

T-CLA: Taxonomy of Cultural-Linguistic Artifacts for Modeling Misalignment in NLP

Kezia Oketch,^{1,2} John P. Lalor,^{1,2} Denghui Zhang,³ Ahmed Abbasi^{1,2}

¹Human-centered Analytics Lab, University of Notre Dame

²Department of IT, Analytics, and Operations, University of Notre Dame

³School of Business, Stevens Institute of Technology

koketch@nd.edu, john.lalor@nd.edu, dzhang42@stevens.edu, aabbasi@nd.edu

Abstract

We introduce T-CLA, a taxonomy for analyzing cultural-linguistic misalignment in NLP systems. We define misalignment as a mismatch between sociolinguistic variation in user language and model representations, and argue that such variation can be represented and analyzed at the feature level. We construct T-CLA through a bottom-up analysis of 8,680 Swahili health-related responses from Kenyan speakers, identifying 170 artifacts spanning code-mixing, Sheng usage, tribal lexicons, and loanwords. We validate the taxonomy through an annotation study with 1,560 annotations, showing that these artifacts are broadly recognizable but unevenly detectable across speakers. Using T-CLA as an analytic lens, we find that cultural-linguistic artifacts are associated with model prediction error across tasks and models, while subgroup analyses reveal fairness disparities across tribal affiliations. While T-CLA is specific to Kenyan Swahili, the methodology provides a transferable approach for identifying, validating, and analyzing cultural-linguistic artifacts across linguistic and cultural contexts.

1 Introduction

Contemporary NLP systems are increasingly deployed in contexts where language use is socially and culturally situated (Hershcovich et al., 2022; Halliday, 2004). However, most approaches to modeling language continue to treat variation, particularly sociolinguistic variation, as peripheral noise rather than as constitutive of meaning (Blodgett et al., 2020; Joshi et al., 2020). This creates a fundamental mismatch between standardized representations of language and the ways in which linguistic forms are shaped by cultural practice and social context, a gap that has been linked to failures and inequities in NLP systems (Bender et al., 2021; Blodgett et al., 2020). Recent work in cross-cultural NLP further argues that such mismatches

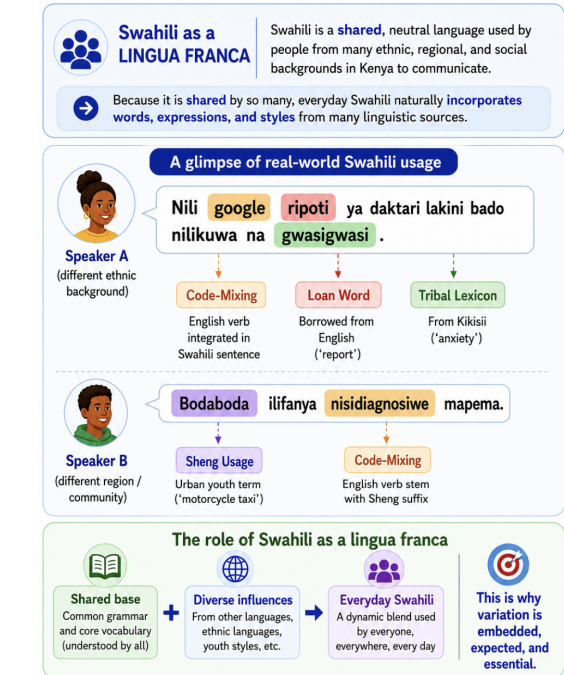


Figure 1: Illustration of how sociolinguistic variations are embedded within common language use. Examples of variation in Swahili as a lingua franca.

reflect deeper cultural and societal variation in language use and interpretation (Hershcovich et al., 2022; Cao et al., 2023).

In multilingual and multitribal societies, these variations occur within commonly used lingua franca such as Swahili (Figure 1), indicating that they are embedded in everyday language use. Speakers from different ethnolinguistic communities contribute distinct cultural expressions to a shared lingua franca. We refer to this sociolinguistic diversity as *tribal heterogeneity*. We call the recurrent feature-level manifestations of this variation *cultural-linguistic artifacts (CLAs)*. We use the term *cultural-linguistic misalignment* to describe the mismatch between these artifacts and the representations learned by NLP systems. Prior work has documented that NLP systems struggle

with many of these non-standard language varieties, including dialects, code-switching, and informal registers (Blodgett et al., 2016; Jørgensen et al., 2015; Joshi et al., 2020; Kantharuban et al., 2023) such as Sheng, an urban Swahili-based sociolect. Existing explanations largely attribute these failures to data scarcity or domain mismatch (Ghosal et al., 2025; Joshi et al., 2020; Kantharuban et al., 2023), providing limited insight into how culturally grounded linguistic variation contributes to model behavior. These limitations reveal important gaps. First, multilingual NLP lacks feature-level frameworks for operationalizing such variation as measurable linguistic artifacts, limiting analysis of cultural-linguistic misalignment. Second, when constructing lexicons of cultural-linguistic artifacts to help measure misalignment, tribal heterogeneity creates unequal familiarity with sociolinguistic variation. As a result, culturally grounded forms may not be uniformly recognizable across speakers, causing some forms of sociolinguistic variation to remain undetected across annotators. Third, the downstream implications of cultural-linguistic misalignment remain poorly understood, particularly its relationship to model error, interpretability, and fairness disparities in NLP systems. These gaps point to the need for a framework that treats sociolinguistic variation as a measurable source of model behavior.

To address these gaps, we introduce T-CLA, a taxonomy of cultural-linguistic artifacts in Swahili text. We conceptualize T-CLA as a representational artifact that formalizes culturally situated linguistic knowledge for use in AI evaluation and design (Abbasi et al., 2024). Construction of T-CLA is guided by Systemic Functional Linguistics (Halliday, 2004; Abbasi and Chen, 2008), which conceptualizes language as meaning-making in social context. We then validate T-CLA through an annotation study using Ensemble Annotation Accuracy and Adjusted Ensemble Annotation Accuracy, which distinguish failures of detection from failures of categorization and enable analysis of unequal recognizability across annotators. Finally, we use T-CLA to analyze the downstream implications of cultural-linguistic misalignment through model prediction error and fairness disparities across tribal groups. Our analyses reveal that: (1) annotation disagreement arises primarily from unequal detection of sociolinguistic variation; (2) cultural-linguistic artifacts exhibit relationships with downstream model prediction error across tasks and mod-

els; (3) subgroup analyses reveal fairness disparities across tribal groups.

This paper makes three contributions. First, we introduce T-CLA, a feature-level taxonomy that operationalizes culturally grounded sociolinguistic variation in Swahili as measurable linguistic artifacts. Second, we propose an annotation-based validation framework that distinguishes failures of artifact detection from those of artifact categorization. Third, we demonstrate that CLA categories are associated with model prediction error across tasks and models and show how T-CLA can be used alongside subgroup fairness analyses to identify culturally situated variation in model behavior.¹

2 Related Work

2.1 Sociolinguistic Variation in NLP

A growing body of work shows that NLP systems struggle with sociolinguistic variation, including dialects, code-switching, and informal or non-standard varieties (Blodgett et al., 2020; Jørgensen et al., 2015; Joshi et al., 2020; Kantharuban et al., 2023). Work on code-switched NLP has further shown that multilingual and mixed-language inputs remain challenging even when the underlying languages are individually supported (Khanuja et al., 2020; Piergallini et al., 2016; Hamed et al., 2025; Borisov et al., 2025; Kellert and Tyagi, 2025; Yoo et al., 2025). Research on language variation argues that such variation is shaped by typological, cognitive, and social factors (Ponti et al., 2019). Prior NLP research has long shown that models can rely on surface artifacts in data as predictive signals, producing poorly understood behavior (Gururangan et al., 2018). We build on this literature by treating sociolinguistic variation not only as a source of performance degradation, but also as an interpretable object of analysis via a taxonomy.

2.2 Cross-cultural NLP & Cultural Alignment

Recent work in cross-cultural NLP argues that language technologies must account for culture as a constitutive factor in meaning, interaction, and evaluation (Hershcovich et al., 2022). This perspective has motivated a growing body of work on how model behavior varies across cultural settings and human societies (Cao et al., 2023). In parallel, work on African-centered NLP has emphasized

¹Our code is available on GitHub: <https://github.com/nd-hal/T-CLA>.

that language technologies for African languages require approaches that reflect local linguistic and cultural realities rather than treating them as peripheral to mainstream NLP development (Adebara and Abdul-Mageed, 2022). Our work contributes to this line of research by operationalizing cultural-linguistic misalignment at the level of observable linguistic features in Swahili text.

