiv*, abs/2311.07463.

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Martin Chadwick, Gaurav Singh Tomar, Xavier Garcia, Evan Senter, Emanuel Taropa, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Yujing Zhang, Ravi Addanki, Antoine Miech, Annie Louis, Laurent El Shafey, Denis Teplyashin, Geoff Brown, Elliot Catt, Nithya Attaluri, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaly Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, Hanzhao Lin, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yong Cheng, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan,

Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, YaGuang Li, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Gamaleldin Elsayed, Ed Chi, Mahdis Mahdieh, Ian Tenney, Nan Hua, Ivan Petrychenko, Patrick Kane, Dylan Scandinaro, Rishub Jain, Jonathan Uesato, Romina Datta, Adam Sadovsky, Oskar Bunyan, Dominik Rabiej, Shimu Wu, John Zhang, Gautam Vasudevan, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Betty Chan, Pam G Rabinovitch, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Sahitya Potluri, Jane Park, Elnaz Davoodi, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Chris Gorgolewski, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Paul Suganthan, Evan Palmer, Geoffrey Irving, Edward Loper, Manaal Faruqui, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Michael Fink, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marin Georgiev, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Minnie Lui, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Lam Nguyen Thiet, Daniel Andor, Pedro Valenzuela, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Sarmishta Velury, Sebastian Krause, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Tejasi Latkar, Mingyang Zhang, Quoc Le, Elena Allica Abellan, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Sid Lall, Ken Franko, Egor Filonov, Anna Bulanova, Rémi Leblond, Vikas Yadav, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Hao Zhou, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Jeremiah Liu, Mark Omernick, Colton Bishop, Chintu Kumar, Rachel Sterneck, Ryan Foley, Rohan Jain, Swaroop Mishra, Jiawei Xia, Taylor Bos, Geoffrey Cideron, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Petru Gurita, Hila Noga, Premal Shah, Daniel J. Mankowitz, Alex Polozov, Nate Kushman, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Anhad Mohananey, Matthieu Geist, Sidharth Mudgal, Sertan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Quan Yuan, Sumit Bagri, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Aliaksei Severyn, Jonathan Lai, Kathy Wu, Heng-Tze Cheng, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Mark Geller, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Andrei Sozanschi, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Abhimanyu Goyal, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Sabaer Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Tao Zhu, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Dustin Tran, Yeqing Li, Nir Levine, Ariel Stolovich, Norbert Kalb, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Balaji Lakshminarayanan, Charlie Deck, Shyam Upadhyay, Hyo Lee, Mike Dusenberry, Zonglin Li, Xuezhi Wang,

Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Summer Yue, Sho Arora, Eric Malmi, Daniil Mirylenka, Qijun Tan, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Steven Zheng, Francesco Pongetti, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Ragha Kotikalapudi, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Chenkai Kuang, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Pei Sun, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Ishita Dasgupta, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Yuan Liu, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Ivo Penchev, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Adam Kurzrok, Lynette Webb, Sahil Dua, Dong Li, Preethi Lahoti, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Taylan Bilal, Evgenii Eltyshev, Daniel Balle, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Adams Yu, Christof Angermueller, Xiaowei Li, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Kevin Brooks, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Komal Jalan, Dinghua Li, Ginger Perng, Blake Hechtman, Parker Schuh, Milad Nasr, Mia Chen, Kieran Milan, Vladimir Mikulik, Trevor Strohman, Juliana Franco, Tim Green, Demis Hassabis, Koray Kavukcuoglu, Jeffrey Dean, and Oriol Vinyals. 2023. Gemini: A family of highly capable multimodal models.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 2020-December.

Yu-Chu Chang, Xu Wang, Jindong Wang, Yuanyi Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Weirong Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qian Yang, and Xingxu Xie. 2023. A survey on evaluation of large language models. *ArXiv*, abs/2307.03109.

Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. 2023. Unleashing the potential of prompt engineering in large language models: a comprehensive review. *arXiv preprint arXiv: 2310.14735*.

Yu Ying Chiu, Liwei Jiang, Maria Antoniak, Chan Young Park, Shuyue Stella Li, Mehar Bhatia, Sahithya Ravi, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2024. Culturalteaming: Ai-assisted interactive red-teaming for challenging llms' (lack of) multicultural knowledge. *ArXiv*, abs/2404.06664.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Common Crawl. 2023. Common crawl language statistics. `https://commoncrawl.github.io/cc-crawl-statistics/plots/languages`. Accessed: 10-02-2024.

