Afri-MCQA : Multimodal Cultural Question Answering Dataset for African Languages

Anonymous ACL submission

Abstract

Multimodal large language models (MLLMs) have demonstrated impressive capabilities in reasoning across visual, audio, and text modalities. However, their development has been primarily centered on high-resource, Western contexts, leaving significant gaps in understanding diverse global cultures. This difference is especially noticeable in Africa, which is home to more than one-third of the world's languages. However, it remains severely underrepresented in existing multimodal benchmarks. In this paper, we introduce Afri-MCQA , the first large-scale multilingual cultural question-answering dataset specifically designed for 15 African languages with questions in text and audio modalities. Our comprehensive evaluation across both, open and closed models, show significant performance gaps when processing African languages and their cultural contexts, with notable differences between text and speech modalities. These findings underscore the urgent need for more inclusive multimodal AI development. By open-sourcing Afri-MCQA, we aim to support research toward building AI systems that can understand and serve Africa's linguistically and culturally diverse communities, ultimately contributing to more equitable global AI technologies.

1 Introduction

Multimodal large language models (MLLMs) have shown strong performance across tasks that require reasoning over visual, spoken, and textual inputs to follow user instructions (Liu et al., 2023; Team et al., 2025b; Abdin et al., 2024). However, these systems are predominantly developed in high-resource, mostly Western contexts, or as Mihalcea et al. put it, they have a 'weird' coverage (Mihalcea et al., 2024). Therefore, their knowledge and reasoning abilities tend to benefit the languages and cultures represented in their training data. While recent benchmarks have begun to assess global cultural knowledge through question answering, their coverage of African languages and contexts remains limited (Romero et al., 2024; Vayani et al., 2025a; Parida et al., 2023).

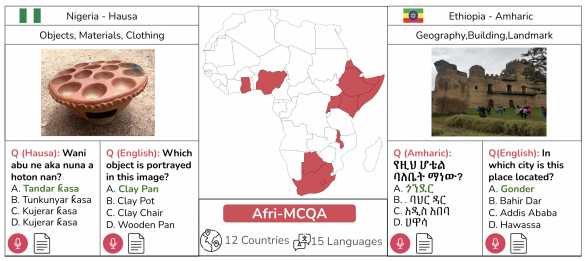


Figure 1: Afri-MCQA, a multilingual and multimodal cultural question answering benchmark.

Africa is one of the most culturally diverse and rapidly growing regions in the world. It is home to more than one-third of the worlds languages (Hammarström, 2018) and a population exceeding 1.3 billion, projected to surpass 2.5 billion by 2050 (Simane et al., 2025). Similarly, African languages have linguistic features that are very different from many high-resource languages represented in LLMs including rich morphology, use of noun classes, tonality, serial verb constructions, among others (Nurse and Philippson, 2006; Adebara and Abdul-Mageed, 2022). Additionally, many African languages are primarily spoken, with literacy often occurring in a colonial or foreign language. This makes speech-based applications particularly relevant for enabling access to technology.

Although most evaluation datasets for low-resource languages are translated (NLLB Team, 2022; Adelani et al., 2024a,b), recently researchers have made promising progress in creating culturally relevant evaluation benchmarks for African languages across various downstream NLP tasks (Adelani et al., 2023; Azime et al., 2025; Muhammad et al., 2025; Yu et al., 2025). However, most of

these evaluations are designed for text-only downstream tasks. In the multimodal setting, only few African languages have been included. For example, CVQA (Romero et al., 2024) probes visual cultural knowledge and includes three African languages, whileWorldCuisines (Winata et al., 2025), a food related benchmark only included one.

For AI to be a truly global and equitable technology, it must be designed to serve all populations. This requires a focused effort to address the unique linguistic and cultural diversity of African communities, which are often overlooked by current systems. An important step toward this goal is to measure how well current MLLMs can understand and reason about African cultural knowledge in a multimodal setting. As we have noted, current benchmarks have several limitations: **(1)** the majority support only text, **(2)** some of them are created by translating English benchmarks, and **(3)** few languages from Africa are represented in current multimodal benchmarks.

To address the above limitations, we introduce Afri-MCQA , **the first multilingual cultural visual question-answering dataset supporting text and speech modalities**. It consists of $\approx 7.5k$ questions and answer pairs (in English and native languages) in 15 African languages collected from 12 countries. Our data collection involves native speakers as annotators, who reside in the countries where their respective languages are spoken. To ensure high-quality data, we included native speakers as coordinators for each language, who are familiar with the culture of the countries involved in our work. Each language in Afri-MCQA contains 500 Q&A grounded in images. Each pair has both written text and spoken audio, in both native and English languages. We conduct a detailed evaluation of different MMLMs in both native and English data across different experimental setups, querying with both audio and text Q&A. In Table 1, we summarize and compare our work with previous studies that include African languages in the VQA datasets. By combining linguistic diversity, multimodality, and cultural grounding, Afri-MCQA fills a critical gap in African language VQA research. We outline below the contributions of our work:

- We introduce Afri-MCQA, the first large-scale multilingual visual cultural question answering dataset with questions *both in text and speech* for *15 African languages* and their translations in English, spoken in 12 countries, collected and evaluated by native speakers.
- We demonstrate that **text-based multilingual ability does not transfer to speech understanding**, emphasizing the importance of increasing African language representation in multimodal model training.
- We open-source Afri-MCQA to promote research on multimodal systems that can more effectively understand and represent Africas diverse languages and cultures.

2 Related work

With the exponential growth in popularity of LLMs and MLLMs across the world, cultural and multilingual multimodal evaluation has received more attention. For instance, Liu et al. (2021) tested models on verifying statements about pairs of culturally-related images, but this was framed merely as a binary classification task. CVQA (Romero et al., 2024) and CulturalVQA (Nayak et al., 2024) explicitly target cultural knowledge through human-written questions, yet they are limited to a small set of languages per continent. ALM-bench (Vayani et al., 2025a) expands the language coverage, but relies heavily on LLM-generated questions and web-sourced images. Additionally, a common limitation across these benchmarks is that they query models only through text, overlooking the importance of speech, an important modality for communities where language is mostly spoken.

On the other hand, work on African languages has focused mainly on text-based benchmarks. While a few multilingual resources, such as GlobalMMLU (Singh et al., 2024), include a small subset of African languages, other region-specific benchmarks provide broader coverage. Examples of these include MasakhaNEWS (Adelani et al., 2023) that evaluate text classification, AfriQA (Ogundepo et al., 2023) for cross-lingual question answering, or larger suites such as AfroBench (Ojo et al., 2025) and IrokoBench (Adelani et al., 2024b) that cover multiple *text* tasks. Across these efforts, a consistent finding is that models perform poorly on African languages, with open-weight models showing a significant gap with proprietary systems.

Despite this progress, all of these evaluations remain text-only. Visual knowledge and multimodal reasoning for African contexts are still largely untested. Existing multi-region multimodal datasets (Romero et al., 2024; Vayani et al., 2025b)

Datasets	# African Lang	# Countries	QA categories	# QA per langs	Audio QA	Parallel Data
CVQA (Romero et al., 2024)	5	4	10	200	×	✓
WC-VQA (Winata et al., 2025)	1	1	1	-	×	✓
M5 (Schneider and Sitaram, 2024)	4	4	-	-	×	×
HaVQA (Parida et al., 2023)	1	1	-	6,200	×	✓
Afri-MCQA (ours)	15	12	10	500	✓	✓

Table 1: Data statistics for Afri-MCQA compared to existing VQA datasets that include African languages.

Language	Family/branch	Region	# spk	#QA	# eng (h)	# nat (h)
Ethiopia-amh (amh)	Afro-Asiatic / Ethio-Semitic	East	57M	500	1.56	1.51
Nigeria-hau (hau)	Afro-Asiatic / Chadic	West	77M	496	2.8	3.04
Nigeria-ibo	Niger-Congo / Volta-Niger	West	31M	501	1.61	1.59
Ethiopia-orm (orm)	Afro-Asiatic / Cushitic	East	37M	512	2.2	2.3
Rwanda-kin	Niger-Congo / Bantu	East	18M	501	2.73	2.67
Kenya-kik (ki)	Niger-Congo /Bantu	East	8.1M	495	1.66	1.72
Somali-som(so)	Afro-Asiatic / Cushitic	East	22M	501	1.96	1.99
Eritrea-tir (tig)	Afro-Asiatic / Ethio-Semitic	East	9M	537	2.13	2.31
Ghana-twi	Niger-Congo / Volta-Niger	West	9M	537	2.41	2.43
Nigeria-yor (yor)	Niger-Congo / Volta-Niger	West	46M	498	1.98	2.06
Botswana-tsn(tn)	Niger-Congo / Volta-Niger	South	14M	502	1.89	2.39
Malawi-nya	Niger-Congo /Bantu	S & E	12M	501	1.41	1.5
Lesotho-sot(sot)	Niger-Congo /Bantu	South	13.5M	533	-	1.9
S.Africa-zul(zul)	Niger-Congo / Bantu	South	28M	528	-	1.37

Table 2: **Overview of the languages represented in our dataset.** For each language, we list its family and branch classification, geographic region, estimated L1 & L2 speaker population(#spk), number of question-answer pairs collected (#QA), and the total duration of audio collected. #eng (h) is the total hours of accented English audio, and #nat (h) is the total hours of audio in that language.

include only a few African languages and do not adequately capture the regions diversity. To address this, we introduce Afri-MCQA, the first human-created benchmark designed to evaluate multimodal cultural knowledge in Africa, querying through both text and speech. Our dataset covers 15 African languages, includes native-collected questions grounded in local imagery, and enables evaluation of vision-language and audio-visual LLMs. Unlike prior work, Afri-MCQA directly addresses the needs of African languages that are primarily spoken and underrepresented in existing multimodal benchmarks.

3 Dataset

To create Afri-MCQA , we selected the 15 widely spoken languages in sub-Saharan Africa (by number of speakers, as per Ethnologue[1]). While prior multimodal evaluation frameworks cover only a few African languages, Afri-MCQA spans multiple language families and 12 countries, and represents an aggregate speaker population of approximately 392.6 million (see Table 2).

We hired annotators who are native language speakers through the Upwork[2] freelance platform. We hired annotators based on (i) fluency in the English language, (ii) prior annotation experience or data collection work, (iii) high project completion rate on Upwork, and (iv) residence in a country where the target language is spoken. After the selection process, we divided the annotation into two phases designed to ensure quality.

Below, we (i) describe our two-phase data collection pipeline, outlining the steps followed in the annotation process and our quality assurance procedures, and (ii) describe the structure of each data point (image, text, and audio).

3.1 Dataset collection

Guideline and platform Preparation We followed the annotation guideline of Romero et al. (2024), including categories of images, question templates, and distractors, while extending them to include audio recording instructions (for both local-language and English audio) and adding a two-step review and verification process to collect quality data. We provide full annotation guidelines in Appendix D.

Training and Screening Phase The first phase consisted of a small-scale submission designed to train and screen annotators to ensure the quality of the data collection. Therefore, we began by providing them with guidelines and training them to understand the tasks, ensuring they understood the criteria and quality standards of the task. After training, they were tasked with submitting 50 QA samples (image, bilingual QA, and audio). We reviewed these submissions to verify that the task understanding and quality were in line with the provided guidelines. After the review, we provided detailed feedback and, when necessary, scheduled follow-up meetings to review their submissions and clarify any issues. Only annotators whose submissions met our quality criteria and incorporated the feedback successfully were selected for the next phase.