2.3 Systemic Functional Linguistics & NLP

Systemic Functional Linguistics (SFL) conceptualizes language as a socially situated meaning-making system in which linguistic choices simultaneously encode experiential, interpersonal, and textual functions (Halliday, 2004). Unlike purely structural approaches to language, SFL emphasizes how discourse reflects social context, speaker identity, communicative intent, and cultural meaning. These properties make SFL well-suited for analyzing sociolinguistic variation, multilingual discourse, and culturally embedded forms of language use. Prior work has applied SFL in computational linguistics and NLP to discourse analysis (Costetchi, 2013; Ocdenaria et al., 2025), sentiment analysis (Tian et al., 2023), genre classification (Whitelaw and Patrick, 2004), authorship analysis (Cortal and Finkel, 2025), and social media language modeling (Hovy and Yang, 2021). Building on this perspective, we use SFL as a guiding kernel theory for constructing T-CLA, allowing us to organize cultural-linguistic artifacts by the functional roles they play in expressing experience, negotiating social relations, and structuring discourse.

2.4 Taxonomies, Artifacts, & Representations

Taxonomies are widely used in NLP to organize complex phenomena and support systematic analysis of model behavior. Prior work has shown that structured evaluation frameworks can make model failures more interpretable in behavioral and stress testing studies (Ribeiro et al., 2020; Naik et al., 2018). Related work in fairness and bias has similarly relied on structured representations of demographic and social attributes to diagnose disparities in predictions (Friedler et al., 2019; Lalor et al., 2022; Guo et al., 2022). Empirical studies demonstrate that language use varies across cultural contexts, for example in mental health discourse, where expressions of distress and support differ across regions and communities (Rai et al., 2025). Despite these advances, existing approaches largely focus on identifying cross-cultural differ-

ences at the dataset or task level. More recent work in culturally-aware NLP has proposed broader taxonomies of cultural dimensions and adaptation strategies, but these approaches remain comparatively coarse-grained and are not directly tied to observable linguistic features (Liu et al., 2025). Our work builds on these directions and differs by introducing a feature-level taxonomy of sociolinguistic variation that links culturally grounded language use to model error and fairness disparities. Thus, we move from documenting cross-cultural differences to operationalizing cultural-linguistic misalignment as a measurable and analyzable phenomenon.

3 T-CLA Coding, Construction, Validation, & Applications

3.1 Coding Procedure of T-CLA

We construct T-CLA as a taxonomy of sociolinguistic variation grounded in observed language use. It is designed to be both extensible and usable within steering and safety frameworks because it maps linguistic artifacts to tokens. Beyond identifying culturally situated linguistic variation, such representations can support inference-time interventions in which culturally grounded exemplars are used to steer model behavior toward more contextually aligned outputs (Zhang et al., 2026). Prior work demonstrates that explicitly teaching models subtle linguistic distinctions through prompting can reduce downstream errors (Gligorić et al., 2024), motivating feature-level representations that can support similar targeted interventions. As shown in Figure 2, T-CLA is developed using an inductive coding procedure (Parfenova et al., 2025). Details of our data are presented in Appendix A. We begin by examining all 8,680 free-text responses in our Swahili dataset, spanning four health-related psychometric tasks. We focus on health-related responses because healthcare provides a consequential setting in which language use can reflect culturally grounded expressions, local terminology, code-mixing, lexical borrowing, and community-specific ways of communicating health experiences. These forms of variation make the domain particularly useful for examining how multilingual models respond to culturally situated language use. The four psychometric tasks also provide downstream outcomes for evaluating associations between these artifacts and model behavior.

Each response was manually reviewed by a flu-

1 Open-Ended Responses

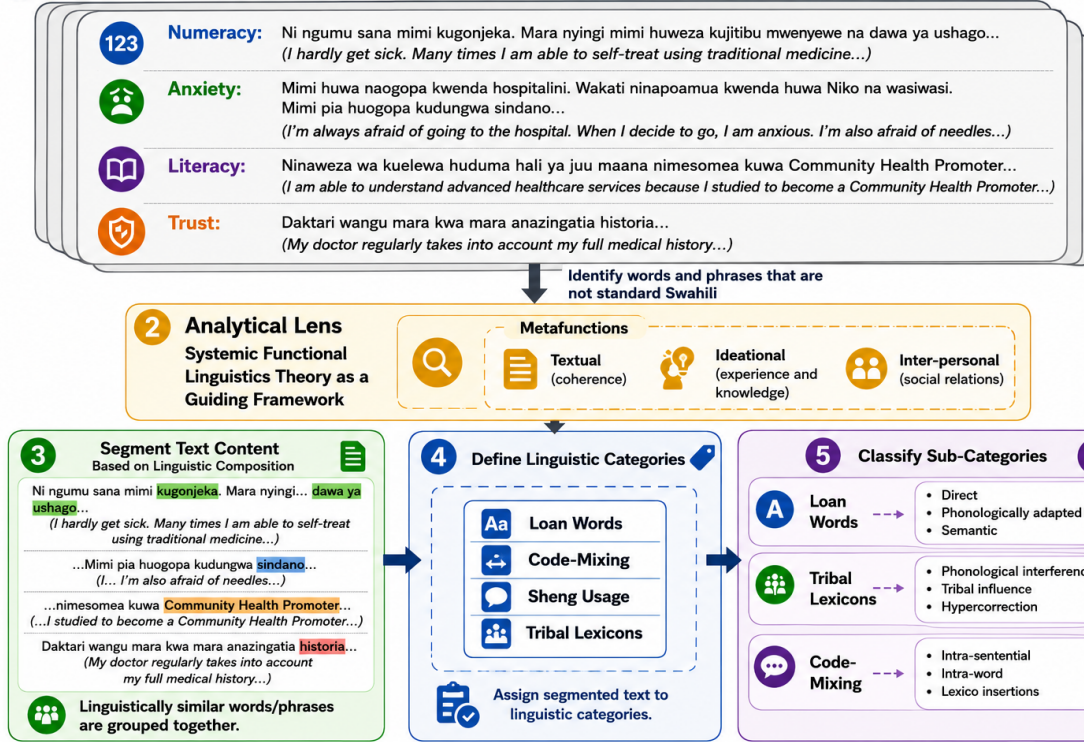


Figure 2: Coding Open-Ended text responses to questions on four psychometric tasks (Health Numeracy, Anxiety Visiting Doctors, Subjective Literacy, and Trust in Physicians). This inductive coding process is based on *grounded theory* (Glaser and Strauss, 2017) and the linguistic categorization is informed by *systemic functional linguistics theory* (Halliday, 2004).

ent Swahili speaker and annotated for the presence of CLAs. The process followed an iterative grounded approach (Charmaz, 2006; Baumer et al., 2017; Nelson, 2020), where candidate artifacts were identified, refined, and consolidated through multiple passes over the data. During this process, similar or overlapping artifacts were merged, and category boundaries were refined to ensure consistency and interpretability. Coding continued until no new artifact types were identified in subsequent passes. To ensure consistency, the annotator iteratively refined category definitions and revisited previously coded responses to maintain coherence across the dataset.

3.2 Constructing T-CLA

The resulting taxonomy consists of 170 individual CLAs, which are organized into four higher-level categories (Figure 3): **Code-Mixing**, referring to the integration of elements from multiple languages within a single utterance (Khanuja et al., 2020; Piergallini et al., 2016); **Sheng Usage**, an urban Swahili-based sociolect characterized by lexical innovation and hybridization (Githiora, 2002);

Tribal Lexicons, capturing vocabulary influenced by speakers' ethnolinguistic backgrounds (Mazrui, 1998); **Loan Words**, reflecting historical and ongoing lexical borrowing from other languages (Kang, 2011). These categories correspond to distinct sociolinguistic processes, including language contact, identity construction, ethnolinguistic influence, and lexical borrowing. While not exhaustive, they represent the most salient and recurrent forms of variation observed in our dataset, particularly those associated with model prediction errors. Importantly, T-CLA operates at the feature level, linking culturally grounded language use to observable linguistic markers in text. This enables direct integration with computational models, including token-level analysis and downstream evaluation of model behavior.

