Afri-MCQA: Multimodal Cultural Question Answering for African Languages

Anonymous ACL submission

Abstract

Africa is home to over one-third of the world's languages, yet remains underrepresented in AI research. We introduce Afri-MCQA, the first **M**ultilingual **C**ultural **Q**uestion-**A**nswering benchmark covering 7.5k Q&A pairs across 15 African languages from 12 countries. The benchmark offers parallel English-African language Q&A pairs across text and speech modalities and was entirely created by native speakers. Benchmarking large language models (LLMs) on Afri-MCQA shows that open-weight models show poor performance across evaluated cultures, with near-zero accuracy on open-ended VQA when queried using native language or speech. To evaluate linguistic competence, we include control experiments meant to assess this specific aspect separate from cultural knowledge, and we observe significant performance gaps between native languages and English for both text and speech. These findings underscore the need for speech-first approaches, culturally grounded pretraining, and cross-lingual cultural transfer. We will release Afri-MCQA to support more inclusive multimodal AI development.

1 Introduction

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 world's languages (Hammarström, 2018) and a population exceeding 1.3 billion, projected to surpass 2.5 billion by 2050 (Simane et al., 2025). 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, and 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 such as Multimodal large language models (MLLMs) particularly relevant for enabling access to technology.

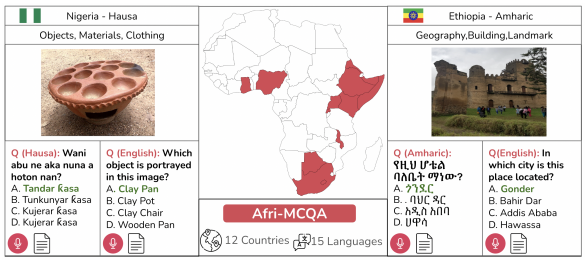


Figure 1: Examples of Afri-MCQA datapoints, containing parallel text and speech QA pairs grounded in culturally relevant images across English and native African languages.

MLLMs perform well on tasks that require reasoning over visual, spoken, and textual inputs to follow user instructions (Liu et al., 2021a; Kamath et al., 2025; Abdin et al., 2024). However, these systems are developed primarily for high-resource, mostly Western-centric, or as Mihalcea et al. (2025) put it, they have a 'WEIRD' coverage. As a result, their knowledge and reasoning abilities tend to favor the languages and cultures represented in their training data (Mihalcea et al., 2025).

Although most evaluation datasets for low-resource languages are translated (NLLB Team, 2022; Adelani et al., 2024, 2025; Azime et al., 2024; Alabi et al., 2025), researchers have recently made promising progress in creating culturally relevant benchmarks for African languages across various NLP tasks (Adelani et al., 2023; Azime et al., 2025; Muhammad et al., 2025; Yu et al., 2025; Tonja et al., 2024b,a). However, these evaluations are designed for text-only tasks. In the language-vision space, recent benchmarks assess global cultural knowledge through question answering (Romero et al., 2024; Vayani et al., 2025a; Winata et al., 2025), yet their coverage of African languages and contexts remains limited.

Africa's rich cultural diversity demands AI sys-

tems that can effectively serve its communities. A critical step toward this goal is evaluating how well current MLLMs understand and reason about African cultural knowledge in multimodal settings. To enable this, we introduce Afri-MCQA , **the first multilingual cultural VQA dataset supporting text and speech modalities** (as shown in Figure 1). The dataset consists of $\approx 7.5k$ Q&A pairs (in English and native languages) across 15 African languages from 12 countries. Our data collection involves native speakers as annotators, who reside in the countries where their respective languages are spoken. Each language covers ≈ 500 image-grounded Q&A samples, with both text and spoken audio in the native language and English. We evaluate multiple MLLMs across various setups to answer the following research questions:

RQ1: How well do MLLMs understand African cultural contexts in visually-grounded QA?

RQ2: How does input modality (text vs. speech) affect performance?

RQ3: How does query language (native vs. English) affect performance, and do differences reflect language understanding or cultural knowledge gaps?

RQ4: How does task format (Multiple-Choice vs. Open QA) affect accuracy?

Our contributions are: **(1)** We introduce Afri-MCQA, the first large-scale multilingual visual cultural QA benchmark for 15 African languages across 12 countries, with parallel text and speech QA created by native speakers. **(2)** We demonstrate that text-based multilingual ability does not transfer to speech understanding, emphasizing the need for more African language representation in multimodal training. **(3)** We will open-source Afri-MCQA to advance multimodal research for Africa's diverse languages and cultures.

2 Related work

With the growth in popularity of LLMs and MLLMs, cultural and multilingual multimodal evaluation has received more attention. For instance, Liu et al. (2021b) 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.

Previous work on African languages has focused mainly on text-based benchmarks. While some 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 evaluates 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., 2025) 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 to proprietary systems.

Despite this progress, all of these evaluations remain text-only, while 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) include only a few African languages and do not adequately capture the region's diversity. Afri-MCQA addresses these gaps with three key contributions: (**1**) broader coverage with 15 languages across 12 countries, (**2**) inclusion of speech modality with parallel native and African-accented English audio, and (**3**) diagnostic control experiments that separate linguistic competence from cultural knowledge limitations.

3 Dataset

To create Afri-MCQA, we selected 15 widely spoken languages in sub-Saharan Africa (by number of speakers, according to Ethnologue[1]) across 12 countries, representing an aggregate speaker population of approximately 392.6 million (see Table 2). We hired native language speakers as annotators through Upwork.[2] The selection criteria were based on (i) fluency in English, (ii) prior annotation or data collection experience, (iii) high project completion rate on Upwork, and (iv) res-

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

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

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.

Lang-Country	Family / Branch	Reg	#spk	#QA	#eng(h)	#nat(h)
Akan/Twi-Ghana	Niger-Congo / Volta-Niger	West	9M	537	2.41	2.43
Amharic-Ethiopia	Afro-Asiatic / Ethio-Semitic	East	57M	500	1.56	1.51
Chichewa-Malawi	Niger-Congo / Bantu	S & E	12M	501	1.41	1.50
Hausa-Nigeria	Afro-Asiatic / Chadic	West	77M	496	2.80	3.04
Igbo-Nigeria	Niger-Congo / Volta-Niger	West	31M	501	1.61	1.59
Kikuyu-Kenya	Niger-Congo / Bantu	East	8.1M	495	1.66	1.72
Kinyarwanda-Rwanda	Niger-Congo / Bantu	East	18M	501	2.73	2.67
Luganda-Uganda	Niger-Congo / Bantu	East	10M	500	2.30	2.42
Oromo-Ethiopia	Afro-Asiatic / Cushitic	East	37M	512	2.20	2.30
Setswana-Botswana	Niger-Congo / Bantu	South	14M	502	1.89	2.39
Somali-Somalia	Afro-Asiatic / Cushitic	East	22M	501	1.96	1.99
Tigrinya-Eritrea	Afro-Asiatic / Ethio-Semitic	East	9M	537	2.13	2.31
Yoruba-Nigeria	Niger-Congo / Volta-Niger	West	46M	498	1.98	2.06
Sesotho-Lesotho	Niger-Congo / Bantu	South	13.5M	533	–	1.90
Zulu-S.Africa	Niger-Congo / Bantu	South	28M	528	–	1.37

Table 2: **Overview of languages in Afri-MCQA**. #spk = estimated L1 & L2 speakers, #eng (h) = hours of accented English audio, #nat (h) = hours of native language audio. – indicates English audio not collected for that language.

idence in a country where the target language is spoken. After the selection process, we divided the annotation into two phases designed to ensure quality.

