Reasoning Beyond Labels: Measuring LLM Sentiment in Low-Resource, Culturally Nuanced Contexts

Millicent Ochieng[1], **Anja Thieme**[1], **Ignatius Ezeani**[2], **Risa Ueno**[1], **Samuel Maina**[1], **Keshet Ronen**[3], **Javier González**[1], **Jacki O'Neill**[1]

[1]Microsoft Research, [2]Lancaster University, [3]University of Washington

Abstract

Sentiment analysis in low-resource, culturally nuanced contexts challenges conventional NLP approaches that assume fixed labels and universal affective expressions. We present a diagnostic framework that treats sentiment as a context-dependent, culturally embedded construct, and evaluate how large language models (LLMs) reason about sentiment in informal, code-mixed WhatsApp messages from Nairobi youth health groups. Using human-annotated data, sentiment-flipped counterfactuals, and rubric-based explanation evaluation, we probe LLM interpretability, robustness, and alignment with human reasoning. Framing our evaluation through a social science measurement lens, we operationalize LLM outputs as an instrument for measuring the abstract concept of sentiment. Our findings reveal significant variation in model reasoning quality, with top-tier LLMs demonstrating greater interpretive stability, while smaller open-weight models in our study show reduced stability under ambiguity or sentiment shifts. This work highlights the need for culturally sensitive, reasoning-aware AI evaluation in complex, real-world communication.

1 Introduction

Sentiment analysis is a prevalent NLP technique used to obtain meaningful information and semantics from text (Onyenwe et al., 2020). It is often conflated with emotion detection (Nandwani and Verma, 2021); or opinion mining such as consumer sentiment (Burnham, 2024). Instead, sentiment analysis primarily determines polarity in the intent behind a written message often characterized as positive, negative, or neutral (Nandwani and Verma, 2021).

Speech Act Theory by Austin (1975) further highlights that what a reader understands from a message depends on words choices; their individual meanings; ordering; as well as lexical or syntactic variations. Combined, these introduce significant ambiguity on how inferences about a message are drawn (Corvi et al., 2025). Moreover, the social semiotics theory by Halliday (2014) emphasizes that sentiment is not just a linguistic phenomenon; it is also deeply embedded in social and cultural contexts, which highlights how emotions are conveyed and interpreted based on cultural norms and values (Zhang, 2024).

In this paper, we acknowledge that interpreting or measuring sentiment can be difficult —particularly in informal, multilingual, under-resourced, and culturally nuanced communication contexts. Expressions of emotion and attitude are shaped by local language practices, shared cultural knowledge, and interactional context (Matsumoto, 1990; Lindquist, 2021; Fang et al., 2022). In real-world communications such as youth chat, social media, or hyperlocal exchanges among multilingual speakers, language is frequently code-mixed[1], fluid, and shaped by the moment—that is, influenced by who is speaking and who is listening, the topic being discussed, the speaker's emotional tone, or intentions at that time, and the setting (e.g., online chat vs. in-person talk). Meanings are negotiated, implicit, and frequently ambiguous—making sentiment difficult to interpret, even for humans, espe-

[1]The practice where multilingual speakers fluidly shift between languages in conversation

cially when removed from the original platform or context of exchange (O'Neill and Martin, 2003). These complexities do not just complicate classification—they challenge the very *measurement* of sentiment. As argued by Wallach et al. (2025), evaluating GenAI models requires treating such tasks as a social science measurement challenge, where abstract, culturally-contested concepts must be systematically defined and carefully connected to observable indicators.

In our work, we treat **sentiment** not as a fixed label, but as a context-dependent expression of intent. It may be explicit (e.g., "I'm so angry right now"), but more often in our dataset, it appears through muted cues (e.g., "You're always online")—subtle, culturally and contextually situated, and open to interpretation. We define **ambiguity** as cases where the intended sentiment is unclear, underspecified, open to multiple readings, or leads to disagreement even among culturally fluent, context-aware annotators—not because the language is misunderstood, but due to differing interpretations of tone or social context (see Table 1).

We use **cultural nuance** to describe how language practices, religious or affective expressions, and shared social knowledge shape how sentiment is conveyed and perceived. In our dataset, such nuance is embedded within: practices of *code-mixing* (e.g., "kama hauko school shindaapo!!")[2]; *local shorthand* (e.g., mm for mimi)[3]; *emoji-only or emoji-enhanced* messages via graphical symbols (e.g.," 😅") or their textual counterpart the *emoticon* (e.g., ":)") (Liu et al., 2021; Yoo and Rayz, 2021); *irony*; and *youth-specific slang* (Sheng)[4]. These elements are often combined to produce rich, but difficult-to-classify, sentiment signals; and these cultural complexities are evident throughout our dataset (see Table 1 and Appendix Table 11)

So far, standard sentiment analysis treats sentiment as a fixed classification problem with a single, context-independent "ground truth" (Mohammad, 2017; Wankhade et al., 2022; Sharma et al., 2024). Recent exploratory work on similar multilingual, code-mixed WhatsApp data has examined LLM sentiment classification and qualitative reasoning (Ochieng et al., 2025), but without a diagnostic framework, robustness testing, or measurement-oriented analysis. Despite advances in LLMs (Brown et al., 2020; Touvron et al., 2023; Jiang et al., 2023; OpenAI, 2023), sentiment evaluation remains label-centric, with metrics like accuracy and F1 obscuring how models reason and whether their decisions align with human interpretation.

We propose a diagnostic approach to sentiment analysis that treats LLMs not only as classifiers but as tools for structuring and probing sentiment in complex, real-world communication. Informed by Wallach et al.'s measurement framework (Wallach et al., 2025) that separates the *conceptualization* of sentiment from its *operationalization*. Our goal is to shift how sentiment is measured in LLMs—from fixed label prediction toward a more interpretive, ambiguity-aware framework. We ask: How do LLMs reason about sentiment in real-world, culturally grounded messages?

To achieve this, we investigate how LLMs reason about sentiment, how they explain their judgments, handle ambiguity, and echo human disagreement. For instance, while a traditional classifier might label "sawa tu 😐"[5] as neutral, our framework surfaces the emotional nuance by analyzing emoji, tone, and context, revealing how such utterances can signal quiet frustration or withdrawal. We analyze model explanations, confidence scores, and token-level highlights indicating which parts of the message influenced the model's judgment, across three evaluation settings: messages with annotator agreement (Gold), disagreement (Ambiguous), and **sentiment-flipped counterfactuals**[6] **(Synthetic)**. These *synthetic* examples are automatically generated by prompting GPT-4 to rewrite real WhatsApp messages in our dataset with their *sentiment flipped (positive to negative or vice versa)*—while preserving meaning, cultural tone and informal language. We guide this process using a structured taxonomy of sentiment-bearing components (e.g., negation, emoji, tone, key phrases; see Appendix A.8, Table 12). These counterfactual flips serve as our operationalization of the sentiment concept. Applied in testing whether models respond appropriately to affective changes, we use a dual evaluation protocol with human annotators and LLM-as-a-judge to assess counterfactual plausibility and explanation

[2]Swahili-English: "If you're not in school, stay there." While casual, this often conveys a dismissive stance, reflecting cultural norm that link education to intellectual legitimacy.

[3]*mimi* means "me" in Swahili

[4]An urban slang spoken by youth in Kenya, blending Swahili, English, and local languages.

[5]*sawa tu* means "just okay," but can imply resignation or frustration depending on tone and context.

[6]A *counterfactual* is a sentiment-flipped variant of a real message.

