Synthetic Data Generation Pipeline for Low-Resource Swahili Sentiment Analysis: Multi-LLM Judging with Human Validation

Samuel Gyamfi[1], Alfred Malengo Kondoro[1,2], Yankı Öztürk[1]
Richard H. Schreiber[1], Vadim Borisov[1]
[1]tabularis.ai [2]Hanyang University
samuel.gyamfi@tum.de

Abstract

Despite serving over 100 million speakers as a vital African lingua franca, Swahili remains critically under-resourced for Natural Language Processing, hindering technological progress across East Africa. We present a scalable solution: a controllable synthetic data generation pipeline that produces culturally grounded Swahili text for sentiment analysis, validated through automated LLM judges. To ensure reliability, we conduct targeted human evaluation with a native Swahili speaker on a stratified sample, achieving 80.95% agreement between generated sentiment labels and human ground truth, with strong agreement on judge quality assessments. This demonstrates that LLM-based generation and quality assessment can transfer effectively to low-resource languages. We release a dataset and provide a reproducible pipeline in tandem, providing ample knowledge and working material for NLP researchers in low-resource contexts. Our dataset and the full reproducible generation pipeline are publicly available in the Swahili Sentiment Dataset repository and the GitHub repository.

1 Introduction

Advances in Natural Language Processing (NLP) have not been distributed evenly across the world's languages, leaving many "low-resource" languages without sufficient data for modern model development (Joshi et al., 2020; Nekoto et al., 2020; Hedderich et al., 2021). Swahili (Kiswahili), a major lingua franca of East and Central Africa and an official African Union language, remains one of these under-served languages despite its more than 100 million speakers (Orife et al., 2020; Ndimbo et al., 2025). Data scarcity limits the development of high-quality NLP tools for Swahili and risks widening the digital divide. The language's agglutinative morphology and complex noun-class system further amplify the challenge, since such linguistic structures are often underrepresented in corpora dominated by English and other high-resource languages (Gutman and Avanzati, 2013; Arnett and Bergen, 2024).

Synthetic data (sometimes artificial data (Borisov and Schreiber, 2024)) generation using foundation models, particularly Large Language Models (LLMs) (Bommasani et al., 2022), offers a scalable alternative to manual annotation (Liu et al., 2024; Nadas et al., 2025). However, producing high-quality, diverse, and culturally grounded synthetic text for low-resource languages requires careful control to ensure linguistic fidelity (Kirk et al., 2024; Li et al., 2023).

This work introduces a controllable synthetic data pipeline for Swahili that prompts LLMs to generate culturally grounded text, employs LLM-as-a-judge (Zheng et al., 2023) for multi-dimensional quality scoring, and filters aggressively to retain only high-quality samples. We validate this approach through rigorous human evaluation, achieving 80.95% agreement between generated sentiment labels and native speaker ground truth. On the AfriSenti benchmark (Muhammad et al., 2023), models fine-tuned with our synthetic data show consistent macro-F1 gains over zero-shot baselines, demonstrating that judged synthetic supervision can reliably transfer sentiment capability to a low-resource setting.

Our main contributions are as follows:

1. A controllable generation pipeline for Swahili that incorporates sentiment, domain, aspect, tone, and cultural relevance constraints.
2. An automated LLM-based judging mechanism validated against human ground truth.
3. Empirical evidence that judged synthetic data improves Swahili sentiment classification performance on AfriSenti.

2 Related Work

Prior work on low-resource languages addresses data scarcity through transfer learning, cross-lingual and semi-supervised methods, and distantly supervised labeling, often trading scalability for noise (Hedderich et al., 2021). Recent work increasingly favors data augmentation, including back-translation and LLM-based synthetic data generation, as scalable alternatives to manual annotation.

Traditional Data Augmentation. Early approaches relied on rule-based transformations such as synonym replacement and token-level perturbations (Feng et al., 2021), which are easy to implement but offer limited diversity and can introduce grammatical errors (Wei and Zou, 2019). Related work has also explored simpler augmentation techniques, such as contextual augmentation for low-resource English-Swahili MT, yielding moderate gains (Gitau and Marivate, 2023).

Sentiment Analysis in Low-Resource Settings. Sentiment analysis is a widely studied NLP task for modeling opinions expressed in text (Medhat et al., 2014; Wankhade et al., 2022), but approaches developed for resource rich languages often transfer poorly to LRLs (Joshi et al., 2020).

For Swahili, the scarcity of large, high-quality labeled corpora comparable to English benchmarks, such as IMDb or Yelp, remains a primary bottleneck (Tunga and David, 2025). This challenge is compounded by the language's agglutinative morphology, which embeds sentiment-bearing morphemes within complex verb forms and complicates standard tokenization and feature extraction (Arnett and Bergen, 2024; Mathayo and Kondoro, 2025), as well as by cultural nuances expressed through idioms, proverbs, and context-dependent phrasing that multilingual models trained on Western-centric data often fail to capture (Muhammad et al., 2023). In response, recent efforts have introduced dedicated benchmarks such as AfriSenti (Muhammad et al., 2023) and explored synthetic data generation to alleviate labeled data scarcity (Sundarreson and Kumarapathirage, 2024), a direction this work further advances.

Generative AI Models for Synthetic Data. Earlier generative approaches such as GANs and VAEs were explored for text generation (Goodfellow et al., 2014; Kingma and Welling, 2022), but they offered limited controllability for task-specific data synthesis (Liu et al., 2024).

The emergence of large language models has substantially advanced synthetic data generation (Nadas et al., 2025; Davidson et al., 2025; Sundarreson and Kumarapathirage, 2024), enabling fluent and contextually grounded text generation through prompting (Brown et al., 2020). Prior work leverages LLMs for synthetic data creation across tasks such as text classification, instruction following, question answering, and information extraction (Li et al., 2023; Tan et al., 2024; IR-LLM Community, 2023), using techniques including zero- and few-shot prompting, self-instruction pipelines (Wang et al., 2023), and retrieval-augmented generation (Lewis et al., 2020). Recent studies further show that fine-tuned teacher LLMs can generate large synthetic corpora for training smaller student models, a paradigm that is particularly effective in low-resource settings (Kaddour and Liu, 2024).

Synthetic Data for Low-Resource Languages. Although high-quality human-annotated datasets exist for some low-resource languages, including Swahili (Tunga and David, 2025; Zawuya et al., 2025), they are often limited in scale, making synthetic data generation a central strategy for resource expansion (Doshi and Bhattacharyya, 2024). Back-translation remains a foundational approach in low-resource neural machine translation (Li et al., 2020; Bojar et al., 2016), while more recent work explores LLMs for synthetic data generation. However, LLM performance frequently declines in LRL settings due to pre-training data imbalance (Wang et al., 2024; Qin et al., 2025), motivating adaptation strategies such as multilingual fine-tuning (Moskvoretskii et al., 2024), cross-lingual transfer (Latouche et al., 2024), and specialized prompting (Deshpande et al., 2024). Related studies also investigate machine translation-based synthetic data (translationese) for language-model pre-training in Indic languages (Doshi et al., 2024). For Swahili specifically, retrieval-augmented generation has been applied to conversational AI, demonstrating the potential of combining LLMs with external knowledge sources in this low-resource context (Ndimbo et al., 2025).

Handling Morphological Complexity. Morphologically rich, agglutinative languages such as Swahili pose well-known challenges for NLP due to dense sequences of inflectional and derivational morphemes. Standard subword tokenization methods, like byte-pair encoding (BPE), frequently misalign with morpheme boundaries of such languages, introducing sparsity and obscuring grammatical structure (Arnett and Bergen, 2024). Prior work demonstrates that morphology-aware approaches to subword segmentation, such as deploying the Prefix-Root-Postfix-Encoding (PRPE) algorithm for machine translation (Chen and Fazio, 2021), can substantially benefit low-resource settings.

Linguistic and computational studies of Swahili highlight a highly structured, multi-slot verbal morphology that contributes to segmentation errors in neural models (Wahome et al., 2023; Mathayo and Kondoro, 2025), motivating rule-based approaches such as SwaRegex, a finite-state lexical transducer that achieves high-accuracy in verb segmentation (Muthee et al., 2022). Recent work further shows that Swahili morphology is highly productive in digital contexts, with novel derivations and reduplication patterns emerging under code-mixing and social media use (Makulilo, 2025; Gabriel et al., 2018). Although our approach does not perform explicit morphological segmentation, these findings motivate the evaluation of the morphological plausibility of generated Swahili text.

Quality Control and Evaluation. A critical aspect of synthetic data generation is assessing its quality and utility. Automated metrics like BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005) are widely used but often show limited correlation with human judgment (Schmidtová et al., 2024; Pillutla et al., 2021). Evaluating downstream task performance remains the gold standard (Guo et al., 2024). Recent work explores using LLMs themselves as evaluators or "judges" (Zheng et al., 2023), similar to the approach adopted in our pipeline. Ensuring diversity and avoiding distribution shift are also key concerns (Nadas et al., 2025). The subjectivity of tasks can also impact the effectiveness of synthetic data (Li et al., 2023).