3.1 Dataset collection

Guideline and Platform Preparation We followed the annotation guidelines of Romero et al. (2024), including image categories, question templates, and distractors, and extended them to include audio recording instructions (for both native language and English) and adding a two-step review and verification process. Full annotation guidelines are provided in Appendix I.

Training and Screening Phase The first phase consisted of a small-scale pilot designed to train and screen annotators. We provided detailed guidelines and training to ensure annotators understood the task criteria and quality standards. After training, each annotator submitted 50 QA samples. We reviewed these submissions to verify alignment with the guidelines. Based on this review, we provided detailed feedback and, when necessary, scheduled follow-up meetings to clarify issues. Only annotators whose submissions met our quality criteria and successfully incorporated feedback were selected for the next phase. Approved samples from this screening phase were included in the final dataset.

Main Annotation Phase In this phase, the remaining 450 items per language were collected by annotators selected in the first phase. To ensure quality across the large volume of submissions, we involved *language coordinators*, experienced native speakers with strong linguistic and cultural knowledge for each language. Coordinators reviewed all submissions for linguistic accuracy, cultural appropriateness, adherence to guidelines, and audio quality. Review guidelines are provided in Appendix N. When issues arose, coordinators discussed them directly with annotators to resolve them. Unresolved disagreements were then raised with the project team for a final decision. *Language coordinators* are co-authors of this paper and played a major role in the quality assurance process. After their approval, 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 language. Below, we describe each of these components.

Image / Category Selection To build the image set for Afri-MCQA, we encouraged annotators to contribute their own images whenever possible. When self-sourcing was not feasible, we permitted the use of open-license images from websites

we provided (see Appendix I for list of websites). All collected images were categorized into the 10 classes defined by Romero et al. (2024). Figure 2 shows the distributions of images per category (See Appendix A for category distributions across languages).

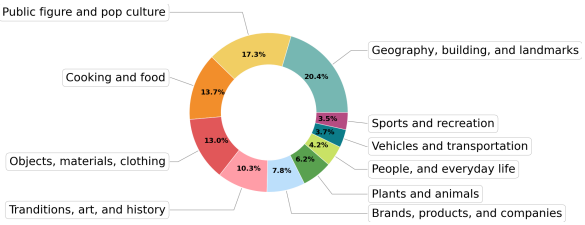


Figure 2: Image categories in our dataset and their distributions.

Question & Answer Generation For each image, annotators wrote up to 3 multiple-choice QA triplets (question + 1 correct answer + 3 distractors) in both their native language and English. To ensure the benchmark remains challenging, we instructed annotators to design complex questions that require reasoning to answer correctly (See annotation guidelines in Appendix I).

Audio Recording To investigate spoken language understanding capabilities, we instructed annotators to record audio for each question and its corresponding answers, reading them clearly in both the native language and English. Therefore, Afri-MCQA includes audio recordings for both questions and answers in native language and African-accented English.

4 Experimental setup

Our experimental design is organized to address each research question as follows:

Models: To assess how well current MLLMs understand African cultural contexts (RQ1), we selected models based on two criteria: a) support for both image and audio input, enabling multimodal evaluation, and (b) availability of different model sizes within each family to assess scaling effects. From open-weight MLLMs, we selected Qwen 2.5-Omni (3B &7B) (Jin Xu, 2025) and Gemma-3n-(2B & 4B)-it (Kamath et al., 2025). For text-only baselines, we include Gemma3 (12B & 27B)-it (Kamath et al., 2025). For comparison with closed-source models, we include Gemini-2.5 Pro (Comanici et al., 2025), which supports audio, text, and vision inputs.

Query Modality: To explore how input modality affects performance (RQ2), we evaluate models using both text and audio modalities. 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 both in African languages and accented English. This comparison shows whether model performance generalizes from text to speech. Since we evaluate VQA, all settings include the image related to the question.

Query Languages: To explore how query languages affect models' performance and whether gaps reflect linguistic or cultural limitations (RQ3), each question is presented in *native language* and *English*. This setup allows us to compare model behavior between English and native languages.

Task Format: To understand how task format affects model performance (RQ4), we evaluate models on both Multiple-Choice VQA (MC-VQA) and Open-ended VQA. MC-VQA provides answer options, while Open-ended VQA requires answer generation. Comparing these formats shows whether strong MC-VQA performance reflects actual cultural understanding or simply selecting from provided options. For each task format, we use the same prompt templates across models and languages.

Prompt: We evaluated all settings and models using a *Location-aware* prompt (adding location/-country as a context). We chose this setting as it performed better than image-only prompts (without providing context). We provide results for Image-only prompts and language-wise results in Appendix C, and the prompts used are in Appendix B.

5 Evaluation

This section describes our two evaluation setups and the metrics used in this study.

5.1 Cultural VQA Evaluation

We evaluate models on visually-grounded cultural QA using Afri-MCQA in both modalities. For all tasks, models are provided with an image and must use visual information to answer the question.

Text-based VQA: For the text modality, we use two evaluation formats: **(1) MC-VQA**: Models are

given an image, a text question, and four answer options. They must select the correct option by reasoning over the image and question. **(2) Open-ended VQA**: Models receive an image and a text question without answer options, and are required to generate the correct answer. This tests their ability to retrieve and reason about cultural knowledge without potential hints in the answer set.

Audio-based VQA: Similarly, we evaluate audio modality using two formats: **(1) Audio MC-VQA**: Same setting as MC-VQA described above, but models are queried through African-accented English and native language speech. **(2) Audio open-ended VQA**: Given an image and the question in speech format without answer options and models required to generate the correct answer in text.

5.2 Control Experiments

While it is technically challenging to determine whether prediction failures on Afri-MCQA arise from limitations in language understanding or gaps in cultural knowledge, we conduct control experiments on easy tasks that primarily require language understanding in either text or speech form. These evaluations provide evidence to understand the extent to which language-understanding limitations may contribute to the observed failures.

Text-based experiments: To probe text understanding, we evaluate on two benchmarks: **(1) AfriXNLI** (Adelani et al., 2025): natural language inference, and **(2) AfriMMLU** (Adelani et al., 2025): general knowledge QA. By analyzing performance on these text-only tasks and comparing it with results on Afri-MCQA, we obtain an approximate measure of the models' baseline linguistic competence on the studied languages.

Audio-based experiments: To probe audio understanding, we conduct two tasks: **(1) ASR:** transcribing spoken African language audio to text, assessing whether models can accurately capture spoken content as a prerequisite for answering questions. **(2) Language Identification (LID):** identifying which of the 15 languages is spoken, testing the model's ability to recognize spoken languages. These tasks reveal whether poor audio VQA results from speech-processing failures or cultural reasoning limitations.

5.3 Evaluation metrics

We evaluate models in a zero-shot setting using automatic metrics across all tasks, with additional human evaluation for open-ended VQA (text).

Automatic Evaluation: We report accuracy scores for MC-VQA and classification tasks, and use GPT-4o-mini (Hurst et al., 2024) as a judge for Open-ended VQA. For Open-ended QA, we additionally compute chrF++ (Popović, 2015) scores and present them in Appendix C. We report Word Error Rate (WER) for ASR.

