HaVQA: A Dataset for Visual Question Answering and Multimodal Research in Hausa Language

Shantipriya Parida[1], **Idris Abdulmumin**[2,4], **Shamsuddeen Hassan Muhammad**[3,4], **Aneesh Bose**[5,10], **Guneet Singh Kohli**[6,10], **Ibrahim Said Ahmad**[3,4], **Ketan Kotwal**[7], **Sayan Deb Sarkar**[8,10], **Ondřej Bojar**[9], **Habeebah Adamu Kakudi**[3]

[1]Silo AI, Finland, [2]Ahmadu Bello University, Zaria, Nigeria, [3]Bayero University Kano, Nigeria, [4]HausaNLP, [5]Microsoft, India, [6]Thapar University, India, [7]Idiap Research Institute, Switzerland, [8]ETH Zurich, Switzerland, [9]Charles University, MFF ÚFAL, Czech Republic, [10]CORD.ai, India

correspondence: shantipriya.parida@silo.ai, iabdulmumin@abu.edu.ng

Abstract

This paper presents HaVQA, the first multimodal dataset for visual question-answering (VQA) tasks in the Hausa language. The dataset was created by manually translating 6,022 English question-answer pairs, which are associated with 1,555 unique images from the Visual Genome dataset. As a result, the dataset provides 12,044 gold standard English-Hausa parallel sentences that were translated in a fashion that guarantees their semantic match with the corresponding visual information. We conducted several baseline experiments on the dataset, including visual question answering, visual question elicitation, text-only and multimodal machine translation.

1 Introduction

In recent years, multidisciplinary research, including Computer Vision (CV) and Natural Language Processing (NLP), has attracted many researchers to the tasks of image captioning, cross-modal retrieval, visual common-sense reasoning, and visual question answering (VQA, Antol et al., 2015; Goyal et al., 2017). In the VQA task, given an image and a natural language question related to the image, the objective is to produce a correct natural language answer as output (Kafle and Kanan, 2017; Ren et al., 2015a). VQA is one of the challenging tasks in NLP that requires a fine-grained semantic processing of both the image and the question, together with visual reasoning for an accurate answer prediction (Yu et al., 2019b). The general approaches followed by existing VQA models include *i)* extracting features from the questions and the images, and *ii)* utilizing the features to understand the image content to infer the answers.

Recently, research on improving visual question-answering systems using multimodal architectures and sentence embeddings (Kodali and Berleant, 2022; Urooj et al., 2020; Gupta et al., 2020; Pfeiffer et al., 2021) has seen tremendous growth. However, most of the VQA datasets used in the VQA research are limited to the English language (Kafle and Kanan, 2017). Although the accuracy of the VQA systems for English improved significantly with the advent of Transformer-based models (e.g., BERT, Devlin et al., 2018), the lack of VQA datasets has restricted the development of such systems in most languages, especially the low-resource ones (Kumar et al., 2022).

The availability of original datasets for state-of-the-art natural language processing tasks has since been appreciated, especially on the African continent. While some of the efforts to create such datasets for African languages are supported by funding such as Facebook's Translation Support for African Languages, the Lacuna Fund[1], and many others, including the dataset in this work are driven by the enthusiasm for developing quality NLP solutions for African languages that are useful to the local communities.

Contributions: The main contribution of this work is building a multimodal dataset (HaVQA) for the Hausa language, consisting of question-answer pairs along with the associated images and is suitable for many NLP tasks. As per our knowledge, HaVQA is the first VQA dataset for Hausa language, and will enrich Hausa natural language processing (NLP) resources, allowing researchers to conduct VQA and multimodal research in Hausa.

2 Related work

2.1 Datasets for African NLP

African languages are low-resourced; many do not have any datasets for everyday NLP tasks. While some datasets exist for some African languages, they are often limited in scope or lack the necessary quality (Kreutzer et al., 2022). For Visual

[1]https://lacunafund.org/

Findings of the Association for Computational Linguistics: ACL 2023, pages 10162–10183
July 9-14, 2023 ©2023 Association for Computational Linguistics

Question Answering, no publicly available dataset exists in any African language. Luckily, there has been a recent surge in efforts by researchers to create datasets for African languages. In this section, we provide an overview of some of the most recent examples of such efforts.

Abdulmumin et al. (2022) created the multi-modal Hausa Visual Genome dataset for machine translation and image captioning. Adelani et al. (2022a) created the MAFAND-MT[2] collection of parallel datasets between 16 African languages and English or French. The HornMT[3] dataset was created for machine translation in languages in the Horn of Africa. Akera et al. (2022) created about 25,000 parallel sentences between 5 Ugandan languages and English, covering topics such as agriculture, health and society.

Muhammad et al. (2022) classified about 30,000 tweets in each of the four major Nigerian languages as either positive, negative, neutral, mixed or indeterminate for sentiment analysis task. Subsequently, the dataset was expanded to include 14 African languages, resulting in the largest sentiment dataset for African languages (Muhammad et al., 2023a,b). Aliyu et al. (2022) collected about 4,500 partially code-switched tweets for detecting hate against the Fulani herdsmen in Nigeria.