Andy Crabtree, David M. Nichols, Jon O'Brien, Mark Rouncefield, and Michael B. Twidale. 2000. Ethnomethodologically informed ethnography and information system design. *Journal of the American Society for Information Science and Technology*, 51(7):666–682.

Andy Crabtree, Jacki O'Neill, Peter Tolmie, Stefania Castellani, Tommaso Colombino, and Antonietta Grasso. 2006. The practical indispensability of articulation work to immediate and remote help-giving. In *Conference on Computer Supported Cooperative Work*.

A. Seza Doğruöz, Sunayana Sitaram, Barbara E. Bullock, and Almeida Jacqueline Toribio. 2023. A survey of code-switching: Linguistic and social perspectives for language technologies. In *Annual Meeting of the Association for Computational Linguistics*.

Rishav Hada, Varun Gumma, Adrian de Wynter, Harshita Diddee, Mohamed Ahmed, Monojit Choudhury, Kalika Bali, and Sunayana Sitaram. 2023. Are large language model-based evaluators the solution to scaling up multilingual evaluation?

Amr Hendy, Mohamed Gomaa Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *ArXiv*, abs/2302.09210.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, L'elio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *ArXiv*, abs/2401.04088.

Arshad Kaji and Manan Shah. 2023. Contextual code switching for machine translation using language models.

Naveena Karusala, David Odhiambo Seeh, Cyrus Mugo, Brandon L Guthrie, Megan Andreas Moreno, Grace C John-Stewart, Irene Inwani, Richard J. Anderson, and Keshet Ronen. 2021. "that courage to encourage": Participation and aspirations in chat-based peer support for youth living with hiv. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*.

Viet Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Nguyen. 2023. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, Xinyi Wu, Enrico Shippole, Kurt Bollacker, Tongshuang Wu, Luis Villa, Sandy Pentland, and Sara Hooker. 2023. The data provenance initiative: A large scale audit of dataset licensing & attribution in ai.

Bodhisattwa Prasad Majumder, Oana-Maria Camburu, Thomas Lukasiewicz, and Julian McAuley. 2021. Knowledge-grounded self-rationalization via extractive and natural language explanations. In *International Conference on Machine Learning*.

David Martin, Benjamin V. Hanrahan, Jacki O'Neill, and Neha Gupta. 2014. Being a turker. In *CSCW 2014 - Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing*, Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW, pages 224–235. Association for Computing Machinery. Copyright: Copyright 2014 Elsevier B.V., All rights reserved.; 17th ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW 2014 ; Conference date: 15-02-2014 Through 19-02-2014.

Gemma Team Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, L. Sifre, Morgane Riviere, Mihir Kale, J Christopher Love, Pouya Dehghani Tafti, L'eonard Hussenot, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Am'elie H'eliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Cl'ement Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikula, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Pier Giuseppe Sessa, Rahma Chaabouni, Ramona

Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vladimir Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Brian Warkentin, Ludovic Peran, Minh Giang, Cl'ement Farabet, Oriol Vinyals, Jeffrey Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology. *ArXiv*, abs/2403.08295.

Ishani Mondal, Kalika Bali, Mohit Jain, Monojit Choudhury, Ashish Sharma, Evans Gitau, Jacki O'Neill, Kagonya Awori, and Sarah Gitau. 2021. A linguistic annotation framework to study interactions in multilingual healthcare conversational forums. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 66–77, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. Wt5?! training text-to-text models to explain their predictions. *ArXiv*, abs/2004.14546.

Jessica Ojo, Kelechi Ogueji, Pontus Stenetorp, and David I Adelani. 2023. How good are large language models on african languages? *arXiv preprint arXiv:2311.07978*.

Jacki O'Neill and David Martin. 2003. Text chat in action. In *Proceedings of the 2003 ACM International Conference on Supporting Group Work*, GROUP '03, page 40–49, New York, NY, USA. Association for Computing Machinery.

OpenAI. 2023a. Gpt-4 technical report.

OpenAI. 2023b. Gpt4 technical report.

James Orao. 2012. The kenyan indigenous languages and the mass media: Challenges and opportunities.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155.

Shana Poplack. 2001. *Code Switching: Linguistic*, pages 2062–2065.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.