To characterize how these artifacts occur in the corpus, we additionally examine their prevalence and co-occurrence patterns. Appendix A.1 reports the frequency of responses containing each CLA category, the distribution and summary statistics of CLAs per response, and pairwise co-occurrence frequencies. Responses contain between 0 and 8

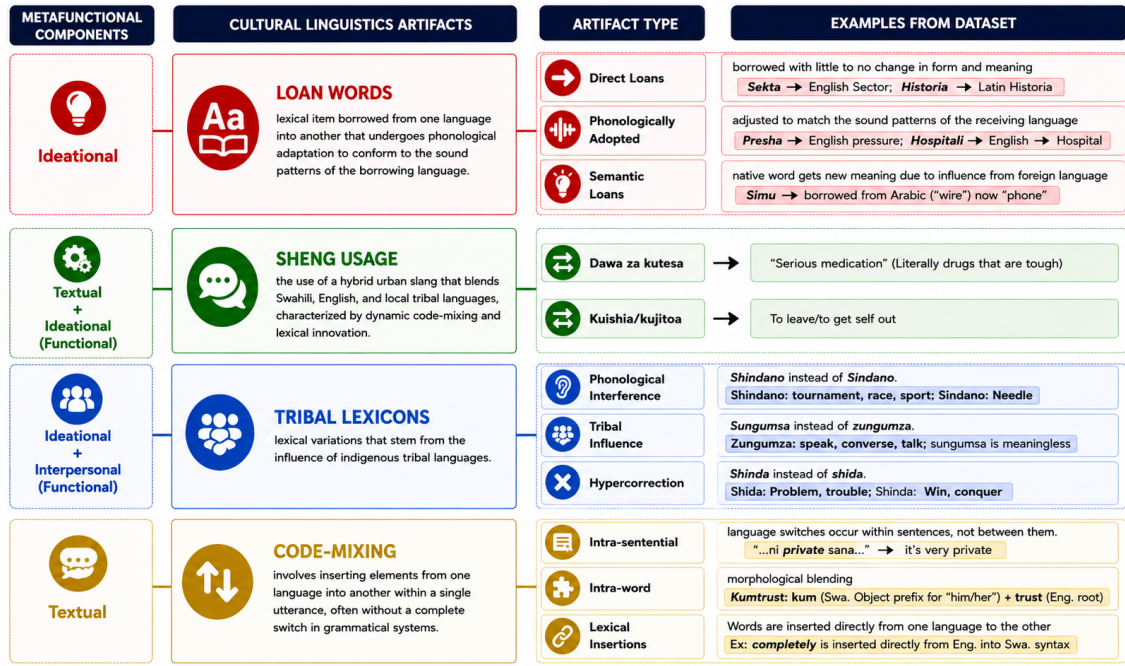


Figure 3: **T-CLA**, a taxonomy of cultural-linguistic artifacts categorized as Loan Words (Kang, 2011), *Sheng* (Githiora, 2002), Tribal Lexicons (Amol et al., 2024), and Code-Mixing (Piergallini et al., 2016)

CLAs, with a mean of 1.64 and a median of 1. Loan Words are the most prevalent category, followed by Tribal Lexicons and Code-Mixing, while Sheng Usage is comparatively rare. Co-occurrence is most common between Loan Words and Tribal Lexicons, followed by Code-Mixing and Loan Words. These patterns indicate that culturally grounded linguistic variation frequently appears in combination and not as isolated features.

3.3 T-CLA Validation: Annotation Study

To evaluate the perceptibility and consistency of T-CLA, we designed a survey-based annotation study implemented in Qualtrics and recruited 182 annotators via Prolific. The annotation task is constructed over the 170 distinct CLAs. To ensure sufficient coverage, we partition these CLAs into 17 sets, each containing 10 sentences. Each set includes a mix of the four CLA categories. Annotators were randomly assigned to an item set to obtain multiple independent annotations for each CLA. The initial annotation dataset contained 1,830 annotations across the 170 CLAs. For each item, annotators are presented with a sentence and asked to assign the most appropriate category based on their knowledge and interpretation. In addition to the four CLA categories, we included Standard Swahili as a response option. This lets us test whether annotators perceive an artifact at all, not just which one. To

maintain data quality, we included attention-check items and recorded completion times. Before the annotation task, participants were presented with category definitions and worked examples with explanations of the correct label. This calibration was designed to align annotators' interpretation and reduce variability due to subjective understanding of category boundaries. Consistent with this design, we observe relatively high agreement and reliable identification of several categories in the results (§4). Additional experimental design information is included in Appendix B.1. Due to differences in the number of CLAs per category, the resulting dataset exhibits some imbalance. After filtering for quality, the resulting dataset consists of 1,560 annotations.

3.3.1 Validation Metrics

To evaluate whether T-CLA defines perceptible and distinguishable forms of sociolinguistic variation, we measure both annotation accuracy and chance-corrected agreement using ensemble-based annotation metrics. Our validation distinguishes between two sources of annotation error: (1) failures to correctly classify a perceived artifact and (2) failures to perceive sociolinguistic variation altogether, where annotators instead default to standard Swahili. This distinction is important because several forms of cultural-linguistic artifacts in Swahili are subtle, so-

cially embedded, and unevenly perceptible across speakers due to differences in socio- and ethno-linguistic exposure (Ponti et al., 2019; Blodgett et al., 2020; Hovy and Yang, 2021).

Ensemble Annotation Accuracy (EAA). Let i index a cultural-linguistic artifact item and let y_i denote the expert-assigned category for item i . For each item, we compute an ensemble annotation label $\hat{y}_i$ as the majority-vote response across annotators. EAA is then defined as: $EAA_i = \mathbb{I}(\hat{y}_i = y_i)$ where $\mathbb{I}(\cdot)$ is the indicator function. Category-level EAA is computed by averaging across all items belonging to category c : $EAA_c = \frac{1}{|I_c|} \sum_{i \in I_c} EAA_i$, where I_c denotes the set of items whose expert-assigned label belongs to category c .

Adjusted Ensemble Annotation Accuracy (AEAA). To distinguish failures of detection from failures of classification, we additionally compute AEAA. In this setting, responses labeled as standard Swahili are excluded prior to computing the ensemble label. The adjusted ensemble label $\hat{y}_i^{adj}$ is therefore defined as the majority-vote label among remaining non-standard responses. AEAA is computed analogously: $AEAA_i = \mathbb{I}(\hat{y}_i^{adj} = y_i)$. This adjustment isolates category classification performance conditional on successful detection of sociolinguistic variation, thereby separating failures of detection from failures of categorization. Comparing EAA and AEAA therefore allows us to determine whether annotation errors arise primarily from cross-category confusion or from failures to detect culturally grounded non-standard linguistic forms altogether.

Chance-Corrected Agreement and Class-Sensitive Performance. We compute Cohen’s κ (Cohen, 1960) to measure chance-corrected agreement between ensemble annotation labels and expert-assigned labels. Cohen’s κ is defined as: $\kappa = \frac{p_o - p_e}{1 - p_e}$, where p_o denotes observed agreement and p_e denotes agreement expected by chance. We compute κ separately for EAA and AEAA in order to evaluate annotation reliability before and after accounting for default standard Swahili responses. Because aggregate accuracy can obscure differences across CLA categories, we additionally report macro-averaged F1 and category-level F1 scores. These measures give equal weight to each CLA category and provide a more informative assessment when category frequencies are uneven.

Uncertainty Estimation. We quantify uncertainty in the validation estimates using nonparametric bootstrap confidence intervals. We repeatedly resample CLA items with replacement and recompute the corresponding validation statistics for each bootstrap sample. We report 95% bootstrap confidence intervals for EAA, AEAA, Cohen’s κ , and the normalized confusion-matrix estimates.

3.4 T-CLA Applications

3.4.1 As an Analytic Instrument

Having established T-CLA as a perceptible and empirically distinguishable taxonomy, we next examine whether these CLAs are associated with variation in model prediction error. We analyze model behavior at the response level to complement aggregate performance metrics. Prior work in this setting demonstrates that culturally grounded forms may be associated with prediction errors (Oketch et al., 2025a). Because many PLMs and LLMs are trained predominantly on standardized languages, culturally grounded forms may introduce prediction errors even within a shared language. We therefore use T-CLA as an analytic tool for studying how different forms of sociolinguistic variation are associated with downstream model performance. We define prediction error as the difference between the human-assigned and model-generated score ($\Delta_i = \text{Human}_i - \text{Model}_i$) and examine its variation by CLA. The procedure for generating model scores is described in Appendix C.

To quantify the relationship between individual CLAs and model error, we first fit separate univariate ordinary least squares (OLS) regression models for each CLA, psychometric task, and language model. For response i and CLA j , we estimate $\Delta_i = \beta_0 + \beta_j X_{ij} + \varepsilon_i$, where X_{ij} is a binary indicator for whether CLA j is present in response i . The coefficient β_j estimates the association between the presence of a given CLA and model prediction error, while R^2 quantifies the proportion of variance in prediction error explained by that CLA. To account for multiple comparisons, we apply the Benjamini–Hochberg false discovery rate (FDR) correction across the 16 tests associated with each language model (four psychometric tasks $\times$ four CLA categories). We next fit multivariable models containing all four CLA categories simultaneously, $\Delta_i = \beta_0 + \sum_{j=1}^4 \beta_j X_{ij} + \varepsilon_i$. These models estimate the association between each CLA and prediction error while accounting for the remaining CLA

categories, allowing us to assess whether associations observed in the univariate analyses persist after accounting for co-occurring artifacts.