Our work builds upon these insights by employing a conditional LLM prompting strategy specifically designed for Swahili, incorporating culturally relevant prompts, and utilizing a multi-faceted LLM-based judging mechanism for quality control, and then evaluating the efficacy of this pipeline on sentiment classification tasks. We contribute a large-scale dataset generated using this pipeline and evaluate its quality based on automated judgments, providing insights into the effectiveness of and challenges facing this approach for Swahili.

3 Synthetic Data Generation Pipeline

In this section, we present our synthetic data generation pipeline, which consists of five steps.

3.1 Step 1: Criteria Definition

To ensure diversity without relying on seed samples, we define a comprehensive set of controllable generation parameters spanning sentiment, structure, quality, and semantic context. **Sentiment** is modeled on a 5-point scale with intermediate values (e.g., 1.5, 4.5) to capture nuance, with a higher sampling weight assigned to *neutral* to reflect natural distributions. **Target Length** ranges from *micro* (10–25 words) to *extensive* (250–300 words), acknowledging that LLMs do not always adhere strictly to word-count constraints. To mimic real-world noise, we introduce a **Desired Quality** parameter ranging from *abysmal* to *exceptional*, explicitly prompting for imperfections such as typos. Semantic variation is controlled through curated sets of **Domains**, **Aspects**, and **Tones**: domains cover over 160 categories relevant to East African digital experiences (e.g., mobile money, politics, hospitality, technology), aspects span over 180 evaluation dimensions (e.g., quality, customer service, cultural relevance), and tones include more than 50 nuanced emotional states such as *sarcastic* or *disappointed but hopeful*. This fine-grained control enables the generation of highly structured and diverse synthetic data for robust downstream finetuning; the full list of criteria is provided in Appendix E.

3.2 Step 2: Prompt Construction

Our pipeline utilized two distinct prompt templates: one for generating samples and another for evaluating these samples. These were provided to the generator models and the judge models, respectively, and were dynamically filled with randomized criteria during the generation process. We explicitly evaluated the use of a generator prompt written in the target language to promote closer stylistic adherence. Additional details regarding the prompt design and structure are provided in Appendix B.

3.3 Step 3: Synthetic Text Generation

To increase linguistic and stylistic diversity in the generated data, we employ two distinct generator models: **Llama 3 70B** (AI@Meta, 2024) and **Gemini Flash 2.0** (Google DeepMind). The models are used in an interleaved manner, allowing com-

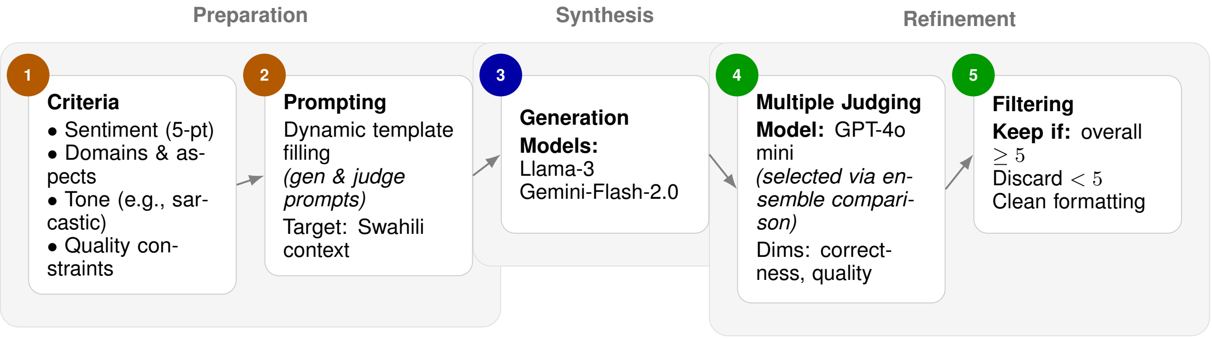


Figure 1: The proposed pipeline for robust synthetic data generation.

plementary generation behaviors to surface across domains, tones, and sentiment configurations. Both models were selected based on evidence of strong multilingual capabilities and the likelihood of exposure to substantial Swahili data during pre-training, as suggested by model scale, provider documentation, and reported multilingual performance. The pipeline itself is agnostic to the total number of generated samples and can be scaled to arbitrarily large datasets as computational resources permit.

3.4 Step 4: Automated Judging with Multi-Model Ensemble

After generating all 50,000 synthetic Swahili samples using our two generator models, we implemented a comprehensive automated evaluation strategy to assess sample quality and determine the most reliable judge for large-scale filtering.

Judge Model Selection. We evaluated three distinct LLM judges to score the generated samples: **GPT-4o mini** (OpenAI, 2024), **Claude 4.5 Haiku** (Anthropic, 2025), and **Grok 4.1 Fast** (xAI, 2025). These models were selected based on their performance on public leaderboards, cost-effectiveness for large-scale evaluation, and recent state-of-the-art releases. Each judge evaluated samples across five dimensions: Language Correctness (0–5), Cultural Relevance (0–5), Sentiment Alignment (0–5), Instruction Following (0–5), and Overall Quality (0–10).

Ensemble Evaluation. All three judges independently scored the complete set of 50,000 generated samples. We then conducted ensemble experiments, averaging scores across the three judges and comparing ensemble performance against individual judge performance. We observed strong Pearson correlations between judges, particularly between GPT-4o mini and the ensemble average.

Human Validation Study. To validate the reliability of our automated judges, we conducted targeted human evaluation on a stratified random subset of 126 samples. A native Swahili speaker with expertise in East African linguistics reviewed each sample, providing ground-truth sentiment labels and independent quality judgments across the same five evaluation dimensions used by the automated judges. Comparing human annotations against automated judge scores revealed strong agreement across all judges, with performance metrics detailed in Appendix C.

Final Judge Selection. Based on these validation results, we selected **GPT-4o mini** (OpenAI, 2024) as the sole judge for our final filtering pipeline. This decision prioritizes cost-effectiveness and reproducibility: by demonstrating that a single, affordable model can serve as a reliable evaluator when properly validated, we provide a practical blueprint for researchers in low-resource settings who may lack budgets for expensive multi-model ensembles. GPT-4o mini's performance was nearly identical to the full ensemble while offering significantly lower inference costs, making it the optimal choice for scaling to our complete dataset.

3.5 Step 5: Automated Filtering and Dataset Finalization

Using GPT-4o mini as the selected judge, we filtered the complete dataset of 50,000 generated samples based on their Overall Quality scores. We applied a threshold of Overall Quality ≥ 5 (out of 10), automatically removing any sample that fell below this standard. This threshold was chosen to balance dataset scale with quality assurance, retaining samples that met minimum acceptable standards while removing only egregiously poor outputs.

This filtering process retained 47,980 samples. As a final preprocessing step, we cleaned the retained samples by removing any formatting artifacts (such as brackets or placeholders used to guide generation), ensuring the text was production-ready for downstream training.

4 Analysis of Synthetic Data

4.1 Intrinsic Quality and Generation Analysis

From an initial set of 50,000 generated samples, 47,980 were retained after filtering for an Overall Quality score ≥ 5 (out of 10), as assigned by the GPT-4o mini judge.

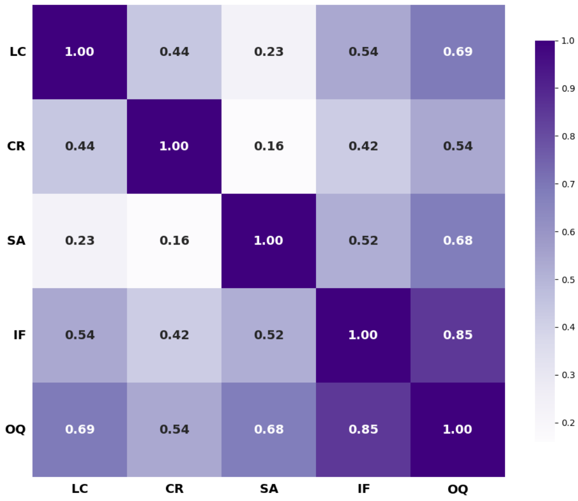


Figure 2: Pearson correlation heatmap between the five judgment metrics: Language Correctness (LC), Cultural Relevance (CR), Sentiment Alignment (SA), Instruction Following (IF), Overall Quality (OQ).