Koustav Rudra, Shruti Rijhwani, Rafiya Begum, Kalika Bali, Monojit Choudhury, and Niloy Ganguly. 2016. Understanding language preference for expression of opinion and sentiment: What do Hindi-English speakers do on Twitter? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1131–1141, Austin, Texas. Association for Computational Linguistics.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.

Together Computer. 2023. Redpajama: an open dataset for training large language models.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Haryo Akbarianto Wibowo, Erland Hilman Fuadi, Made Nindyatama Nityasya, Radityo Eko Prasojo, and Alham Fikri Aji. 2023. Copal-id: Indonesian language reasoning with local culture and nuances. *ArXiv*, abs/2311.01012.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2022. Reframing human-AI collaboration for generating free-text explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 632–658, Seattle, United States. Association for Computational Linguistics.

Sarah Wiegreffe, Ana Marasović, and Noah A. Smith. 2021. Measuring association between labels and free-text rationales. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10266–10284, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ruochen Zhang, Samuel Cahyawijaya, Jan Christian Blaise Cruz, Genta Winata, and Alham Aji. 2023. Multilingual large language models are not (yet) code-switchers. In *Proceedings of the 2023*

Conference on Empirical Methods in Natural Language Processing, pages 12567–12582, Singapore. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P. Xing, Joseph E. Gonzalez, Ion Stoica, and Haotong Zhang. 2023. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. *ArXiv*, abs/2309.11998.

A Appendix

A.1 The Prompt

```
You are a helpful NLP assistant, specializing in Sentiment Analysis. You are provided with a WhatsApp chat message (QUERY) in English, Swahili,
Sheng, or in more than one language (code-mixed), along with the defintions about the sentiment classes. Your task is to analyze the message
and categorize it as Positive, Negative, or Neutral based on the sentiment expressed, along with a justification. Make sure to highlight the
words/span of text in the query that you used to make your decision in your justification.

⟨DEFINITIONS⟩
**Negative Sentiment**: It expresses some sort of negative feeling or view or opinion about someone or something.

**Neutral Sentiment**: It neither expresses a positive nor a negative sentiment of the speaker. It could be a general comment, acknowledgement,
chitchat or any factual advice or a simple greeting.

**Positive Sentiment**: The sentiment needs to be classified as positive if the speaker feels strong and positive at any particular utterance,
except the normal aspects such as any form of greetings.
⟨/DEFINITIONS⟩


QUERY: "{query}"


{output_format_instructions}
**DO NOT OUTPUT ANYTHING OTHER THAN THE JSON OBJECT**
```

Figure 3: LangChain prompt for Sentiment Analysis. We randomize the order of the definitions to alleviate position bias.

A.2 Description of Human Evaluation Criteria

In Table 4, we provide a brief description of each of the five rubrics for human evaluation we adopted as outlined by (Chang et al., 2023) on *'how to evaluate'*.

Evaluation Criteria	Description
Linguistic accuracy	LLM's capacity for precise linguistic interpretation and generation, covering grammar, vocabulary, idioms, and language-specific nuances, while ensuring factual accuracy.
Contextual and cultural relevance	LLM's ability to provide contextually and culturally relevant justifications in sentiment analysis, ensuring responses are appropriate and significant to the given context.
Fluency in maintaining consistency	LLM's fluency in producing consistent and logical justifications across various sentiment analysis cases, ensuring smooth content flow and uniform tone.
Alignment with human expectations	LLM's ability to produce justifications aligned with human reasoning ensures ethically appropriate predictions, reflecting human values and societal norms, fostering trust in sensitive applications like sentiment analysis.
Transparency in LLM's decision-making process	LLM's ability to clearly and openly communicate its decision-making process, enabling users to understand the rationale behind responses and gain insights into its inner workings.

Table 4: Description of Human Evaluation Criteria.