Finally, we examine whether the association between CLAs and model error varies when multiple CLA categories co-occur by including all pairwise interactions: $\Delta_i = \beta_0 + \sum_{j=1}^4 \beta_j X_{ij} + \sum_{k>j} \beta_{jk} (X_{ij} \times X_{ik}) + \varepsilon_i$. Here, $X_{ij} \times X_{ik}$ represents the co-occurrence of CLA categories j and k within the same response, and β_{jk} captures whether their joint presence is associated with prediction error beyond their individual associations. Across all specifications, positive coefficients indicate underprediction by the model, while negative coefficients indicate overprediction.

3.4.2 As a Sociolinguistic Fairness Lens

Because CLAs are socially and regionally patterned, their uneven distribution across tribal groups may be associated with differences in model behavior. We therefore use T-CLA as an analytic lens for studying subgroup-level fairness disparities. We selected a mix of multilingual (mBERT, XLM-RoBERTa), regional (AfriBERTa), and Swahili-specific (SwahBERT) models commonly used in African NLP tasks. Our dataset includes participants from more than 20 tribal affiliations, providing a setting for examining whether model performance varies across culturally and linguistically heterogeneous groups. Beyond aggregate model performance, prediction errors can vary across demographic and cultural groups, creating disparities that are obscured by population-level evaluation (Lalor et al., 2024; Oketch et al., 2025b; Bevilacqua et al., 2025). For the subgroup analyses reported here, we focus on two highly represented tribes in the dataset, Kikuyu and Kamba. For each analysis, we binarized the tribe variable separately, comparing the focal tribe against all other tribes. We quantify fairness via two measures, xAUC and Disparate Impact (DI). xAUC (Kallus and Zhou, 2019) measures disparities in the ranked ordering of predicted scores across groups: $\Delta \text{xAUC} = p(R_1^b \leq R_0^a) - p(R_1^a \leq R_0^b)$. A positive ΔxAUC indicates that individuals from group a in the reference class $T = 1$ are ranked higher by the model than those from group b in the comparison class $T = 0$. DI quantifies the relative rate at which two groups, reference and comparison, receive predicted positive outcomes, serving as a proxy for group-level equity (Lalor et al., 2022; Friedler et al., 2019; Barocas and Selbst, 2016)

$DI = \frac{p(\hat{y}=1|T=0)}{p(\hat{y}=1|T=1)}$. A DI value below 1 indicates that the comparison group receives positive predictions at a lower rate than the reference group, signaling potential group-level disparity.

4 Results and Discussion

4.1 T-CLA Validation Results

Table 1 summarizes category-wise human annotation performance across the four CLA categories. Overall, the results indicate strong agreement between human ensemble annotations and expert-assigned categories.

CLA Category	EAA	AEAA	κ_E	κ_A	F1 _E	F1 _A
Code-Mixing	0.92	0.92	0.91	0.91	0.94	0.94
Loan Words	0.85	0.97	0.87	0.93	0.89	0.96
Sheng Usage	0.80	1.00	0.89	1.00	0.89	1.00
Tribal Lexicons	0.87	1.00	0.92	0.98	0.93	0.98
Overall	0.88	0.96	0.82	0.94	0.91	0.97

Table 1: Category-wise human annotation performance. EAA and AEAA denote Ensemble Annotation Accuracy and Adjusted Ensemble Annotation Accuracy, respectively. κ_E and κ_A denote Cohen’s κ under EAA and AEAA. F1_E and F1_A denote category-level F1 scores; overall F1 values are macro-averaged across CLA categories.

The comparison between EAA and AEAA reveals an important distinction between artifact detection and classification. Performance improves after responses labeling an artifact as Standard Swahili are excluded prior to ensemble aggregation, indicating that disagreement is driven primarily by failures to detect sociolinguistic variation in lieu of confusion among CLA categories once an artifact is perceived. Bootstrap estimates further support the stability of the aggregate results, with 95% confidence intervals of 82.35–92.35% for EAA and 92.26–98.81% for AEAA. Category-level uncertainty estimates for the confusion matrices are reported in Appendix D.1.

Figure 4 provides a more detailed view of this distinction through row-normalized confusion matrices. Under EAA, most disagreement is concentrated in the Standard Swahili column rather than across CLA categories. This pattern is particularly evident for Loan Words and Tribal Lexicons, whereas Code-Mixing exhibits comparatively little confusion with Standard Swahili. Under AEAA, cross-category confusion is limited, further indicating that some artifacts are more difficult to perceive as sociolinguistically distinct than to classify once

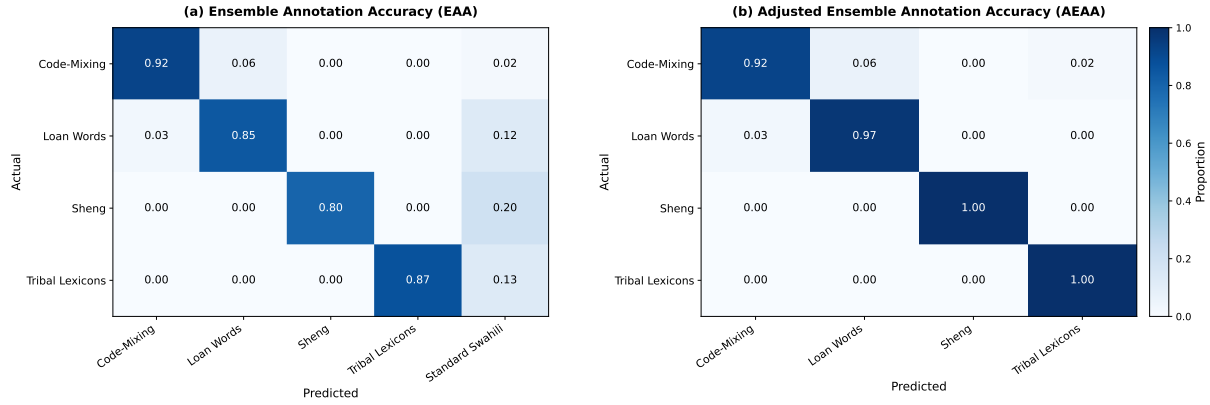


Figure 4: Confusion matrices for Ensemble Annotation Accuracy and Adjusted Ensemble Annotation Accuracy.

detected. Sheng Usage exhibits a similar pattern in the observed annotations, but this result should be interpreted cautiously given the small number of Sheng artifacts in the validation set.

4.2 LLM-Based Validation

As an additional robustness check, we evaluated whether a frontier LLM could independently apply the T-CLA taxonomy using the same category definitions and annotation examples provided to human annotators. Table 2 summarizes GPT-5.4 performance relative to the expert-assigned labels.

CLA Category	EAA	AEAA	κ_E	κ_A	F1 _E	F1 _A
Code-Mixing	0.63	0.63	0.68	0.67	0.77	0.77
Loan Words	0.75	0.93	0.51	0.65	0.73	0.80
Sheng Usage	0.20	0.20	0.27	0.27	0.29	0.29
Tribal Lexicons	0.97	1.00	0.85	0.87	0.88	0.90
Overall	0.73	0.80	0.60	0.70	0.67	0.69

Table 2: Category-wise GPT-5.4 annotation performance. EAA and AEAA denote Ensemble Annotation Accuracy and Adjusted Ensemble Annotation Accuracy, respectively. κ_E and κ_A denote Cohen’s κ under EAA and AEAA. F1_E and F1_A denote category-level F1 scores; overall F1 values are macro-averaged across CLA categories.

GPT-5.4 performed strongly on Tribal Lexicons and Loan Words but showed considerably lower performance on Sheng Usage and Code-Mixing. This pattern is consistent with evidence from other annotation settings showing that instances that are more difficult for humans to interpret can also be more difficult for language models to classify (Alipour et al., 2026). These results show that T-CLA can be operationalized by an LLM while reinforcing the value of human annotation as the primary validation standard.

4.3 T-CLA-guided Error Analysis Results

We first examine the association between each CLA category and model prediction error using the univariate regression models defined in §3.4.1. Table 7 reports the estimated coefficients and corresponding R^2 values across psychometric tasks and language models. After Benjamini–Hochberg correction, significant associations remain across multiple CLA categories, tasks, and models, although their magnitude and direction vary considerably. The generally modest R^2 values indicate that individual CLA categories explain only a limited proportion of prediction-error variance on their own. We next examine whether these associations persist when the four CLA categories are modeled jointly. The multivariable estimates in Table 8 show that most associations identified in the univariate analyses remain significant after accounting for the presence of other CLA categories. Model fit also varies across tasks and language models, with R^2 values ranging from 0.021 to 0.239. These results suggest that the observed associations are not explained solely by co-occurrence among CLA categories. Figure 5 further examines the multivariable main effects and pairwise interactions. In the left panel, individual categories contribute unevenly to error, with Sheng usage and tribal lexicons exhibiting the largest coefficient magnitudes. These effects are pronounced for mBERT and SwahBERT. The right panel shows that the relationship between CLAs and prediction error can also vary when multiple CLA categories co-occur. These results indicate that culturally grounded linguistic variation is associated with model prediction error in ways that depend on both the type of linguistic artifact and the model-task setting.