Example	Complexities	Annotator 1	Annotator 2
🤣 🤣 🤣*uyu sasa anachoma manzee* "🤣🤣🤣 this guy is now messing up, bro"	**Shorthand:** "uyu" instead of "huyu" **Urban slang (Sheng):** "anachoma", "manzee" **Tone:** *friendly teasing, mockery, or social critique* **Emoji use:** 🤣 **Code-mixing:** Swahili-Sheng blend **Cultural reference:** *Assumes shared understanding of local slang, social behaviors, and norms*	**Label:** Negative **Notes:** *We see the ridicule and embarrassment from the persona and the audience despite the laugh.*	**Label:** Positive **Notes:** *Expresses criticism with amusement portrayed with laughing emojis.*
Nmeacha izea "I'm sorry, I have stopped"	**Code-mixing:** Swahili-Sheng blend **Urban slang (Sheng):** "izea" **Shorthand:** "nmeacha" **Tone:** flat or understated **Ambiguity:** lacks strong emotional cues	**Label:** Neutral **Notes:** *We see a casual apology that doesn't express strong emotion.*	**Label:** Positive **Notes:** *Speaker is apologetic and remorseful.*
U can't see the future but God can "You can't see the future but God can"	**Shorthand:** "U" for "you" **Religious expression:** appeals to divine foresight **Tone:** factual or reassuring **Cultural context:** common in faith-based communication **Ambiguity:** sentiment depends on interpretation of tone/intention	**Label:** Neutral **Notes:** *A remark without strong personal emotion.*	**Label:** Positive **Notes:** *Speaker expresses trust in God, offering reassurance.*
Hello, guys yani mko tu na mmenyamaza?? "Hello, guys are online and you are quiet?"	**Code-mixing:** Swahili-English blend **Tone:** questioning, possibly sarcastic **Social cue:** expectation of group participation **Ambiguity:** tone varies between concern and frustration	**Label:** Negative **Notes:** *We see disappointment and negative shock from the persona on why people are so quiet.*	**Label:** Neutral **Notes:** *Expresses concern and curiosity on the silence of the group.*
Yes I eat too much iz it normal "Yes I eat too much is it normal"	**Shorthand:** "iz" for "is", informal tone **Self-disclosure:** reveals possible worry **Ambiguity:** phrased as a question, unclear tone; genuine concern vs casual comment	**Label:** Negative **Notes:** *We see worry and distress about too much eating, suggests a negative sentiment.*	**Label:** Neutral **Notes:** *Question seeking clarification.*

Table 1: Examples of annotator disagreement illustrating cultural and linguistic complexities.

quality.

This paper makes the following contributions:

- We adapt a social science measurement lens to evaluate model reasoning about language, reframing sentiment analysis as a problem of concept systematization and measurement.
- We introduce a diagnostic framework to analyze how LLMs reason about sentiment in informal, code-mixed, and culturally embedded communication. This involves creating synthetic data using a counterfactual approach based on a taxonomy of sentiment components (e.g., negation, emoji, tone).
- We propose a dual evaluation protocol with human annotators and an LLM-as-a-judge to assess explanation quality and counterfactual plausibility. Through this, we identify reasoning inconsistencies in LLMs, distinguishing between reducible errors and irreducible ambiguity across evaluation settings.

2 Related Work

Sentiment Analysis in Informal and Multilingual Communication: While sentiment analysis has largely focused on English-language data from structured domains such as reviews or news, real-world communication in informal, multilingual, and code-mixed contexts presents deeper challenges (Choudhary et al., 2018). Prior work on code-mixed sentiment (e.g., Swahili-English, Hindi-English) has highlighted the need for inclusive resources (Zhang et al., 2023; Doğruöz et al., 2023a,b; Kaji and Shah, 2023), yet low-resource, conversational data in health or community settings remains underexplored. Recent exploratory work on multilingual, code-mixed WhatsApp data from Nairobi youth examined LLM-based sentiment classification and qualitative reasoning using standard metrics and manual inspection (Ochieng et al., 2025). Building on this line of work, we move beyond exploratory analysis by introducing a structured diagnostic framework that evaluates how LLMs reason about sentiment across diverse evaluation settings, explicitly models ambiguity, and probes robustness via sentiment-flipped counterfactuals assessed with shared human and LLM-based rubrics.

Evaluating LLM Reasoning: Traditional sentiment evaluation relies on metrics like accuracy and F1, which fail to capture how models reason, especially in ambiguous or culturally situated cases (Lyu et al., 2024). To address this, recent work has explored explanation-based evaluation through token attribution, rationales, and confidence scores (Joshi et al., 2023; Dhaini et al., 2025). Other work

has shown that LLMs like GPT-4 can serve as evaluators, often approximating human ratings in generation tasks (Liu et al., 2023). However, a missing component in this literature is the use of dual evaluation protocols that involve both human and LLM judges applying shared rubrics. Such approaches are particularly valuable in settings with annotator disagreement, where interpretive alignment matters more than single-label accuracy. Our work builds on and extends this direction by systematically comparing model and human evaluations across diverse examples, including ambiguous and counterfactually altered messages.

Counterfactuals and Contrastive Evaluation in NLP: Counterfactuals offer a powerful tool for probing model reasoning by introducing minimal, targeted changes to input data (Yang et al., 2021). In sentiment analysis, these typically flip polarity through shifts in tone, negation, or word choice. While prior work often relied on rule-based or synthetic constructions (Yang et al., 2021), we use GPT-4 to generate sentiment-flipped versions of messages—shifting from positive to negative and vice versa—grounded in a taxonomy of transformation types such as emoji use, phrase substitution, and tone modulation. More broadly, our approach aligns with recent work on problem variation as a diagnostic for reasoning (Xu et al., 2025), which emphasizes the need for systematic, multi-level perturbations, including counterfactuals to reveal model limitations beyond memorization.

3 Evaluation as Measurement: Experimental Setup

3.1 Dataset and Annotation

We build on the WhatsApp Chat Dataset originally collected by Karusala et al. (2021) and annotated by Mondal et al. (2021), which comprises multilingual conversations among young people living with HIV in informal settlements in Nairobi, Kenya. These discussions, drawn from two health-focused WhatsApp groups moderated by a medical facilitator, are informal, context-rich, and code-mixed across English, Swahili, and Sheng. All messages were anonymized, and ethical protocols from the original collection were strictly followed. The dataset is not publicly released due to sensitivity, but researchers may request access for academic use.

For this study, we developed a structured annotation protocol focused on culturally grounded sentiment, interpretive ambiguity, and context-specific expression. Designed through iterative pilot testing and calibrator discussions (see Appendix A.1). Two trained annotators — Kenyan youth aged 20–24 — labeled each message for sentiment (positive, negative, neutral), provided English translations where needed, and tagged word-level language identifiers. Messages with annotator disagreement were retained for targeted evaluation. From the full dataset of 6,197 messages, we define three evaluation subsets: the **Gold Set** (6,121 messages with full annotator agreement), the **Ambiguous Set** (76 messages with disagreement), and the **Synthetic Set** (sentiment-flipped messages generated from a pool of 1,547 non-neutral messages using GPT-4), see Table 5. No post-processing is applied to normalize emojis, punctuation, or shorthand expressions, as these elements are integral to the communicative and emotional tone of the data.

3.2 Task and Model Setup

We frame sentiment analysis as a multi-class classification task over informal, multilingual WhatsApp messages. Given an input message, the model is prompted to predict a sentiment label (positive, negative, or neutral) and to generate a natural language explanation (max 200 words). The task is performed via in-context learning using few-shot prompting, with manually selected examples from the Gold Set that reflect the natural mix of Swahili, Sheng, and English, including both clear and mildly ambiguous cases (see Table 7 in the Appendix). Pilot comparisons of three prompting strategies (definitions with examples, definitions only, and no definitions) showed that prompts combining definitions and examples yielded the most stable and interpretable outputs. We evaluate a range of LLMs varying in architecture and size, including GPT-4-Turbo and GPT-4-32k (OpenAI, 2023), Gemma-3-27B (Team et al., 2025), LLaMA-3-8B (Touvron et al., 2023), Mistral-7B (Jiang et al., 2023), OpenChat-3.5 (Wang et al., 2024), and Phi-4 (Gunasekar et al., 2023). We selected models to reflect realistic usage in low-resource settings, contrasting locally deployable lightweight open-weight models with large proprietary models accessed via APIs. Phi-4 was included despite its English-only marketing due to strong pilot performance on code-mixed Swahili–English text. Each model outputs a sentiment label, a natural language explanation, token-level highlights, and a confidence score[7] scaled

[7] Model-reported confidence where available.

from 0 to 5. Evaluation is conducted across three data partitions: Gold, Ambiguous, and Synthetic.