[1] https://www.ethnologue.com/insights/ethnologue200/

[2] https://www.upwork.com/

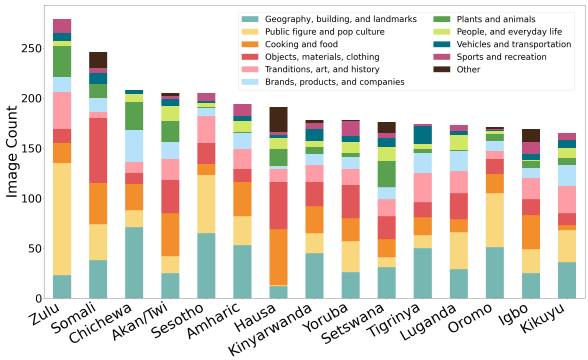


Figure 2: Language-wise distribution of the categories.

Main Annotation Phase This was the main data collection stage, during which the remaining 450 items were collected per language. This was done by the annotators selected in the first phase. To ensure quality in the large volume of submissions, we involved *language coordinators*, experienced native speakers with strong linguistic and cultural knowledge for each language. They were tasked with reviewing all submissions for linguistic accuracy, cultural appropriateness, adherence to guidelines, and audio quality. We provided review guidelines in Appendix F. *Language coordinators* are co-authors of this paper and played a major role in the quality assurance process. After they approved the submissions, the project team conducted a final review to ensure overall data quality.

3.2 Dataset Composition

Each data point in Afri-MCQA includes an image and a set of carefully constructed multiple-choice questions in both text and speech modalities, both in English and the native-region language. Below, we describe each of these aspects.

Image / Category Selection Images provide the visual context from which models are expected to answer questions. To build the image set for Afri-MCQA , we encouraged annotators to contribute their own images whenever possible. However, we also permitted the use of open-license images from websites we provided when self-sourcing was not feasible. All collected images were categorized according to the 10 classes defined by Romero et al. (2024). Figure 2 presents these categories together with their distribution across languages.

Question & Answer Generation For each image, annotators are asked to write up to three multiple-choice QA triplets (question + 1 correct answer + 3 distractors) in both the local language and English. To create a strong benchmark dataset, we requested annotators to create more complex questions that require reasoning (see annotation guidelines in Appendix D).

Audio Recording To investigate the *spoken* language understanding capabilities of the models, we instructed the annotators to record audio for each question and its corresponding answers, clearly reading them aloud in both the local language and English. Therefore, Afri-MCQA includes audio recordings for both questions and answers (in both local and accented English languages).

4 Experiment setup

Models We evaluate a variety of multimodal open-weight models for a comprehensive evaluation of Afri-MCQA. Specifically, we selected Qwen 2.5-Omni(3 &7) B (Jin Xu, 2025), Gemma (12 & 27) B-it (Team et al., 2025a) and Gemma-3n-(2 & 4) B-it (Team et al., 2025a). We additionally include Gemini-2.5 Pro (Comanici et al., 2025), a state-of-the-art closed-weights model, as a strong baseline for performance comparison.

Query Modality We evaluate models by querying them across two input modalities: *text* and *audio*. For text evaluation, models receive the written version of the question. For audio evaluation, we use native speaker recordings of the same questions and options, allowing us to assess how well current MLLMs handle spoken language inputs in African languages.

Query Languages For each query modality, each question is presented in both its *native language* and *English*. This setup allows us to compare model behavior between English and native languages.

Visual Question Answer (VQA) Evaluation Formats We evaluate models across two modalities (text and audio) using two distinct formats for each, enabling a comprehensive assessment of multimodal understanding capabilities.

Text-based VQA: We use two evaluation formats: (1) *Multiple-Choice VQA (MC-VQA)*: Models select the correct answer from four options (A−D), providing direct accuracy measurement through option selection. (2) *Open-ended VQA*: Models generate answers without access to multiple-choice options, testing their ability to retrieve and reason about cultural knowledge without

structural cues.

Audio-based VQA: We similarly evaluate using two formats: (1) *Audio Multiple-Choice VQA (Audio MC-VQA)*: Models receive spoken audio containing the question followed by all four options and must select the correct option (AD), evaluating their ability to process African-accented English and native language speech. (2)*Audio Answer Generation*: Models receive the complete audio (question and all options) but must generate the actual answer text rather than selecting an option letter, requiring both speech comprehension and answer production capabilities.

This allows us to compare performance across modalities (text vs. audio) and response formats (selection vs. generation), revealing how models handle different input types and whether they can understand African-accented English and native African languages across various task formulations.

Prompt For each QA format, we use the same prompt templates to ensure consistency across models and languages. We evaluated models under three distinct prompt conditions to assess the impact of visual and contextual grounding:

- **Text-only:** `Question: {q} Options: {opts}`
- **Image-grounded:** `[Image] Question: {q} Options: {opts}`
- **Image + Location:** `[Image] Location: {country}. Question: {q} Options: {opts}`

In the main paper, we report results for the location-aware setup as our primary evaluation; comprehensive results for Text-only and Image-grounded are provided in Appendix A.

Evaluation metrics We evaluate models in a zero-shot setting using automatic metrics for both Multiple-Choice and Open QA questions. For the latter, we additionally conduct human evaluations on a subset of samples. The following paragraphs detail our evaluation metrics.

(I) Automatic metrics We report the accuracy score for MC-VQA, whereas for Open-ended VQA, we use LLM as a judge (we also report chrf++ (Popović, 2015) score in the Appendix A as additional metrics for Open-ended VQA). We use GPT-4o-mini (Hurst et al., 2024) as a judge for LLMs-as-judge evaluation, where each model prediction is compared against the reference and labeled under a two-tier schema: *Match* (semantically equivalent to the gold answer), or *Wrong* (factually incorrect or irrelevant). Only responses labeled as *Match* are considered valid responses to compute accuracy.

(II) Human Evaluation. Given the cultural and multilingual nature of Afri-MCQA , we additionally conducted human evaluations on 50 randomly sampled questions per language. These were conducted on the best-performing model in each family for Open-ended VQA with Location Aware prompt.

Each language was assigned one rater, a bilingual native speaker, under the assumption that the task requires cultural familiarity. Raters were presented with the image, the question, nine answer candidates, and the ground-truth answer (explicitly marked). They were instructed to select all answers that were either exact matches to the gold or valid alternative correct answers. We report human accuracy as the fraction of gold answers correctly identified, and compute correlations with automated scores on the same subset to assess alignment.

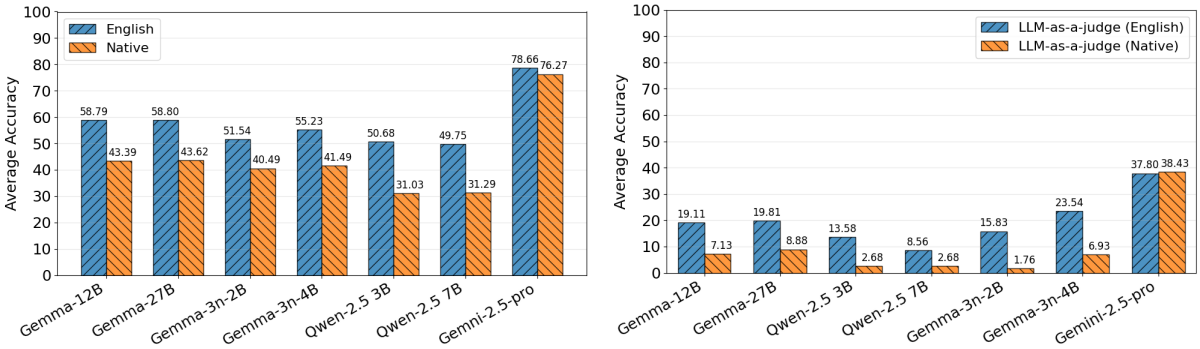
5 Results

We present our results by modality, beginning with the text-based tasks and later extending to audio-based evaluations. For both, we report results for Multiple-Choice (MC-VQA) and Open-ended VQA in English and native languages.

5.1 Text-based QA

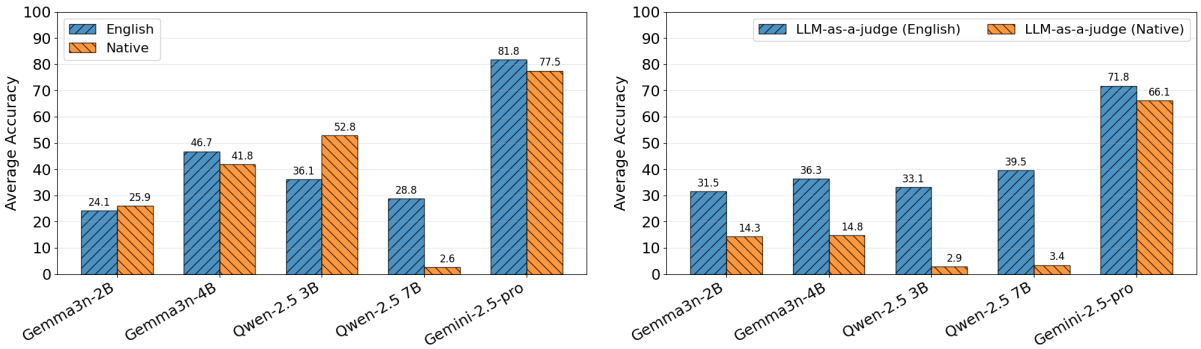
In Figure 3, we report results for seven multimodal LLMs evaluated in both English and native African languages: accuracy for MC-VQA(3(a)) and LLM-as-a-judge scores(3(b)) for open-ended generation, both using location-aware prompts.

As expected, we find that open-weight models consistently perform better when the question is in English than in the native language. For MC-VQA, we observe a performance gap of 11-15% for Gemma model family and 18-19% for Qwen's. Gemini-2.5 Pro outperformed all open-weight models by a significant margin in both native and English QA, while showing a much smaller performance gap (2.5%) between them. We found that increasing the open-weight models' parameter count does not necessarily translate to a higher accuracy in low-resource languages. For example, when we compare `gemma-27B` shows no meaningful improvement over its smaller `gemma-12B` in both En-



(a) MC-VQA(Text) (b) Text-based Open-end VQA

Figure 3: Average MC-VQA and Open-QA accuracies for seven multimodal LLMs evaluated in English and in native African languages for Text



(a) MC-VQA(Audio) (b) Audio Answer Generation

Figure 4: **Average MC-VQA and Answer Generation.** Accuracy for five multimodal LLMs evaluated in English and in native African languages for Audio

glish and native QA.

While MC-VQA shows moderate performance differences between languages, Open-Ended VQA, shown in Figure 3(b), demonstrates much wider gaps between the query languages. English outputs substantially outperform native language outputs across both evaluation metrics for all open-weight models. For instance, Gemma-3n-4B model yielded an average score of 23.5 for English but only 6.9 score for native language, while Gemma-27B showed scores of 19.8 and 8.9 respectively. Interestingly, we found that Gemini-2.5-Pro performed slightly better in the native language with a much smaller gap between both languages.

Human evaluations on Open-Ended VQA confirm these trends (Figure 5). Conducted on a random subset of 50 samples across eight languages, they show that all open-weight models perform better in English, with some smaller models even achieving near-zero accuracy in native languages. Consistent with the automatic evaluations, Gemini

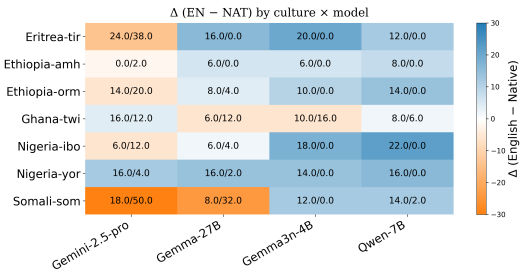


Figure 5: **Human Evaluation for Text Open-ended VQA.** Accuracy across best-performing models for English and Native. We observed that, while most models perform best in the English setting, Gemini 2.5 Pro seems to perform better in the Native language.