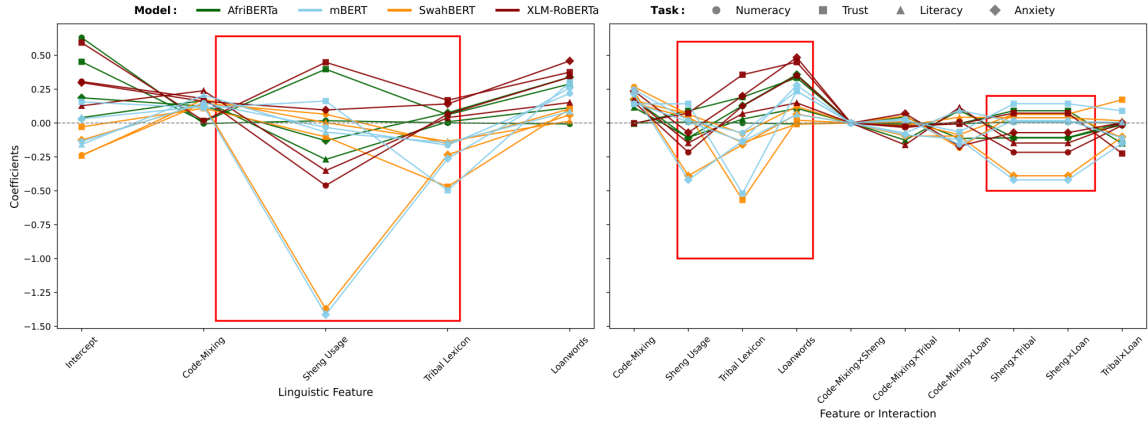


Figure 5: Regression results for multivariable main effects (left) and interaction effects (right) on prediction error ($\hat{\Delta}$). Red boxes highlight coefficients with larger deviations from zero.

4.4 T-CLA & Fairness Results

Figure 6 shows differences in ranking quality for two highly represented tribes in our dataset, Kikuyu and Kamba. Models with strong overall performance can still exhibit large $\Delta xAUC$ deviations, indicating ranking disparities across groups. For example, SwahBERT achieves high AUC on Trust but shows a large $\Delta xAUC$ for Kamba speakers.

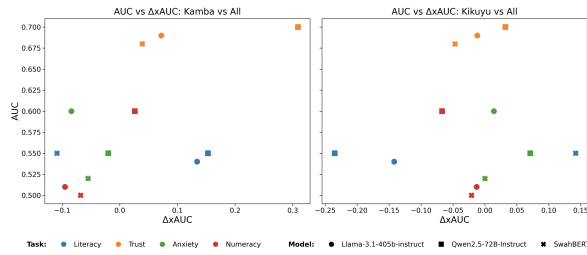


Figure 6: Plot of model performance (AUC) against fairness $\Delta xAUC$ (Kamba vs. All on the left, Kikuyu vs. All on the right)

The error and fairness analyses show that model performance varies both across forms of culturally grounded linguistic variation and across tribal groups. Figure 5 identifies CLA categories and combinations associated with variation in prediction error, while Figure 6 shows that model ranking performance also differs across tribal groups. These findings demonstrate the value of examining culturally grounded linguistic variation alongside subgroup fairness metrics when evaluating multilingual NLP systems.

5 Conclusion

We introduced T-CLA, a taxonomy of cultural-linguistic artifacts grounded in observed sociolin-

guistic variation in Swahili. We defined these artifacts as feature-level expressions of culturally situated language use and framed cultural-linguistic misalignment as the mismatch between these artifacts and model representations. Through a large-scale annotation study, we showed that the T-CLA categories are perceptible and empirically distinguishable, while also identifying differences in their perceptual salience. Using T-CLA as an analytic instrument, we found that culturally grounded linguistic variation is associated with model prediction error. These associations vary in magnitude and direction across CLA categories, models, and tasks, and most persist when the CLA categories are modeled jointly. Pairwise interaction analyses further show that associations with model error can vary when multiple forms of sociolinguistic variation co-occur. Our fairness analysis additionally identifies differences in model ranking performance across tribal groups, demonstrating the importance of considering culturally grounded linguistic variation alongside subgroup performance when evaluating multilingual NLP systems.

T-CLA provides a framework for moving beyond aggregate measures of multilingual model performance toward feature-level analysis of culturally situated language use. Although the taxonomy developed here is specific to Kenyan Swahili and the health-related context studied, the methodology for identifying, validating, and analyzing cultural-linguistic artifacts can be adapted to other languages, cultural contexts, domains, and downstream tasks. This provides a basis for culturally informed dataset auditing, model evaluation, and the development of interventions that better account for socially embedded linguistic variation.

Limitations

While this work takes an initial step toward formalizing and empirically validating cultural-linguistic misalignment through a feature-level taxonomy, it is not without limitations. First, T-CLA is constructed from Swahili health-related responses from Kenyan speakers, and may not generalize directly to other languages, domains, or cultural contexts. Second, T-CLA is developed through an inductive coding process led by a single annotator, although it is subsequently validated through a large-scale annotation study. Third, our analysis considers a limited set of pretrained and instruction-tuned models, and findings may vary with different training regimes. Finally, our regression-based analysis identifies associations between cultural-linguistic artifacts and model error, but does not establish causal relationships.

Acknowledgments

This work was supported in part by the National Science Foundation (IIS-2403438). Any opinions, findings, conclusions, or recommendations expressed are those of the authors and do not necessarily reflect the views of the sponsor.

References

- Ahmed Abbasi and Hsinchun Chen. 2008. Cybergate: A design framework and system for text analysis of computer-mediated communication1. *Mis Quarterly*, 32(4):811–837.
- Ahmed Abbasi, David Dobolyi, John P Lalor, Richard G Netemeyer, Kendall Smith, and Yi Yang. 2021. Constructing a psychometric testbed for fair natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3748–3758.
- Ahmed Abbasi, Jeffrey Parsons, Gautam Pant, Olivia R Liu Sheng, and Suprateek Sarker. 2024. Pathways for design research on artificial intelligence. *Information Systems Research*, 35(2):441–459.
- Ife Adebara and Muhammad Abdul-Mageed. 2022. [Towards afrocentric NLP for African languages: Where we are and where we can go](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3814–3841, Dublin, Ireland. Association for Computational Linguistics.
- Shayan Alipour, Shruti Phadke, Seyed Shahabeddin Mousavi, Amirhossein Afsharrad, Morteza Zihayat, and Mattia Samory. 2026. The gray area: Characterizing moderator disagreement on reddit. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 20, pages 58–75.
- Cynthia Jayne Amol, Everlyn Asiko Chimoto, Rose Delilah Gesicho, Antony M Gitau, Naome A Etori, Carington Kinyanjui, Steven Ndung’u, Lawrence Moruye, Samson Otieno Ooko, Kavengi Kitonga, and 1 others. 2024. State of nlp in kenya: A survey. *arXiv preprint arXiv:2410.09948*.
- Solon Barocas and Andrew D Selbst. 2016. Big data’s disparate impact. *Calif. L. Rev.*, 104:671.
- Eric PS Baumer, David Mimno, Shion Guha, Emily Quan, and Geri K Gay. 2017. Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the Association for Information Science and Technology*, 68(6):1397–1410.
- Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.
- Marielena Bevilacqua, Kezia Oketch, Ruiyang Qin, Will Stamey, Xinyuan Zhang, Yi Gan, Kai Yang, and Ahmed Abbasi. 2025. When automated assessment meets automated content generation: Examining text quality in the era of gpts. *ACM Transactions on Information Systems*, 43(2):1–36.
- Su Lin Blodgett, Solon Barocas, Hal Daumé Iii, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of “bias” in nlp. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5454–5476.
- Su Lin Blodgett, Lisa Green, and Brendan O’Connor. 2016. Demographic dialectal variation in social media: A case study of african-american english. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1119–1130.
- Maksim Borisov, Zhanibek Kozhirbayev, and Valentin Malykh. 2025. Low-resource machine translation for code-switched kazakh-russian language pair. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 66–76.
- Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. [Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study](#). In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 53–67, Dubrovnik, Croatia. Association for Computational Linguistics.