3.3 Counterfactual Generation Framework

We use sentiment-flipped counterfactuals as a diagnostic stress test, aligned with hypothesis validity testing (Wallach et al., 2025), to evaluate whether models detect and explain controlled shifts in sentiment rather than to construct a new gold-standard dataset. Starting from 1,547 non-neutral messages in the Gold Set, we prompt GPT-4 to generate three sentiment-flipped variants per message, reversing the original polarity (positive to negative or vice versa) while preserving meaning, tone, and conversational style. Some messages require minimal lexical changes, for example, *Napenda wazo lako* ("I like your idea") → *Sipendi wazo lako* ("I dislike your idea"). Others, however, demand deeper shifts in intent or tone, such as *Sema tuu niache kukuaibisha* ("Just tell me to stop embarrassing you") → *Sema tu niendelee kukusifu* ("Just say it, so I continue praising you"). To guide generation across this range of complexity, we developed a taxonomy of sentiment-bearing components, including negation, tone, emoji, and sentiment phrases, which informs the generation prompt (Table 12). Rather than manually selecting outputs, we apply a second GPT-4 prompt that selects the strongest candidate based on plausibility, fluency, and contextual fit (Table 9). This two-step process allows for richer variation at generation time while promoting interpretability and consistency in the resulting counterfactuals. The selected flips constitute the Synthetic Set used for robustness evaluation. Human assessment of a subset of these counterfactuals is conducted later to audit quality and surface limitations (Section 3.4). We discuss the implications of relying on LLM-based generation and filtering in the Limitations Section.

3.4 Human and LLM-as-a-judge Evaluation Protocol

We evaluate model explanations and Synthetic counterfactuals using a structured, rubric-based protocol involving human annotators and GPT-4 as an automated judge. This dual evaluation is designed to assess interpretive quality and counterfactual plausibility, and to compare human and model judgments. For model explanations, two annotators independently rated 480 explanations drawn from the Gold (180), Ambiguous (120), and Synthetic (180) sets. All six models were included where explanations were available. For the Ambiguous set, only four models (`LLaMA-3-8B`, `GPT-4-Turbo`, `GPT-4-32k`, and `Gemma-3-27B`) consistently produced usable outputs, reflecting the difficulty of these cases. Explanations were scored on faithfulness, contextual or cultural appropriateness, logical coherence, and clarity or completeness using a binary (0/1) scale. To audit the quality of the Synthetic Set, six annotators evaluated a sample of 50 sentiment-flipped messages on fluency, naturalness, sentiment flip clarity, and meaning preservation, also using a binary (0/1) scale. This human assessment is intended as a diagnostic quality check that surfaces limitations such as semantic drift or stylistic mismatch, rather than as an exhaustive validation of all generated counterfactuals. GPT-4 was prompted to apply the same rubrics using standardized evaluation instructions (Tables 10 and 8), allowing direct comparison between human and LLM judgments. Within the measurement framework of Wallach et al. (2025), this protocol corresponds to the *interrogation* step, enabling analysis of content validity (alignment between explanations and the sentiment concept) and consequential validity (how explanation quality affects interpretation and use). Full rubric definitions and example annotations are provided in Appendix A.7.

4 Results and Analysis

4.1 Overall Model Performance

For predicting sentiment labels as a baseline, we observe that model coverage[8] varies substantially across settings, especially under counterfactual perturbation revealing a key axis of performance variation (see Table 6). While top-tier models like `GPT-4-Turbo` and `GPT-4-32k`, consistently provide labels for all examples (100% coverage), several open-weight models—most notably `LLaMA-3-8B`—show sharp declines, especially in the Synthetic set, where coverage drops as low as 37.6%. This sharp drop suggests that even fluent, sentiment-flipped rewrites can disrupt model processing, exposing model fragility to subtle language changes in tone, emoji, or phrasing.

We further observe that on the Gold Set, all models achieve strong average F1 scores. The best performance is observed from `GPT-4-32k` (0.90), `Mistral-7B` (0.90), and `Gemma-3-27B` (0.89). Most models maintain balance across sen-

[8]Coverage reflects the percentage of examples for which a model returned a valid sentiment label.

timent classes, but class-specific performance still varies. LLaMA-3-8B underperforms markedly on negative sentiment (0.51), pointing to difficulty detecting more implicit or culturally nuanced negativity. Neutral sentiment is generally the most challenging class, echoing prior findings on underspecified affect and implicit tone. These results establish strong baselines while highlighting gaps in both robustness and class sensitivity that motivate further analysis of model reasoning and explanation quality.

4.2 Reasoning Quality in LLM Explanations

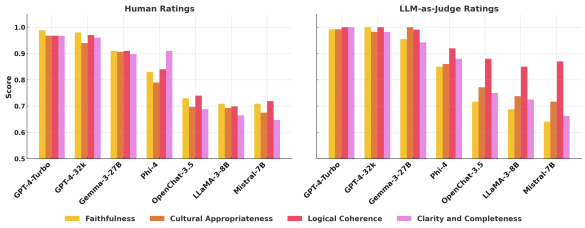

Figure 1: Rubric-based average explanation scores across models.

We evaluated explanation quality of the models reasoning about a message's sentiment using rubric-based scores from both human annotators and GPT-4-based LLM-as-a-judge systems, see Figure 1. Across all models and dimensions, we observe broad agreement in relative rankings between the two rating sources, though LLM-as-a-judge ratings tend to be more generous overall. GPT-4-32k and GPT-4-Turbo consistently achieve top scores across all rubrics, with near-perfect ratings from both humans and LLMs. Gemma-3-27B also performs well, with high ratings for faithfulness, logical coherence and cultural appropriateness, though with modest drops in clarity. By contrast, Phi-4, OpenChat-3.5, LLaMA-3-8B, and Mistral-7B show significantly lower performance, particularly on faithfulness and clarity—dimensions most sensitive to hallucination and underspecification. Human raters were notably stricter in these areas, especially for open-weight models, revealing that LLM-based evaluations may overestimate explanation quality. Despite these differences in score magnitude, the rubric-level trends are consistent: Logical Coherence is the strongest dimension across most models, while Faithfulness, Cultural Appropriateness and Clarity & Completeness expose key weaknesses in less capable systems. Interestingly, Mistral-7B, which led in classification F1, ranks lowest in explanation quality by both rating sources, highlighting a persistent disconnect between predictive accuracy and reasoning quality. Conversely, the strongest models (GPT-4 variants and Gemma) exhibit both high classification performance and robust explanatory reasoning. These findings emphasize the importance of explanation-focused evaluation, as high task accuracy alone may mask serious limitations in model understanding and reasoning.

4.3 Probing LLM Robustness to Synthetic Set (Counterfactual Flips)

Criterion	Human	LLM-as-a-judge
Fluency	0.89	1.00
Naturalness	0.68	0.97
Flip Clarity	0.79	0.98
Meaning Preservation	0.78	0.58

Table 2: Average rubric-based scores for synthetic flips.