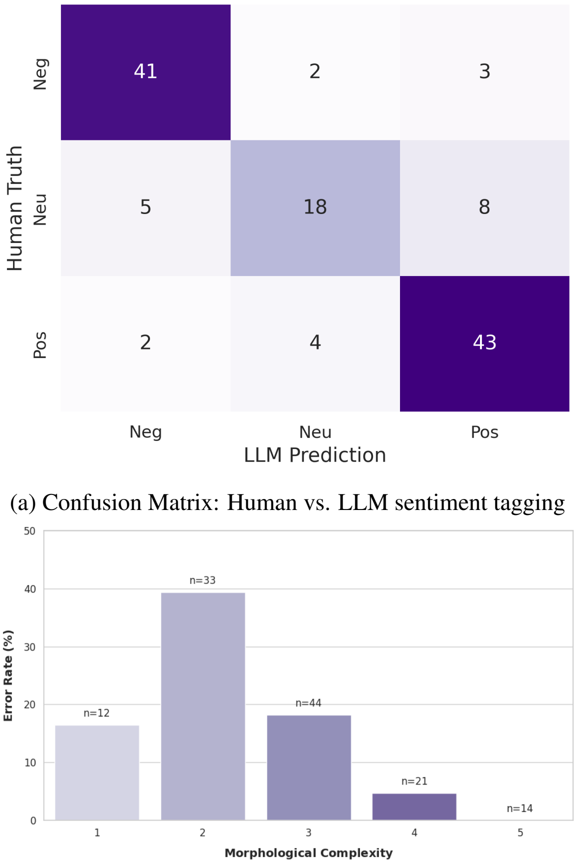
Judge Score Distributions: The selected judge, GPT-4o mini, assigned scores across five dimensions. The distribution of these scores for the initial 50,000 samples is shown in Table 1. The heatmap in Figure 2 indicates positive correlations between all judgment metrics, with Overall Quality strongly correlating with Instruction Following (r=0.85) and moderately with Language Correctness (r=0.69) and Sentiment Alignment (r=0.68).

Generator Comparison: Gemini Flash 2.0 generally outperformed **Llama3-70B** in judged quality, receiving higher average scores across all dimensions (Table 1). Notably, **Gemini Flash 2.0** achieved an average Overall Quality of 7.81 compared to 6.55 for **Llama3-70B**. These differences are consistent across all evaluation dimensions.

4.2 Human Verification and Morphological Analysis

To validate the reliability of our automated generation and judging pipeline, we conducted targeted human evaluation on a stratified random subset of 126 generated samples (described in Section 3.4). A native Swahili speaker with expertise in East African linguistics reviewed each sample, providing ground-truth sentiment labels and independent quality judgments.

Generator Sentiment Accuracy: As illustrated in Figure 3a, we observed **80.95% accuracy** (Cohen's Kappa of 0.706) between the sentiment labels assigned by our generator LLMs (Llama 3 70B and Gemini Flash 2.0) and human ground-truth sentiment. The confusion matrix shows that the vast majority of samples were correctly tagged, with only 2 false negative instances and 3 false positive instances, confirming that our generators reliably produce sentiment-aligned text.



(a) Confusion Matrix: Human vs. LLM sentiment tagging

(b) Error Rate by Morphological Complexity

Figure 3: Human validation results showing strong sentiment agreement and the impact of morphological complexity on error rates.

Morphological Complexity Analysis: Swahili is an agglutinative language in which sentiment cues are often encoded within complex verbal morphology (e.g., *si-ta-ku-pend-a*; NEG-FUT-2SG-love-FV, "I will not love you"). To examine whether

Generator Model	Lang. Correct.	Cult. Relevance	Sent. Align.	Instr. Follow.	Overall Qual.
Gemini Flash 2.0	4.17 ± 0.49	4.55 ± 0.67	4.19 ± 1.30	4.02 ± 0.81	7.81 ± 1.37
Llama3-70B	3.81 ± 0.53	3.45 ± 0.93	3.70 ± 1.39	3.25 ± 0.80	6.55 ± 1.26

Table 1: Average judged scores and their standard deviation by generator model (Pre-filtering, N=50,000). The Gemini model show superior performance on all criteria.

morphological complexity influences error patterns, each sample was assigned a *Morphological Complexity Score* (1–5) based on the number of realized suffix slots within verbs (subject agreement, tense/aspect, negation, object marking, verbal extensions). Scores were assigned through manual inspection by the native speaker following consistent slot-based criteria.

As shown in Figure 3b, the highest error rate (approximately 39%) occurred at moderate complexity levels (score 2), while samples with highly complex morphology (score 5) showed no sentiment errors in this subset. Qualitative analysis indicates that misclassifications arise less from morphological form itself and more from pragmatic factors such as hedging, politeness markers, and evaluative ambiguity. Highly inflected but structurally canonical verb forms align more closely with training data patterns, yielding more stable sentiment judgments. These results suggest that conversational-level morphology combined with pragmatic nuance poses greater challenges than morphologically dense but structurally regular constructions, while confirming the pipeline's robustness to Swahili's morphological richness.

5 Experimental Validation of Synthetic Data Pipeline

In addition to evaluating three pre-trained multilingual sentiment models[1], we trained and evaluated a simple DistilBERT base model[2] (Sanh et al., 2019) on our generated data. Their performance was assessed on the Swahili test set ('swa') of the AfriSenti benchmark (Muhammad et al., 2023), comparing their zero-shot performance against performance after fine-tuning solely on our generated data. The metric of choice was macro F1 as a way to better handle class imbalances, as well as providing a fair assessment of performance. Fine-tuning used the following hyperparameters: learning rate of 5×10^{-5}, 5 epochs, AdamW optimizer (Loshchilov and Hutter, 2019)) with the HuggingFace Transformers library (Wolf et al., 2020), the rest of the hyperparameters are default to the model or selected package.

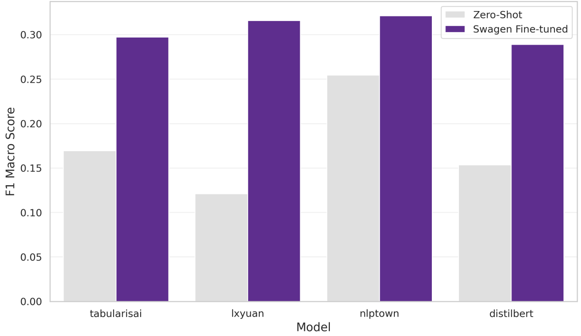


Figure 4: Macro F1 scores on AfriSenti-Swahili: Zero-Shot vs. Fine-tuned on Synthetic Data.

5.1 Results

Fine-tuning on our synthetic dataset substantially improved the performance of all three multilingual sentiment classification models, as well as the trained distilbert-base-uncased model on the AfriSenti-Swahili test set compared to their zero-shot capabilities (Figure 4, Table 2). All models benefited from fine-tuning with our synthetic data. Even smaller models like our Tabularis multilingual-sentiment-analysis model and the distilbert base model singularly trained for swahili sentiment analysis show competitive performance with much larger models. Models like the nlptown/bert-base-multilingual-uncased-sentiment trained on a massive corpus of data still benefited from training.

6 Discussion and Future Work

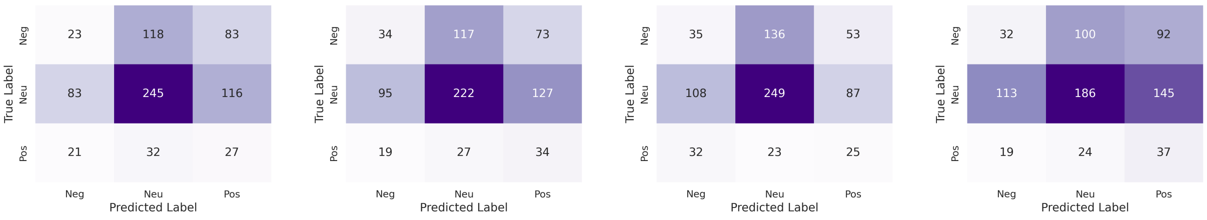
Our synthetic data generation pipeline demonstrates a viable strategy for augmenting data for Swahili. The notable improvement in sentiment analysis performance post-fine-tuning on this LLM-generated data underscores its value for adapting existing multilingual models. According to auto-

[1] tabularisai/multilingual-sentiment-analysis (tabularisai et al., 2025), lxyuan/distilbert-base-multilingual-cased-sentiments-student, and nlptown/bert-base-multilingual-uncased-sentiment

[2] distilbert/distilbert-base-cased

Model	Zero-Shot Macro F1	Synthetic Data Fine-tuned Macro F1
tabularisai	0.1695	0.2972
lxyuan	0.1211	0.3158
nlptown	0.2545	0.3213
swahili distilbert base	0.1536	0.2891

Table 2: Sentiment Model Performance (Macro F1) on AfriSenti-Swahili Test Set



(a) **tabularisai**

(b) **lxyuan**

(c) **nlptown**

(d) **swahili**

Figure 5: Confusion matrices for models fine-tuned with synthetic data, evaluated on the AfriSenti-Swahili test set.

mated judgments, **Gemini Flash 2.0** was a more effective generator than **Llama3-70B** for this task.

A recurring failure mode in low-resource language data augmentation is the proliferation of “translationese,” text that is grammatically correct but culturally hollow. By explicitly prompting for and judging *Cultural Relevance*, such as the use of *methali* (proverbs) and local entities, our pipeline moves beyond mere translation. The high agreement between the human evaluator and the LLM judge confirms that modern models, when prompted correctly, can distinguish between generic translation and culturally resonant text. The performance gains on AfriSenti, a benchmark known for difficult and colloquial samples, suggest that cultural grounding is a critical factor in transfer learning success for African languages.