Human Evaluation: We evaluate 50 randomly sampled questions per language on the best-performing model per family. Bilingual native speakers rated whether model outputs matched the gold answer or were valid alternatives.

6 Results

In this Section, we present results organized by evaluation type, beginning with cultural VQA performance across MC-VQA and Open-ended VQA tasks, followed by control experiments.

6.1 Cultural VQA Results

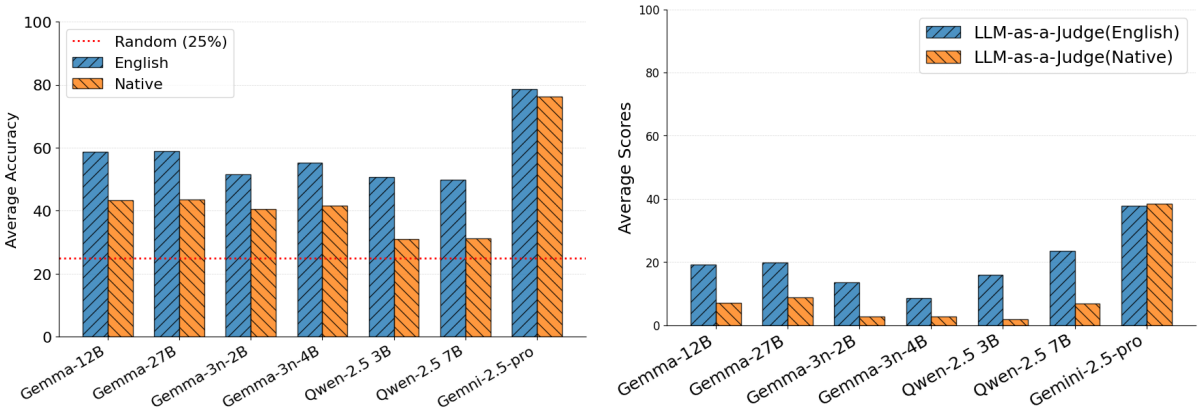
We show evaluations on MC-VQA and Open-ended VQA tasks, comparing performance across text and audio modalities in both English and native languages.

6.1.1 Text-based QA

Figure 3 depicts MC-VQA average accuracy (3(a)) and LLM-as-a-judge scores (3(b)) for open-ended QA in both English and native African languages using location-aware prompts.

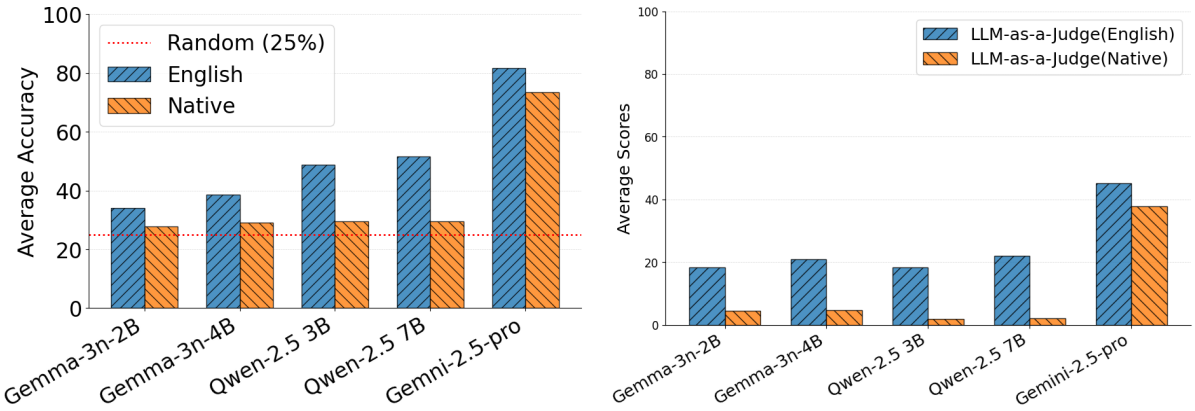
Open-weight models consistently perform better when the question is in English compared to native languages across both tasks. We also observe a significant performance gap between MC-VQA and Open-ended QA, where all models, including Gemini-2.5 Pro, show major drops when tasked with answering in an open-ended setting, even in English. This suggests that *generating culturally grounded responses is more challenging than selecting from predefined options*, particularly for native-language queries, where performance degrades significantly compared to English queries.

We also observe that increasing model size among open-weight models does not necessarily translate to improved performance in low-resource languages. Larger variants show limited or no improvement over smaller counterparts in native language QA, indicating that model scaling alone is



(a) MC-VQA (Text) (b) Open-ended VQA (Text)

Figure 3: Performance comparison of models on text-based question answering tasks: (a) Text MC-VQA (Multiple Choice) and (b) Text Open-ended QA in English and Native languages.



(a) MC-VQA (Audio) (b) Open-ended VQA (Audio)

Figure 4: Performance comparison of models on audio-based question answering tasks: (a) Audio MC-VQA (Multiple Choice) and (b) Audio Open-ended VQA in English and Native languages.

insufficient to address low-resource challenges. Notably, some smaller models achieve near-zero accuracy for native languages for Open-ended QA. In contrast, Gemini-2.5-Pro outperforms all open-weight models across both tasks while maintaining comparable performance between English and native languages, *highlighting the current gap between proprietary and open-weight models*. Human evaluations on Open-Ended VQA (in Figure 5), conducted on a random subset of 50 samples across eight languages, are consistent with the trends observed in automatic evaluations, with Gemini-2.5 Pro performs the best when queried in native languages compared to English.

6.1.2 Audio-based QA

Figure 4 shows MC-VQA accuracy (4(a)) and LLM-as-a-judge scores for Open-ended QA (4(b)) on audio inputs in both English and native African languages. Similar to text-based evaluation, open-weight models achieve higher performance when queried in English compared to native languages across both tasks. For open-weight models, *audio modality is significantly more difficult than text modality*, with notable performance degradation across both tasks. Gemini-2.5-Pro, however, demonstrates robust multimodal capabilities, with consistent performance across modalities. The performance drop from MC-VQA to Open-ended QA is also noticeable in the audio modality, with nearly zero accuracy in spoken native languages. These results align with our LID and ASR analyses (Section 6.2.2). Open-weight models demonstrate poor LID capabilities, especially the Qwen variants, which exhibit near-random accuracy and significantly compromised native ASR performance compared to English. These failures also affect performance models in downstream tasks, such as

Open-ended VQA. Similar to text evaluations, we find that scaling up model size shows little improvement in native language understanding.

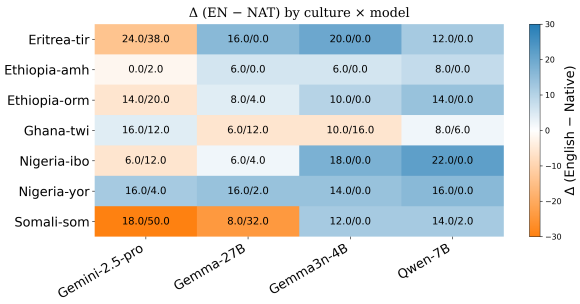


Figure 5: **Human Evaluation for Text Open-ended VQA.** Accuracy across best-performing models for English and Native. We observed that, while most models perform best in the English setting, Gemini-2.5 Pro seems to perform better in the Native language.

6.2 Results for Control Experiments

We present the results on control experiments on established benchmarks alongside our cultural QA task, aiming to probe the linguistic competence of the tested models and provide pointers on why models fail on Afri-MCQA.