We categorized each counterfactual by its main transformation and found that flips most commonly altered sentiment-bearing keywords, phrases, tone, and emoji—components central to both explicit and stylistic sentiment signaling, see Figure 3. Less frequent were transformations involving negation, intent framing, or valence modulation, which require more interpretive reasoning. From our analysis, GPT-4 often produced plausible synthetic flips (see examples in Table 14). We assessed the quality of the synthetic flips using rubric-based judgments from both human annotators and LLMs-as-Judges (GPT-4-Turbo and GPT-4-32k), see Table 2. LLM ratings were uniformly high—near-perfect in fluency, naturalness, flip clarity, and slightly lower for meaning preservation. Human annotators, however, were notably stricter, especially in naturalness and meaning preservation, revealing significant gaps in how surface-level and semantic quality are perceived. In particular, humans flagged many cases as semantically incorrect or stylistically unnatural, despite their formal fluency. Manual analysis revealed that *positive-to-negative* flips posed greater challenges. LLMs frequently overcorrected, introducing harsh or exaggerated tone, especially in code-mixed inputs (see Table 15). Conversely, *negative-to-positive* flips tended to be smoother and more culturally appropriate. While human raters penalized positive-to-negative flips for harsh tone or topic drift, LLMs-as-Judges often gave high marks even in such cases—suggesting they were

less sensitive to subtle shifts in meaning or register. While the flipped sentiment was often correct, the model struggled with non-English and code-mixed inputs, frequently normalizing local shorthand, translating content into English, or rewriting messages in Standard Swahili (Kiswahili Sanifu), thereby altering the original language composition (see third example in Table 14).

Model	Eff. F1 (Pre-CF)	Eff. F1 (Post-CF)	Δ Post–Pre
GPT-4-Turbo	0.960	0.980	+0.020
GPT-4-32k	0.970	0.980	+0.010
Phi-4	0.940	0.786	–0.154
Gemma-3-27B	0.940	0.466	–0.474
Mistral-7B	0.892	0.466	–0.425
OpenChat-3.5	0.910	0.441	–0.469
LLaMA-3-8B	0.783	0.349	–0.434

Table 3: Effective F1 before and after counterfactual sentiment flips.

To quantify model robustness under transformation, we compute *Effective F1*—the product of F1 and coverage. As shown in Table 3, both GPT-4-Turbo and GPT-4-32k maintained high or improved post-flip performance (up to 0.980). In contrast, mid-sized and open models suffered significant drops (0.40–0.47), driven by both misclassification and partial outputs. Notably, Phi-4 preserved coverage but underperformed on positive flips, indicating brittle generalization. Beyond label accuracy, explanation quality further reveals this fragility. On the Synthetic Set, only the GPT-4 variants produced consistently faithful, coherent, culturally grounded, and context-sensitive reasoning. Other models often generated fluent but incorrect explanations after sentiment was flipped, with sharp drops in faithfulness and completeness—especially for Mistral-7B, OpenChat-3.5, and LLaMA-3-8B (Table 16).

4.4 How does model confidence and alignment reflect interpretive ambiguity?

Model	Avg. Conf.	Coverage (%)	Eff. Conf.
GPT-4-Turbo	4.639	100.0	4.64
GPT-4-32k	4.440	100.0	4.44
Phi-4	4.711	99.5	4.69
Gemma-3-27B	4.698	47.6	2.24
OpenChat-3.5	4.249	47.4	2.01
Mistral-7B	4.132	47.6	1.97
LLaMA-3-8B	3.981	37.6	1.50

Table 4: Effective Confidence on the Synthetic Set.

To assess confidence calibration, we report average model confidence and coverage across the Gold and Synthetic Sets (Table 17). While most models maintain high confidence on the Gold Set, only the GPT-4 variants and Phi-4 sustain both high confidence and near-complete coverage on counterfactual inputs. In contrast, models like Gemma-3-27B and OpenChat-3.5 appear overconfident despite skipping over half of the flipped messages. To quantify this further, we compute an Effective Confidence score (confidence × coverage), reported in Table 4, revealing a sharp drop for open models—underscoring their brittleness under minimal sentiment shifts. Although the Gold Set contains messages with full human agreement, models show only moderate alignment with one another. As shown in Figure 2, the highest agreement is observed between Gemma-3-27B and Phi-4 ($\kappa = 0.73$), and between GPT-4-Turbo and GPT-4-32k ($\kappa = 0.70$). However, other pairings show weaker agreement—such as GPT-4-Turbo and Mistral-7B ($\kappa = 0.48$)—despite similar average F1 scores. This suggests that even on "clear" cases, LLMs diverge in interpretation, reflecting differences in how they weigh tone, cues, and cultural context.

5 Discussion

LLMs-as-Generators: Crafting Cultural Counterfactuals Using GPT-4 to generate sentiment-flipped counterfactuals revealed both the model's strengths and its limitations. Often, it produced fluent, contextually appropriate flips that successfully reversed sentiment while preserving tone and informal style. However, our diagnostic analysis surfaced key weaknesses. Flips from positive to negative frequently introduced exaggerated emotional intensity, suggesting the model struggles to calibrate negative sentiment in subtle, conversational contexts. Additionally, while GPT-4 provided self-reported labels for the components it modified (e.g., tone, emoji, phrasing), these attributions were often imprecise or inconsistent. These findings underscore both the potential and fragility of using LLMs to generate culturally grounded synthetic data—and highlight the continued need for more iteration in prompt instructions as well as human oversight when precision over tone, meaning, and linguistic structure is essential.

LLMs-as-Judges: Evaluating Counterfactuals We used GPT-4 as a judge to assess the quality of sentiment-flipped messages—selecting the best rewrite among three generated variants and then scoring the selected flip for fluency, naturalness,

meaning preservation, and successful sentiment reversal. This approach streamlined evaluation and scaled the generation pipeline. In many cases, GPT-4's selections aligned with the human judgments. However, because these decisions rely entirely on the model's internal criteria, we observed inconsistencies—especially for non-English messages with culturally layered meaning. For instance, some selected flips introduced subtle shifts in tone or more formal phrasing, reducing cultural fidelity even when sentiment was accurately reversed. In other cases, we observe that GPT-4 successfully produced plausible flips that changed a message's perceived sentiment, this was achieved in different ways, which do not necessarily reflect the most *minimal* changes to achieve that effect. For example, flipping "Hahaha" (+) could be achieved by "Not funny" (-) or "Ughhh," (-) or "This is not funny at all" (-). These insights suggests that additional checks should be put into place already at the filtering step to assess if flips are indeed consistent with the tone, phrasing, language composition or cultural meaning of the original message to ensure chosen variants are truly the most faithful transformations. These findings point to the need for human-in-the-loop validation at each stage—particularly when using LLMs to adjudicate nuanced, multilingual language in low-resource settings.

Prediction is not 'understanding' Models such as `Mistral-7B`, `Phi-4`, and `OpenChat-3.5` score competitively on standard metrics, yet generate explanations that often lack coherence, faithfulness, or cultural grounding—especially in cases where sentiment is subtle, indirect, or stylistically embedded, as revealed by human evaluation. These reasoning gaps become even more pronounced under sentiment counterfactuals, with flipped affect lead to sharp performance drops—up to 0.47 F1 for open-weight models—exposing brittle generalization to plausible shifts in tone, emoji, or phrasing. In contrast, `GPT-4-Turbo` and `GPT-4-32k` demonstrate greater robustness in both prediction and reasoning, suggesting that scale and stronger instruction tuning support more stable reasoning.

LLMs vary not just in accuracy, but in worldview Agreement scores between models remain low, even on the Gold Set, where human annotators were unanimous. This divergence reflects not just model sensitivity to surface cues, but deeper differences in how LLMs encode sentiment priors, cultural nuance, and conversational style. That `GPT-4-Turbo` and `Mistral-7B` can yield similar F1 yet diverge in label agreement ($\kappa = 0.48$) illustrates that we are not simply comparing better vs. worse models, but different interpretive frameworks. However, we do not understand the models underlying interpretive frameworks, and how well it maps to existing theory, and consistency in reasoning varies significantly across models, especially open-weight models.