Limitations include the inherent differences between synthetic and real-world data, contradictions in generated samples (sometimes asking for contradictory tones and sentiments as a result of randomized criteria insertion), and the intrinsic challenges of the AfriSenti benchmark. Additionally, while we validated our automated judge against a human baseline, we utilized the raw judgment scores for filtering. As noted by recent statistical frameworks for LLM evaluations (Lee et al., 2025), naive estimates from imperfect judges can introduce bias, potentially overestimating performance at low accuracy levels or underestimating it at high levels. Future iterations of this pipeline would benefit from applying bias-adjusted estimators and constructing confidence intervals that explicitly account for the judge’s sensitivity and specificity.

The meaning of the term “low resource language” is not unanimously agreed upon among researchers of language technologies (Cieri et al., 2016). Differences in available information and the goals of such research inevitably lead to some obscurity in what can be understood from the term: *low density*, *less commonly taught*, *under-resourced*, *less computerized*, and *less privileged* are all among possible connotations (Singh, 2008). We interpret a low-resource language as one that is unable to directly benefit from state-of-the-art statistical methods directly due to its scarcity of structured data.

Although an understudied language in NLP tasks, Swahili enjoys the privileges of official language status in several countries, a written literary tradition, and well documented dialects (Miachina, 1981). We recognize that the methods described in this paper presume state-of-the-art LLMs’ability to readily generate text in the target language and might not be reproducible for the vast majority of the world’s languages.

Future work should prioritize expanded human evaluation of synthetic data, specifically addressing semantic cohesion. Our native speaker expert noted that instances with generated proverbs, while linguistically accurate, were incongruent with the context in which they appeared. This misalignment likely arises from the stochastic combination of disparate generation criteria (e.g., requesting a proverb within a technical error report). Further investigation is warranted to refine prompt constraints, ensuring that cultural markers are applied only where they are contextually appropriate.

7 Conclusion

This work demonstrates that carefully controlled synthetic data generation, when paired with automated LLM-based quality assessment and targeted human verification, provides a viable and scalable pathway for advancing sentiment analysis in morphologically complex, low-resource languages such as Swahili. By integrating fine-grained generation constraints with multi-dimensional judging and aggressive filtering, our pipeline produces linguistically coherent and culturally grounded synthetic text that transfers effectively to downstream sentiment classification tasks. Empirical validation shows strong agreement between automated judges and native-speaker annotations, supporting the reliability of LLM-as-a-judge approaches beyond high-resource settings. Fine-tuning multilingual models on the resulting dataset yields consistent macro F1 improvements over zero-shot baselines on the AfriSenti-Swahili benchmark, confirming that high-quality synthetic supervision can meaningfully bridge labeled data gaps. Beyond Swahili, our findings suggest that combining controllable generation, automated evaluation, and selective human oversight offers a reproducible blueprint for building NLP resources in under-resourced languages where manual annotation is costly or infeasible.

References

AI@Meta. 2024. Llama 3 model card.

Anthropic. 2025. Claude haiku 4.5. https://www.anthropic.com/news/claude-haiku-4-5. Released: 2025-10-15.

Catherine Arnett and Benjamin K. Bergen. 2024. Why do language models perform worse for morphologically complex languages?

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie N. Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, and 2 others. 2016. Findings of the 2016 conference on machine translation (WMT16). In *Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, and 95 others. 2022. On the opportunities and risks of foundation models. *Preprint*, arXiv:2108.07258.

Vadim Borisov and Richard H Schreiber. 2024. Open artificial knowledge. *arXiv preprint arXiv:2407.14371*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. 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, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

William Chen and Brett Fazio. 2021. Morphologically-guided segmentation for translation of agglutinative low-resource languages. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, pages 20–31, Virtual. Association for Machine Translation in the Americas.

Christopher Cieri, Mike Maxwell, Stephanie Strassel, and Jennifer Tracey. 2016. Selection criteria for low resource language programs. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4543–4549, Portorož, Slovenia. European Language Resources Association (ELRA).

Tim Davidson, Benoit Seguin, Enrico Bacis, Cesar Magalhaes, and Hamza Harkous. 2025. Orchestrating synthetic data with reasoning. In *SynthData @ ICLR2025*.

Tejas Deshpande, Nidhi Kowtal, and Raviraj Joshi. 2024. Chain-of-translation prompting (cotr): A novel prompting technique for low resource languages. *Preprint*, arXiv:2409.04512.

Meet Doshi and Pushpak Bhattacharyya. 2024. Synthetic data for multilingual nlp: A survey. CFILT, IIT Bombay.

Meet Doshi, Raj Dabre, and Pushpak Bhattacharyya. 2024. Pretraining language models using translationese. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5843–5862, Miami, Florida, USA. Association for Computational Linguistics.

Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for nlp. *arXiv preprint arXiv:2105.03075*.

B. Gabriel, Bisamaza Emilien, and Ndayishimiye Jean Léonard. 2018. Morphological doubling theory to two bantu languages reduplication: A comparative perspective of kinyarwanda and swahili. *International Journal of English and Literature*, 3:31–40.

Catherine Gitau and Vukosi Marivate. 2023. Textual augmentation techniques applied to low resource machine translation: Case of swahili.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial networks. *Preprint*, arXiv:1406.2661.

Google DeepMind. Gemini flash. `https://deepmind.google/models/gemini/flash/`.

Yue Guo, Tal August, Gondy Leroy, Trevor Cohen, and Lucy Lu Wang. 2024. APPLS: Evaluating evaluation metrics for plain language summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9194–9211, Miami, Florida, USA. Association for Computational Linguistics.

Alejandro Gutman and Beatriz Avanzati. 2013. Swahili. The Language Gulper.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online. Association for Computational Linguistics.

IR-LLM Community. 2023. Awesome information retrieval in the age of large language model. GitHub repository.

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.

Jean Kaddour and Qi Liu. 2024. Synthetic data generation in low-resource settings via fine-tuning of large language models. *Preprint*, arXiv:2310.01119.

Diederik P Kingma and Max Welling. 2022. Auto-encoding variational bayes. *Preprint*, arXiv:1312.6114.

Hannah Rose Kirk, Jatinder Singh, and Bertie Vidgen. 2024. Transparency in the wild: Navigating transparency in a deployed ai system to broaden needfinding approaches. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*, Rio de Janeiro, Brazil. ACM.

Gaetan Lopez Latouche, Marc-André Carbonneau, and Benjamin Swanson. 2024. Zero-shot cross-lingual transfer for synthetic data generation in grammatical error detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3002–3016, Miami, Florida, USA. Association for Computational Linguistics.

Chungpa Lee, Thomas Zeng, Jongwon Jeong, Jy yong Sohn, and Kangwook Lee. 2025. How to correctly report llm-as-a-judge evaluations. *Preprint*, arXiv:2511.21140.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Hongliang Li, Yang Gao, Jingbo Zhao, and Philip S. Yu Zhou. 2020. Revisiting back-translation for low-resource machine translation between chinese and vietnamese. *IEEE Access*, 8:115032–115041.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic data generation with large language models for text classification: Potential and limitations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10443–10461, Singapore. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, and Andrew M. Dai. 2024. Best practices and lessons learned on synthetic data. *Preprint*, arXiv:2404.07503.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Prisca Boniphace Makulilo. 2025. Morphological productivity and lexical innovation in swahili: Digital communication and language transformation in social media spaces. *Language, Technology, and Social Media*.

Irene Masiringi Mathayo and Alfred Malengo Kondoro. 2025. Unveiling swahili verb conjugations: A comprehensive dataset for low-resource nlp. In *Proceedings of the 2024 8th International Conference on Natural Language Processing and Information Retrieval*, NLPIR '24, page 149–156, New York, NY, USA. Association for Computing Machinery.

Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113.

E.N. Miachina. 1981. The swahili language: a descriptive grammar.

Viktor Moskvoretskii, Nazarii Tupitsa, Chris Biemann, Samuel Horváth, Eduard Gorbunov, and Irina Nikishina. 2024. Low-resource machine translation through the lens of personalized federated learning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8806–8825, Miami, Florida, USA. Association for Computational Linguistics.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Seid Muhie Yimam, David Ifeoluwa Adelani, Ibrahim Sa'id Ahmad, Nedjma Ousidhoum, Abinew Ali Ayele, Saif M. Mohammad, Meriem Beloucif, and Sebastian Ruder. 2023. AfriSenti: A Twitter sentiment analysis benchmark for african languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13968–13981, Singapore. Association for Computational Linguistics.

Mutwiri George Muthee, Mutua Makau, and Omamo Amos. 2022. Swaregex: a lexical transducer for the morphological segmentation of swahili verbs. *African Journal of Science, Technology and Social Sciences*.

Mihai Nadas, Laura Diosan, and Andreea Tomescu. 2025. Synthetic data generation using large language models: Advances in text and code.