2.5 Pro again performed better when queried in native languages compared to English.

5.2 Audio-based QA

Figure 4 presents the performance of five multimodal LLMs on audio MC-VQA (4(a)) and audio answer generation (4(b)) tasks, evaluated in both

accented African English and native African languages using location-aware prompts. As shown in Figure 4(a), Gemini 2.5 Pro performs 81.8% accuracy on English audio and 77.5% on native languages, demonstrating a 4.3% point gap. This minimal performance gap indicates robust multilingual audio processing capabilities and fair representation of African languages in the training data. Among open-source models, Gemma-3n-4B achieves the highest English accuracy at 46.7%, maintaining relatively stable performance at 41.8% in native languages (4.9% gap). Qwen-2.5-Omni 7B demonstrates 28.8% English accuracy with a huge drop in native language to 2.7% native (91.4% gap). We have noticed in most cases Qwen-2.5-Omni 7B selected option B as the correct answer for all the questions; this indicates the absence of African language audio training data.

For audio answer generation, Gemini 2.5 Pro achieves 71.8% on accented English audio and 66.1% on native languages, demonstrating 7.9% gap. This result indicates the models multilingual capability in audio-to-text generation in low-resource languages. Open-source models show significant performance gaps for the generation task. For instance, Qwen-2.5-Omni-7B achieves an LLM-as-a-judge accuracy of 39.5% on accented English audio, but only 3.4% on native audio, indicating a significant drop in performance for non-English inputs, which suggests the absence of African language audio in the training corpora.

6 Discussion

How well do MLLMs *know and understand* African cultural contexts in visually-grounded QA tasks? Across text-based MC-VQA, the closed model (Gemini-2.5 Pro) performs significantly better than open-weight systems, reaching near 80% accuracy while the best open models hover at or below 60%. These results show that current open models are, at best, average on this benchmark (far from parity with the closed model). Open-ended VQA further exposes this gap: Gemini attains higher scores than open models, which often fall to single-digit accuracy in native languages. These results highlight the *need to democratize performance in open-weight systems to match proprietary ones in African cultural understanding*.

How does the query language affect the model's response on visual cultural knowledge? We find that query language effects are model-dependent. For open-weight models, English consistently outperforms the native languages on both MC-VQA and open-ended tasks, suggesting that their cultural knowledge may be more readily accessible in English, even for models advertised as multilingual. On the other hand, Gemini shows a minimal English-Native gap in MC-VQA and even a slight advantage for Native in open-ended VQA, a pattern supported by both automatic and human evaluation on a subset. This finding is interesting, as it may indicate that *querying visual cultural knowledge through native languages may be more effective than English when the MLLMs have sufficient language fluency and cultural knowledge*.

What is the current performance gap in state-of-the-art MMLMs when queried with text vs. speech in African languages? We observe a significant performance gap when processing African languages in text versus speech formats between the closed-source Gemini 2.5 Pro and various open-source models. While Gemini 2.5 Pro maintains stable performance between text and audio on native languages, open-source models experience severe degradation, with some collapsing almost entirely on spoken input. This gap is clearly noted in audio answer generation, where models fail to reproduce the correct answer even after hearing it, reflecting limitations in both speech processing and cross-modal transfer. The difference between Geminis and the open-source models underscores that *text-based multilingual competence does not reliably extend to speech understanding, leaving many African language communities where communication is primarily oral, unsupported by current open-source MLLMs*.

How does accuracy differ when evaluated with Multiple-Choice QA compared to Open QA? Multiple-choice substantially overestimates performance relative to open-ended generation across models and languages, confirming that recognizing the correct answer among options is much easier than generating culturally precise answers. Even Gemini's strong MC-VQA accuracy drops to modest open-ended scores, while open models degrade much more, sometimes scoring near-zero accuracy in native languages. We hypothesize that these results are not only due to the lack of options in the model input but also to language generation limitations, where smaller open-weight models struggle to generate responses in the relatively low-resource languages covered in Afri-MCQA. Consequently,

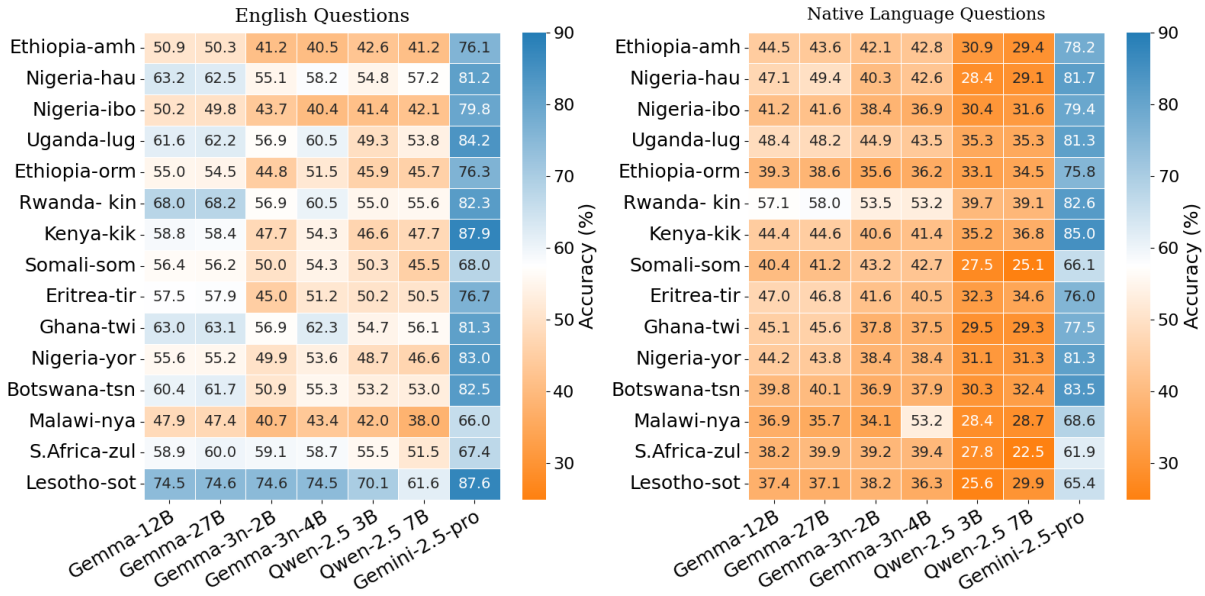


Figure 6: Language-wise accuracy of seven multimodal LLMs on the text-based VQA task, evaluated separately on English and native African language questions.

MC-VQA overestimates actual cultural knowledge, whereas open-ended evaluation offers a stricter and more informative test, but may also be limited by the multilingual language generation abilities of tested models.

7 Conclusion

In this work, we introduced Afri-MCQA, a multilingual and multimodal benchmark designed to assess MLLMs understanding of culturally-relevant African visual contexts. The dataset contains over 7.5k image-grounded Q&A pairs in 15 languages from 12 countries, covering both text and speech inputs. All data were human-annotated by native speakers through a multi-step quality pipeline, including annotator training, review rounds by language coordinators, and final verification for linguistic and cultural accuracy. Through these steps, we ensured that Afri-MCQA provides a trustworthy and representative resource for evaluating multimodal reasoning in underrepresented languages from sub-Saharan Africa.

Our evaluation shows several takeaways. The evaluated closed-weight model (Gemini-2.5 Pro) consistently outperformed open-weight models, achieving up to 79% accuracy on MC-VQA compared to below 60% for the best open systems. Interestingly, performance across query languages is also dependent on models: open-weight systems perform best in English, while the closed model maintains or even improves accuracy in native languages. Finally, open-ended QA remains significantly harder than multiple-choice, possibly due to both the lack of hints provided as multiple-choice options or a bottleneck in the language generation abilities of the models in low-resource languages.

Overall, we highlight the need to narrow the performance gap between open and closed models. We hope Afri-MCQA provides both a benchmark and a foundation for building more inclusive, culturally aware multimodal systems that better represent African languages and cultures.

8 Limitation

We believe Afri-MCQA represents an important step toward more inclusive evaluation by foregrounding African languages and cultural contexts that have long been overlooked in existing benchmarks. Although the dataset spans 15 languages across 12 countries, Africa is home to thousands of languages and cultural groups, many of which remain unrepresented. Furthermore, while our question categories aim to reflect culturally grounded knowledge, culture itself is fluid, subjective, and deeply contextual. Our formulation inevitably abstracts away from finer-grained variations such as regional, generational, or community-specific differences that shape cultural understanding. As with most human-curated datasets, potential biases in data collection remain. Annotators backgrounds and interpretations of cultural relevance may influence the formulation of questions or the selection

of images. Additionally, due to computational and financial constraints, we evaluate only a limited set of open- and closed-source models, so the reported performance gaps may not fully capture the broader landscape. Finally, Afri-MCQA is a primary, human-curated dataset created without the involvement of LLMs and with minimal reliance on web-sourced images. Because this process is inherently time-consuming, the dataset is of moderate size and intended as an evaluation benchmark rather than a pretraining or fine-tuning resource, for which it would likely cause overfitting.

9 Ethical Considerations

Our work involves the collection of culturally grounded questionanswer pairs in 15 African languages, annotated, spoken and reviewed by native speakers. All annotators participated voluntarily and were compensated fairly for their work in accordance with local wage standards on the Upwork platform. Before beginning annotation, contributors were informed about the goals of the project, the intended use of the dataset for research and evaluation purposes, and their right to withdraw from participation at any stage.

We took several steps to ensure cultural sensitivity and respect throughout the data creation process. Question formulation guidelines were designed to avoid harmful stereotypes, offensive content, or culturally inappropriate framing. All annotations were reviewed by language coordinators who are themselves native speakers to check for accuracy, contextual appropriateness, and respectful representation. Despite these efforts, we acknowledge that culture is deeply complex and subjective, and that our dataset may still reflect certain biases or oversimplifications.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

Ife Adebara and Muhammad Abdul-Mageed. 2022. Towards afrocentric nlp for african languages: Where we are and where we can go. *arXiv preprint arXiv:2203.08351*.

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. 2024a. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta. Association for Computational Linguistics.

David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure FP Dossou, Akintunde Oladipo, Doreen Nixdorf, and 1 others. 2023. Masakhanews: News topic classification for african languages. *arXiv preprint arXiv:2304.09972*.

David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba O Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, and 1 others. 2024b. Irokobench: A new benchmark for african languages in the age of large language models. *arXiv preprint arXiv:2406.03368*.