- Kathy Charmaz. 2006. *Constructing grounded theory: A practical guide through qualitative analysis*. sage.
- Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.
- Gustave Cortal and Alain Finkel. 2025. Formalizing style in personal narratives. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7322–7337.
- Eugeniu Costetchi. 2013. A method to generate simplified systemic functional parses from dependency parses. In *Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013)*, pages 68–77.
- Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 329–338.
- Soumya Suvra Ghosal, Soumyabrata Pal, Koyel Mukherjee, and Dinesh Manocha. 2025. Promptrefine: Enhancing few-shot performance on low-resource indic languages with example selection from related example banks. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 351–365.
- Chege Githiora. 2002. Sheng: peer language, swahili dialect or emerging creole? *Journal of African cultural studies*, 15(2):159–181.
- Barney Glaser and Anselm Strauss. 2017. *Discovery of grounded theory: Strategies for qualitative research*. Routledge.
- Kristina Gligorić, Myra Cheng, Lucia Zheng, Esin Durmus, and Dan Jurafsky. 2024. Nlp systems that can’t tell use from mention censor counterspeech, but teaching the distinction helps. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5942–5959.
- Yue Guo, Yi Yang, and Ahmed Abbasi. 2022. Autodebias: Debiasing masked language models with automated biased prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1012–1023.
- Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. [Annotation artifacts in natural language inference data](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.
- Michael AK Halliday. 2004. [An introduction to functional grammar](#). revised by cmim matthiessen. London: Arnold.
- Injy Hamed, Caroline Sabty, Slim Abdennadher, Ngoc Thang Vu, Thamar Solorio, and Nizar Habash. 2025. A survey of code-switched arabic nlp: Progress, challenges, and future directions. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4561–4585.
- Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. [Challenges and strategies in cross-cultural NLP](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.
- Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies*, pages 588–602.
- Anna Jørgensen, Dirk Hovy, and Anders Søgaard. 2015. Challenges of studying and processing dialects in social media. In *Proceedings of the workshop on noisy user-generated text*, pages 9–18.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. [The state and fate of linguistic diversity and inclusion in the NLP world](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.
- Nathan Kallus and Angela Zhou. 2019. The fairness of risk scores beyond classification: Bipartite ranking and the xauc metric. *Advances in neural information processing systems*, 32.
- Yoonjung Kang. 2011. Loanword phonology. *The Blackwell companion to phonology*, 4:1003–1026.
- Anjali Kantharuban, Ivan Vulić, and Anna Korhonen. 2023. [Quantifying the dialect gap and its correlates across languages](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7226–7245, Singapore. Association for Computational Linguistics.
- Olga Kellert and Nemika Tyagi. 2025. Where and how do languages mix? a study of spanish-guaraní code-switching in paraguay. In *Proceedings of the 7th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 26–31.

- Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020. Gluecos: An evaluation benchmark for code-switched nlp. *arXiv preprint arXiv:2004.12376*.
- John P Lalor, Ahmed Abbasi, Kezia Oketch, Yi Yang, and Nicole Forsgren. 2024. Should fairness be a metric or a model? a model-based framework for assessing bias in machine learning pipelines. *ACM Transactions on Information Systems*, 42(4):1–41.
- John P Lalor, Yi Yang, Kendall Smith, Nicole Forsgren, and Ahmed Abbasi. 2022. Benchmarking intersectional biases in nlp. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 3598–3609.
- Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2025. [Culturally aware and adapted NLP: A taxonomy and a survey of the state of the art](#). *Transactions of the Association for Computational Linguistics*, 13:652–689.
- Ali A Mazrui. 1998. Neo-dependency and africa’s fragmentation.
- Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353.
- Laura K Nelson. 2020. Computational grounded theory: A methodological framework. *Sociological methods & research*, 49(1):3–42.
- Jose Marie Ocdenaria, Kenneth Lisbo, and 1 others. 2025. Speaking to the inner child: A discourse analysis of healing narratives in digital therapeutic culture. In *Proceedings of the 39th Pacific Asia Conference on Language, Information and Computation*, pages 803–812.
- Kezia Oketch, John P Lalor, and Ahmed Abbasi. 2025a. Benchmarking sociolinguistic diversity in swahili nlp: A taxonomy-guided approach. *arXiv preprint arXiv:2508.14051*.
- Kezia Oketch, John P. Lalor, Yi Yang, and Ahmed Abbasi. 2025b. [Bridging the LLM accessibility divide? performance, fairness, and cost of closed versus open LLMs for automated essay scoring](#). In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, Vienna, Austria and virtual meeting. Association for Computational Linguistics.
- Angelina Parfenova, Andreas Marfurt, Jürgen Pfeffer, and Alexander Denzler. 2025. Text annotation via inductive coding: Comparing human experts to llms in qualitative data analysis. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6456–6469.
- Mario Piergallini, Rouzbeh Shirvani, Gauri Shankar Gautam, and Mohamed Chouikha. 2016. Word-level language identification and predicting codeswitching points in swahili-english language data. In *Proceedings of the second workshop on computational approaches to code switching*, pages 21–29.
- Edoardo Maria Ponti, Helen O’horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3):559–601.
- Sunny Rai, Khushi Shelat, Devansh Jain, Ashwin Kishen, Young Min Cho, Maitreyi Redkar, Samindara Hardikar-Sawant, Lyle Ungar, and Sharath Chandra Guntuku. 2025. Cross-cultural differences in mental health expressions on social media. In *Proceedings of the 3rd Workshop on Cross-Cultural Considerations in NLP (C3NLP 2025)*, pages 132–142.
- Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4902–4912.
- Lin Tian, Xiuzhen Jenny Zhang, Myung Hee Kim, and Jennifer Biggs. 2023. Task and sentiment adaptation for appraisal tagging. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1960–1970.
- Casey Whitelaw and Jon Patrick. 2004. Selecting systemic features for text classification. In *Proceedings of the Australasian Language Technology Workshop 2004*, pages 93–100.
- Haneul Yoo, Cheonbok Park, Sangdoo Yun, Alice Oh, and Hwaran Lee. 2025. Code-switching curriculum learning for multilingual transfer in llms. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7816–7836.
- Denghui Zhang, Rushi Wang, Jiateng Liu, Kezia Oketch, Yiyu Shi, Heng Ji, and Ahmed Abbasi. 2026. Pavlov’s dog and large language models: The double-edged power of context conditioning. *IEEE Intelligent Systems*, 41(2):56–62.

A Appendix: Data

We use a Swahili dataset that consists of four psychometric tasks: Health Literacy, Health Numeracy, Anxiety Visiting Doctors, and Trust in Doctors, dimensions that have been found to be key factors in health intervention settings (Abbasi et al., 2021). The tribal affiliations represented in the dataset span multiple ethnolinguistic groups in Kenya. Bantu groups represented in the sample include communities such as the Kikuyu, Luhya, and Kamba, whose heritage languages belong to

the Bantu branch of the Niger-Congo language family. Nilotic groups include communities such as the Luo and Kalenjin, whose heritage languages belong to the Nilotic branch of the Nilo-Saharan language family. These groupings describe broad linguistic lineages and do not imply linguistic or cultural homogeneity within each group. Figure 7 shows the distribution of participants across the tribal affiliations represented in the dataset.

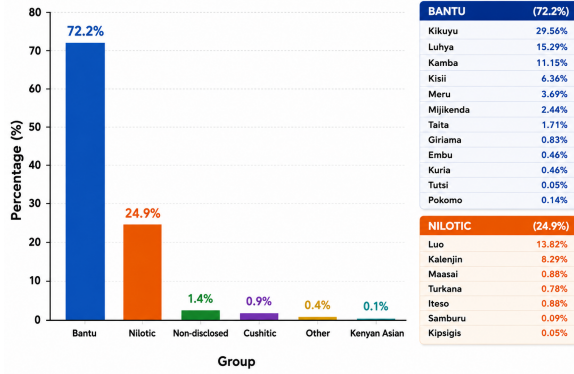


Figure 7: Distribution of study participants across language groups.

To obtain construct-level labels, and hence the human-assigned scores described in §3.4.1, responses from the corresponding survey items were aggregated into a single psychometric score and subsequently normalized to a 0–1 range. Table 3 shows representative examples of participant responses alongside their normalized Anxiety scores.

Score	Swahili Response	English Translation
0.9524	Pesa kwa matibabu hu- nifanya niwe na wasi- Ripoti ya daktari hunipa wasiwasi mwingi maana sijui atasema nau- gua ugonjwa upi.	Money for treatment makes me anxious. A doctor’s report gives me a lot of worry because I do not know which illness they will say I have.
0.3095	Wasiwasi hutokea wakati usipojua shida au tatizo la afya nililionalo au kutojua gharama ya matibabu pamoja na dawa kwa ujumla baada ya vipimo.	Anxiety arises when you do not know the health condition I have or the overall cost of treat- ment and medication af- ter tests.