6.2.1 Text-based Experiments

In Table 3, we report the performance of the models on the control-experiment benchmarks in addition to Afri-MCQA. All models showed performance degradation from English to native languages across all benchmarks. Gemini-2.5-Pro maintains the smallest gaps, particularly on Afri-MCQA, where the drop is minimal. Open-weight models show a big drop, with Qwen variants showing the most severe gaps on AfriXNLI and AfriMMLU.

When comparing AfriMMLU with Afri-MCQA, open-weight models show considerably higher AfriMMLU scores in English than Afri-MCQA scores, suggesting that *while models possess general factual knowledge, they lack Africa-specific cultural understanding*. Gemini-2.5-Pro shows a smaller gap between these benchmarks, indicating it has acquired more African cultural knowledge during training. AfriXNLI exposes the most severe cross-lingual gaps, particularly for Qwen models, suggesting that linguistic reasoning tasks are more sensitive to language resource availability than factual retrieval tasks.

We additionally compute Spearman rank correlations between Afri-MCQA and our text control datasets (see Appendix D). We observe strong and statistically significant correlations between AfriXNLI and AfriMMLU performance for several models. In contrast, correlations between Afri-MCQA and AfriXNLI or AfriMMLU are generally weaker and not statistically significant. Hence, while limitations in language understanding may play a large role in explaining *why* models fail in Afri-MCQA, other factors related to African cultural and visual knowledge may be important as well.

	AfriXNLI			AfriMMLU			Afri-MCQA		
Model	Eng	Nat	Δ	Eng	Nat	Δ	Eng	Nat	Δ
Gemini-2.5-Pro	89.66	76.3	-13.36	94	83.46	-10.54	78.68	76.27	-2.4
Gemma3n-4B	78.33	51.24	-27.09	62.2	37.56	-24.64	55.23	41.49	-13.7
Gemma3n-2B	80.66	51.47	-29.19	53.6	35.74	-17.86	51.54	40.32	-11.2
Qwen-7B	81.11	34.33	-46.78	73.8	36.10	-37.70	49.75	31.30	-18.46
Qwen-3B	65.66	36.7	-28.96	65.8	34.72	-31.08	50.68	31.00	-19.65

Table 3: Text-based performance (%). Eng = English accuracy, Nat = Native accuracy, Δ = English − Native gap (negative = drop).

6.2.2 Audio-based Experiments

Figure 6 shows three evaluations on native African language audio. **LID:** Gemini-2.5-Pro achieves near-perfect LID accuracy, while Gemma variants show moderate performance. Qwen models perform at near random levels, indicating minimal exposure to African language audio during pretraining. **ASR Performance:** WER shows a similar pattern. Gemini-2.5-Pro maintains reasonable native ASR, whereas Gemma models exhibit substantial degradation. Qwen models produce high error rates, indicating hallucinations rather than meaningful transcription. This aligns with observed results for open-ended VQA (Audio). Gemini-2.5-Pro achieves moderate accuracy, whereas open-weight models score near zero. In light of these results, *it is likely that models fail on open-ended VQA (Audio) because they cannot properly identify or transcribe the tested spoken languages*.

Hence, poor Open-ended VQA (Audio) results come from errors at each step: **(1)** Open models often fail at identifying African languages, reflecting fundamental gaps in audio representation, **(2)** they show high ASR errors, suggesting that most spoken content is not perceived adequately for understanding and answering a question **(3)** these errors can carry over to open-ended VQA when models receive wrong transcriptions, as they cannot answer correctly even if they know the cultural content. These findings demonstrate that *foundational speech processing capabilities are prerequisites for meaningful evaluation of cultural reasoning in African languages*.

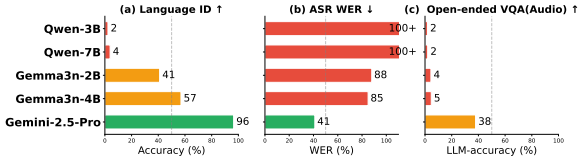


Figure 6: Audio probing results on native African languages: (a) LID (↑ higher is better), (b) ASR (↓ lower is better), and (c) Open-ended VQA (↑ higher is better). (+) means WER is more than 100%

7 Discussion

We now organize our findings to address each research question, summarizing key patterns observed across models, modalities, and languages.

RQ1: How well do MLLMs understand African cultural contexts in visually-grounded QA? MLLMs show limited understanding of African cultural contexts. As shown in Figures 3 and 4, even the best performing model (Gemini-2.5-Pro) achieves only 78% on MC-VQA and 38% on Open-VQA (text-based, English). Smaller models (Gemma3, Qwen2.5) perform substantially worse, ranging from 50–59% on MC-VQA and 9–24% on Open-VQA, indicating significant room for improvement.

RQ2: How does input modality (text vs. speech) affect performance? Performance degrades when switching from text to speech input, particularly for smaller models. As shown in Figures 3 and 4, on MC-VQA, Qwen models drop by approximately 1–2% while Gemma models show mixed results. The gap is more visible in Open-VQA, where audio-based native language queries yield near-zero accuracy (2–5%) for smaller models. Control experiments for audio show that this comes from poor language identification (2–4% for Qwen) and high ASR word error rates (85–100%+ for non-Gemini models).

RQ3: How does query language (native vs. English) affect performance, and do differences reflect language understanding or cultural knowledge gaps? English queries consistently outperform native language queries across all models and settings. As shown in Figure 3, the gap ranges from 2% (Gemini-2.5-Pro on MC-VQA) to 19% (Qwen on MC-VQA). Control experiments on AfriXNLI and AfriMMLU (Table 3) show that language understanding gaps (Δ = 13–47%) are substantially larger than cultural knowledge gaps (Δ = 2–19% on Afri-MCQA), suggesting language understanding is the dominant limitation. However, models also struggle with cultural QA in English, indicating that both linguistic and cultural limitations contribute to poor performance.

RQ4: How does task format (Multiple-Choice vs. Open QA) affect accuracy? As shown in Figures 3 and 4, models perform significantly better on MC-VQA than Open-VQA. Gemini-2.5-Pro achieves 78% on MC-VQA vs. 38% on Open-VQA (a 40% gap). Smaller models show even larger relative drops, with some scoring less than 10% on Open-VQA. This performance gap suggests that MC-VQA benefits from simplified answer selection, while Open-VQA exposes true limitations. Models struggle to generate culturally grounded responses even in English, with performance degrading further for native languages.

8 Conclusion

We introduced Afri-MCQA, the first large-scale multilingual and multimodal benchmark for African cultural visual QA, covering 15 languages across 12 countries with over 7.5k Q&A pairs in text and speech modalities. Our evaluation shows: 1) MLLMs struggle significantly with African cultural knowledge, 2) speech processing presents a critical bottleneck, and 3) gaps persist across languages and task formats.

These findings motivate several research directions: (1) **speech-first approaches:** Many African languages are primarily oral, yet current open-weight models lack basic LID and ASR capabilities for these languages; (2) **culturally-grounded pretraining:** The gap between AfriMMLU and Afri-MCQA performance suggests language data alone is insufficient; models need explicit exposure to African cultural content; and (3) **cross-lingual cultural transfer:** Models may "know" cultural facts in English but cannot access them through native language queries, motivating research into cross-lingual knowledge retrieval.

We will release Afri-MCQA to provide both a benchmark and a foundation for building more inclusive, culturally aware multimodal systems that better represent African languages and cultures.