Israel Abebe Azime, Atnafu Lambebo Tonja, Tadesse Destaw Belay, Yonas Chanie, Bontu Fufa Balcha, Negasi Haile Abadi, Henok Biadglign Ademtew, Mulubrhan Abebe Nerea, Debela Desalegn Yadeta, Derartu Dagne Geremew, Assefa Atsbiha Tesfu, Philipp Slusallek, Thamar Solorio, and Dietrich Klakow. 2025. ProverbEval: Exploring LLM evaluation challenges for low-resource language understanding. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6250–6266, Albuquerque, New Mexico. Association for Computational Linguistics.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Harald Hammarström. 2018. A survey of african languages.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Jinzheng He Hangrui Hu Ting He Shuai Bai Keqin Chen Jialin Wang Yang Fan Kai Dang Bin Zhang Xiong Wang Yunfei Chu Junyang Lin Jin Xu, Zhifang Guo. 2025. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond

Elliott. 2021. Visually grounded reasoning across languages and cultures. *arXiv preprint arXiv:2109.13238*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Rada Mihalcea, Oana Ignat, Longju Bai, Angana Borah, Luis Chiruzzo, Zhijing Jin, Claude Kwizera, Joan Nwatu, Soujanya Poria, and Thamar Solorio. 2024. Why ai is weird and should not be this way: Towards ai for everyone, with everyone, by everyone. *Preprint*, arXiv:2410.16315.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, David Ifeoluwa Adelani, Ibrahim Said Ahmad, Saminu Mohammad Aliyu, Paul Röttger, Abigail Oppong, Andiswa Bukula, Chiamaka Ijeoma Chukwuneke, Ebrahim Chekol Jibril, Elyas Abdi Ismail, Esubalew Alemneh, Hagos Tesfahun Gebremichael, Lukman Jibril Aliyu, Meriem Beloucif, Oumaima Hourrane, Rooweither Mabuya, Salomey Osei, and 8 others. 2025. AfriHate: A multilingual collection of hate speech and abusive language datasets for African languages. 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 1854–1871, Albuquerque, New Mexico. Association for Computational Linguistics.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Van Steenkiste, Lisa Anne Hendricks, Aishwarya Agrawal, and 1 others. 2024. Benchmarking vision language models for cultural understanding. *arXiv preprint arXiv:2407.10920*.

James Cross Onur Çelebi Maha Elbayad Kenneth Heafield Kevin Heffernan Elahe Kalbassi Janice Lam Daniel Licht Jean Maillard Anna Sun Skyler Wang Guillaume Wenzek Al Youngblood Bapi Akula Loic Barrault Gabriel Mejia Gonzalez Prangthip Hansanti John Hoffman Semarley Jarrett Kaushik Ram Sadagopan Dirk Rowe Shannon Spruit Chau Tran Pierre Andrews Necip Fazil Ayan Shruti Bhosale Sergey Edunov Angela Fan Cynthia Gao Vedanuj Goswami Francisco Guzmán Philipp Koehn Alexandre Mourachko Christophe Ropers Safiyyah Saleem Holger Schwenk Jeff Wang NLLB Team, Marta R. Costa-jussà. 2022. No language left behind: Scaling human-centered machine translation.

Derek Nurse and Gerard Philippson, editors. 2006. *The Bantu Languages*. Routledge Language Family Series. Routledge, London, England.

Odunayo Ogundepo, Tajuddeen R Gwadabe, Clara E Rivera, Jonathan H Clark, Sebastian Ruder, David Ifeoluwa Adelani, Bonaventure FP Dossou, Abdou Aziz Diop, Claytone Sikasote, Gilles Hacheme, and 1 others. 2023. Afriqa: Cross-lingual open-retrieval question answering for african languages. *arXiv preprint arXiv:2305.06897*.

Jessica Ojo, Odunayo Ogundepo, Akintunde Oladipo, Kelechi Ogueji, Jimmy Lin, Pontus Stenetorp, and David Ifeoluwa Adelani. 2025. Afrobench: how good are large language models on african languages? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19048–19095.

Shantipriya Parida, Idris Abdulmumin, Shamsuddeen Hassan Muhammad, Aneesh Bose, Guneet Singh Kohli, Ibrahim Said Ahmad, Ketan Kotwal, Sayan Deb Sarkar, Ondřej Bojar, and Habeebah Adamu Kakudi. 2023. Havqa: A dataset for visual question answering and multimodal research in hausa language. *arXiv preprint arXiv:2305.17690*.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, and 1 others. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *arXiv preprint arXiv:2406.05967*.

Florian Schneider and Sunayana Sitaram. 2024. M5–a diverse benchmark to assess the performance of large multimodal models across multilingual and multicultural vision-language tasks. *arXiv preprint arXiv:2407.03791*.

Belay Simane, Thandi Kapwata, Natasha Naidoo, Guéladio Cissé, Caradee Y. Wright, and Kiros Berhane. 2025. Ensuring africas food security by 2050: The role of population growth, climate-resilient strategies, and putative pathways to resilience. *Foods*, 14(2):262.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, and 1 others. 2024. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *arXiv preprint arXiv:2412.03304*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025a. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025b. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar

Thawakar, Henok Biadglign Ademtew, Yahya Hmaiti, Amandeep Kumar, Kartik Kukreja, and 1 others. 2025a. All languages matter: Evaluating lmms on culturally diverse 100 languages. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19565–19575.

Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, Henok Biadglign Ademtew, Yahya Hmaiti, Amandeep Kumar, Kartik Kukreja, and 1 others. 2025b. All languages matter: Evaluating lmms on culturally diverse 100 languages. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19565–19575.

Genta Indra Winata, Frederikus Hudi, Patrick Amadeus Irawan, David Anugraha, Rifki Afina Putri, Wang Yutong, Adam Nohejl, Ubaidillah Ariq Prathama, Nedjma Ousidhoum, Afifa Amriani, Anar Rzayev, Anirban Das, Ashmari Pramodya, Aulia Adila, Bryan Wilie, Candy Olivia Mawalim, Cheng Ching Lam, Daud Abolade, Emmanuele Chersoni, and 32 others. 2025. WorldCuisines: A massive-scale benchmark for multilingual and multicultural visual question answering on global cuisines. 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 3242–3264, Albuquerque, New Mexico. Association for Computational Linguistics.

Hao Yu, Jesujoba Oluwadara Alabi, Andiswa Bukula, Jian Yun Zhuang, En-Shiun Annie Lee, Tadesse Kebede Guge, Israel Abebe Azime, Happy Buzaaba, Blessing Kudzaishe Sibanda, Godson Koffi Kalipe, Jonathan Mukiibi, Salomon Kabongo Kabenamualu, Mmasibidi Setaka, Lolwethu Ndolela, Nkiruka Odu, Rooweither Mabuya, Shamsuddeen Hassan Muhammad, Salomey Osei, Sokhar Samb, and 2 others. 2025. INJONGO: A multicultural intent detection and slot-filling dataset for 16 African languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9429–9452, Vienna, Austria. Association for Computational Linguistics.

A Additional Experiments results

Text-based QA - here we report additional results for different prompts and per country-language results.

A.1 MC-VQA results

Language	G-12B	G-27B	G-3n-2B	G-3n-4B	Q-2.5 3B	Q-2.5 7B	Gemni-2.5-pro
Ethiopia-amh	50.34	50.51	42.09	44.5	42.09	40.72	72.23
Nigeria-hau	64.92	65.44	56.36	57.06	57.59	58.87	77.31
Nigeria-ibo	48.43	47.05	44.11	47.64	37.64	40.98	77.45
Uganda-lug	61.65	60.88	56.84	58.76	49.51	53.56	80.73
Ethiopia-orm	52.32	50.58	46.51	50.96	44.18	44.37	77.32
Rwanda-kin	68.53	67.6	57.86	60.48	54.86	56.92	78.27
Kenya-kik	56.56	55.75	49.09	54.14	42.02	47.47	86.46
Somali-som	56.36	57.99	50.81	54.06	49.86	46.06	64.09
Eritrea-tir	57.01	56.49	41.01	47.89	48.94	50	73.15
Ghana-twi	63.31	63.15	59.01	63.95	55.34	56.45	81.49
Nigeria-yor	55.8	57.11	52.14	53.44	47.37	47	82.2
Botswana-tsn	61.74	61.17	53.21	55.11	51.7	52.46	80.8
Malawi-nya	46.15	46.79	43.75	45.19	42.14	37.82	65.54
S.Africa-zul	64.63	66.06	65.23	64.99	48.5	53.52	88.76
Lesotho-sot	75.28	74.95	75.28	74.63	53.82	60.16	92.5
Average	58.86	58.76	53.28	55.52	48.37	49.75	78.56

Table 3: Prompt - Image only–English(text-based)

Language	G-12B	G-27B	G-3n-2B	G-3n-4B	Q-2.5 3B	Q-2.5 7B	Gemni-2.5-pro
Ethiopia-amh	43.81	41.76	43.29	43.47	29.73	29.89	76.8
Nigeria-hau	46.94	46.77	43.45	44.5	30.89	32.16	81.32
Nigeria-ibo	40.58	41.76	38.43	37.64	30	32.94	79.01
Uganda-lug	50.86	50.48	45.08	44.31	35.07	36.22	81.11
Ethiopia-orm	39.53	38.56	34.68	37.2	30.04	31	76.55
Rwanda-kin	56.92	57.67	55.24	55.24	34.83	37.82	80.71
Kenya-kik	45.85	45.45	40.6	41.61	33.74	37.17	84.84
Somali-som	41.19	40.78	43.9	45.12	27.77	24.93	65.62
Eritrea-tir	47.01	47.01	39.82	41.05	28.07	33.5	73.28
Ghana-twi	45.45	45.61	36.52	31.57	29.18	29.66	75.91
Nigeria-yor	44.19	44	39.29	30.71	32.02	32.95	81.23
Botswana-tsn	42.04	40.34	38.06	31.43	28.4	31.81	81.06
Malawi-nya	38.94	37.17	36.21	35.09	28.68	28.84	70.19
S.Africa-zul	39.66	39.54	39.78	39.3	29.39	22.46	40
Lesotho-sot	37.07	37.72	38.37	37.2	28.34	30.56	47.64
Average	44. 00	43.64	41.05	41.14	30.27	31.46	73.01

Table 4: Image only-native (text-based)

A.2 Open-ended VQA results

Language	G-12B	G-27B	G-3n-2B	G-3n-4B	Q-2.5 3B	Q-2.5 7B	Gemni-2.5-pro
Ethiopia-amh	40.37	40.72	34.53	37.28	26.63	33.67	44.5
Nigeria-hau	46.59	47.46	45.2	45.02	32.11	33.2	52.53
Nigeria-ibo	40.1	39.41	36.47	41.96	26.86	30.98	44.11
Uganda-lug	47.01	46.24	41.81	44.12	32.56	39.69	53.37
Ethiopia-orm	40.11	40.31	38.37	42.82	34.3	33.72	42.24
Rwanda-kin	52.43	52.05	48.31	47.56	43.07	48.12	57.67
Kenya-kik (ki)	46.06	45.65	37.97	40.6	31.51	34.74	58.79
Somali-som(so)	40.51	41.05	38.01	37.94	22.22	34.68	43.49
Eritrea-tir(ti)	46.84	44.91	40.35	44.73	30.52	35.96	54.38
Ghana-twi	44.01	44.65	41.62	40.35	32.21	36.68	49.44
Nigeria-yor (yo)	40.63	39.88	41.34	39.47	30.52	39.13	44.75
Botswana-tsn(tn)	42.8	43.75	33.52	37.87	36.36	39.55	49.81
Malawi-nya	32.85	32.21	34.13	37.01	27.56	35.41	39.1
S.Africa-zul	31.35	31.79	32.5	32.87	48.5	48.69	47.25
Lesotho-sot	34.35	65.02	36.76	36.87	53.32	54.8	55
Average	41.74	41.67	38.72	40.43	33.91	38.40	49.09

Table 5: No Image–English (text-based)