Table 3: Sample participant responses for the Anxiety task.

A.1 Distribution and Co-occurrence of Cultural-Linguistic Artifacts

The corpus contains 8,680 task-specific responses from 2,170 participants, with each participant pro-

viding one response for each of the four psychometric tasks (2,170 participants $\times$ 4 tasks = 8,680 responses). For the descriptive analyses below, CLA prevalence and co-occurrence are summarized at the participant level, yielding 2,170 participant-level observations. Table 4 reports the prevalence of each CLA category in the corpus.

CLA Category	Responses	Percent (%)
Sheng Usage	17	0.78
Code-Mixing	115	5.30
Loan Words	1,505	69.35
Tribal Lexicons	213	9.82

Table 4: Frequency of responses containing each CLA category. Responses may contain multiple CLA categories; therefore, frequencies are not mutually exclusive and percentages do not sum to 100%.

CLAs per Response	Responses	Percent (%)
0	580	26.73
1	576	26.54
2	427	19.68
3	334	15.39
4	173	7.97
5	54	2.49
6	16	0.74
7	9	0.41
8	1	0.05

Summary Statistic	Value
Minimum	0
Maximum	8
Mean	1.64
Median	1

Table 5: Distribution and summary statistics of the number of CLAs per response.

CLA Category 1	CLA Category 2	Responses	(%)
Loan Words	Tribal Lexicons	156	7.19
Code-Mixing	Loan Words	88	4.06
Loan Words	Sheng Usage	15	0.69
Code-Mixing	Tribal Lexicons	9	0.41
Code-Mixing	Sheng Usage	3	0.14
Sheng Usage	Tribal Lexicons	2	0.09

Table 6: Pairwise co-occurrence frequencies of CLA categories in the annotated corpus.

B Appendix: T-CLA Annotation Information

Prior to beginning the annotation task, all participants viewed an informed consent agreement describing the study purpose, voluntary nature of

participation, expected duration, anonymity of responses, potential risks and benefits, and data-protection procedures. Participants proceeded to the survey only after consenting to participate and were paid \$4 for completing the 12–15 minute survey. A screenshot of the consent page is provided in Figure 8. The data collection protocol was reviewed by the University of Notre Dame Institutional Review Board (IRB; Protocol ID: 26-02-9913) and determined to be exempt under applicable human-subjects research guidelines.

B.1 Appendix: Annotation Guidelines and Examples

This section presents the definitions and examples used in the annotation task to distinguish between cultural-linguistic artifact (CLA) categories. These materials were shown to annotators prior to completing the task.

B.2 Category Definitions

Annotators were provided with the following definitions of cultural-linguistic artifact (CLA) categories:

- **Loanword**: A word borrowed from another language and adapted into Swahili.
- **Sheng**: Informal language that mixes Swahili, English, and other local languages, commonly used in urban Kenyan contexts.
- **Tribal lexicon**: Lexical variations influenced by local or indigenous languages, often reflected in pronunciation or spelling.
- **Code-mixing**: The use of words or phrases from another language within Swahili grammar in the same sentence.
- **Standard Swahili**: Words and constructions that conform to widely accepted, formal Swahili usage, without influence from other languages.

These definitions were presented to annotators prior to the task to guide category selection. They are intentionally kept simple to ensure clarity and consistency in a crowdsourced setting, while still capturing the core distinctions between categories.

B.3 Annotation Examples

Example 1 (Loanword) Sentence: *Dereva* wa gari la wagonjwa aliendesha gari kwa haraka sana

nikawa na wasiwasi.

Explanation: “Dereva” is borrowed from the English word “driver” and adapted to Swahili pronunciation and spelling.

Example 2 (Code-mixing) Sentence: Ni-namwamini daktari wangu *because* ananisikiliza kwa makini na huwa *anaexplain* matibabu yangu vizuri.

Explanation: The sentence mixes English words (“because”, “explain”) with Swahili grammar, forming a code-mixed construction.

Example 3 (Sheng) Sentence: *Mtoi* wangu alipougua usiku wa manane tukiwa *ushago*.

Explanation: “Mtoi” and “ushago” are Sheng words commonly used in Kenyan informal speech.

Example 4 (Tribal lexicon) Sentence: Kwa asilimia hamsini kwa mia nahisi *ninawesapata* taarifa kwa kimsingi za afya na pia kuzielewa kwa usahihi kwa *usaindizi*.

Explanation: These forms reflect influence from local languages on Swahili pronunciation and spelling.

Example 5 (Standard Swahili) Sentence: Kina-chonifanya nihisi wasiwasi zaidi ninapotembelea ofisi ya daktari wangu ni kwamba wakati mwingine sina uhakika na afya yangu.

Explanation: The sentence uses standard Swahili vocabulary and grammar. It does not contain loanwords, code-mixing, Sheng, or tribal lexical influence.

Validation of the Taxonomy of Cultural-Linguistic Artifacts

Informed Consent Agreement

ABOUT THIS RESEARCH

This study examines how Swahili is used in Kenya, including different styles such as standard Swahili, borrowed words, Sheng, regional influence, and mixed language use. We are interested in how native Swahili speakers naturally recognize and interpret these different forms of language.

TAKING PART IN THIS STUDY IS VOLUNTARY

You may choose not to take part in the study or may choose to leave the study at any time. Deciding not to participate, or deciding to leave the study later, will not result in any penalty and will not affect your relationship with the University of Notre Dame.

As an alternative to participating in the study, you may choose not to take part.

WHY IS THIS STUDY BEING DONE?

This study is being conducted to better understand how Swahili speakers perceive different types of language variation. The results will help researchers understand how language changes, how borrowing and mixing occur, and how speakers interpret different styles of Swahili in everyday communication. The findings may contribute to academic research in linguistics and language use in East Africa and beyond.

You were selected as a possible participant because you fulfilled participant geographic criteria. The study is being conducted by **Kezia Oketch** and **Ahmed Abbasi** of the **University of Notre Dame**. The study is funded by **University of Notre Dame**.

WHAT WILL HAPPEN DURING THE STUDY?

If you agree to be in the study, you will be asked to do the following things:

- Read descriptions associated with the project.
- Complete an on-line survey, consisting of several questions written in Swahili as it is used in Kenya. You will be asked to respond to several questions and statements potentially pertaining to these sentences.

The study will require about 12–15 minutes of your time in total. Your responses will be anonymous.

WHAT ARE THE RISKS OF TAKING PART IN THE STUDY?

There are no known risks associated with this study beyond those encountered in normal reading and online survey participation. Some sentences may contain informal or mixed language forms, but no offensive or sensitive content will be included. Additionally, all collected data will be stored on a secure server.

WHAT ARE THE POTENTIAL BENEFITS OF TAKING PART IN THE STUDY?

There are no direct personal benefits to you for participating. However, your responses will help researchers better understand language use and variation in Kenyan Swahili, which may contribute to academic knowledge and future language research.

HOW WILL MY INFORMATION BE PROTECTED?

Efforts will be made to keep your personal information confidential. We cannot guarantee absolute confidentiality. Your personal information may be disclosed if required by law. No information which could identify you will be shared in publications about this study.

Organizations that may inspect and/or copy your research records for quality assurance and data analysis include groups such as the study investigator and his/her research associates, University of Notre Dame's Institutional Review Board (IRB; Protocol ID: 26-02-9913) and its designees, and (as allowed by law) state or federal agencies, especially the Office for Human Research Protections (OHRP), who may need to access the research records.

By clicking "I Agree" below, you acknowledge that you have read the information above, have had the opportunity to ask questions, and voluntarily agree to participate in this study.

I Do Not Agree

I Agree

Figure 8: Part of the informed consent page shown to annotation-study participants prior to the tasks.