9 Limitations

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 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.

10 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. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3814–3841, Dublin, Ireland. Association for Computational Linguistics.

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. 2024. 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 F. P. Dossou, Akintunde Oladipo, Doreen Nixdorf, Chris Chinenye Emezue, Sana Al-azzawi, Blessing Sibanda, Davis David, Lolwethu Ndolela, Jonathan Mukiibi, Tunde Ajayi, Tatiana Moteu, Brian Odhiambo, and 46 others. 2023. MasakhaNEWS: News topic classification for African languages. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 144–159, Nusa Dua, Bali. Association for Computational Linguistics.

David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba Oluwadara Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, Chiamaka Ijeoma Chukwuneke, Happy Buzaaba, Blessing Kudzaishe Sibanda, Godson Koffi Kalipe, Jonathan Mukiibi, Salomon Kabongo Kabenamualu, Foutse Yuehgoh, Mmasibidi Setaka, Lolwethu Ndolela, and 8 others. 2025. IrokoBench: A new benchmark for African languages in the age of large language models. 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 2732–2757, Albuquerque, New Mexico. Association for Computational Linguistics.

Jesujoba Oluwadara Alabi, Michael A. Hedderich, David Ifeoluwa Adelani, and Dietrich Klakow. 2025. Charting the landscape of African NLP: Mapping progress and shaping the road ahead. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27807–27841, Suzhou, China. Association for Computational Linguistics.

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.

Israel Abebe Azime, Atnafu Lambebo Tonja, Tadesse Destaw Belay, Mitiku Yohannes Fuge, Aman Kassahun Wassie, Eyasu Shiferaw Jada, Yonas Chanie, Walelign Tewabe Sewunetie, and Seid Muhie Yimam. 2024. Walia-LLM: Enhancing Amharic-LLaMA by integrating task-specific and generative datasets. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 432–444, Miami, Florida, USA. 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*.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, and 196 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021a. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021b. Visually grounded reasoning across languages and cultures. *arXiv preprint arXiv:2109.13238*.

Rada Mihalcea, Oana Ignat, Longju Bai, Angana Borah, Luis Chiruzzo, Zhijing Jin, Claude Kwizera, Joan Nwatu, Soujanya Poria, and Thamar Solorio. 2025. Why ai is weird and shouldnt be this way: Towards ai for everyone, with everyone, by everyone.

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 Wibowo, Santiago Góngora, Aishik Mandal, Sukannya Purkayastha, Jesus-German Ortiz-Barajas, Emilio Cueva, Jinheon Baek, Soyeong Jeong, and 1 others. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *Advances in Neural Information Processing Systems*, 37:11479–11505.

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*.

Atnafu Lambebo Tonja, Israel Abebe Azime, Tadesse Destaw Belay, Mesay Gemeda Yigezu, Moges Ahmed Ah Mehamed, Abinew Ali Ayele, Ebrahim Chekol Jibril, Michael Melese Woldeyohannis, Olga Kolesnikova, Philipp Slusallek, and 1 others. 2024a. Ethiollm: Multilingual large language models for ethiopian languages with task evaluation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6341–6352.

Atnafu Lambebo Tonja, Bonaventure FP Dossou, Jessica Ojo, Jenalea Rajab, Fadel Thior, Eric Peter Wairagala, Anuoluwapo Aremu, Pelonomi Moiloa, Jade Abbott, Vukosi Marivate, and 1 others. 2024b. Inkubalm: A small language model for low-resource african languages. *arXiv preprint arXiv:2408.17024*.

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 Image Categories

A) Language wise category distribution

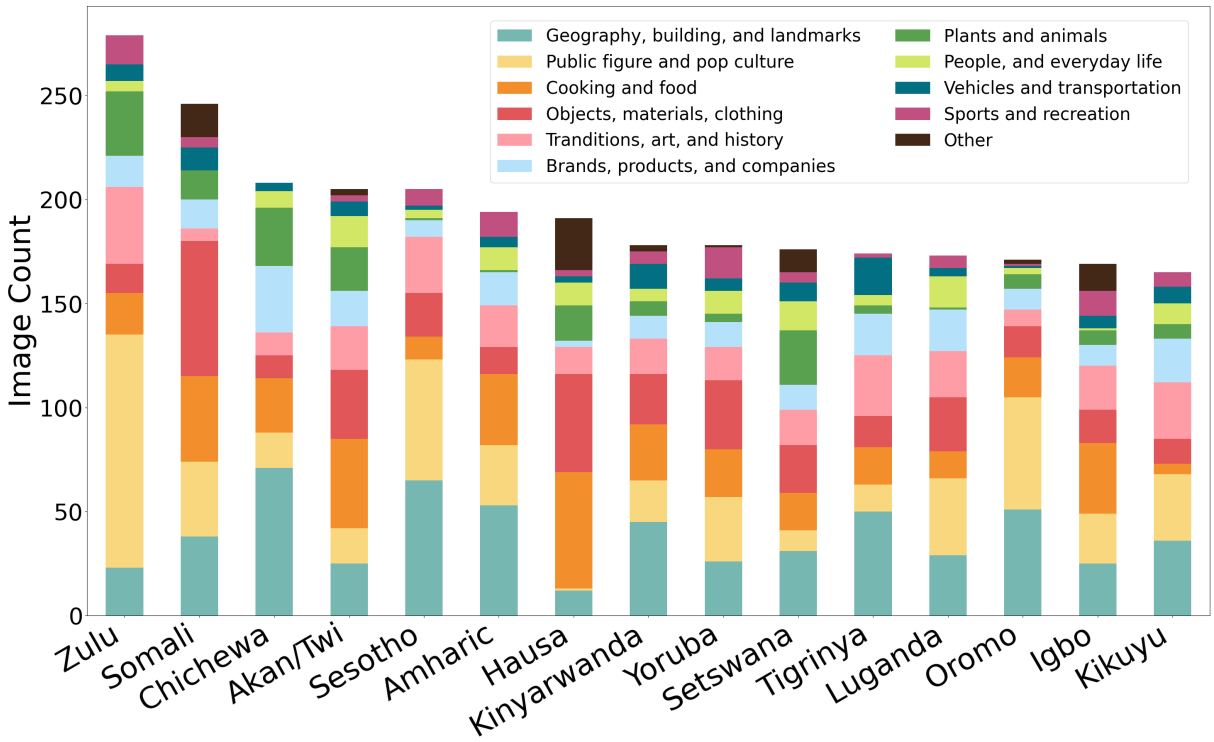


Figure 7: Language-wise distribution of the categories.