Language	G-12B	G-27B	G-3n-2B	G-3n-4B	Q-2.5 3B	Q-2.5 7B	Gemni-2.5-pro
Ethiopia-amh	35.05	35.56	32.98	36.76	27.49	25.77	54.12
Nigeria-hau	31.17	37.52	34.2	42.58	25.3	26.89	50.75
Nigeria-ibo	34.9	36.27	33.72	36.27	27.25	31.96	52.94
Uganda-lug	33.91	34.29	35.83	37.37	32.36	32.36	52.4
Ethiopia-orm	31.97	31.58	26.68	32.17	29.84	29.84	50.38
Rwanda-kin	43.44	43.82	38.76	42.32	31.08	32.58	60.48
Kenya-kik (ki)	33.13	32.72	30.1	28.48	29.29	30.3	55.55
Somali-som(so)	28.72	29.26	28.04	33.46	19.24	29.67	47.69
Eritrea-tir(ti)	38.25	37.84	35.26	38.94	31.4	34.03	58.77
Ghana-twi	32.85	32.85	30.3	32.37	26.79	29.98	45.93
Nigeria-yor (yo)	32.39	34.64	30.14	34.07	28.08	32.29	52.05
Botswana-tsn(tn)	32.95	32.38	31.62	34.46	29.73	28.4	54.16
Malawi-nya	34.45	33.97	32.05	32.85	23.71	29	46.47
S.Africa-zul	30.94	31	31.66	31.78	25.2	25.89	40.02
Lesotho-sot	32.84	32.9	32.23	32.23	19.23	23.5	47.64
Average	33.79	34.44	32.32	35.07	27.06	29.50	51.29

Table 6: No Image–native(text-based)

Language	G-12B		G-27B		Q-2.5 3B		Q-2.5 7B		G-3n-2B		G-3n-4B		Gemni	
	chrf++	LLM	chrf++	LLM	chrf++	LLM	chrf++	LLM	chrf++	LLM	chrf++	LLM	chrf++	LLM
Ethiopia-amh	27	26.2	27.55	26.8	17.69	15	20.7	19.2	20.73	16.4	31.53	30.4	47.88	60.2
Nigeria-hau	7.28	12.63	7.47	13.43	7.73	7.41	7.7	7.82	10.82	8.45	13.43	11.02	14.63	25.5
Nigeria-ibo	10.41	10.54	11.04	12.13	10.31	10.14	11.8	8.55	9.41	7.75	15.65	15.9	25.61	36.18
Uganda-lug	20.64	25.45	22.37	27.04	15.38	16.1	17.51	18.29	17.1	18.89	27.91	33.2	35.7	44.33
Ethiopia-orm	18.44	17.19	19.24	19.34	16.86	11.72	20.08	17.58	19.2	16.41	27.38	25	34.01	38.09
Rwanda-kin	12.51	18.76	12.06	18.36	11.37	10.18	12.51	12.38	10.82	16.38	16.37	22.92	18.92	36.53
Kenya-kik (ki)	17.77	21.82	18.2	23.64	14.42	11.31	16.54	12.12	16.73	15.15	30.33	31.11	41.2	49.49
Somali-som(so)	13.77	15.94	13.4	16.93	12.59	13.55	15.03	17.73	13.55	16.33	17.04	18.73	18.62	25.5
Eritrea-tir(ti)	20.11	28.75	20.47	29.29	14.82	19.17	17.87	22.24	14.64	19.23	24.22	30.02	35.75	49.01
Ghana-twi	20.37	29.05	21.01	29.24	16.83	23.46	18.87	23.09	19.41	24.58	25.71	32.4	25.71	34.64
Nigeria-yor (yo)	10.2	17.27	9.38	17.87	9.13	13.86	9.92	12.25	10.12	15	13.13	20.6	20.61	43.98
Botswana-tsn(tn)	16.14	20.96	15.16	19.66	14.93	17.37	16.34	20.56	15.55	18.56	20.59	27.94	22.98	28.14
Malawi-nya	16.41	15.17	15.56	15.57	11.08	17.37	13.39	12.77	13.42	18.56	20.72	22.75	25.24	30.34
S.Africa-zul	18.48	15.72	17.94	16.1	11.48	8.7	12.05	9.28	17.49	15.72	18.2	16.67	34.87	39.58
Lesotho-sot	12.56	11.26	14.01	11.82	11.49	8.44	10.69	8.63	13.37	11.44	14.16	12.01	24.7	25.7
Average	16.13	19.14	16.32	19.81	13.07	13.58	14.73	8.56	14.61	15.83	20.75	23.54	28.43	37.80

Table 7: Location/Language Aware-English(Open-endded VQA -text based)

Language	G-12B		G-27B		Q-2.5 3B		Q-2.5 7B		G-3n-2B		G-3n-4B		Gemni	
	chrf++	**LLM**	**chrf++**	**LLM**	**chrf++**	**LLM**	**chrf++**	**LLM**	**chrf++**	**LLM**	**chrf++**	**LLM**	chrf++	LLM
Ethiopia-amh	2.47	22.8	2.91	22.6	0.11	6.6	1.01	4.2	0	1	0.71	20.4	42.17	55.4
Nigeria-hau	4.23	3.01	4.31	3.21	2.79	0.6	2.9	0.6	1.41	2.4	3.83	2.6	15.02	30.26
Nigeria-ibo	5.93	5.96	5.93	5.96	3.65	0.99	4.97	0.99	0.44	0.99	5.42	2.98	24.11	33
Uganda-lug	9.48	9.96	9.22	11.16	5.69	3.19	4.27	3.19	0.5	1.79	8.45	8.76	30.33	40.44
Ethiopia-orm	6.56	6.45	6.35	5.86	5.26	6.05	3.26	1.76	0.68	0.59	6.27	4.69	37.92	39.84
Rwanda-kin	8.31	9.38	8.23	9.58	4.8	3.39	3.65	3.02	0.39	0.2	6.24	7.19	25.27	34.13
Kenya-kik	7.47	3.03	7.82	30.3	5.81	1.21	4.03	0.61	1.84	1.21	9.31	3.03	39.17	47.37
Somali-som	7.85	7.57	7.9	7.37	3.72	1.39	4.42	2.19	2.17	1.99	5.64	4.78	22.59	30.48
Eritrea-tir	0.85	3.98	0.95	4.34	0.04	5.79	0.43	5.79	0.01	0.18	0.13	14.1	27.23	37.43
Ghana-twi	6.19	5.77	6.3	6.89	4.92	2.42	4.43	2.61	1.88	0.74	6.89	5.21	31.63	33.89
Nigeria-yor	3.91	6.43	3.72	5.62	2.25	1	2.3	1.2	1.22	0.6	4.3	7.8	22.33	46.18
Botswana-tsn	7.91	4.39	7.11	4.19	4.4	2.79	4.15	2	2.04	0.6	7.26	5.99	34.41	39.75
Malawi-nya	8.99	5.19	8.44	4.99	4.23	0.4	3.43	0.8	1.71	0.4	9.33	3.59	31.5	34.13
S.Africa-zul	13.54	10.04	12.72	8.33	4.68	1.33	5.21	4.88	14.32	10.42	14.75	10.42	41.13	44.51
Lesotho-sot	7.6	3	7.15	2.81	4.73	3	8.08	4.88	7.77	3.38	8	3	28.67	29.83
Average	6.75	7.13	6.60	6.88	3.80	2.67	3.73	2.68	2.42	1.76	6.43	6.93	30.22	38.43

Table 8: Location/Language Aware-native(Open-ended VQA -text based)

Language	G-12B		G-27B		Q-2.5 3B		Q-2.5 7B		G-3n-2B		G-3n-4B		Gemni	
	chrf++	**LLM**	**chrf++**	**LLM**	**chrf++**	**LLM**	**chrf++**	**LLM**	**chrf++**	**LLM**	**chrf++**	**LLM**	chrf++	LLM
Ethiopia-amh	19.86	15.8	20.04	15.8	10.38	6.6	11.76	9.2	15.42	10	24.62	20.8	40.41	43.8
Nigeria-hau	6.28	12.63	6.53	11.62	7.98	6.8	7.98	7.01	7.36	8.02	7.35	12.02	9.37	16.83
Nigeria-ibo	10.57	8.55	9.95	8.15	9.78	3.76	10.5	6.76	8.97	5.96	13.27	11.13	24.63	30.82
Uganda-lug	20.59	24.25	19.68	22.47	13.85	13.92	15.89	15.9	16.43	18.49	27.36	30.2	49.04	48.21
Ethiopia-orm	18.79	15.04	19.01	14.45	12.29	6.84	14.68	11.33	14.08	8.01	23.32	19.14	37.92	36.52
Rwanda-kin	11.41	15.97	11.22	15.37	10.29	7.39	11.38	10.18	9.41	10.78	15.15	18.16	23.07	32.93
Kenya-kik	18.55	20.81	18.2	20.2	12.03	7.68	15.97	11.11	15.12	13.94	28.99	28.48	49.78	59.19
Somali-som	14.79	18.53	15.4	19.12	8.03	14.94	12.45	17.13	12.97	15.34	18.04	21.71	26.79	32.07
Eritrea-tir	14.71	16.09	14.64	15.19	8	6.87	9.84	7.59	10.71	7.78	18.09	12.84	31.98	40.14
Ghana-twi	21.95	29.8	20.66	29.24	8.85	18.44	15.78	21.23	26.57	24.95	25.57	33.33	36.76	47.87
Nigeria-yor	10.87	17.87	10.61	17.47	7.02	11.45	8.89	11.24	9.25	14.4	13.48	21	23.2	45.38
Botswana-tsn	15.75	22.95	16.14	22.75	9.27	17.76	14.95	22.16	13.77	18.36	19.43	25.35	30.27	39.72
Malawi-nya	12.25	9.78	12.97	11.38	8.7	6.19	9.88	6.19	15.33	8.18	15.33	12.77	26.35	26.35
S.Africa-zul	16.12	13.25	15.77	11.93	10.61	7.1	10.7	8.52	15.65	12.69	16.36	13.07	40.02	40.34
Lesotho-sot	11.99	9.19	11.75	9.57	10.59	6.94	10.55	9.01	12.71	9.94	12.33	9.38	23.31	23.83
Average	14.97	16.7	14.84	16.31	9.84	9.51	12.08	11.64	13.58	12.46	19.1	19.29	31.53	37.6

Table 9: Image only-english(Open-endded VQA -text based)

Language	G-12B		G-27B		Q-2.5 3B		Q-2.5 7B		G-3n-2B		G-3n-4B		Gemni	
	chrf++	**LLM**	**chrf++**	**LLM**	**chrf++**	**LLM**	**chrf++**	**LLM**	**chrf++**	**LLM**	**chrf++**	**LLM**	chrf++	LLM
Ethiopia-amh	4.65	17.2	3.78	17.6	3.5	3.2	0.68	2.4	1.51	14.6	1.21	13.8	42.59	54.8
Nigeria-hau	3.42	3.01	3.83	2	3.12	3.41	3.12	1.4	3.77	3.81	3.44	3.81	14.67	29.66
Nigeria-ibo	5.21	4.17	5.31	2.98	2.53	2.78	5.32	0.99	5.41	2.58	5.67	2.19	23.2	29.62
Uganda-lug	9.17	10.96	8.14	9.96	4.28	4.38	4.72	4.58	8.1	8.96	7.95	7.57	31.54	39.04
Ethiopia-orm	6.28	7.03	5.94	6.45	2.01	2.15	4.68	2.93	5.67	3.32	5.54	2.73	29.32	38.48
Rwanda-kin	7.66	9.98	7.59	9.78	2.36	2.59	3.98	2.2	6.69	8.58	6.19	7.39	25.51	35.33
Kenya-kik (ki)	7.58	3.23	7.57	3.03	4.53	4.04	6.11	1.62	8.7	2.22	8.9	2.63	39.23	48.08
Somali-som(so)	7.56	8.17	7.3	8.76	3.29	2.39	6.21	3.78	5.73	5.78	5.7	4.98	25.02	32.07
Eritrea-tir(ti)	1.14	3.62	1.21	3.8	0.12	1.63	0.19	0.72	0.74	4.16	0.69	4.7	27.03	35.62
Ghana-twi	6.52	6.15	6.24	7.08	5.03	2.42	5.74	4.28	6.44	4.66	6.21	4.66	29.44	32.22
Nigeria-yor	3.76	6.22	3.94	6.92	2.3	0.6	3.33	0.8	4.3	8.2	4.22	8	20.6	43.73
Botswana-tsn	6.74	4.99	6.48	5.59	2.97	2.99	5.86	8.58	6.6	4.19	6.95	4.59	34.32	36.93
Malawi-nya	8.78	5.99	8.3	1.2	3.31	1	5.36	2.79	8.35	3.79	8.33	3.19	29.83	31.94
S.Africa-zul	14.11	9.66	14.68	10.61	3.44	2.65	3.72	1.52	13.78	10.23	14.06	9.66	40.69	42.8
Lesotho-sot	7.28	3.75	7.76	4.69	3.13	2.25	6.88	10.69	7.55	3.19	7.04	3.19	24.95	25.7
Average	6.66	6.94	6.54	6.7	2.97	2.57	4.39	3.29	6.22	5.88	6.14	5.54	29.20	37.07