Task	Model	Sheng	R^2	Tribal Lexicon	R^2	Loanword	R^2	Code-Mixed	R^2
Anxiety	AfriBERTa	-0.130	0.000	0.075**	0.005	0.339***	0.141	0.124**	0.004
	mBERT	-1.414**	0.005	-0.262***	0.059	0.261***	0.079	0.123**	0.004
	SwahBERT	-1.372**	0.005	-0.234***	0.050	0.060**	0.004	0.140***	0.006
	XLM-RoBERTa	0.096	0.000	0.141***	0.011	0.458***	0.157	0.160**	0.004
Subjective Literacy	AfriBERTa	-0.269	0.000	0.009	0.000	0.113***	0.021	0.166***	0.011
	mBERT	-0.068	0.000	-0.145***	0.018	0.085***	0.008	0.207***	0.012
	SwahBERT	0.003	0.000	-0.135***	0.016	0.014	0.000	0.174***	0.008
	XLM-RoBERTa	-0.352	0.000	0.040	0.001	0.151***	0.021	0.238***	0.012
Numeracy	AfriBERTa	0.017	0.000	0.001	0.000	-0.007***	0.020	-0.004	0.001
	mBERT	-0.032	0.000	-0.164***	0.016	0.218***	0.038	0.148**	0.004
	SwahBERT	0.065	0.000	-0.153***	0.016	0.105***	0.010	0.145**	0.004
	XLM-RoBERTa	-0.461	0.000	0.066	0.002	0.338***	0.062	0.180**	0.004
Trust	AfriBERTa	0.397	0.000	0.061**	0.004	0.288***	0.124	0.010	0.000
	mBERT	0.162	0.000	-0.497***	0.176	0.300***	0.086	0.112*	0.003
	SwahBERT	-0.102	0.000	-0.470***	0.166	0.104***	0.011	0.108*	0.003
	XLM-RoBERTa	0.448	0.000	0.168***	0.018	0.375***	0.119	0.016	0.000

Table 7: Univariate regression coefficients and corresponding R^2 values for associations between individual cultural-linguistic artifacts and model prediction error across psychometric tasks. Statistical significance is based on Benjamini–Hochberg false discovery rate (FDR) adjusted p -values across the 16 tests associated with each language model (four psychometric tasks $\times$ four CLA categories): * $p_{\text{adj}} < 0.05$, ** $p_{\text{adj}} < 0.01$, and *** $p_{\text{adj}} < 0.001$.

C Appendix: Model Error Analysis Experiments

C.1 Pretrained Language Models

As visualized in Figure 9, we fine-tuned four transformer-based pretrained language models (PLMs)—mBERT, XLM-RoBERTa, AfriBERTa, and SwahBERT—using the Hugging Face Transformers library for psychometric construct prediction. Participant text responses were tokenized and encoded into contextualized representations, with the final hidden representation of the [CLS] token passed to a task-specific classification head. Continuous psychometric labels were discretized via median split to produce binary training targets. Models were optimized using weighted cross-entropy loss with a batch size of 16, learning rate of 1×10^{-5} , and 5 training epochs. During inference, the classification head produced logits for the binary classes, which were transformed via the softmax function into posterior class probabilities. The probability assigned to the positive class was treated as the model-derived psychometric prediction score, yielding a continuous value in the range $[0, 1]$, and corresponds to the model-generated score described in §3.4.1. All experiments were evaluated using 5-fold cross-validation on an NVIDIA A100 GPU.

C.2 Regression Coefficients

While Figure 5 visualizes the directional effects of sociolinguistic features on model prediction error, Table 7 provides the complete regression coefficients

and associated significance estimates across models and tasks.

C.3 Large Language Models

We evaluated two multilingual large language models (LLMs), Qwen2.5-72B-Instruct and Llama3.1-405B-Instruct, using few-shot prompting format via the Replicate and DeepInfra APIs. Each prompt first assigned the model a psychometric evaluation role and specified the target construct to be assessed. Figure 10 details the prompt structure used during LLMs inference. All API calls used consistent decoding parameters across models: temperature = 0.3, max_tokens = 100, and top_p = 1.0, with no stop sequence specified. The same prompt structure was maintained across all folds and psychometric tasks.

D Appendix: Additional Results

D.1 Bootstrap Confidence Intervals for Annotation Confusion Matrices

Table 9 reports 95% bootstrap confidence intervals for the row-normalized confusion matrices shown in Figure 4. Confidence intervals were obtained using 10,000 stratified item-level bootstrap samples.

Task	Model	Sheng	Tribal Lexicon	Loanword	Code-Mixed	R^2
Anxiety	AfriBERTa	-0.334	0.121***	0.349***	0.098*	0.157
	mBERT	-1.303**	-0.228***	0.238***	0.106**	0.132
	SwahBERT	-1.196**	-0.226***	0.037	0.138***	0.061
	XLM-R	-0.212	0.204***	0.476***	0.124*	0.183
Subjective Literacy	AfriBERTa	-0.316	0.024	0.113***	0.157***	0.032
	mBERT	0.030	-0.136***	0.068**	0.203***	0.035
	SwahBERT	0.121	-0.136***	-0.003	0.175***	0.025
	XLM-R	-0.438	0.061*	0.153***	0.227***	0.035
Numeracy	AfriBERTa	0.019	0.000	-0.007***	-0.003	0.021
	mBERT	0.020	-0.138***	0.203***	0.134**	0.052
	SwahBERT	0.157	-0.142***	0.089***	0.139**	0.027
	XLM-R	-0.653	0.113**	0.347***	0.153*	0.070
Trust	AfriBERTa	0.222	0.099***	0.298***	-0.012	0.135
	mBERT	0.456	-0.465***	0.253***	0.097*	0.239
	SwahBERT	0.255	-0.464***	0.057**	0.107*	0.173
	XLM-R	0.145	0.219***	0.397***	-0.014	0.149

Table 8: Multivariable regression analyses examining the joint associations of the four CLA categories with model prediction error across psychometric tasks and language models. All four CLA categories are entered simultaneously in each model. Statistical significance is based on Benjamini–Hochberg false discovery rate (FDR) adjusted p -values: * $p_{\text{adj}} < 0.05$, ** $p_{\text{adj}} < 0.01$, and *** $p_{\text{adj}} < 0.001$.

Metric	Actual	Code-Mixing	Loan Words	Sheng	Tribal Lexicons	Standard Swahili
EAA	Code-Mixing	0.92 [0.84, 0.98]	0.07 [0.02, 0.13]	0	0	0.02 [0.00, 0.05]
	Loan Words	0.03 [0.00, 0.07]	0.85 [0.76, 0.93]	0	0	0.13 [0.06, 0.21]
	Sheng	0	0	0.80 [0.40, 1.00]	0	0.20 [0.00, 0.60]
	Tribal Lexicons	0	0	0	0.87 [0.74, 0.97]	0.13 [0.03, 0.26]
AEAA	Code-Mixing	0.92 [0.84, 0.98]	0.07 [0.02, 0.13]	0	0.02 [0.00, 0.05]	–
	Loan Words	0.03 [0.00, 0.07]	0.97 [0.93, 1.00]	0	0	–
	Sheng	0	0	1.00 [1.00, 1.00]	0	–
	Tribal Lexicons	0	0	0	1.00 [1.00, 1.00]	–

Table 9: Row-normalized confusion-matrix estimates with 95% bootstrap confidence intervals for Ensemble Annotation Accuracy (EAA) and Adjusted Ensemble Annotation Accuracy (AEAA). Confidence intervals are based on 10,000 stratified item-level bootstrap samples. Standard Swahili is excluded under AEAA by definition.

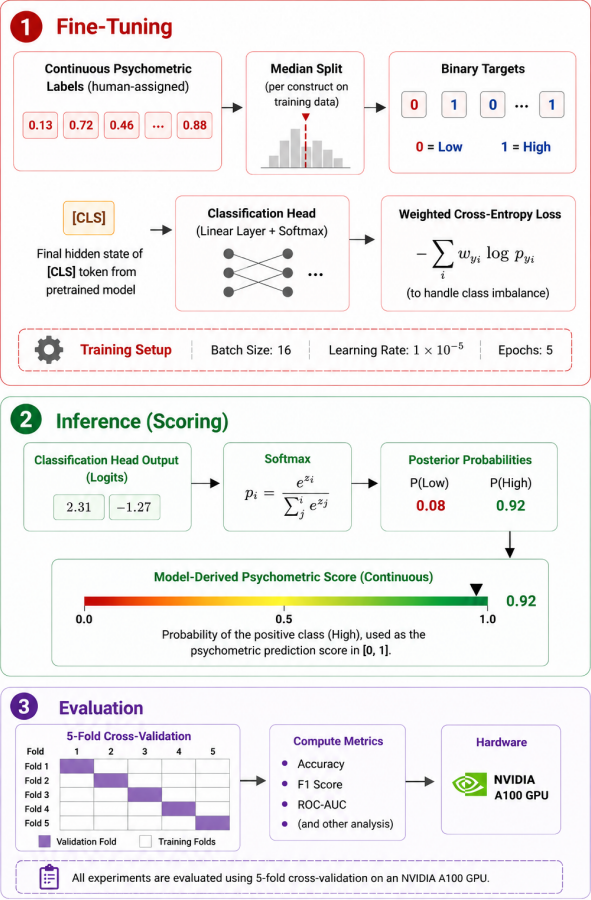


Figure 9: Overview of the psychometric prediction pipeline using pretrained language models.



Figure 10: Prompt structure used for psychometric classification with multilingual large language models.