B) Image category distribution

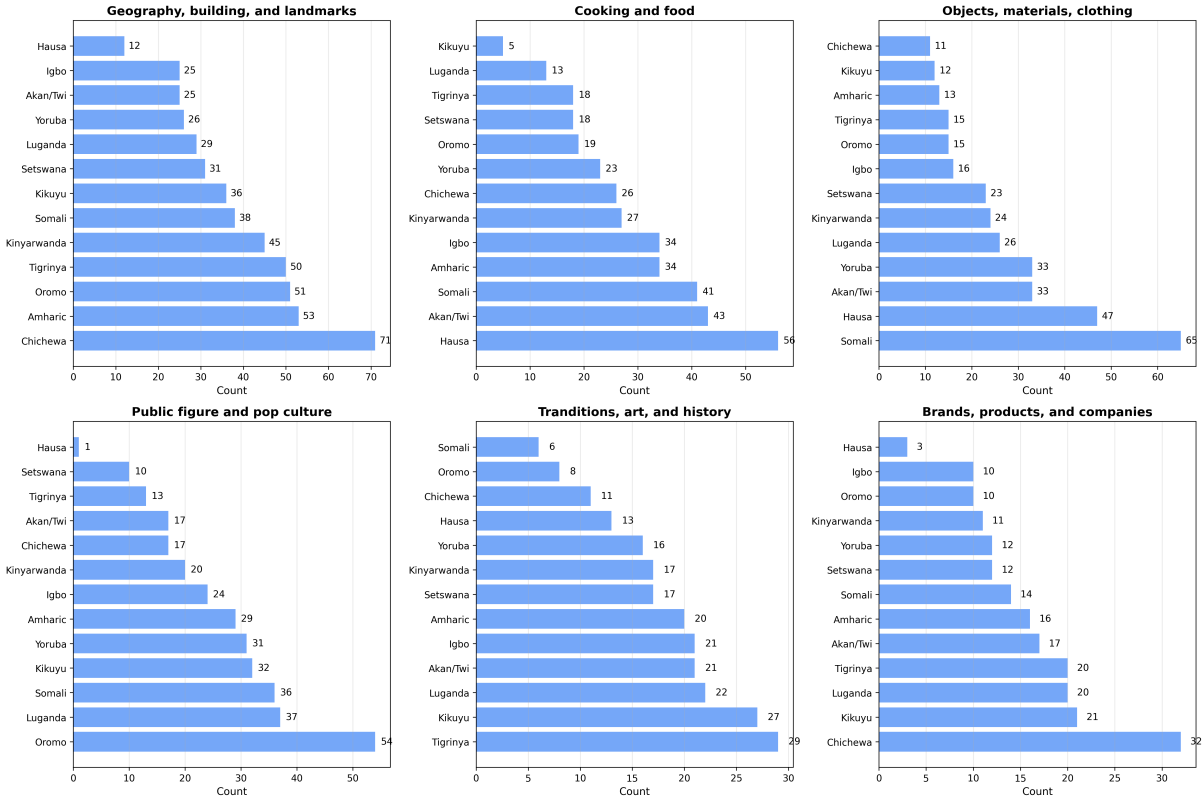


Figure 8: Top 6 Image category distribution

B Prompt

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

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

Figures 9 & 10 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 9: 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 10: Location-aware audio prompts.

C Additional Experiments results

C.1 Text-based QA

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

C.2 Language wise heatmap for MCQA-text

D Correlation statistics

Country-Lang	G-12B	G-27B	G-3n-2B	G-3n-4B	Q-2.5 3B	Q-2.5 7B	Gemni-2.5-pro
Ethiopia-amh	50.85	50.34	41.23	40.54	42.61	41.23	76.11
Nigeria-hau	63.17	62.47	55.14	58.21	54.79	57.2	81.16
Nigeria-ibo	50.19	49.8	43.72	40.39	41.37	42.15	79.8
Uganda-lug	61.65	62.23	56.92	60.48	49.32	53.75	84.15
Ethiopia-orm	55.03	54.48	44.76	51.5	45.93	45.73	76.35
Rwanda-kin	67.97	68.16	56.92	60.48	55.05	55.61	82.27
Kenya-kik	58.83	58.38	47.67	54.34	46.63	47.67	87.87
Somali-som	56.36	56.23	50	54.33	50.27	45.52	68.02
Eritrea-tir	57.54	57.89	45.05	51.22	50.17	50.52	76.66
Ghana-twi	62.99	63.15	56.93	62.27	54.7	56.14	81.33
Nigeria-yor	55.61	55.24	49.9	53.63	48.68	46.62	82.95
Botswana-tsn	60.41	61.74	50.94	55.3	53.21	53.03	82.54
Malawi-nya	47.91	47.43	40.7	43.42	41.98	37.98	66.02
S.Africa-zul	58.9	59.97	59.13	58.66	55.55	51.49	67.38
Lesotho-sot	74.47	74.63	74.63	74.47	70.08	61.62	87.64
Average	58.79	58.80	51.54	55.23	50.68	49.75	78.66

Table 4: Text Prompt - Location/Language Aware–English

Country-Lang	G-12B	G-27B	G-3n-2B	G-3n-4B	Q-2.5 3B	Q-2.5 7B	Gemni-2.5-pro
Ethiopia-amh	44.5	43.64	42.09	42.78	30.93	29.38	78.17
Nigeria-hau	47.12	49.38	40.31	42.58	28.44	29.12	81.67
Nigeria-ibo	41.17	41.56	38.43	36.86	30.39	31.56	79.41
Uganda-lug	48.36	48.16	44.89	43.54	35.26	35.26	81.31
Ethiopia-orm	39.34	38.56	35.65	36.24	33.13	34.49	75.77
Rwanda-kin	57.11	58.05	53.55	53.18	39.7	39.13	82.58
Kenya-kik	44.42	44.64	40.6	41.41	35.25	36.76	85.05
Somali-som	40.37	41.19	43.22	42.68	27.5	25.06	66.12
Eritrea-tir	47.01	46.84	41.57	40.52	32.28	34.56	75.96
Ghana-twi	45.13	45.61	37.79	37.48	29.5	29.34	77.51
Nigeria-yor	44.19	43.82	38.36	38.36	31.08	31.27	81.27
Botswana-tsn	39.77	40.15	36.93	37.87	30.3	32.38	83.52
Malawi-nya	36.85	35.73	34.15	53.18	28.36	28.68	68.58
S.Africa-zul	38.23	39.9	39.18	39.42	27.8	22.46	61.88
Lesotho-sot	37.39	37.07	38.21	36.26	25.56	29.91	65.36
Average	43.39	43.62	40.49	41.49	31.03	31.29	76.27

Table 5: Text Prompt - Location/Language Aware–native

Country-Lang	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 6: Prompt - Image only–English(text-based)

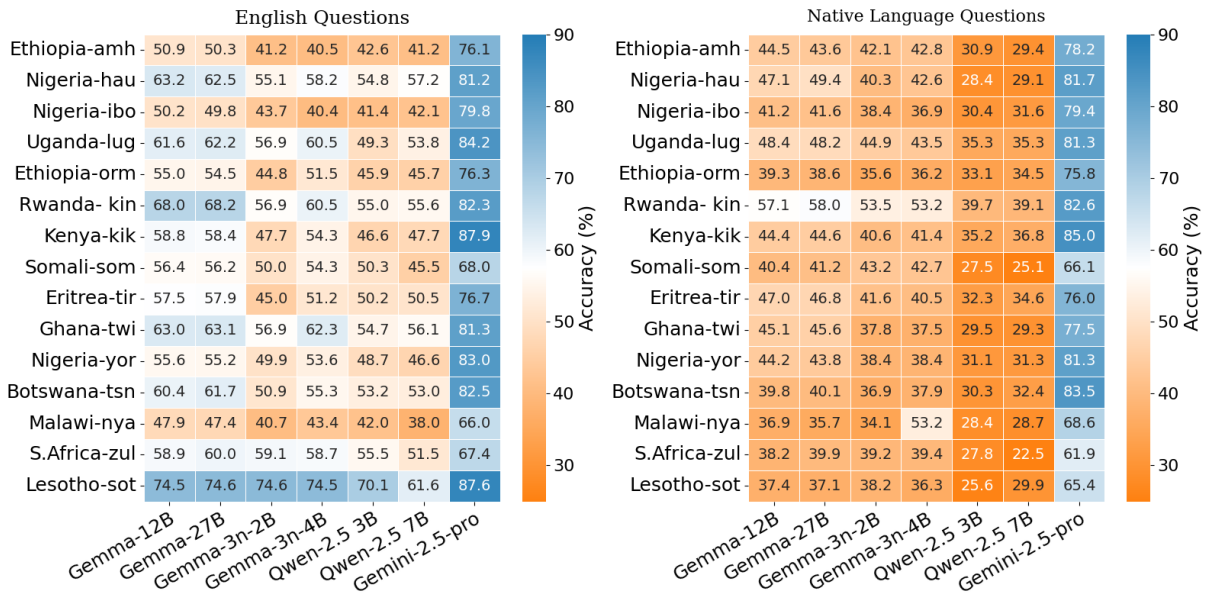


Figure 11: Language-wise accuracy of seven multimodal LLMs on the text-based VQA task, evaluated separately on English and native African language questions.