Confidence is not calibration While average confidence scores remain high across models, only OpenAI's models (`GPT-4-Turbo` and `GPT-4-32k`) consistently maintain high confidence, full coverage on perturbed data, accurate predictions, and reliable reasoning. In contrast, models such as `Phi-4` also exhibits high confidence and broad coverage, but manual inspection reveals frequent reasoning errors, highlighting a gap between confidence and correctness.

Annotation as a site of interpretive complexity Our study highlights the complexities of designing robust annotation protocols for nuanced, real-world data. Annotators frequently encountered edge cases that exposed ambiguity in how sentiment should be labeled, especially when affect was culturally or contextually embedded. This reinforces growing recognition in human-centered NLP that annotation is an interpretive process requiring iteration, theoretical grounding, and thoughtful handling of disagreement.

Sentiment is structured by context Our work challenges simplified views of sentiment as binary or fixed, framing it instead as context-dependent and semantically layered. While our initial definition in the annotation protocol and component taxonomy aimed to capture more nuance, more specification is needed. For example, ***context-dependency*** emerged as central to interpretation, as seen in our annotation examples. There are many aspects that can shape what context-dependency as an element of sentiment means. As illustrated through our study, context can be informed by: the conversation topic (e.g., health advice); cultural norms (e.g., in Kenya); or religious cues; as well as other interpersonal dynamics (e.g., what the recipient of a message assumes or knows about its writer) that can be harder to capture or specify. Yet, future work will need to expand efforts to further systematize and formalize those components of sen-

timent to be able to achieve more robust evaluation approaches.

6 Conclusion

We reframe sentiment analysis in low-resource, culturally nuanced contexts as a problem of reasoning, not just classification. Using a diagnostic framework grounded in social science measurement, we evaluate how LLMs interpret sentiment in multilingual, code-mixed WhatsApp messages from Nairobi youth health groups. Our findings reveal that while top-tier LLMs demonstrate interpretive robustness, open models often fail under ambiguity and cultural nuance, highlighting deep gaps in reasoning quality. As sentiment increasingly becomes a benchmark task for real-world NLP, our work urges a shift from fixed-label accuracy to context-aware, culturally grounded evaluation. Future sentiment systems must be judged not only by what label they assign, but how and why they reason that way.

Limitations

While our diagnostic framework offers a deeper lens into sentiment reasoning, several limitations remain:

(1) Sentiment itself remains an inherently subjective construct. Our LLM-guided systematization of text components like negation, tone, emojis, keywords and phrase rewordings look reasonable (face validity) and may capture the most salient aspects of the sentiment concept (content validity). However, further research is needed to inspect whether this systematization fully specifies all observable criteria connected to sentiment (substantive validity) (Wallach et al., 2025); as well as how the components may relate to one-another; and whether its operationalization via LLM-as-a-judge is consistent and coherent with the LLMs internal interpretation of these components.

(2) Our counterfactual generation pipeline uses a two-stage prompting process: GPT-4 first generates three flipped variants of a message, then selects the most plausible one for inclusion. While this filtering step improves fluency and contextual fit, it relies entirely on the model's internal criteria, which we do not independently validate. Future work should investigate how this selection process affects flip quality, what may be lost or altered during filtering, and incorporate human-in-the-loop checks to ensure that selected flips accurately reflect the intended sentiment transformation and preserve linguistic and contextual fidelity.

(3) Our study focuses on a single, culturally specific dataset of health-related WhatsApp messages from Nairobi youth. While this setting is intentionally chosen to surface ambiguity and contextual nuance, it limits direct generalization to other populations or sociolinguistic contexts. We view the framework itself as transferable, but its application to other code-mixed languages, age groups, or cultural settings remains an important direction for future research.

Ethical Consideration

This study uses anonymized WhatsApp messages from Nairobi youth health groups, collected with consent under prior research protocols. All data were reviewed to remove identifying information and sensitive content. Our use of LLMs to generate synthetic sentiment data in a code-mixed, culturally grounded setting raises important ethical considerations. Language reflects identity, and synthetic rewrites, especially in informal, multilingual contexts must be handled with care to avoid erasing nuance or reinforcing stereotypes. While we designed prompts to preserve tone and intent, LLMs may still encode harmful biases. We emphasize the importance of cultural sensitivity, context-aware evaluation, and collaboration with local experts to ensure respectful and responsible analysis.

References

John Langshaw Austin. 1975. *How to do things with words*. Harvard university press.

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.

Michael Burnham. 2024. What is sentiment meant to mean to language models? *Research & Politics*, 11(4):20531680241307941.

Nurendra Choudhary, Rajat Singh, Ishita Bindlish, and Manish Shrivastava. 2018. Sentiment analysis of

code-mixed languages leveraging resource rich languages.

Emily Corvi, Hannah Washington, Stefanie Reed, Chad Atalla, Alexandra Chouldechova, P Alex Dow, Jean Garcia-Gathright, Nicholas Pangakis, Emily Sheng, Dan Vann, et al. 2025. Taxonomizing representational harms using speech act theory. *arXiv preprint arXiv:2504.00928*.

Mahdi Dhaini, Kafaite Zahra Hussain, Efstratios Zaradoukas, and Gjergji Kasneci. 2025. Evalxnlp: A framework for benchmarking post-hoc explainability methods on nlp models. *Preprint*, arXiv:2505.01238.

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

A. Seza Doğruöz, Sunayana Sitaram, and Zheng-Xin Yong. 2023b. Representativeness as a forgotten lesson for multilingual and code-switched data collection and preparation. In *Conference on Empirical Methods in Natural Language Processing*.

Xia Fang, Magdalena Rychlowska, and Jens Lange. 2022. Cross-cultural and inter-group research on emotion perception. *Journal of Cultural Cognitive Science*, 6:1–7.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks are all you need. *Preprint*, arXiv:2306.11644.

Michael Alexander Kirkwood Halliday. 2014. Language as social semiotic. *The Discourse Studies Reader. Amsterdam: John Benjamins*, pages 263–272.

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. *Preprint*, arXiv:2310.06825.

Brihi Joshi, Aaron Chan, Ziyi Liu, Shaoliang Nie, Maziar Sanjabi, Hamed Firooz, and Xiang Ren. 2023. Er-test: Evaluating explanation regularization methods for language models. *Preprint*, arXiv:2205.12542.

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*.

Kristen A. Lindquist. 2021. Language and emotion: Introduction to the special issue. *Affective Science*, 2:91–98.

Chuchu Liu, Fan Fang, Xu Lin, Tie Cai, Xu Tan, Jianguo Liu, and Xin Lu. 2021. Improving sentiment analysis accuracy with emoji embedding. *Journal of Safety Science and Resilience*, 2(4):246–252.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2024. Towards faithful model explanation in nlp: A survey. *Computational Linguistics*, 50:657–723.

David Matsumoto. 1990. Cultural similarities and differences in display rules. *Motivation and Emotion*, 14:195–214.

Saif M. Mohammad. 2017. Challenges in sentiment analysis. pages 61–83.

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.

Pansy Nandwani and Rupali Verma. 2021. A review on sentiment analysis and emotion detection from text. *Social network analysis and mining*, 11(1):81.

Millicent Ochieng, Varun Gumma, Sunayana Sitaram, Jindong Wang, Vishrav Chaudhary, Keshet Ronen, Kalika Bali, and Jacki O'Neill. 2025. Beyond metrics: Evaluating LLMs effectiveness in culturally nuanced, low-resource real-world scenarios. In *Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025)*, pages 230–247, Vienna, Austria. Association for Computational Linguistics.

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.

Ikechukwu Onyenwe, Samuel Nwagbo, Njideka Mbeledogu, and Ebele Onyedinma. 2020. The impact of political party/candidate on the election results from a sentiment analysis perspective using# anambradecides2017 tweets. *Social Network Analysis and Mining*, 10(1):55.