Edmund V. Ndimbo, Qin Luo, Gimo C. Fernando, Xu Yang, and Bang Wang. 2025. Leveraging retrieval-augmented generation for swahili language conversation systems. *Applied Sciences*, 15(2):524.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, and Laura Martinus. 2020. Participatory research for low-resourced machine translation: A case study in african languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Iroro Fred O. Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade Abbott, Vukosi Marivate, Salomon Kabongo, Musie Meressa, Espoir Murhabazi, Orevaoghene Ahia, Elan van Biljon, Arshath Ramkilowan, Adewale Akinfaderin, Alp Öktem, Wole Akin, Ghollah Kioko, and 7 others. 2020. MASAKHANE – machine translation for africa. In *Proceedings of the AfricaNLP Workshop 2020 co-located with ICLR 2020*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. MAUVE: Measuring the gap between neural text and human text using divergence frontiers. In *Advances in Neural Information Processing Systems*, volume 34, pages 25154–25167. Curran Associates, Inc.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. 2025. A survey of multilingual large language models. *Patterns (N Y)*, 6(1):101118.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *NeurIPS* $EMC^2Workshop$.

Patrícia Schmidtová, Saad Mahamood, Simone Balloccu, Ondřej Dušek, Albert Gatt, Dimitra Gkatzia, David M Howcroft, Ondřej Plátek, and Adarsa Sivaprasad. 2024. Automatic metrics in natural language generation: A survey of current evaluation practices. *arXiv preprint arXiv:2408.09169*.

Anil Kumar Singh. 2008. Natural language processing for less privileged languages: Where do we come from? where are we going? In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*.

Pushpika Sundarreson and Sapna Kumarapathirage. 2024. Sentigen: Synthetic data generator for sentiment analysis. *Journal of Computing Theories and Applications*, 1(4):461–477.

tabularisai, Samuel Gyamfi, Vadim Borisov, and Richard H. Schreiber. 2025. multilingual-sentiment-analysis (revision 69afb83). `https://huggingface.co/tabularisai/multilingual-sentiment-analysis`.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957, Miami, Florida, USA. Association for Computational Linguistics.

Mahadia Tunga and Davis David. 2025. Introducing a swahili social media sentiment analysis dataset for the telecom industry. *Language Resources and Evaluation*.

Maina Wahome, Agus Subiyanto, and Oktiva Herry Chandra. 2023. An analysis of swahili verbal inflection and derivational morphemes: An item and arrangement approach. *Journal of Languages, Linguistics and Literary Studies*.

Daniel Wang, Tim Bakkenes, and Anton Johansson. 2024. Fine tuning methods for low-resource language. OpenReview Submission.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

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

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

xAI. 2025. Grok 4.1 fast. `https://x.ai/blog/grok-4-1-fast`. Released: 2025-11-19.

Chaddy Anthony Zawuya, Alfred Malengo Kondoro, Diana Rwegasira, and Juma H. Lungo. 2025. Maneno yetu: Dynamic corpus construction and pretraining for swahili nlp. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, CIKM '25, page 6564–6569, New York, NY, USA. Association for Computing Machinery.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

A Judge Model Selection Rationale

The selection of the "Judge" models for the initial ensemble experiments (described in Section 3.4) was based on a combination of performance capability, cost, and availability at the time of the study.

1. **GPT-4o mini**: This model was selected primarily due to its availability, speed, and cost-effectiveness. As a model from OpenAI (OpenAI, 2024), we hypothesized it had been trained on a massive and diverse corpus of internet data, likely including a non-trivial amount of Swahili text, making it a strong candidate for linguistic validation. Its performance in our initial pilot showed high correlation with larger, more expensive models, making it the ideal choice for scaling the judging process to the full dataset.

2. **Grok 4.1 Fast**: This model (xAI, 2025) was included in our initial ensemble based on its high ranking on OpenRouter and competitive pricing. We aimed to diversify the "judge panel" by including a model from xAI to mitigate potential biases inherent in using only one provider's model architecture.

3. **Claude 4.5 Haiku**: Chosen as a representative of the Anthropic model family (Anthropic, 2025), Claude 4.5 Haiku was selected due to its recent release as a state-of-the-art (SOTA) lightweight model. It offered a compelling balance of high-quality instruction following and lower inference costs compared to the larger Opus or Sonnet models.

B Prompting

This section details the prompt templates used for generating Swahili text with specific characteristics and for subsequently evaluating the generated text. These templates were designed to guide the language models effectively.

B.1 Generation Prompt Template

The following template was used to instruct the generator language model (e.g., Llama3-70B, Gemini-Flash) to produce Swahili text. The placeholders (e.g., criteria['sentiment']) are dynamically filled based on the specific criteria for each desired sample.

> Swahili Text Generation Prompt Template
>
> Tafadhali tengeneza maandishi ya Kiswahili yanayokidhi vigezo vifuatavyo kwa umakini mkubwa:
>
> 1. **Mwelekeo wa Hisia (Sentiment)**: {criteria['sentiment']}
> 2. **Kikoa/Mada (Domain)**: {criteria['domain']}
> 3. **Kipengele Maalum (Aspect)**: {criteria['aspect']}
> 4. **Mtindo wa Lugha (Tone)**: {criteria['tone']}
> 5. **Urefu Unaolengwa (Target Length)**: {criteria['target_length']}
> 6. **Ubora Unaotarajiwa (Desired Quality)**: {criteria['desired_quality']}
>
> **Maagizo Muhimu Zaidi:**
>
> - **Utamaduni wa Afrika Mashariki:** Jumuisha methali, misemo, au marejeleo yanayofahamika katika utamaduni wa Kiswahili/Afrika Mashariki inapowezekana na kwa njia ya asili. Lenga muktadha wa kitamaduni unaoeleweka.
> - **Lugha Bora:** Tumia Kiswahili sanifu na sahihi kisarufi. Zingatia matumizi ya visawe na msamiati unaofaa.
> - **Upekee wa Kienyeji:** Ongeza maelezo au vidokezo vinavyoonyesha uelewa wa mazingira ya Afrika Mashariki (k.m., majina ya maeneo, bidhaa za kawaida, hali za kijamii).
> - **Fuata Maagizo:** Hakikisha maandishi yanayotokana yanaakisi kwa uaminifu vigezo vyote vilivyotolewa (hisia, kikoa, kipengele, mtindo, urefu, ubora).
>
> **Muundo wa Majibu:** Tafadhali toa maandishi yaliyotengenezwa PEKEE, bila maelezo ya ziada, utangulizi, au hitimisho lako mwenyewe. Weka maandishi yote ndani ya mabano ya mraba '[ ]'.
>
> Mfano wa muundo unaotarajiwa:
>
> [<Weka maandishi yako ya Kiswahili yaliyotengenezwa hapa>]

B.2 Judging Prompt Template

To evaluate the generated Swahili text, the following prompt template was provided to the judging model. The `{criteria_str}` placeholder is replaced with a JSON string detailing the original generation criteria, `{text}` is replaced with the Swahili text to be evaluated, and `{criteria['sentiment']}` is replaced with the target sentiment for specific alignment checking.

Swahili Text Evaluation Prompt Template

You are an expert evaluator of Swahili text, focusing on linguistic quality, cultural relevance, and adherence to instructions. Evaluate the following Swahili text based on the provided criteria.

Criteria: `{criteria_str}`

Swahili Text to Evaluate: `{text}`

Evaluation Tasks: Please provide scores based on the following dimensions:

- **Language_Correctness:** (Scale 0-5) How grammatically correct, fluent, and natural is the Swahili used? (0=Very Poor, 5=Excellent)
- **Cultural_Relevance:** (Scale 0-5) How well does the text incorporate East African/Swahili cultural context, proverbs, or nuances appropriately? (0=Not relevant/Inappropriate, 5=Highly relevant and natural)
- **Sentiment_Alignment:** (Scale 0-5) How accurately does the text reflect the target sentiment (‘{criteria[‘sentiment’]}’)? (0=Completely misaligned, 5=Perfectly aligned)
- **Instruction_Following:** (Scale 0-5) How well does the text adhere to all other specified criteria (Domain, Aspect, Tone, Length, Quality)? (0=Ignores most criteria, 5=Follows all criteria closely)
- **Overall_Quality:** (Scale 0-10) Considering all aspects, what is the overall quality of this generated text as a representative Swahili sample? (0=Very low, 10=Outstanding)

Output Format: Provide ONLY a JSON object containing the scores. Do not include any explanations or surrounding text.

Example Output:

```
{
"Language_Correctness": 4,
"Cultural_Relevance": 3,
"Sentiment_Alignment": 5,
"Instruction_Following": 4,
"Overall_Quality": 8
}
```

Your JSON Output:

C Ensemble Experiments

In this section, we present the quantitative results of our inter-judge ensemble experiments. These visualizations illustrate the level of agreement between the selected LLM judges and their individual vs. collective performance when compared to human-annotated ground truth.