Country-Lang	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 7: Image only-native (text-based)

Country-Lang	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	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	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	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	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 8: Location/Language Aware-English(Open-endded VQA -text based)

Country-Lang	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 9: Location/Language Aware-native(Open-ended VQA -text based)

Country-Lang	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 10: Image only-english(Open-endded VQA -text based)

Country-Lang	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	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	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	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 11: Image only-native(Open-endded VQA -text based)

	English					Native				
Country-Lang	**Gemma3n-2B**	**Gemma3n-4B**	**Qwen-3B**	**Qwen-7B**	**Gemini-2.5 Pro**	**Gemma3n-2B**	**Gemma3n-4B**	**Qwen-3B**	**Qwen-7B**	**Gemini-2.5 Pro**
Ghana-twi	37.78	44.35	58.96	62.83	87.01	28.83	32.11	31.48	31.27	69.12
Ethiopia-amh	34.38	36.17	42.43	43.60	83.52	29.48	32.97	30.03	28.63	82.46
Malawi-nya	27.60	33.17	42.96	44.30	80.09	26.52	27.98	29.62	32.08	70.80
Nigeria-hau	36.96	41.46	62.74	62.53	83.61	27.95	32.29	29.93	27.51	82.32
Nigeria-ibo	26.85	30.09	40.60	44.50	77.15	27.18	26.18	27.15	26.08	56.50
Kenya-kik	27.12	28.72	44.41	45.98	83.28	26.86	25.00	31.90	33.15	72.34
Rwanad-kin	45.78	51.40	59.07	63.06	85.16	30.69	32.07	29.31	31.44	78.13
Uganda-lug	39.64	46.10	47.65	50.56	81.73	33.87	31.81	34.39	38.38	72.48
Ethiopia-orm	37.55	39.20	41.23	42.35	74.52	29.11	28.87	26.49	27.06	71.46
Botswana-tsn	35.86	40.92	51.06	54.60	83.15	25.05	26.11	27.17	19.70	74.46
Somali-som	26.24	30.73	49.17	56.05	82.14	24.76	26.42	24.88	24.14	75.41
Eritrea-tir	36.55	44.30	46.98	52.30	75.48	28.27	29.76	31.26	34.33	72.35
Nigeria-yor	31.17	35.49	45.56	47.58	84.13	23.64	25.53	26.31	26.68	76.54
Lesotho-sot	–	–	–	–	–	29.77	30.00	34.00	31.12	59.15
S.Africa-zul	–	–	–	–	–	27.04	26.63	28.89	31.06	74.18
Average	**34.11**	**38.62**	**48.68**	**51.56**	**81.61**	**27.86**	**29.01**	**29.52**	**29.51**	**73.41**

Table 12: Audio-Loc-Lang Aware(MC-VQA)

	English					Native				
Country-Lang	**Gemma3n-2B**	**Gemma3n-4B**	**Qwen-3B**	**Qwen-7B**	**Gemini-2.5 Pro**	**Gemma3n-2B**	**Gemma3n-4B**	**Qwen-3B**	**Qwen-7B**	**Gemini-2.5 Pro**
Ghana-twi	33.67	46.40	57.23	62.19	87.75	27.61	24.74	32.55	32.51	65.16
Ethiopia-amh	31.01	35.05	45.18	44.58	79.72	28.17	28.38	31.66	29.91	79.69
Malawi-nya	25.18	32.44	41.81	48.50	77.72	28.71	29.44	37.07	26.86	68.87
Nigeria-hau	36.49	42.18	63.30	65.41	84.28	27.95	30.60	25.24	24.80	81.36
Nigeria-ibo	23.37	27.08	37.65	39.71	76.62	27.93	23.44	21.85	27.17	55.50
Kenya-kik	25.26	30.31	45.65	45.96	83.73	25.53	25.00	27.74	35.04	71.81
Rwanda-kin	44.24	48.59	57.54	61.33	81.79	28.28	28.97	30.83	31.00	79.24
Uganda-lug	36.74	44.09	45.49	45.94	78.65	32.49	31.12	35.64	35.20	69.27
Ethiopia-orm	34.50	32.62	37.40	42.50	71.15	27.68	28.16	26.01	26.78	71.22
Botswana-tsn	35.86	40.29	50.85	54.36	85.31	23.99	25.69	27.89	25.00	70.82
Somali-som	24.11	29.78	49.64	55.34	80.42	26.65	22.17	25.30	24.64	72.64
Eritrea-tir	33.11	36.34	47.29	47.67	74.51	26.98	27.19	32.10	31.97	70.39
Nigeria-yor	23.50	35.25	47.07	48.92	85.99	24.82	25.77	28.01	26.68	77.96
Lesotho-sot	–	–	–	–	–	–	–	31.12	28.29	59.11
S.Africa-zul	–	–	–	–	–	–	–	29.04	32.09	72.07
Average	**31.31**	**36.96**	**48.16**	**50.02**	**80.59**	**27.45**	**26.98**	**29.47**	**29.20**	**71.01**

Table 13: MC-VQA Audio - Image Only

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	19.21	29.36	21.61	31.01	19.16	29.98	21.11	33.06	31.92	45.59
Ethiopia-amh	17.86	18.65	27.61	24.94	18.5	19.05	21.14	22.56	49.39	64.04
Malawi-nya	13.05	18.16	14.55	20.82	13.05	21.22	18.04	24.81	27.9	35.35
Nigeria-hau	13.02	14.89	11.34	18.68	13.26	14.42	17.28	19.76	17.99	31.21
Nigeria-ibo	12.31	6.94	12.09	9.03	11.26	7.89	14.48	12.6	30.9	42.13
Kenya-kik	17.87	14.47	18.89	18.16	15.86	14.47	19.21	18.23	44.58	55.26
Rwanda-kin	15.99	17.14	15.09	17.39	15.59	16.25	16.49	19.48	25.82	43.99
Uganda-lug	16.37	18.71	19.13	21.60	16.14	19.6	17.09	20.33	30.9	52.34
Ethiopia-orm	18.72	17.80	23.03	22.25	17.02	15.49	20.62	20.54	38.46	45.43
Botswana-tsn	18.15	24.10	19.27	25.16	16.48	23.33	17.61	23.56	26.79	36.81
Somali-som	14.39	21.51	16.39	23.17	15.02	18.29	17.54	23.99	23.05	32.39
Eritrea-tir	16.3	21.08	17.89	21.94	15.93	21.59	19.39	27.61	35.64	53.33
Nigeria-yor	11.55	16.31	11.91	18.71	11.3	16.63	12.81	18.99	25.07	49.64
Average	15.75	18.39	17.58	20.99	15.27	18.31	17.91	21.96	32.11	45.19