A.3 LLMs' F1 Scores Across Languages

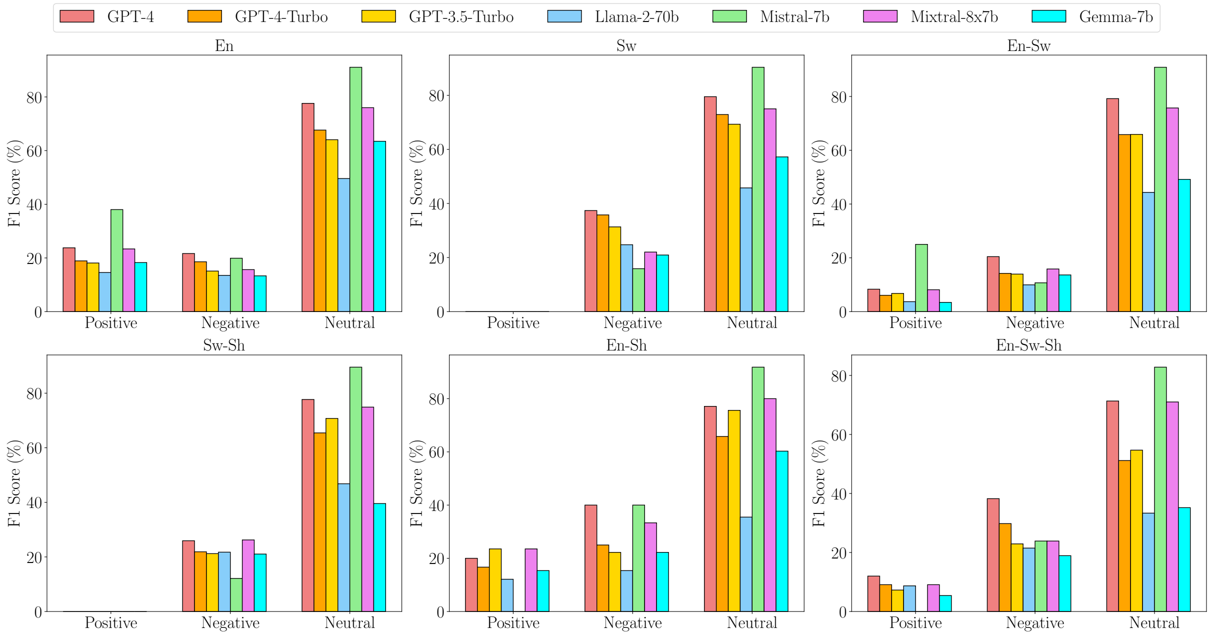

Figure 4: Models' F1 scores across languages for Positive, Negative, and Neutral sentiments.

A.4 Detailed Qualitative Analysis of Model Explanations Across Multilingual Examples

Message	GPT-4	GPT-4-Turbo	GPT-3.5-Turbo	Llama-2-70b	Mistral-7b	Mixtral-8x7b	Gemma-7b
Original: *"Mm nimeboeka acha ni lale gud 9t"* **Translation:** *"Am bored let me sleep ..goodnight"* **Actual Sentiment:** Negative	The phrase 'nimeboeka' translates to 'I'm bored' in English, which indicates a negative sentiment. The speaker is expressing dissatisfaction or a lack of interest, which is a negative feeling. **Prediction:** Negative	The phrase 'nimeboeka' which means 'I'm bored' indicates a negative feeling or mood. The speaker's intention to go to sleep ('acha ni lale') could be seen as a way to escape this boredom, which further emphasizes the negative sentiment. The addition of 'gud 9t' (good night) does not change the overall sentiment as it is a common phrase used before going to sleep and does not express a strong positive feeling as per the definitions provided. **Prediction:** Negative	The sentiment is categorized as Neutral. The message does not express a strong positive or negative feeling. It seems to be a casual comment about feeling tired and going to sleep. The use of 'acha ni lale gud 9t' indicates the speaker's intention to go to sleep, which is a neutral statement. **Prediction:** Neutral	The message expresses a positive sentiment due to the use of positive words such as 'Mm', 'ni lale', and 'gud'. **Prediction:** Positive	The given query is in Swahili language and translates to 'I have not found it yet good 9t'. The sentiment expressed is neutral as the speaker is simply stating a fact about their current situation. **Prediction:** Neutral	The query 'Mm nimeboeka acha ni lale gud 9t' is in Sheng, a slang language spoken in Kenya. The phrase roughly translates to 'I have given up, it's late now, 9 o'clock'. The sentiment expressed in this message is Negative as the speaker expresses giving up and the lateness of the time could indicate a negative feeling or situation. The word 'acha' (give up) is particularly indicative of a negative sentiment. **Prediction:** Negative	The phrase 'Mm nimeboeka acha ni lale gud 9t' is a Sheng phrase that translates to 'I am not feeling well'. The speaker is expressing negative sentiment due to the use of the word 'acha' and 'gud', which are both words used to describe feeling unwell. The phrase also includes a reference to the number '9t', which is commonly used in Sheng to express feeling sick. Therefore, I classify the sentiment of the message as Negative. **Prediction:** Negative