OpenAI. 2023. Gpt-4 technical report.

Neeraj Anand Sharma, A. B.M.Shawkat Ali, and Muhammad Ashad Kabir. 2024. A review of sentiment analysis: tasks, applications, and deep learning techniques. *International Journal of Data Science and Analytics 2024 19:3*, 19:351–388.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

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.

Hanna Wallach, Meera Desai, A Feder Cooper, Angelina Wang, Chad Atalla, Solon Barocas, Su Lin Blodgett, Alexandra Chouldechova, Emily Corvi, P Alex Dow, et al. 2025. Position: Evaluating generative ai systems is a social science measurement challenge. *arXiv preprint arXiv:2502.00561*.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2024. Openchat: Advancing open-source language models with mixed-quality data. *Preprint*, arXiv:2309.11235.

Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review 2022 55:7*, 55:5731–5780.

Xinnuo Xu, Rachel Lawrence, Kshitij Dubey, Atharva Pandey, Fabian Falck, Risa Ueno, Aditya Nori, Rahul Sharma, Amit Sharma, and Javier González. 2025. Re-imagine: Symbolic benchmark synthesis for reasoning evaluation. In *ICLR 2025 - Workshop on Reasoning and Planning for LLMs*.

Linyi Yang, Jiazheng Li, Padraig Cunningham, Yue Zhang, Barry Smyth, and Ruihai Dong. 2021. Exploring the efficacy of automatically generated counterfactuals for sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 306–316, Online. Association for Computational Linguistics.

Byungkyu Yoo and Julia Taylor Rayz. 2021. Understanding emojis for sentiment analysis. In *The international FLAIRS conference proceedings*, volume 34.

Junfeng Zhang. 2024. Sentiment and language: A socio-semiotic analysis. *Philosophy Journal*, 3(1):118–127.

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.

A Appendix

A.1 Annotation Protocol

Annotators and Process. This protocol was developed to guide consistent sentiment annotation of informal, multilingual WhatsApp messages exchanged among youth in Nairobi. Two trained annotators—both fluent in English, Swahili, and Sheng—applied the protocol over the course of one month. Annotation covered 6,197 messages drawn from a dataset of informal, code-mixed conversations among youth living with HIV. Work was conducted in Excel. Annotators labeled each message independently, treating it as a standalone utterance while considering cultural context, code-switching, and emoji use. We began with a jointly labeled calibration set of 100 examples, followed by independent annotation with regular meetings to discuss edge cases and resolve ambiguities.

Sentiment Categories. Messages were labeled as **Negative (-1)**, **Neutral (0)**, or **Positive (1)** based on expressed affect. Annotators were instructed to:

- Label as **Negative** if the message expressed frustration, sadness, criticism, or distress (e.g., *"I am tired !!!", "You have a mental problem"*).
- Label as **Neutral** if the message conveyed information, routine conversation, or general greetings without strong sentiment (e.g., *"When are you coming?", "Good morning 👋"*).
- Label as **Positive** if the message expressed joy, support, pride, or optimism (e.g., *"I'm much happy to interact and share with you guys!"*).

Ambiguity and Cultural Nuance. Annotators flagged ambiguous cases with written justifications. Given the culturally grounded and multilingual nature of the data, particular attention was paid to tone, idioms, emoji use, and context-specific expressions of affect.

A.2 Evaluation Subsets

Subset	Positive	Negative	Neutral	Total
Gold Set	1196	351	4574	6,121
Synthetic Set	351	1196	-	1,547
Ambiguous Set	-	-	-	76

Table 5: Sentiment-wise distribution of messages.

A.3 Overall Model Performance

Model	Pos	Neg	Neu	Avg	Cov. %
Gold Set (annotated Pos/Neg/Neu)					
GPT-4-Turbo	0.98	0.92	0.75	0.88	100.0
GPT-4-32k	0.93	0.90	0.86	0.90	100.0
Gemma-3-27B	0.93	0.96	0.79	0.89	100.0
Phi-4	0.93	0.91	0.80	0.88	100.0
Mistral-7B	0.91	0.88	0.92	0.90	98.9
OpenChat-3.5	0.93	0.77	0.87	0.86	99.9
LLaMA-3-8B	0.94	0.51	0.86	0.77	92.9
Pre-CF (original Pos/Neg examples)					
GPT-4-Turbo	0.98	0.94	—	0.96	100.0
GPT-4-32k	0.99	0.95	—	0.97	100.0
Gemma-3-27B	0.97	0.90	—	0.94	100.0
Phi-4	0.97	0.90	—	0.94	100.0
OpenChat-3.5	0.97	0.85	—	0.91	100.0
Mistral-7B	0.97	0.89	—	0.93	95.9
LLaMA-3-8B	0.97	0.77	—	0.87	90.2
Post-CF (synthetic counterfactuals)					
GPT-4-Turbo	0.97	0.99	—	0.98	100.0
GPT-4-32k	0.97	0.99	—	0.98	100.0
Phi-4	0.67	0.90	—	0.79	99.5
Gemma-3-27B	0.97	0.99	—	0.98	47.6
Mistral-7B	0.97	0.99	—	0.98	47.6
OpenChat-3.5	0.90	0.97	—	0.93	47.4
LLaMA-3-8B	0.91	0.96	—	0.93	37.6

Table 6: F1 scores by sentiment class on the Gold Set, Pre-CF (original positive/negative examples used to generate counterfactuals), and Post-CF (synthetic counterfactuals with flipped sentiment). Coverage rate (%) reflects the proportion of examples for which a model returned a valid sentiment label.

A.4 Prediction Agreement Across Models

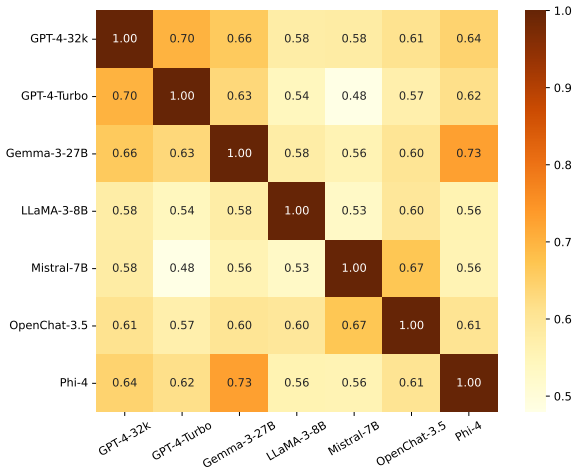

Figure 2: Cohen's κ agreement between model predictions on the Gold Set. Despite full annotator agreement, models show only moderate pairwise consistency—indicating divergence in their underlying reasoning and sensitivity to sentiment cues.

A.5 Model Prompts

Sentiment Classification + Explanation Prompt

```
You are an NLP assistant for sentiment analysis.
Given a WhatsApp message (QUERY), classify its sentiment
as Positive, Negative, or Neutral.
Provide a justification using extracted keywords and a
brief explanation.
Return your confidence score (0-5). Use JSON output format
only.

The prompt includes:
- Sentiment definitions (Positive, Neutral, Negative)
- Examples of clearly and ambiguously labeled messages
- JSON output format with keywords, explanation, label,
and confidence

QUERY: "{query}"

Output Format:
{
  "justification": {
    "keywords": [ ... ],
    "explanation": "..." },
  "sentiment": "...",
  "confidence_score": "..."
}
```

Table 7: Instruction prompt for joint sentiment classification, justification, and confidence scoring.