Table 10: Image only-native(Open-endded VQA -text based)

Language	G-12B		G-27B		Q-2.5 3B		Q-2.5 7B		G-3n-2B		G-3n-4B	
	chrf++	LLM	chrf++	LLM	chrf++	LLM	chrf++	LLM	chrf++	LLM	chrf++	LLM
Ethiopia-amh	7.17	3.2	7.47	4	7.72	2.2	8.11	3.6	7.21	2.2	8.25	2.8
Nigeria-hau	4.42	2	4.51	2	6.69	2.2	6.69	2.2	5.07	2.2	6.16	2.61
Nigeria-ibo	7.2	4.37	7.07	3.78	8.56	3.78	8.92	1.19	6.66	1.39	8.07	4.17
Uganda-lug	11.23	9.15	11.73	9.74	9.76	4.58	10.15	5.57	10.7	7.55	13.3	9.94
Ethiopia-orm	9.84	5.66	9.84	5.86	9.59	3.32	9.92	4.3	9.8	3.91	11.46	5.86
Rwanda-kin	7.42	6.19	7.35	5.39	8.75	3.59	8.61	5.39	6.8	4.39	8.92	4.99
Kenya-kik	9.11	4.44	8.2	4.24	9.51	2.42	11.47	2.83	9.88	4.24	12.81	6.26
Somali-som	7.3	3.78	7.4	3.98	8.03	1.2	8.65	2.59	7.13	1.59	9.55	3.59
Eritrea-tir	8.4	4.34	8.97	5.06	8	3.25	10.15	3.62	9.56	3.8	12.16	5.24
Ghana-twi	8.77	6.89	8.59	7.08	8.85	4.47	9.74	5.4	8.54	5.4	11.22	6.89
Nigeria-yor	4.62	2.61	4.83	3.01	7.02	3.21	7.56	3.82	5.28	3.8	6.48	4.2
Botswana-tsn	9.81	9.98	9.88	9.98	9.27	4.39	9.84	6.79	9.03	5.99	12.3	9.18
Malawi-nya	7.17	2.4	7.01	2.79	8.7	1.4	8.44	1.8	7.17	1.4	8.75	2.2
S.Africa-zul	7.47	2.46	6.31	2.27	6.66	1.14	6.91	0.57	6.35	2.27	6.51	2.46
Lesotho-sot	7.47	3.94	7.53	4.32	9.72	4.5	8.83	4.13	7.56	4.13	7.17	3.75
Average	7.76	4.76	7.77	4.9	8.45	3.04	8.93	3.46	7.77	3.61	9.54	4.94

Table 11: No Image-english(Open-endded VQA -text based)

Language	G-12B		G-27B		Q-2.5-3B		Q-2.5-7B		G-3n-2B		G-3n-4B	
	chrf++	LLM	chrf++	LLM	chrf++	LLM	chrf++	LLM	chrf++	LLM	chrf++	LLM
Ethiopia-amh	7.03	10.6	5.46	9.4	0.6	1.9	0.58	1.4	6.75	10.6	5.25	11
Nigeria-hau	3.76	1	3.55	0.8	3.1	1.2	3.1	0.2	5.07	2.6	4.31	1.8
Nigeria-ibo	4.48	2.98	4.14	2.98	4.09	1.12	4.05	0.6	6.66	1.49	4.47	1.59
Uganda-lug	6.81	4.98	6.4	9.74	4.23	2.57	4.69	2.99	10.7	10.7	6.38	3.98
Ethiopia-orm	4.6	7.03	4.84	1.95	6.2	3.6	5.91	2.93	9.8	1.6	5.6	0.78
Rwanda-kin	5.76	4.19	5.73	3.39	3.56	2	3.24	1.8	6.8	3.1	5.75	2
Kenya-kik	3.99	0.61	4.1	0.61	3.26	4.01	3.84	4.04	9.88	1.72	5.1	0.81
Somali-som	5.61	2.79	6.1	3.19	1.78	0.6	15.43	1.59	7.13	0.23	5.36	0.6
Eritrea-tir	1.52	1.81	1.57	1.08	0.61	3.98	0.33	0	9.56	1.73	1.27	2.35
Ghana-twi	3.48	1.12	3.63	1.12	4.18	0	5.99	4.84	8.54	1.21	4.3	1.3
Nigeria-yor	2.37	2.21	2.44	2.6	1.42	0	2.86	1.2	5.28	1.43	2.74	2
Botswana-tsn	5.87	2	5.69	2	3.32	2.4	8	6.79	9.03	2.1	6.84	2
Malawi-nya	5.12	1	5.3	1.2	1.75	0.8	4.39	4.59	7.17	0.5	5.95	0.8
S.Africa-zul	6.21	2.27	6.44	2.27	4.75	2.08	4.81	2.27	6.64	2.27	6.63	2.08
Lesotho-sot	5.83	2.44	5.36	1.13	2.01	1.13	6.09	7.69	5.35	0.94	5.47	1.13
Average	4.82	3.13	4.71	2.89	2.99	1.82	4.22	2.86	7.62	2.26	5.02	2.28

Table 12: No Image-native(Open-endded VQA -text based)

B Audio results

B.1 Audio MC-VQA

Language	English					Native				
	G-3n-2B	G-3n-4B	Q-2.5-3B	Q-2.5-7B	Gemini	G-3n-2B	G-3n-4B	Q-2.5-3B	Q-2.5-7B	Gemini
Ghana-twi	30.04	58.58	35.07	34.89	89.43	28.36	47.01	17.35	2.05	74.44
Ethiopia-amh	33.47	43.89	28.86	22.04	84.33	37.27	48.3	29.66	2	84.84
Malawi-nya	20.76	39.12	39.72	16.37	76.86	24.35	36.93	72.26	2.4	69.08
Nigeria-hau	21.37	62.1	42.74	39.11	86.77	17.34	45.56	41.33	0.2	86.49
Nigeria-ibo	24.85	41.41	26.06	16.77	78.41	37.98	48.28	45.66	2.42	67.92
Kenya-kik	20.29	37.47	25.47	16.56	84.4	21.12	35.4	52.59	1.24	78.8
Rwanda-kin	29.4	53	40.6	30	84.62	23.6	33	58.4	5	83.6
Uganda-lug	23.46	42.94	37.18	32.21	82	22.31	36.85	82.07	6.37	76.25
Ethiopia-orm	0	28.57	57.14	57.14	71.43	21.01	47.34	47.34	2.93	71.43
Botswana-tsn	26.57	49.29	35.7	32.25	84.76	22.52	35.7	64.91	3.25	78.5
Somali-som	28.34	49.1	33.33	20.16	82.62	29.54	49.9	40.92	0.6	79.9
Eritrea-tir	25.7	51.77	36.69	29.05	75.14	28.31	43.02	36.13	1.12	74.71
Nigeria-yor	29.6	50.2	31.2	28	82.65	35	45.4	61.2	6.6	81.65
Lesotho-sot	-	-	-	-		19.41	36.71	64.98	2.11	64.29
S.Africa-zul	-	-	-	-		21.01	37.25	77.03	1.4	78.71
Average	24.14	46.73	36.14	28.81	0	25.94	41.78	52.79	2.65	77.51

Table 13: Audio-Loc-Lang Aware(MC-VQA)

Language	English				Native			
	Gemma3n-2B	Gemma3n-4B	Qwen-3B	Qwen-7B	Gemma3n-2B	Gemma3n-4B	Qwen-3B	Qwen-7B
Ghana-twi	28.92	59.7	38.06	40.3	30.97	39.55	5.22	1.49
Ethiopia-amh	25.85	46.09	24.25	26.85	32.46	47.09	2.81	1.6
Malawi-nya	20.16	42.32	16.57	19.36	27.74	37.92	3.39	1.4
Nigeria-hau	29.23	62.1	38.71	43.15	20.16	46.77	1.41	0.4
Nigeria-ibo	26.67	41.21	17.98	16.77	37.17	41.01	3.64	0.81
Kenya-kik	22.57	38.3	23.81	19.67	21.95	30.64	3.11	2.28
Rwanda-kin	27.6	56	32.4	35.6	22.6	33.6	4.4	1.8
Uganda-lug	24.85	43.54	34.19	34.39	20.52	34.26	9.76	4.58
Ethiopia-orm	0	42.86	28.57	57.14	24.2	45.74	1.86	0.53
Botswana-tsn	27.59	53.75	33.67	36.92	23.94	31.03	10.95	2.64
Somali-som	26.75	50.3	27.15	31.54	28.74	48.9	0.6	0.8
Eritrea-tir	23.65	51.4	26.26	33.33	24.02	40.41	0.37	1.68
Nigeria-yor	29.6	51.4	28.8	28.8	34.4	39.2	4.6	2
Lesotho-sot	-	-	-	-	19.83	36.71	7.17	0.53
S.Africa-zul	-	-	-	-	20.45	32.77	7.56	1.68
Average	24.11	49.15	28.49	32.6	25.94	39.04	4.46	1.61

Table 14: MC-VQA Audio - Image Only

Audio answer generation

Language	English				Native			
	Gemma3n-2B	Gemma3n-4B	Qwen-3B	Qwen-7B	Gemma3n-2B	Gemma3n-4B	Qwen-3B	Qwen-7B
Ghana-twi	20.71	43.28	6.9	13.06	18.47	38.43	6.9	1.87
Ethiopia-amh	28.66	48.5	8.82	11.22	26.65	47.49	8.82	0.2
Malawi-nya	19.56	44.11	7.39	10.18	21.16	47.9	7.39	2.79
Nigeria-hau	22.58	54.44	11.69	12.7	17.94	49.6	11.69	0
Nigeria-ibo	25.05	37.17	6.46	4.85	26.06	41.82	6.46	0.4
Kenya-kik	17.18	30.64	7.45	9.94	21.33	40.37	7.45	1.24
Rwanda-kin	25	56.8	16.2	12.4	17.8	38.2	16.2	1.2
Uganda-lug	12.13	33.6	9.15	11.13	16.14	33.67	9.15	2.59
Ethiopia-orm	0	0	0	0	15.16	39.36	13.3	0.53
Botswana-tsn	17.85	41.58	8.52	10.55	14.4	36.71	8.52	1.42
Somali-som	18.56	44.71	6.99	4.19	19.96	57.29	6.99	1
Eritrea-tir	25.33	42.46	4.1	10.24	20.3	45.07	4.1	0.37
Nigeria-yor	21.4	43.6	8.2	11.2	24.4	39.6	8.2	3.6
Lesotho-sot	-	-	-	-	5.19	44.3	20.25	0.42
S.Africa-zul	-	-	-	-	15.97	46.78	43.98	1.4
Average	19.54	40.07	7.84	9.36	18.73	43.11	11.96	1.27