Table 5: This example involves an English-Swahili message with various textual irregularities including grammatical errors (*'ni lale'* instead of *'nilale'*), local abbreviations (*'Mm'* instead of *'Mimi'*), and misspellings (*'gud 9t'* instead of *'good night'*, *'acha'* instead of *'wacha'*). GPT-4-Turbo excelled in this challenging context by demonstrating its robust multilingual capabilities and resilience to noise, leading to both correct prediction and interpretation. GPT-4 similarly managed to accurately identify the key sentiment indicator in the message. However, the remaining models including GPT-3.5-Turbo, Llama-2-70b, Mistral-7b, Mixtral-8x7b, and Gemma-7b showed weaker contextual and linguistic understanding, as well as a lack of robustness to noise. As a result, GPT-3.5-Turbo and Mistral-7b made incorrect predictions, while Mixtral-8x7b and Gemma-7b, despite correct predictions, failed in their interpretation.

Message	GPT-4	GPT-4-Turbo	GPT-3.5-Turbo	Llama-2-70b	Mistral-7b	Mixtral-8x7b	Gemma-7b
Original: *"tuache iyo story ju ishaanza kuniboo."* **Translation:** *"Let's leave that story it is starting to bore me."* **Actual Sentiment:** Negative	The phrase 'ishaanza kuniboo' in the message indicates that the speaker is getting bored or annoyed with the story, which is a negative sentiment. **Prediction:** Negative	The phrase 'ishaanza kuniboo' indicates that the speaker is getting bored or is already bored with the story being discussed. The word 'kuniboo' is a Sheng term that translates to 'boring me' in English, which conveys a negative sentiment. **Prediction:** Negative	The sentiment expressed in the message is negative. The phrase 'tuache iyo story' indicates a desire to stop or leave a certain situation, and 'ishaanza kuniboo' suggests that something negative has started happening to the speaker. Therefore, the overall sentiment is negative. **Prediction:** Negative	The word 'tuache' means 'forget' in Swahili, and the phrase 'iyo story' means 'that story' or 'that thing'. The word 'ju' is an interrogative pronoun meaning 'what'. Therefore, the overall sentiment of the sentence is Neutral because it is asking about something in a factual manner without expressing any positive or negative emotion. **Prediction:** Neutral	The given query in Swahili language 'tuache iyo story ju ishaanza kuniboo' translates to 'leave that story alone' in English. This is a command or instruction and does not express any positive or negative sentiment. **Prediction:** Neutral	The Swahili phrase 'tuache iyo story' translates to 'let's leave that story' in English. The word 'leave' often indicates a negative sentiment as it suggests abandonment or disinterest. The word 'that story' is a reference to a previous conversation or event, which based on the sentiment expressed, seems to be something the speaker wants to move away from. Therefore, the sentiment of this message is Negative. **Prediction:** Negative	The phrase 'tuache iyo story ju ishaanza kuniboo' is a Sheng phrase that translates to 'you are telling a story that is not true'. The speaker is expressing disbelief and negativity towards the story being told. The words 'ishaanza' and 'kuniboo' are the key words used to convey this sentiment. **Prediction:** Negative

Table 6: This message showcases a complex mix of English, Swahili, and Sheng, with the Swahili-Sheng term *'kuniboo'* serving as a key sentiment indicator. GPT-4 and GPT-4-Turbo excelled in this context, demonstrating strong multilingual capabilities by accurately understanding both the linguistic elements and the contextual nuances, leading to correct predictions and interpretation. GPT-3.5-Turbo and Mixtral-8x7b have made a correct prediction, however, failed to recognize the significance of *'kuniboo'*, likely due to a lack of understanding of this specific term. The remaining models, including Llama-2-70b, Mistral-7b, and Gemma-7b, also missed the critical non-English sentiment indicator. Consequently, Llama-2-70b and Mistral-7b made incorrect predictions with incorrect interpretation, whereas Gemma-7b managed a correct prediction but with flawed interpretation.