Counterfactual Evaluation Prompt

```
You are evaluating a synthetic (GPT-4-generated) version of a WhatsApp
message. The synthetic message is a sentiment-flipped
version of the original.
Assess the quality of the synthetic message along four
criteria using 0 or 1:
1. Fluency – Is the synthetic message grammatically
correct and readable?
2. Naturalness – Does it sound plausible for a human
to write?
3. Sentiment Flip Clarity – Is the sentiment clearly
flipped from the original?
4. Meaning Preservation – Is the core meaning
preserved aside from the sentiment?

Original Message: "{original}"
Synthetic Message: "{flipped}"
Transformation Type: {transformation}
GPT-4 Explanation for the Flip: "{explanation}"

Return ONLY this JSON:
{
  "fluency": 0 or 1,
  "naturalness": 0 or 1,
  "sentiment_flip_clarity": 0 or 1,
  "meaning_preservation": 0 or 1,
  "annotator_comment": "optional comment (string)"
}
```

Table 8: Prompt used to evaluate quality of synthetic counterfactuals across four rubric dimensions.

(a) Counterfactual Generation Prompt

```
You are an NLP assistant helping researchers generate
high-quality counterfactual examples for sentiment
classification.
Given a WhatsApp-style message and its sentiment (Positive or
Negative), generate 3 distinct versions that flip the sentiment.
Only modify necessary components. Preserve fluency and realism.
Respect informal tone.
You may flip sentiment by changing components such as:
- keywords, phrases, negation, intent framing, tone (e.g.,
sarcasm), sentiment valence, emojis/icons, code-mixing
Input:
Original message: "{original_message}"
Original sentiment: "{original_sentiment}"
Output Format (JSON List of 3 Objects):
{
  "cf_text": "...",
  "components_changed": [...],
  "flip_explanation": "..."
}
```

(b) Counterfactual Filtering Prompt

```
You are a sentiment evaluation assistant. Your task is to select
the best counterfactual rewrite of a message.
ORIGINAL MESSAGE
"{original}"
(Sentiment: {original_sentiment})
COUNTERFACTUAL CANDIDATES
1. "{cf1}"
2. "{cf2}"
3. "{cf3}"
INSTRUCTIONS
Your goal is to identify which counterfactual most effectively
flips the sentiment while remaining realistic and fluent.
- Flip sentiment plausibly
- Sound natural in WhatsApp chat
- Preserve meaning/context where possible
RESPONSE FORMAT (JSON only):
{
  "selected_cf": "...",
  "justification": "...",
  "predicted_sentiment": "Positive / Negative"
}
```

Table 9: Combined prompts for generating and selecting counterfactual sentiment flips.

Explanation Evaluation Prompt

```
You are a language model tasked with evaluating the quality of a
sentiment explanation. Evaluate the explanation
for the following: 1. Faithfulness – Does it
reflect the original message and prediction without
hallucinating? 2. Contextual Appropriateness – Is
it culturally and linguistically aware? 3. Logical
Coherence – Is it internally consistent and justified?
4. Clarity and Completeness – Is it clear, specific,
and sufficient?
Message:
"{message}"
Predicted Sentiment: {prediction}
Explanation: "{explanation}"

Return ONLY this JSON:
{
  "faithfulness": 0 or 1,
  "contextual_appropriateness": 0 or 1,
  "logical_coherence": 0 or 1,
  "clarity_and_completeness": 0 or 1,
  "annotator_comment": "optional comment (string)"
}
```

Table 10: Prompt used to evaluate explanation quality across four rubric dimensions.

A.6 Further Examples from Our WhatsApp Dataset: Cultural Nuance and Annotator Disagreement

Example	Explanation
My friends it was heard to take drugs bt i just take heart	• Can be read differently due to *situational context (sympathy)*. • Shows **emotional vulnerability**, which may invite empathy or humor depending on setting. • Use of **"take heart"** is culturally influenced—often heard in African English as a way to express resilience. • The spelling ("heard" instead of "hard") could be **interpreted differently** (innocent typo vs. deeper linguistic variation).
Kama hauko School shindaapo "Even you are not in school just stay there"	• Can be read differently due to *schooling context*. • Often used **sarcastically or dismissively**, especially in online chat. • The phrase can also reflect **class-based or knowledge-based exclusion** ("If you're not educated, stay out of this"). • Code-mixing adds a **layer of urban youth culture** and localized meaning.
He is faithful all the time	• Can be read differently due to *religion*. • Common in **Christian communities**, especially in African contexts—often part of a **call-and-response**. • Can express **faith during suffering**, giving it emotional depth in testimonies or public speeches. • Without context, it may be misread as a general statement about a person rather than a **declaration about God**.

Table 11: Examples of cultural nuance and their context-dependent interpretations.

A.7 Rubrics for Evaluation

A.7.1 Explanation Evaluation Rubric

Each model-generated explanation was evaluated along four binary (0/1) dimensions:

- **Faithfulness:** Does the explanation accurately reflect the input message and how it informed the model's sentiment prediction? Explanations that include hallucinated, fabricated, or unrelated content should be scored **0**.
- **Contextual Appropriateness:** Does the explanation show awareness of cultural, social, or linguistic context? If it fails to address relevant tone, code-mixing, or local expressions, assign **0**. Optional comments may highlight cultural or linguistic mismatches.
- **Logical Coherence:** Is the explanation internally consistent and logically connected to the sentiment label? Contradictory or illogical justifications are scored **0**.
- **Clarity and Completeness:** Is the explanation clear, specific, and sufficient to support the sentiment label? Vague or underspecified rationales receive **0**.

Scoring: 1 = Yes; 0 = No
Note: Annotators were asked to leave optional comments when assigning a score of 0, especially for cultural/contextual errors or hallucinations.

A.7.2 Synthetic Data Evaluation Rubric

Each GPT-4-generated counterfactual message was evaluated using the following binary (0/1) criteria:

- **Fluency:** Is the synthetic message grammatically well-formed and fluent?

- **Naturalness:** Does the message sound plausible or likely to have been written by a real user?

- **Sentiment Flip Clarity:** Is the reversal in sentiment (compared to the original message) clear and consistent?

- **Meaning Preservation:** Aside from sentiment, does the core meaning/topic of the original message remain intact? Large semantic shifts receive **0**.

Scoring: 1 = Yes; 0 = No
Note: Annotators were encouraged to flag particularly good or bad examples, especially where tone, fluency, or cultural grounding were notably off.

A.8 Sentiment Transformation Taxonomy

To guide counterfactual generation, we organize sentiment-altering edits into the following transformation types:

Transformation Type	Definition	Example
Negation	Add or remove negation to reverse sentiment.	*"I like it"* → *"I don't like it"*
Tone / Intent Shift	Change the tone or implied intent of the message.	*"You could do better"* → *"You're doing great"*
Emoji Substitution	Replace emoji's to reflect different sentiment.	"😅" → "😢"
Keyword Substitution	Swap a sentiment-bearing word.	*"Useful advice"* → *"Terrible advice"*
Phrase Rewording	Paraphrase to shift sentiment while preserving meaning.	*"You always help me"* → *"You always get in my way"*

Table 12: Taxonomy of sentiment-altering transformations used in counterfactual generation.

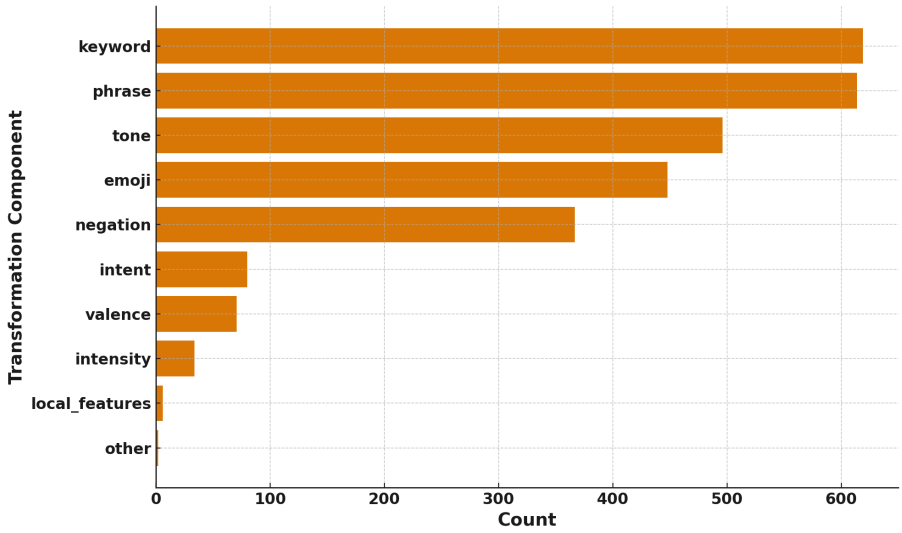

Figure 3: Distribution of transformation components used to generate counterfactual examples. Most sentiment flips involved keyword or phrase substitutions, followed by tone shifts, emoji manipulation, and negation. Less frequent transformations include intent, valence, and intensity adjustments.