Table 15: MC-VQA for Audio - No Image

Language	G3n-2B		G3n-4B		Q-3B		Q-7B		Gemini 2.5 Pro	
	chrf++	acc.	chrf++	acc.	chrf++	acc.	chrf++	acc.	chrf++	acc.
Ghana-twi	40.76	42.91	51.80	49.25	36.85	36.94	58.10	55.04	82.19	84.02
Ethiopia-amh	38.22	30.66	42.21	35.27	37.43	33.07	50.37	42.28	74.81	80.04
Malawi-nya	30.16	29.94	37.50	36.13	30.91	37.72	44.44	38.32	72.28	71.49
Nigeria-hau	24.78	35.28	31.91	42.34	32.18	36.49	50.80	52.02	68.53	79.16
Nigeria-ibo	29.65	23.43	36.67	30.51	23.58	20.00	41.25	32.12	75.63	74.74
Kenya-kik	37.55	29.63	42.72	34.16	33.75	25.88	47.16	38.92	77.30	79.27
Rwanda-kin	32.78	36.40	44.82	44.20	35.51	35.00	54.36	48.00	79.82	78.99
Uganda-lug	35.99	35.19	45.98	39.96	34.00	28.03	47.05	40.56	80.23	77.00
Ethiopia-orm	9.97	0.00	24.52	0.00	62.85	42.86	0.00	0.00	0.00	0.00
Botswana-tsn	40.72	43.29	48.88	46.89	40.89	37.93	50.84	47.87	81.60	82.52
Somali-som	28.56	31.94	33.41	32.73	26.02	25.75	39.82	35.93	73.48	71.02
Eritrea-tir	35.41	34.45	46.55	40.04	36.53	36.13	54.31	46.74	72.21	72.87
Nigeria-yor	26.73	35.80	31.58	41.00	28.09	34.80	35.38	35.20	62.22	81.69
Average	31.64	31.46	39.89	36.34	35.28	33.12	44.15	39.46	69.25	71.75

Table 16: Audio answer generation for English Audio

Language	G3n-2B		G3n-4B		Q-3B		Q-7B		Gemini 2.5 Pro	
	chrf++	acc.	chrf++	acc.	chrf++	**acc.**	chrf++	acc.	chrf++	acc.
Ghana-twi	12.14	11.75	12.24	13.25	5.98	3.92	5.98	4.10	45.72	57.06
Ethiopia-amh	0.58	24.85	1.53	25.85	0.00	2.61	0.00	2.61	47.43	79.30
Malawi-nya	16.03	10.98	17.32	13.77	6.31	5.39	6.31	5.59	58.40	63.05
Nigeria-hau	9.52	13.10	12.53	16.94	2.87	0.40	2.87	0.60	63.06	81.65
Nigeria-ibo	9.42	9.90	9.01	10.71	4.22	2.83	4.22	2.83	41.81	52.08
Kenya-kik	20.13	12.96	21.02	14.40	5.54	1.66	5.54	1.66	54.02	63.09
Rwanda-kin	12.81	14.20	11.84	17.40	2.48	1.20	2.48	1.20	67.48	73.68
Uganda-lug	15.13	18.49	15.41	18.29	4.96	3.59	4.96	3.78	56.59	64.26
Ethiopia-orm	6.22	10.60	5.80	9.40	2.90	2.93	3.27	3.72	51.83	50.55
Botswana-tsn	17.03	15.23	14.58	12.22	4.88	3.04	4.88	3.04	71.22	74.39
Somali-som	6.78	8.58	6.61	11.38	3.00	2.00	3.00	2.00	59.27	69.17
Eritrea-tir	0.10	22.35	0.14	17.32	0.01	7.08	0.01	7.26	47.97	68.63
Nigeria-yor	8.92	19.00	8.30	19.40	2.76	1.60	2.76	1.20	33.51	72.98
Lesotho-sot	6.46	3.56	5.30	4.13	5.14	1.27	6.47	2.11	50.94	50.84
S.Africa-zul	19.12	18.75	22.44	16.86	8.78	4.20	15.36	9.24	69.90	71.43
Average	10.69	14.29	10.94	14.75	3.99	2.91	4.54	3.40	54.61	66.14

Table 17: Audio answer generation for Native Audio

C Image category distribution

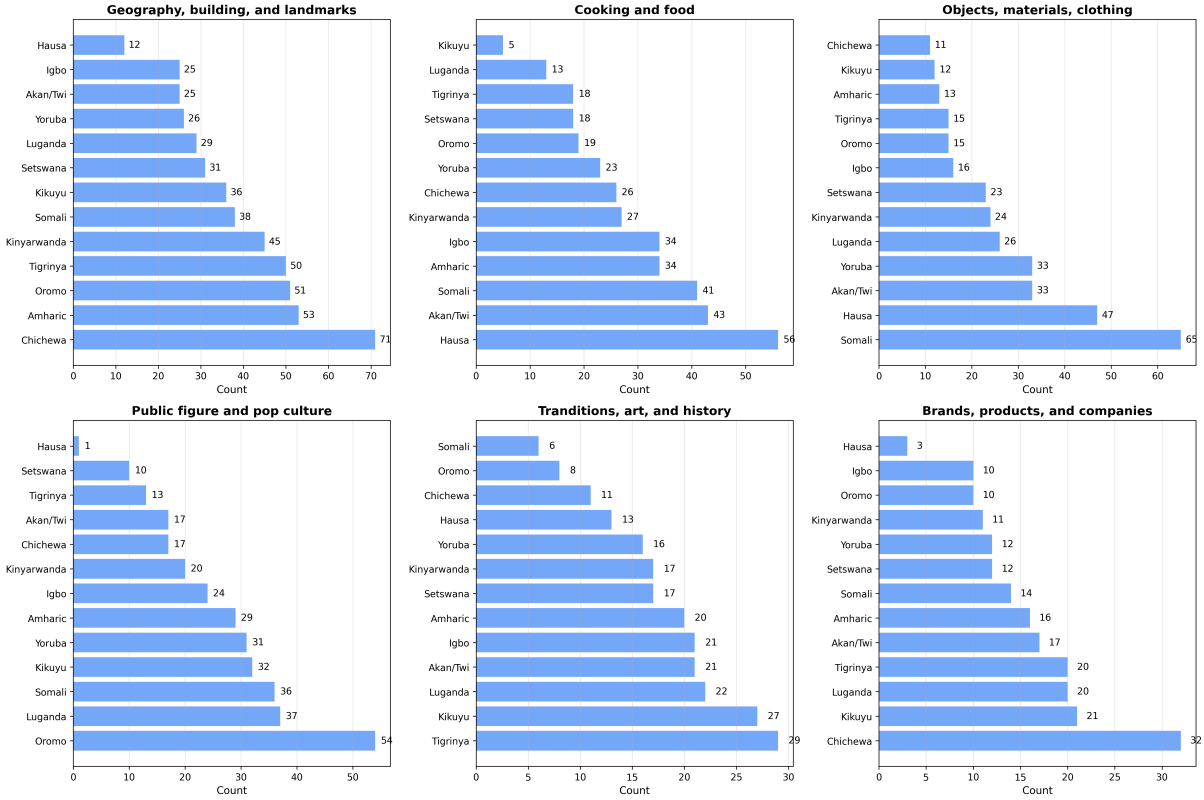


Figure 7: Top 6 Image category distribution

D Annotation guideline

Guidelines for Annotators

The following shows detailed requirements and rules for dataset creation when annotating the dataset.

Image Selection:

- Collect images that represent the diverse cultural aspects and the specific cultural background of your country. The image must fall into one of the categories below. **Pick one of the most relevant categories (more later)**:

- Images should be relevant to your culture or country.
- Ensure that **images are relevant to the questions being posed.** In other words, the image **is needed** to answer the question.
- If the image contains the answer's text, you can blur/crop the image so that it does not contain the answer.
- Image source:
  - Self/personal picture **(highly preferable).** Ask your family or friends to donate their photos, if possible.
  - We also accept external images from:
    - Flickr: https://www.flickr.com/explore (**please make sure the associated license to the image is Creative Commons**), this can be selected at the top left of Flickr ("Any License").
    - Wikimedia Commons: https://commons.wikimedia.org/wiki/Main_Page (**here you do not need to select any license for the images)**,
    - Unsplash: https://unsplash.com/ (**please make sure to search the image first and select the license: Free**). More details (Tutorial) at the end of this document.

- Dollar Street: https://www.gapminder.org/dollar-street (**here you do not need to select any license for the images).** This webpage has images only from some countries; please make sure to select your country to find images if applicable.

  More detailed instructions for each web page are shown at the end of this document.

  - If you use an external image, you'll need to put the **URL** of the original image.
- The image must be of reasonable quality (not pixelated or blurry) and understandable. You can upload images of any ratio as long as it is not too tall or wide (e.g.: don't submit panorama pictures).
- Do not show personally identifiable information (PII) such as faces, car plates, house addresses. Faces of public figures or fictional characters are ok. Also, **please be sure to blur text in the image that will leak the answer**. "PicdeFacer" can be used for blurring: https://picdefacer.com/en/. Tutorial on using PicdeFacer is shown at the end of this document.

Question and Answer Creation:

After finding the image, you must formulate 1-3 questions + answers from that image. Specifically:

- The question must be answerable **only by looking at the image.**
- Ensure that the questions are culturally relevant and specific to the image content.
- Provide answers that are concise, accurate, and directly related to the question.
- You must also provide one correct option and three other incorrect options (distractors). For the distractors, choose relevant options, not obvious wrong answers.
- **The question must be answerable even without multiple-choice questions.** Example of the invalid question: ("What song is not performed by this musician" – not answerable if you don't know the choices)
- Make sure the questions are **written fluently in both the local language and English.** Use a grammar checker if needed i.e. if you are not fluent in English.
- Be mindful of cultural sensitivities and avoid stereotyping or misrepresenting cultural aspects.
- Complexity Levels: Include a variety of question types:
  - Multi-hop reasoning questions
  - Counting or estimation questions
  - Questions requiring local common sense knowledge

        - Cultural inference questions
    - Ensure there are **variations on your question**. Create complex questions (for example, multi-hop reasoning, counting, referencing, or questions that require local commonsense knowledge to be answered) rather than asking Identity questions(e.g., “What is this?” or " Where is this?”).

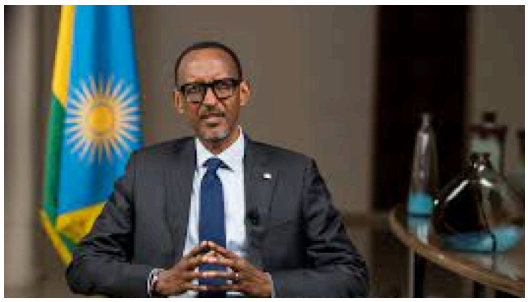
Question complexity types

When formulating questions, include the following complexities

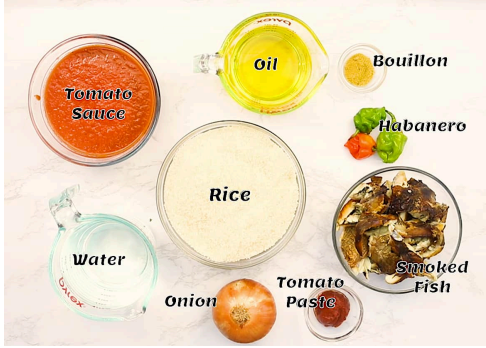
- Superlative (e.g., Q: Who was the youngest among them? A: Cyril Ramaphosa, B:---)

- Multi-hop (e.g., Q: What is the name of the brand that the person in the picture is leading as CEO? A: Nando's B: Shoprite C: Woolworths D: Checkers)

e.g., Q: Where was the person in the picture born? A: Ruanda-Urundi, B…..)