Table 14: Audio Open-ended for English (Loc-Lang Aware)

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 15: Audio Open-ended for Native Audio(Loc-Lang Aware)

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	18.91	28.54	21.52	29.98	19.30	29.36	20.01	29.55	35.14	57.49
Ethiopia-amh	13.41	10.56	15.65	13.26	13.42	11.24	14.90	15.04	40.36	49.44
Malawi-nya	11.83	9.44	11.71	9.69	12.57	14.58	13.30	14.13	25.41	32.69
Nigeria-hau	11.58	13.95	11.02	13.48	12.67	14.89	15.60	15.73	23.14	37.35
Nigeria-ibo	11.36	4.86	11.66	6.25	12.72	5.79	12.72	6.02	26.90	34.95
Kenya-kik	16.33	13.68	16.77	16.58	15.60	13.68	18.32	17.24	44.25	57.11
Rwanda-kin	14.19	13.81	13.96	12.02	15.71	14.87	15.41	14.92	25.34	33.5
Uganda-lug	14.81	16.7	16.32	19.38	15.25	19.15	15.76	18.18	37.54	49.22
Ethiopia-orm	14.36	9.6	17.3	11.94	13.85	10.07	15.75	14.67	38.42	42.62
Botswana-tsn	15.82	20.93	17.02	23.89	14.70	22.03	18.99	24.6	34.18	49.89
Somali-som	13.62	19.39	14.7	22.22	14.63	20.33	16.56	22.8	25.76	37.35
Eritrea-tir	12.36	8.39	12.33	10.11	12.37	8.62	12.62	10.87	29.77	39.57
Nigeria-yor	10.67	13.91	10.65	14.39	10.81	13.43	11.40	17.46	27.21	49.4
Average	13.79	14.13	14.66	15.63	14.12	15.25	15.49	17.01	31.80	43.89

Table 16: Audio Open-ended for English (Image only Audio)

Language	**Native**									
	G3n-2B		**G3n-4B**		**Q-3B**		**Q-7B**		**Gemini-2.5 Pro**	
	chrF++	LLM-acc	chrF++	LLM-acc	chrF++	LLM-acc	chrF++	LLM-acc	chrF++	LLM-acc
Ghana-twi	6.96	2.45	6.89	4.70	9.83	6.75	9.99	5.52	15.72	18.40
Ethiopia-amh	0.08	3.28	0.15	6.11	6.74	1.31	5.27	1.09	28.23	58.52
Malawi-nya	8.44	2.68	7.75	2.43	9.68	3.65	9.38	2.43	20.66	21.90
Nigeria-hau	7.45	4.58	7.35	3.86	10.69	3.86	9.85	4.58	22.09	36.63
Nigeria-ibo	6.39	0.25	5.51	1.00	7.23	0.50	6.64	0.75	11.23	15.25
Kenya-kik	9.07	2.13	9.14	2.93	9.03	2.13	9.09	1.86	20.14	23.73
Rwanda-kin	8.11	5.17	7.55	5.17	9.84	5.86	9.47	3.79	25.57	41.03
Uganda-lug	8.21	4.35	7.51	5.49	5.94	2.75	5.26	1.37	24.11	31.65
Ethiopia-orm	9.04	6.44	8.46	8.11	4.67	0.72	8.15	0.95	27.87	40.19
Botswana-tsn	6.97	2.12	6.89	3.40	9.69	4.67	9.89	3.82	24.55	31.62
Somali-som	6.65	4.25	5.81	7.55	11.33	10.38	11.52	10.14	17.59	29.48
Eritrea-tir	0.02	3.21	0.04	4.28	4.67	0.43	5.35	0.00	15.71	36.70
Nigeria-yor	4.18	1.42	4.00	3.31	6.84	2.36	7.39	2.84	15.73	35.63
Lesotho-sot	9.01	2.89	8.41	4.22	8.70	1.56	8.94	2.00	13.68	20.67
S.Africa-zul	9.65	5.12	10.54	6.76	7.88	2.66	7.67	2.25	34.04	40.37
Average	6.68	3.36	6.40	4.62	8.18	3.31	8.26	2.89	21.13	32.12

Table 17: Audio Open-ended for Native (Image only Audio)

	Gemma3n-2B	Gemma3n-4B	Qwen2.5-3B	Qwen2.5-7B	Gemini-2.5 Pro
XNLI - MMLU	0.684 (p=0.029)	0.88 (p=0.001)	0.254 (p=0.479)	0.426 (p=0.22)	0.831 (p=0.003)
XNLI - Afri-MCQA	0.261 (p=0.498)	0.636 (p=0.065)	-0.388 (p=0.302)	0.316 (p=0.407)	-0.248 (p=0.519)
MMLU - Afri-MCQA	0.373 (p=0.288)	0.463 (p=0.178)	0.236 (p=0.511)	0.479 (p=0.162)	-0.222 (p=0.537)

Table 18

E 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)**:

Image category *

- [ ] Vehicles and transportation
- [ ] Objects, materials, clothing
- [ ] Plants and animals
- [ ] Sports and recreation
- [ ] People, and everyday life
- [ ] Other
- [ ] Cooking and food
- [ ] Geography, building, and landmarks
- [ ] Brands, products, and companies
- [ ] Tranditions, art, and history
- [ ] Public figure and pop culture

- 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.

F Annotation guideline

- 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

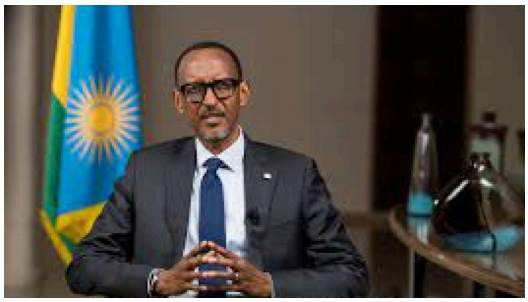
When formulating questions, include the following complexities

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

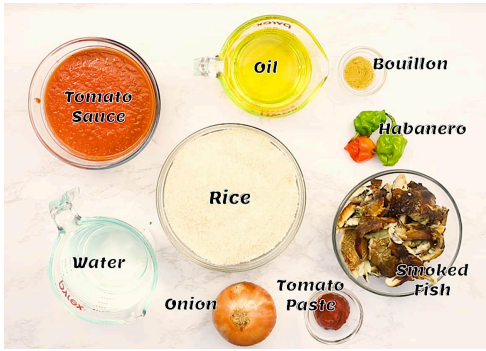
- Multi-hop (e.g., Q: What is the name of the brand that the person in the picture is leading as CEO? A: Nando's B: Shoprite C: Woolworths D: Checkers)

H Annotation guideline

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.

J 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 19: Annotator Demographics

K 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

L Review Guidelines

- **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

N Review Guidelines

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

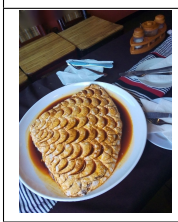
- **Avoid Repetition**: Similar questions across different images
- **Varied Complexity**: Mix of simple and complex questions
- **Different Question Types**: Not all "What is..." questions
- **Topic Distribution**: Balanced across cultural categories

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

O Examples

Table 20: 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