A.9 Outputs of Synthetic Experiments

A.9.1 Component Consistency

Original message	Flipped message	LLM identified components	LLM rationale
Hahaha (+)	This is not funny at all. (–)	phrase, valence	Changing ‘ 🤣 🤣 why?’ to a direct statement expressing displeasure flips the sentiment from positive to negative. Note: hallucinated explanantion
Hahaha (+)	not funny at all (–)	keyword, negation	Changing ‘hahaha’ to ‘not funny at all’ adds negation and alters the key phrase to express dissatisfaction, flipping the sentiment to negative.
Ahahhaa (+)	Not funny at all. (–)	keyword, negation	Adding negation and changing the keyword from a positive laughter expression to a negative critique directly flips the sentiment.
Haha (+)	not funny (–)	phrase, negation	Introducing ‘not’ before ‘funny’ adds a negation to the original message, shifting the sentiment from amusement to disapproval.
Hahaaaaa 😄 😄 😄 (+)	Ughhh 😒 😒 😒 (–)	emoji, tone	Changing the emoji from 😄 (happy) to 😒 (annoyed) and the exaggerated sound from ‘Hahaaaaa’ to ‘Ughhh’ shifts the sentiment from positive to negative.
hahaha 🤣 🤣 🤣 (+)	this is not funny at all 😠 (–)	phrase, emoji [no mentioning of negation]	Adding text that explicitly states discontent (‘this is not funny at all’) and using an angry emoji changes the sentiment from amusement to displeasure.

Table 13: Examples from our synthetic experiment showing varied LLM component sensitivity and consistency in handling laughter expressions.

A.9.2 What Worked Well

Original message	Flipped message	LLM identified components	LLM rationale
Pia mi nko poa (+) “Me too, I’m okay.”	*Pia mi siko poa kabisa leo* (–) “Me too, I’m really not okay today.”	phrase, intensifier	Adding ‘siko’ (not) and ‘kabisa’ (at all) to the original message intensifies the negative sentiment, flipping it from positive. [**Reflection**: Correct flip with Standard Swahili]
Ata me wananiboo leo yote hawajatext (–) Even me, they’re boring me the whole day they haven’t texted.	*Ata me leo nimefurahishwa, wamekuwa wakinitext mara kwa mara.* (–) “Even me, today I’ve been made happy, they’ve been texting me again and again.”	keywords, phrase	By modifying the phrase to express happiness and frequent communication, the sentiment changes from negative to positive. [**Reflection**: Correct flip with Standard Swahili]
I will check on that coz nakohoa sana adi naumwa na kifua (–) “I’ll check on that because I’m coughing a lot - to the point my chest hurts.”	I will check on that coz I’m feeling much better now, the cough and chest pain are gone! (+)	keywords, valence	The switch from describing ongoing symptoms to stating that the symptoms are gone (‘the cough and chest pain are gone’) changes the sentiment from negative to positive by indicating an improvement in health. [**Reflection**: Correct flip with English Translation]

Table 14: Examples from our synthetic experiment showing successful sentiment flips and clear rationale.

A.9.3 What Worked Less Well

	Example 1	Example 2	Example 3
Original message	*but kulinice* "But it was nice"	And thank you 5017 for very **clear** elaboration about the questions at hand. **I love** how you have handled and answered them **to the point** (+)	Dia hope you will be coming visiting I enjoyed your company alot your fun to be with (+)
Flipped message	but *kulinice* was disappointing	And thanks 5017 for very **vague** explanations about the questions at hand. **I hate** how you have **mishandled and dodged them completely**. (–)	Dia, please don't come over. Honestly, hanging out was more exhausting than enjoyable, you're not that fun to be with. (–)
LLM identified components	key phrase	keywords, key phrases, tone	negation, tone, key phrases
LLM rationale	Changing 'kulinice' to 'kulinice was disappointing' directly flips the sentiment by expressing a negative experience.	Changing 'clear elaboration' to 'vague explanations' and 'love' to 'hate' directly flips the positive sentiment to negative. Additionally, 'handled and answered them to the point' was changed to 'mishandled and dodged them completely' to emphasize dissatisfaction.	Introducing a request for absence ('please don't come over') and directly stating negative feelings ('more exhausting than enjoyable', 'not that fun') shifts the sentiment from positive to negative.
Reflections / Insights	From the LLM's explanation sw-en codemixed *kulinice* meaning "it was nice" misunderstood; term "kulinice" was ignored rather than changed because it could not handle the code-mix expression.	Flipped message feels somewhat unnatural: the "thanks" gets in the way of creating weird sarcasm/strong negativity with words like "hate"; hate may be the flip to "love" but it's rare to have such expressions in feedback about questions.	The flipped message sounds strong/extreme; unnatural.

Table 15: Examples from our synthetic experiment showing less effective sentiment flips.

A.10 Model Explanation Quality, Confidence, and Coverage

Dataset	Model	Faithfulness	Contextual Approp.	Logical Coherence	Clarity & Completeness
Ambiguous Set	GPT-4-32k	1.000	0.950	1.000	0.950
	GPT-4-Turbo	1.000	1.000	1.000	1.000
	Gemma-3-27B	1.000	1.000	1.000	0.975
Gold Set	GPT-4-Turbo	0.983	0.983	1.000	1.000
	GPT-4-32k	1.000	0.983	1.000	0.980
	Gemma-3-27B	0.967	1.000	0.983	0.967
	Phi-4	0.917	0.933	1.000	0.950
	OpenChat-3.5	0.783	0.750	0.900	0.783
	LLaMA-3-8B	0.683	0.683	0.817	0.733
	Mistral-7B	0.617	0.650	0.850	0.650
Synthetic Set	GPT-4-32k	1.000	1.000	1.000	1.000
	GPT-4-Turbo	1.000	1.000	1.000	1.000
	Gemma-3-27B	0.906	1.000	1.000	0.875
	Phi-4	0.750	0.750	0.800	0.775
	LLaMA-3-8B	0.700	0.900	0.950	0.700
	Mistral-7B	0.688	0.844	0.906	0.688
	OpenChat-3.5	0.594	0.812	0.844	0.688

Table 16: Explanation quality scores by dataset and model across four dimensions.

Model	Gold Conf.	Gold Cov. %	Synth. Conf.	Synth. Cov. %	Eff. Conf.
`GPT-4-Turbo`	4.174	100.0	4.639	100.0	4.64
`GPT-4-32k`	4.283	100.0	4.440	100.0	4.44
`Phi-4`	4.464	100.0	4.711	99.5	4.69
`Gemma-3-27B`	4.265	100.0	4.698	47.6	2.24
`OpenChat-3.5`	4.204	99.9	4.249	47.4	2.01
`Mistral-7B`	4.237	98.9	4.132	47.6	1.97
`LLaMA-3-8B`	4.311	92.9	3.981	37.6	1.50

Table 17: Average model confidence and coverage across Gold and Synthetic Sets.