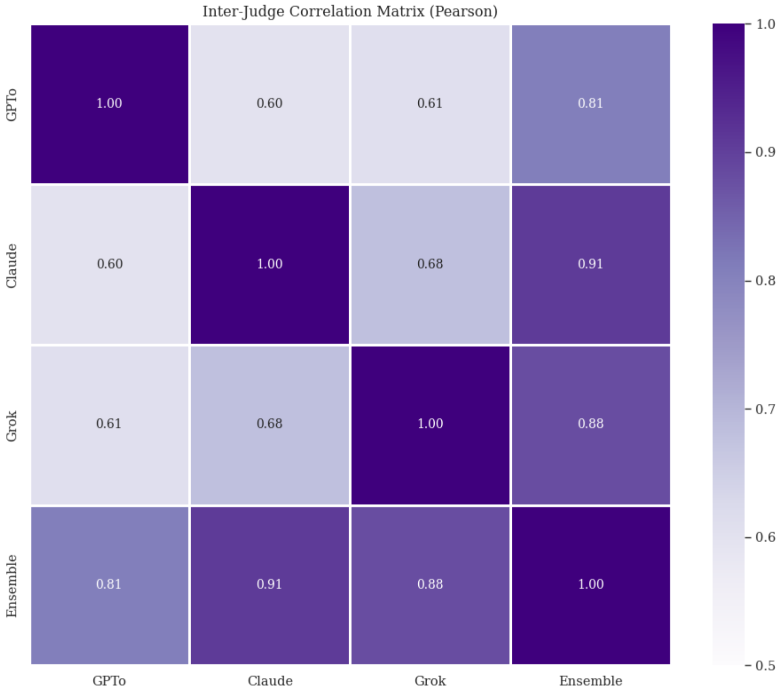


Figure 6: Correlation Heatmap of Judgment Scores: This matrix shows the Pearson correlation between the different evaluation dimensions across the judge ensemble.

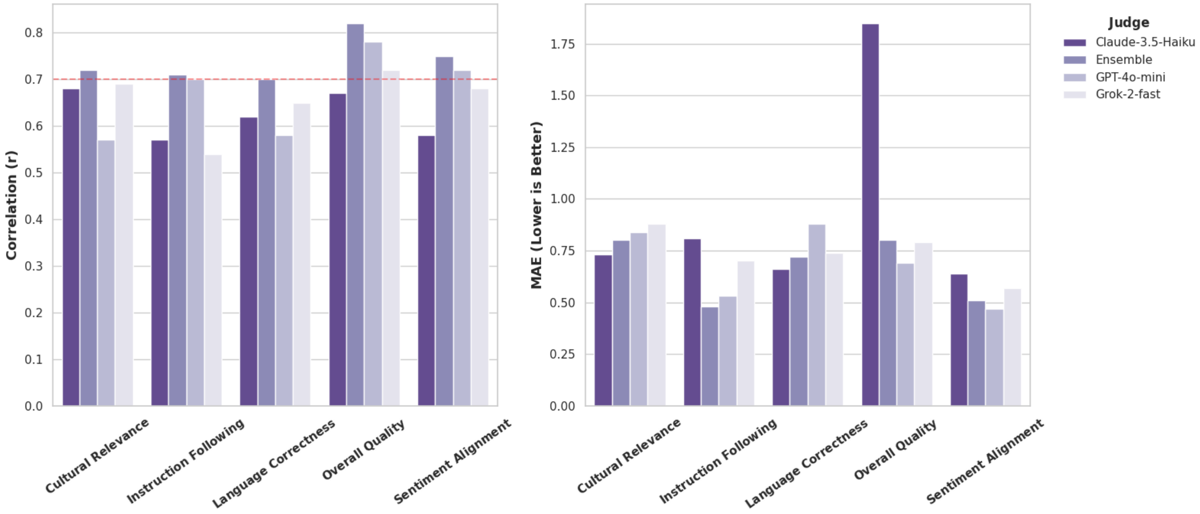


Figure 7: Judge Performance Metrics: Comparison of individual judge models and the final ensemble against human ground truth for Pearson Correlation (left) and Mean Absolute Error (right).

D Generated Samples

This section presents illustrative examples of synthetic data generated during our experiments. We showcase samples that received low overall quality scores, highlighting common failure modes, followed by samples that were judged to be of high quality, demonstrating successful generation according to the specified criteria.

D.1 Examples of Low-Quality Samples (Overall_Quality < 5)

The following samples were rated with an Overall_Quality score below 5, indicating significant deviations from the desired output or failures in adhering to instructions.

Sample ID: swahili_13932

- **Generator Model:** Llama3-70B
- **Overall Quality:** 2.0
- **Text:**

  Tafadhali niwaonyeshe ukarimu wao katika kituo cha fitness cha [Jina la Kituo
- **Criteria:**
  - `sentiment`: '3 - Neutral'
  - `domain`: 'Gym or Fitness Center Review'
  - `aspect`: 'Security Protocols'
  - `tone`: 'enthusiastic'
  - `target_length`: 'medium-long (125-175 words)'
  - `desired_quality`: 'Exceptional'
  - `language`: 'Swahili'
- **Judgment:**
  - `Language_Correctness`: 2.0
  - `Cultural_Relevance`: 1.0
  - `Sentiment_Alignment`: 2.0
  - `Instruction_Following`: 1.0
  - `Overall_Quality`: 2.0

Sample ID: swahili_17332

- **Generator Model:** Gemini-Flash
- **Overall Quality:** 3.0
- **Text:**

  Huyu mkufunzi binafsi ni janga! Anadai anajua sheria, lakini ni uongo mtupu! Kwani yeye ni nani, mjanja asiyejua lolote? "Mganga hajigangi," lakini huyu anatuuzia sumu! Anatuambia anafuata kanuni, lakini hati zake ni kama "kitumbua kimeingia mchanga" - hakuna kinachoeleweka. Anatuibia pesa zetu kwa visingizio vya kijinga, halafu anatuambia "subira yavuta heri." Heri gani, wakati tunazidi kuumia na upuuzi wake? Hata leseni hana! Anafanya kazi kinyume cha sheria kama "panya akimbia bila shati." Nataka pesa zangu zirudishwe! Hana adabu wala heshima! Atajuta!
- **Criteria:**
  - `sentiment`: '5 - Extremely Positive'
  - `domain`: 'Personal Trainer Evaluation'
  - `aspect`: 'Legal Compliance'
  - `tone`: 'angry'
  - `target_length`: 'medium (100-150 words)'
  - `desired_quality`: 'Very Poor'
  - `language`: 'Swahili'
- **Judgment:**
  - `Language_Correctness`: 4.0

  - Cultural_Relevance: 5.0
  - Sentiment_Alignment: 0.0
  - Instruction_Following: 2.0
  - Overall_Quality: 3.0

Sample ID: swahili_7573

- **Generator Model:** Llama3-70B
- **Overall Quality:** 2.0
- **Text:**

  Nilikuwa na matumaini makubwa juu ya huduma ya bima ya [jina la kampuni
- **Criteria:**
  - sentiment: '2 - Negative'
  - domain: 'Insurance Service Feedback'
  - aspect: 'Overall Experience'
  - tone: 'surprised'
  - target_length: 'extensive (250-300 words)'
  - desired_quality: 'Abysmal'
  - language: 'Swahili'
- **Judgment:**
  - Language_Correctness: 2.0
  - Cultural_Relevance: 1.0
  - Sentiment_Alignment: 2.0
  - Instruction_Following: 1.0
  - Overall_Quality: 2.0

Sample ID: swahili_10177

- **Generator Model:** Gemini-Flash
- **Overall Quality:** 3.0
- **Text:**

  Hebu niseme ukweli, sijawahi kukutana na mfumo mbovu kama huu. Wanaita "kitega uchumi," lakini kwangu mimi, ni shimo la kutupia pesa. Lakini, tuseme, nilikuwa najaribu kutafuta hisa za kampuni ya sukari ya Mumias, nikidhani labda kuna mtu bado anaamini katika muujiza. Kazi ya utafutaji? Usiniambie! Ni kama kutafuta sindano kwenye fungu la nyasi, au kama vile wasemavyo, "Kila ndege huruka kwa bawa lake," lakini hapa, mbawa zimevunjika.

  Nimejaribu majina mengi, herufi, hata alama za uandishi, lakini hakuna chochote! Inarudi tupu. Kweli, kama vile msemo unavyosema, "Mficha ugonjwa, kifo humuumbua." Hivyo ndivyo mfumo huu ulivyo: unajaribu kuficha udhaifu wake, lakini udhaifu wenyewe unajitokeza waziwazi. Sijui kama mtu yeyote anafanya kazi hapa, lakini kama wanafanya, wanalipwa bure. Bora niendelee na biashara yangu ya kuuza vitumbua, angalau huko najua ninachokifanya.
- **Criteria:**
  - sentiment: '5 - Extremely Positive'
  - domain: 'Investment Platform Review'
  - aspect: 'Search Functionality'
  - tone: 'pessimistic'
  - target_length: 'medium-long (125-175 words)'
  - desired_quality: 'Very Poor'
  - language: 'Swahili'
- **Judgment:**

  - Language_Correctness: 4.0
  - Cultural_Relevance: 5.0
  - Sentiment_Alignment: 0.0
  - Instruction_Following: 2.0
  - Overall_Quality: 3.0

Sample ID: swahili_26557

- **Generator Model:** Gemini-Flash
- **Overall Quality:** 3.0
- **Text:**

  Hii simu janja inanikera! Wanadai ina uwezo wa uchanganuzi wa kutabiri? Uongo mtupu! Ni kama kumpa nyani shilingi! Nimepoteza pesa zangu bure! Wanasema inatabiri hali ya hewa? Hata mzee wa kienyeji anajua vizuri zaidi! Hawana haya! "Mganga hajigangi!" Nimeghadhabika sana! Watu wanadanganywa kila siku. Hii ni wizi mtupu! Nataka pesa zangu zirudishwe! Hii ni hasira tupu!
- **Criteria:**
  - sentiment: '4.5 - Very Positive'
  - domain: 'Tech Gadget Evaluation'
  - aspect: 'Predictive Analytics'
  - tone: 'angry'
  - target_length: 'short (50-100 words)'
  - desired_quality: 'Exceptional'
  - language: 'Swahili'
- **Judgment:**
  - Language_Correctness: 4.0
  - Cultural_Relevance: 5.0
  - Sentiment_Alignment: 0.0
  - Instruction_Following: 2.0
  - Overall_Quality: 3.0

D.2 Examples of High-Quality Samples (Overall_Quality ≥ 5)

In contrast, the following samples were rated with an Overall_Quality score of 5 or higher, indicating successful adherence to the specified criteria and good generation quality. Out of the total generated samples, 47980 were classified as good (Overall_Quality ≥ 5). We display five representative examples below.