Message	GPT-4	GPT-4-Turbo	GPT-3.5-Turbo	Llama-2-70b	Mistral-7b	Mixtral-8x7b	Gemma-7b
Original: *"Send to everyone you love whether it's real love or friendship love, brotherly love, sisterly love, child love, fatherly love and motherly love. You are lovable if you get FIVE sent back to you !"* **Actual Sentiment:** Neutral	The message is encouraging the recipient to spread love to everyone they care about, and it also assures them that they are lovable. The phrases 'Send to everyone you love' and 'You are lovable' indicate a positive sentiment. **Prediction:** Positive	The message encourages people to send love in various forms and implies that receiving love back is a positive outcome by stating 'You are lovable if you get FIVE sent back to you'. The words 'love', 'lovable', and the concept of sending love to others and receiving it back suggest a positive sentiment. **Prediction:** Positive	The sentiment expressed in the message is Positive. The speaker is expressing love and positivity towards everyone they care about, including family and friends. The phrase 'Send to everyone you love' and the mention of different types of love such as brotherly, sisterly, child, fatherly, and motherly love all contribute to the positive sentiment expressed in the message. **Prediction:** Positive	The message expresses a positive sentiment as it encourages the reader to spread love and mentions receiving love in return. The use of the word 'love' multiple times emphasizes the positive emotion. Additionally, the phrase 'you are lovable' suggests that the speaker feels positively about the recipient. **Prediction:** Positive	The message expresses a positive sentiment towards love and the importance of receiving love in return. The speaker encourages sending love to various relationships and considers being loved as a positive attribute. **Prediction:** Positive	The message expresses a positive sentiment by encouraging the reader to share love with others, whether it be romantic, familial, or platonic. The phrase 'You are lovable if you get FIVE sent back to you' also adds to the positive sentiment as it implies that the recipient is loved and appreciated by others. The use of positive words such as 'love', 'lovable', and 'five' contribute to the overall positive tone of the message. **Prediction:** Positive	The message expresses strong positive sentiment due to the use of words like 'love', 'lovable', and 'FIVE sent back to you'. The message also emphasizes the feeling of inclusivity and affection towards various relationships. **Prediction:** Positive

Table 7: This example focuses on the challenges posed by informational-neutral messages that, despite being predominantly in English, contain ambiguous sentiments that may be interpreted as either positive or negative. This ambiguity leads to discrepancies between the models' predictions and the actual annotations. Despite these challenges, the example also highlights all models' strong linguistic understanding of English, showcasing their ability to handle the language even in nuanced scenarios.

Message	GPT-4	GPT-4-Turbo	GPT-3.5-Turbo	Llama-2-70b	Mistral-7b	Mixtral-8x7b	Gemma-7b
Original: *"Me siyo beste yako"* **Translation:** *"Am not your friend"* **Actual Sentiment:** Negative	The phrase 'Me siyo beste yako' translates to 'I am not your friend' in English. This statement expresses a negative sentiment as it indicates a rejection or negative view towards the person being addressed. **Prediction:** Negative	The phrase 'Me siyo beste yako' in Sheng translates to 'I am not your friend' in English. The use of 'siyo' (not) indicates a denial or negation of friendship, which conveys a negative sentiment towards the relationship with the person being addressed. **Prediction:** Negative	The sentiment is Negative. The phrase 'siyo beste yako' in Swahili translates to 'not your best', which expresses a negative opinion about someone or something. **Prediction:** Negative	The phrase 'Me siyo beste yako' is a Swahili phrase that translates to 'I don't have your best interests at heart' in English. The use of the phrase implies that the speaker does not have the listener's well-being in mind, and may even have negative intentions. Therefore, the sentiment of this phrase is Negative. **Prediction:** Negative	The query 'Me siyo beste yako' in Swahili translates to 'I am not your best' in English. This statement is a simple negation and does not express any strong positive or negative sentiment. It could be a response to a request for compliments or an expression of humility. **Prediction:** Neutral	The query 'Me siyo beste yako' is in Sheng, a code-mixed language in Kenya. The phrase can be translated to 'I am not your best' in English. The use of 'not' indicates a negative sentiment. However, the sentiment is not strong as it's a statement of fact rather than an expression of a negative emotion. Therefore, the sentiment is Negative. **Prediction:** Negative	The phrase 'Me siyo beste yako' is a Swahili phrase that translates to 'I am your best friend' in English. The phrase expresses a strong positive sentiment. The words 'beste' and 'yako' are the key words that I used to make this decision. **Prediction:** Positive

Table 8: This example demonstrates the difficulties all models face when processing the Sheng language, particularly for those with weaker linguistic understanding of non-English languages. Only GPT-4 and GPT-4-Turbo successfully identified and correctly interpreted the language, highlighting the significant variability in multilingual capabilities across the models.