- Intersection (e.g, what national dish is often prepared with the above ingredients in _____

- 

Audio Recording Requirements

Recording Specifications

- **Format**: MP3 or WAV (MP3 preferred for file size)
- **Environment**: Record in a quiet space with minimal background noise

What to Record

For each question-answer pair, record:

1. **Question in Native Language**
2. **Question in English**
3. **Options in Native Language**
4. **Options in English**

Recording Guidelines

- **Speaking Style**:

    - Speak clearly and at a natural pace
    - Use standard pronunciation for your language/dialect
    - Avoid rushed or overly slow speech
- **Consistency**: Maintain consistent volume and tone across recordings
- **Multiple Takes**: Record multiple takes and select the clearest version
- **Native Speakers Only**: Questions should be recorded by native speakers of the language

Category Definition

When selecting a category, pick one of the most relevant. Please follow the guidelines:

- **Vehicles and Transportation:** Local public transport, local vehicles.
- **Objects, Materials, Clothing:** Questions about local/traditional clothes. Unique/local tools or items.
- **Cooking and Food:** Local dishes and food/drink. This category includes native fruits in the context of the image if that fruit is served as a food/drink.
- **Geography, Buildings, Landmarks:** Popular/common landmarks, local architecture/buildings. Local monuments.
- **Plants and Animals:** Plants and animals commonly found in the region.
- **Brands, Products, and Companies:** Questions about understanding local yet popular brands or companies. Even if the brand is about food/transportation, if the main focus of the question is the brand recognition itself, then it should be under this category.
- **Sports & Recreation:** Local sports and fun activities. Focuses on the activity itself rather than the location (in that case, it goes to the 'landmark' category).
- **Tradition, Art, History:** Local ceremonies/festivals/events, local dance/music, folklores. Historical artifacts.
- **People & Everyday Life:** Focuses on the people themselves: i.e., common habits/customs, common occupations and jobs, routine religious activities, everyday activities/routines.
- **Public Figures & Pop Culture:** Questions on the understanding of common public figures (e.g., politicians, artists, musicians, etc.). Common pop culture such as movies and games.

If the category is still ambiguous to you, pick the one you think is the most appropriate.

E Annotator Demography

Annotator ID	Gender	Age Group	Resides in Africa
Ethiopia-amh	M	18-30	✓
Nigeria-hau	F	30-40	✓
Nigeria-ibo	F	0-40	✓
Uganda-lug	F	18-30	✓
Ethiopia-orm	M	30-40	✓
Rwanda-kin	F	18-30	✓
Kenya-kik	F	30-40	✓
Somali-som	F	18-30	✓
Eritrea-tir	M	18-30	✓
Ghana-twi	F	18-30	✓
Nigeria-yor	M	30-40	✓
Botswana-tsn	M	30-40	✓
Malawi-nya	F	18-30	✓
S.Africa-zul	F	18-30	✓
Lesotho-sot	F	30-40	✓

Table 18: Annotator Demographics

F Review Guidelines

Afri-MC-VQA Dataset Review Guidelines

Overview

This document provides comprehensive guidelines for reviewers to ensure quality control and consistency across all submissions for the Afri-MC-VQA dataset. Reviewers must carefully evaluate each submission against these criteria to ensure the integrity of the dataset.

1. Image Review Criteria

1.1 Privacy and Anonymization

All images must protect personal privacy

- **Faces**: All private individuals' faces must be blurred (public figures exempt)
- **Personal Information**: Must be completely obscured:
    - Vehicle registration numbers
    - Phone numbers
    - Home addresses
    - Personal identification documents
    - Credit card information
- **Verification**: Use image zoom to check for any missed PII

1.2 Cultural Relevance

- The image must clearly represent aspects of local culture
- Should be meaningful to the specific country/region
- Must align with one of the designated cultural categories
- Reject generic images that could be from anywhere

1.3 Image Uniqueness

- Check for duplicate images within the same batch
- Verify against existing submissions to avoid repetition
- Similar scenes from different angles are acceptable if they enable different questions
- Flag any stock photos or overly generic content

2. Question Quality Standards

2.1 Language Quality

- **English Version**:
  - Must be grammatically correct
  - Natural phrasing (not awkward translations)
  - Clear and unambiguous
- **Native Language Version**:
  - Fluent and natural in the local language
  - Appropriate register and tone
  - Correct spelling and diacritics

2.2 Translation Accuracy

- **Semantic Equivalence**: Both versions must convey identical meaning
- **Not Word-for-Word**: Translations should be idiomatic, not literal
- **Cultural Adaptation**: Acceptable to adapt phrasing for cultural context while maintaining meaning

2.3 Question Independence

- **Stand-alone**: Question must be answerable from the image alone
- **No Option Dependency**: Remove questions like "Which of these is NOT..." that require seeing options
- **Single Clear Answer**: Only one option should be definitively correct
- **Test**: Cover the options - can the question still be answered?

2.4 Clarity and Specificity

- **No Vagueness**: Questions must be precise
- **Object Specification**: When multiple objects present, clearly indicate which one

2.5 Formatting

- **No Numbering**: Questions should not start with "1.", "Q:", etc.
- **Proper Punctuation**: End with a question mark
- **Capitalization**: Follow standard rules for both languages

3. Multiple Choice Options Review

3.1 Plausibility Check

- **All Options Believable**: Someone unfamiliar with the culture should find all options possible
- **No Obvious Eliminations**: Avoid options that are clearly wrong

- **Context-Appropriate**: All options should fit the question's context

3.2 Format Consistency

- **Uniform Structure**: All options should have similar length and detail
- **Parallel Grammar**: Same grammatical structure across options

3.3 Prohibited Answer Types

- **Never Accept**:
  - "None of the above"
  - "No answer"
  - "Nothing"
  - "All of the above"
  - "Not applicable"

3.4 Objectivity Requirements

- **No Subjective Adjectives**: Avoid beautiful, ugly, nice, bad, etc.
- **Factual Descriptions Only**: Focus on observable characteristics

4. Audio Review Requirements

4.1 Transcript Accuracy

Audio must match text EXACTLY

- **Word-for-Word**: No deviations, additions, or omissions
- **Pronunciation**: Clear and accurate for the language
- **Verification Method**: Follow along with text while listening

4.2 Audio Structure

- **Question Format**: Natural reading with appropriate intonation
- **Option Spacing**: Brief pause (0.5-1 second) between each option
- **Answer Position**: Correct answer must ALWAYS be the first option spoken

4.3 Audio Quality

- **Clarity**: No background noise or distortion
- **Volume**: Consistent throughout recording
- **Speed**: Natural speaking pace (not rushed or too slow)
- **Complete**: Full question and all options included

5. Critical Review Points

5.1 Question-Option Alignment

- **Check**: Ensure question type matches option format
- **Common Issue**: Yes/No question with descriptive options
- **Verify**: Question grammar aligns with option grammar

5.2 Cultural Sensitivity

- **No Stereotypes**: Avoid reinforcing negative cultural stereotypes
- **Respectful Representation**: Ensure dignified portrayal of cultural elements
- **Balanced View**: Include both traditional and modern aspects

5.3 Language Quality Control

- **Spelling**: Check both languages thoroughly
- **Grammar**: Ensure proper sentence structure
- **Punctuation**: Consistent and correct usage
- **Diacritics**: Verify correct marks for languages that require them

5.4 Content Diversity

- **Avoid Repetition**: Similar questions across different images
- **Varied Complexity**: Mix of simple and complex questions
- **Different Question Types**: Not all "What is..." questions
- **Topic Distribution**: Balanced across cultural categories

5.5 Question Complexity

- **Avoid Trivial Questions**:
- **Require Cultural Knowledge**: Questions should test understanding, not just observation
- **Informative Value**: Each question should teach something about the culture

G Prompt examples

Figures 8 & 9 show the location-aware prompts we used in both setups for audio and text modalities.

Text Prompts

Multiple-Choice QA:

SYSTEM You are a helpful AI assistant.

USER You are given an image. Analyze the image and answer the following multiple-choice question. Only one option is correct. Return only the correct option name i.e. A, B, C or D. The question is relevant to {country}.

[Image] Question: {question}

Options: A. {opt_a}, B. {opt_b}, C. {opt_c}, D. {opt_d}

Open-Ended:

SYSTEM You are a helpful AI assistant.

USER You are given an image. Analyze the image and answer the question with a short factual answer. The answer should be a word or a short phrase. Only give the answer, no follow-up or extra information.The question is relevant to {country}

[Image] Question: {question}.

Figure 8: Location-aware text prompts.

Audio Prompts

Multiple-Choice QA:

SYSTEM You are a helpful AI assistant.

USER You are given an image and audio question. Analyse the image and answer the audio Multiple Choice Question. Only one option is correct. Return only the correct option name i.e. A, B, C or D. The question is relevant to {country}.

[Image] [Audio] Question: {question_audio}

Audio Options: A-D {options_audio}

Open-Ended:

SYSTEM You are a helpful AI assistant.

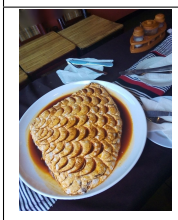
USER You are given an image and audio question. Analyze the image and answer the audio question with a short factual answer. The answer should be a word or a short phrase. Only give the answer, no follow-up or extra information. The question is relevant to {country}.

[Image] Question: {question_audio}

Figure 9: Location-aware audio prompts.

H Examples

Table 19: Examples from the Afri-MCVQA Benchmark

Image	Self-made	Country-Language	Category	Question + Options (English)	Question + Options (Native)
	Yes	Bostwana-Setswana	Objects, materials, clothing	**Q:** What material is used to make this attire? **Options:** A) Reed B) Buttons C) Beads D) Egg Shells	**Q:** Seaparo se se dirilwe ka eng? **Options:** A) Letlhaka B) Dikonopo C) Dibaga D) Dikgapha tsa mae
	No	Uganda-Luganda	Brands, Products and Companies	**Q:** Which Ugandan food company produces the snacks in the picture? **Options:** A) Psalms Food Company B) Maganjo C) Mukwano D) Jesa	**Q:** Kampuni ki eya Uganda ekola ebyokulya ebiri mukifaananyi? **Options:** A) Psalms Food Company B) Maganjo C) Mukwano D) Jesa
	No	Kenya-Kikuyu	Traditions, Art and History	**Q:** Who built this monument? **Options:** A) Burhan Ali Taher and Hussein Alawi B) Hassan Joho and Ali Taib Sheikh Suleiman C) Shahbal and Swaleh Nguru D) Tahir Sheikh Said and Mohammed Jaffer	**Q:** N a maathondekire mcongo y? **Options:** A) Burhan Ali Taher and Hussein Alawi B) Hassan Joho and Ali Taib Sheikh Suleiman C) Shahbal and Swaleh Nguru D) Tahir Sheikh Said and Mohammed Jaffer
	Yes	Ethiopia-Oromo	Cooking and Food	**Q:** What is the name of this traditional food? **Options:** A) Chumbo B) Chechebsa C) Buna kela D) Kori	**Q:** Nyaata aadaa kun maqaan isaa maal jedhama? **Options:** A) Cuumboo B) Caccabsaa C) Buna qalaa D) Qorii
	Yes	Nigeria-Igbo	Sports and Recreation	**Q:** What is the name of the cultural dance group shown in the image? **Options:** A) Mkpkiti B) Egedege C) Atiliogwu D) War dance	**Q:** Ked aha nd otu egwu dnala a egosiptara n'onyonyo? **Options:** A) Egwu mkpkiti B) Egedege C) Atiliogwu D) Ikpirikpe ogu