Sample ID: swahili_889

- **Generator Model:** Llama3-70B
- **Overall Quality:** 6.0
- **Text:**

  Tatizo la kuongeza thamani kwa kodi za uchaguzi zinazotumika kwa kuongeza faida kwa wananchi haijawahi kufikiwa, hali ambayo inaonyesha kuwa kuna matatizo makubwa ya kiuchumi.
- **Criteria:**
  - sentiment: '1 - Extremely Negative'
  - domain: 'Politics'
  - aspect: 'Return on Investment'
  - tone: 'tense'
  - target_length: 'micro (10-25 words)'
  - desired_quality: 'Average'
  - language: 'Swahili'

- **Judgment:**
  - Language_Correctness: 4.0
  - Cultural_Relevance: 3.0
  - Sentiment_Alignment: 4.0
  - Instruction_Following: 2.0
  - Overall_Quality: 6.0

Sample ID: swahili_36367

- **Generator Model:** Llama3-70B
- **Overall Quality:** 7.0
- **Text:**

  Filamu hii imejaa vipengele vyenye kazi kubwa, kama vile uhuishaji mzuri wa wahusika, mandhari mazuri ya kuigizia, na muziki wenye kazi kubwa. Kwa mfano, wahusika wameonyesha uwezo mkubwa wa kucheza, na mandhari yameonyesha utajiri wa Afrika Mashariki. Muziki umetulia kwa kuonyesha hisia za wahusika. Filamu hii inastahili kuitazama kwa sababu ya vipengele vyake vyenye kazi kubwa.
- **Criteria:**
  - sentiment: '4.5 - Very Positive'
  - domain: 'Movie Review'
  - aspect: 'Features'
  - tone: 'formal'
  - target_length: 'short (50-100 words)'
  - desired_quality: 'Poor'
  - language: 'Swahili'
- **Judgment:**
  - Language_Correctness: 4.0
  - Cultural_Relevance: 4.0
  - Sentiment_Alignment: 5.0
  - Instruction_Following: 3.0
  - Overall_Quality: 7.0

Sample ID: swahili_3898

- **Generator Model:** Gemini-Flash
- **Overall Quality:** 5.0
- **Text:**

  Ah, chombo hiki cha 'freemium'! Kweli, 'bure' kama vile kupewa maji ya mvua jangwani. Walisema kinafanya kazi vizuri kwenye simu. Kweli, mimi nasema, 'kazi' ni neno pana sana, kama vile 'maendeleo' nchini kwetu.

  Ukiangalia kwenye simu yangu, inaonekana kama embe lililoanguka kutoka juu ya mnazi – limepondwa kabisa. Labda ni mimi tu, lakini nahisi kama wametumia picha iliyopanuliwa mara elfu moja. Kila kitu kinaonekana kama pixel moja kubwa. Unadhani wamelipa nani kujaribu hii kwenye simu? Labda paka wao?

  'Huu ndio uboreshaji wa hali ya juu,' wanasema. Sawa, na mimi ndiye Rais wa Marekani. Ni kama kujaribu kuendesha lori la mizigo kupitia barabara ya vumbi ya kijijini kwetu. Utajikuta umekwama. 'Haba na haba hujaza kibaba,' lakini haba hizi za uboreshaji zinajaza nini? Machozi yangu? Labda.

  Sijui, labda simu yangu ni ya kale kama kaburi la Farao, lakini nadhani hata kaburi la Farao lina muundo bora kuliko programu hii kwenye simu yangu. Kama wasemavyo, 'Mchumia juani hulia kivulini,' lakini kwa chombo hiki, unalia juani na kivulini. Kweli kabisa.
- **Criteria:**

- sentiment: ’4 - Positive’
  - domain: ’Freemium Tool Evaluation’
  - aspect: ’Mobile Responsiveness’
  - tone: ’sarcastic’
  - target_length: ’long (150-200 words)’
  - desired_quality: ’Poor’
  - language: ’Swahili’
- **Judgment:**
  - Language_Correctness: 4.0
  - Cultural_Relevance: 5.0
  - Sentiment_Alignment: 0.0
  - Instruction_Following: 2.0
  - Overall_Quality: 5.0

Sample ID: swahili_44704

- **Generator Model:** Llama3-70B
- **Overall Quality:** 9.0
- **Text:**

  Tafadhali, nilipata kipindi kizuri sana kwa wakati wote nilipokuwa nikipitia kwenye kituo cha kukodisha magari cha XYZ. Nilipendezwa na utendaji wao wa kufanya kazi kwa wakati, hivyo niliweza kufanya mambo yangu yote kwa wakati. Kama vile methali inavyosema, "Mwanzo mwema, mwisho mwema" - hawa jamaa walifanya kazi kwa mwanzo mwema, hivyo nilipata mwisho mwema. Nilipata magari yangu kwa wakati, na niliweza kufanya safari yangu kwa uhakika. Kwa kweli, nilipendezwa na huduma yao ya kukodisha magari, na nitafanya kazi nao tena.
- **Criteria:**
  - sentiment: ’4 - Positive’
  - domain: ’Car Rental Experience’
  - aspect: ’Time Management’
  - tone: ’ecstatic’
  - target_length: ’medium-short (75-125 words)’
  - desired_quality: ’with typos’
  - language: ’Swahili’
- **Judgment:**
  - Language_Correctness: 5.0
  - Cultural_Relevance: 5.0
  - Sentiment_Alignment: 5.0
  - Instruction_Following: 4.0
  - Overall_Quality: 9.0

Sample ID: swahili_37585

- **Generator Model:** Gemini-Flash
- **Overall Quality:** 7.0
- **Text:**

  Usimulizi wa hadithi katika jukwaa hili una changamoto zake. Kama wahenga walivyosema, "Kila ndege huruka kwa bawa lake." Kuna haja ya kuboresha uzi wa matukio ili kuvutia watazamaji zaidi.
- **Criteria:**
  - sentiment: ’3 - Neutral’
  - domain: ’Streaming Platform User Experience Review’

  - aspect: ’Plot and Storytelling’
  - tone: ’diplomatic’
  - target_length: ’very short (25-50 words)’
  - desired_quality: ’Very Good’
  - language: ’Swahili’
- **Judgment:**
  - Language_Correctness: 4.0
  - Cultural_Relevance: 4.0
  - Sentiment_Alignment: 4.0
  - Instruction_Following: 3.0
  - Overall_Quality: 7.0

E Complete Generation Criteria

This section provides the complete set of values used for each generation criterion in our synthetic data pipeline. During generation, one value from each category was randomly sampled to create diverse, controlled outputs.

E.1 Sentiment Values

The sentiment scale includes 8 values with weighted sampling (neutral sampled more frequently):

- 1 – Extremely Negative
- 1.5 – Very Negative
- 2 – Negative
- 3 – Neutral (2x weight)
- 4 – Positive
- 4.5 – Very Positive
- 5 – Extremely Positive

E.2 Target Length Values

Length specifications include approximate word counts:

- Micro (10–25 words)
- Very short (25–50 words)
- Short (50–100 words)
- Medium-short (75–125 words)
- Medium (100–150 words)
- Medium-long (125–175 words)
- Long (150–200 words)
- Very long (200–250 words)
- Extensive (250–300 words)

E.3 Desired Quality Values

Quality levels range from poor to exceptional, including intentional noise:

- Abysmal
- Very Poor
- Poor
- Below Average
- Fair
- Average
- Above Average
- Good
- Very Good
- Excellent
- Outstanding
- Exceptional
- With typos
- Subpar
- Mediocre
- Decent
- Satisfactory
- Emotional

E.4 Tone Values (50+ categories)

Tones span simple emotions to complex affective states:

- Emotional
- Rational
- Sarcastic (2x)
- Enthusiastic
- Formal
- Casual
- Humorous
- Serious
- Optimistic
- Pessimistic
- Analytical
- Sympathetic
- Critical
- Appreciative
- Indifferent
- Passionate
- Contemplative
- Angry
- Joyful
- Melancholic
- Anxious
- Confident
- Confused
- Curious
- Disappointed
- Ecstatic
- Frustrated
- Grateful
- Nostalgic
- Objective
- Playful
- Reflective
- Skeptical
- Surprised
- Tentative
- Whimsical
- Persuasive
- Disappointed but hopeful
- Diplomatic
- Tense
- Relieved
- Motivational
- Defensive
- Apologetic
- Detached
- Authoritative

- Encouraging
- Inquisitive
- Uncertain
- Directive
- Candid

E.5 Domain Values (160+ categories)

Domains cover diverse review and feedback contexts relevant to East African users:

- Product Review
- Restaurant Review
- Movie Review
- Book Review
- Travel Experience
- Tech Gadget Evaluation
- Hotel Stay
- Concert Experience
- Video Game Critique
- Fitness Equipment Assessment
- Online Course Feedback
- Streaming Service Review
- Software Review
- Podcast Evaluation
- Social Media Platform Review
- E-commerce Feedback
- Mobile App Evaluation
- Subscription Box Review
- Smart Home Device Review
- Fashion Product Review
- Health and Wellness Service
- Car Rental Experience
- Home Appliance Evaluation
- Beauty Product Review
- Food Delivery Service
- Banking Service Evaluation
- Insurance Service Feedback
- Educational Platform Review
- Gym or Fitness Center Review
- Event Venue Review
- Public Transport Feedback
- Airline Experience Review
- Hospital/Medical Service
- Personal Care Product
- Streaming Content Review
- Career Coaching Feedback
- Language Learning Tool
- Online Therapy Service
- Financial Product Review
- Crypto Exchange Feedback
- Telecommunications Service
- Real Estate Platform
- Charity/Non-Profit Service
- Pet Care Product Review
- Furniture Review
- Sports Equipment
- Legal Service Review
- Meal Kit Delivery Service
- Dating App Review
- Gaming Hardware Review
- Investment Platform Review
- Digital Art Tool Evaluation
- Music Streaming Service
- Home Renovation Service
- Medical Device Review
- Fitness Wearable Review
- Ride-Hailing Service
- Electric Scooter Rental
- Grocery Delivery Service
- Event Ticket Booking
- Healthcare App Review
- Car Maintenance Service
- Private Tutor Evaluation
- Childcare Service Review
- Subscription News Service
- Second-Hand Marketplace
- DIY Tool Review
- Local Business Feedback
- Pet Grooming Service
- Wedding Venue Review
- Energy Provider Feedback
- Renewable Energy Service
- Recycling Service Feedback
- Cloud Storage Service
- Job Recruitment Platform
- Freelancer Hiring Feedback
- Time Management App
- Password Management Tool
- Educational Workshop
- Freemium Tool Evaluation
- Luxury Travel Experience
- Adventure Gear Feedback
- Fitness Class Review
- Streaming Original Content
- Meal Prep Service
- Hybrid Car Review
- Urban Planning Feedback
- Noise-Canceling Headphones
- Budget Hotel Review
- Luxury Spa Experience
- Park and Recreation Area
- Real Estate Agent
- Custom Jewelry
- Tailored Suit Service
- Local Festival Feedback

- Community Event Review
- Shared Office Space
- Medical Insurance Plan
- Home Warranty Service
- Satellite TV Review
- Gardening Subscription Box
- Seasonal Product
- Language Certification
- Test Prep Course
- Online Shopping Experience
- Local Government Service
- Public Utility Review
- Mental Health Support App
- Personal Trainer
- Self-Help Book Review
- Cooking Class Feedback
- Art Class Review
- Wine Tasting Event
- Local Coffee Shop Review
- Specialty Food Store
- Luxury Brand Review
- Budget Product Comparison
- Mobile Data Plan
- Artisan Craft Product
- Delivery Driver Service
- Customer Loyalty Program
- Smartwatch App
- Streaming Platform UX
- Small Business E-commerce
- Public Library Feedback
- Museum Experience
- Online Art Gallery
- Fitness Challenge App
- Meal Recipe Kit
- Home Cleaning Service
- Child Safety Product
- Camping Accessory
- Shared Workspace Equipment
- Startup Tool Evaluation
- Food Subscription Service
- Home Decor Platform
- Tech Repair Service
- Budget Airline Review
- Local Restaurant Chain
- Customized Gift Platform
- Eco-Friendly Packaging
- Mobile Game Review
- Travel Insurance
- Startup Accelerator
- Team Collaboration Platform
- Workflow Automation Tool
- Twitter
- Facebook
- Netflix
- Amazon
- Instagram
- Google
- OpenAI
- Politics
- Personalized Gift Service
- Custom Pet Food Subscription
- Money
- Stocks
- Memes
- General (8x)

E.6 Aspect Values (190+ categories)

Aspects define specific evaluation dimensions:

- Overall Experience (4x)
- Quality (5x)
- Value for Money
- Customer Service
- Atmosphere
- User Interface
- Taste and Flavor
- Plot and Storytelling
- Writing Style
- Performance
- Functionality
- Design and Aesthetics
- Comfort
- Durability (2x)
- Sound Quality
- Battery Life
- Ease of Use
- Features
- Speed
- Reliability
- Packaging
- Brand Reputation
- Convenience
- Sustainability (2x)
- Privacy and Security
- Personalization
- Customer Support
- Accessibility
- Responsiveness
- Scalability (3x)
- Learning Curve
- Ease of Setup
- Maintenance Requirements
- Safety
- Return on Investment
- Pricing Transparency (2x)

- Cleanliness
- Location
- Timeliness
- User Engagement
- Compatibility
- Innovativeness
- Environmental Impact
- Community Support (2x)
- Subscription Model
- Availability (2x)
- Credibility
- User Satisfaction (2x)
- Problem-Solving Efficiency
- Data Privacy
- Efficiency
- Trustworthiness
- Content Relevance
- Professionalism
- Customization Options
- Warranty and Support
- Delivery Speed
- Ethical Standards (2x)
- Aesthetic Appeal
- Upgrade Options
- Interactivity
- Accuracy
- Technical Support
- Ease of Integration
- Customer Retention
- Loyalty Programs
- Transparency (2x)
- User Documentation
- Visual Appeal
- Error Handling
- Upgradability
- Consistency
- Novelty
- Cultural Relevance (2x)
- Social Responsibility
- Team Collaboration
- Knowledge Transfer
- Versatility
- Return Policy
- Innovation Pipeline
- Market Adaptability
- Future-Proofing
- Risk Management
- Energy Efficiency
- Resource Optimization
- Supply Chain Transparency
- Community Engagement (2x)
- Data Analytics
- Remote Accessibility
- Cross-Platform Support
- Time Management
- Conflict Resolution
- Emotional Impact
- Social Connectivity
- Network Stability
- Cost Effectiveness
- Integration with Third-Party Tools
- Onboarding Process
- Real-Time Performance
- Precision
- Task Automation
- Multilingual Support
- Gamification Elements
- Team Productivity
- AI Integration
- Intuitiveness
- Knowledge Base Availability
- In-App Guidance
- Search Functionality
- Cloud Syncing
- Mobile Responsiveness
- Refund Policy
- Crisis Management (2x)
- Error-Free Operation
- Technical Depth
- Customer Insights
- Remote Support Availability
- Cultural Appropriateness
- User-Generated Content
- Adaptability to User Needs
- Real-World Applications
- Cohesion
- Operational Complexity
- Heat Management
- Physical Ergonomics
- Carbon Footprint
- Eco-Friendliness
- Repairability
- Noise Levels
- Intellectual Stimulation
- Attention to Detail
- Story Immersion
- Inclusiveness
- Cultural Sensitivity
- Fairness
- Predictive Analytics
- Real-Time Updates
- Latency
- Resource Intensity
- Product Roadmap

- Third-Party Reviews
- Localization Options
- Hardware Compatibility
- Modular Design
- Subscription Management
- Customer Feedback Loop
- Attention-Grabbing Features
- Visual Hierarchy
- Adoption Rate
- Social Proof
- Adaptability to Emerging Trends
- Benchmark Scores
- Industry Standards Compliance
- Security Protocols
- Multi-Device Synchronization
- Visual Continuity
- Comprehensiveness
- Engagement Metrics
- Ease of Replication
- Crisis Adaptability
- Legal Compliance
- Long-Term Usability
- Training Resources
- Ease of Troubleshooting
- Depth of Customization
- Early Access Benefits
- Depth of Analytics
- Interactive Feedback
- Live Support Options
- Proactive Solutions
- Shared Resource Support

E.7 Specificity Values

Specificity controls the level of detail and focus:

- Extremely vague
- Very general
- General
- Somewhat general
- Balanced
- Somewhat specific
- Specific
- Very specific
- Highly detailed
- Comprehensive
- Broad
- Focused
- Narrow
- Precise
- Abstract
- Concrete
- Nuanced
- Superficial
- In-depth
- Thorough
- Cursory
- Elaborate
- Succinct
- Meticulous
- Overview
- Bird’s-eye view
- Microscopic
- Holistic
- Broad but focused
- Granular
- Slightly ambiguous
- Highly technical
- Contextualized
- Situational