Adelani et al. (2022b) created MasakhaNER 2.0, the most extensive corpus for named-entity recognition tasks. Wanjawa et al. (2022) created the multipurpose Kencorpus, a speech and text corpora for machine translation, text-to-speech, question answering, and part-of-speech tagging tasks for three Kenyan languages. KenPOS (Indede et al., 2022) corpus was created for part-of-speech tagging for Kenyan languages. KenSpeech (Awino et al., 2022) is a transcription of Swahili speech created for text-to-speech tasks.

2.2 Visual Question Answering Datasets

Researchers have created several visual question-answering datasets for different purposes, including for medical research (He et al., 2021; Lau et al., 2018; Ben Abacha et al., 2021), improving reading comprehension (Li et al., 2019; Sharma and Jalal, 2022), among others.

DAQUAR (Malinowski and Fritz, 2014) was the first attempt at creating and benchmarking a dataset for understanding and developing models in visual question answering. The dataset consists of 1,449 real-world images and 12,468 synthetic and natural question-answer pairs. Ren et al. (2015b) then created a substantially larger dataset called the COCO-QA dataset using images from the Microsoft COCO dataset (Lin et al., 2014). The dataset consists of 123,287 images with each having a corresponding question-answer pair.

Gao et al. (2015) reused the images from COCO-QA to create the FM-IQA dataset. The authors instruct the annotators to create custom questions and answers, through a crowd-sourcing platform. This resulted in 250,560 question-answer pairs from 120,360 images. Other similar datasets include the Visual Madlibs (Yu et al., 2015), VQA (Antol et al., 2015), Visual7W (Zhu et al., 2016), Visual Genome (Krishna et al., 2016), CLEVR (Johnson et al., 2017a). Others, such as VQA-HAT (Das et al., 2016) and VQA-CP (Agrawal et al., 2018) datasets, extended the VQA to solve specific problems associated with the original dataset.

In the medical area, Ben Abacha et al. (2021) and Lau et al. (2018) created the VQA-Med and VQA-Rad datasets for radiological research, respectively. The VQA-Med dataset was created from 4,200 radiology images and has multiple-choice 15,292 question-answer pairs. The VQA-Rad data consists of 3,515 questions of 11 types that were crafted manually, where clinicians on radiological images gave both questions and answers. He et al. (2021) created the PathVQA, a datasets that consists of 32,795 mostly open-ended questions and their answer pairs that were generated from 4,998 pathology images.

One common theme across all these datasets is that they are all in English. For the majority of other languages, there exists only a few or no datasets for visual question answering tasks. Some of the few in other languages include the original FM-IQA which was created in Chinese before being manually translated into English. Another is the Japanese VQA (Shimizu et al., 2018), where the authors used crowdsourcing to generate. It consists of 99,208 images, each with eight questions, resulting in 793,664 question-answer pairs in Japanese. For African languages, however, there is no dataset for visual question answering tasks. This is the gap that we are trying to fill with the creation of the HaVQA dataset.

[2] https://github.com/masakhane-io/lafand-mt/tree/main/data/text_files

[3] https://github.com/asmelashteka/HornMT

Figure 1: Hausa Visual Question Answering (HaVQA) Annotation Interface

3 Focused Language

Hausa is a Chadic language and is the largest indigenous African language that is spoken as the first or second language by about 79 million people, mainly in northern Nigeria, Niger and northern Cameroon, but also in Benin, Burkina Faso, Central African Republic, Chad, Congo, Côte d'Ivoire, Gabon, Sudan and Togo.[4] The language has well-documented literature and is studied at various local and international institutions. Several international radio stations, including BBC,[5] VoA,[6] and DW[7] run Hausa broadcasting service.

In the early days, the Hausa language was written in "**Ajami**", using Arabic scripts, mainly because of the earlier contacts between the Arabs and the Hausa people (Jaggar, 2001). Nowadays, the language is predominantly written in Latin script. This resulted from the colonial influence that began in the early 19th century. Hausa text is written using the English alphabet except for p, q, v and x, with some additional special letters: ɓ, ɗ, ƙ and ƴ. Nowadays, especially on social media, writing in Hausa has experienced some form of distortion, such as the use of characters p for f, v for b, q or k for ƙ and d for ɗ.

4 The HaVQA Dataset

This section provides a detailed description and analysis of HaVQA Questions and Answers.

4.1 Data Collection and Annotation

We extracted the images along with the question-answer (QA) pairs[8] from the Visual Genome dataset (Krishna et al., 2017). The Visual Genome dataset was created to provide a link between images and natural text, supplying multimodal context in many natural language processing tasks.

We used 7 Hausa native speakers to generate the translations of the QA pair manually. The annotation process was done using a web application developed by integrating the Hausa keyboard to provide easy access to Hausa special characters and restrict access to the unused characters, as shown in Figure 1. During the translation exercise, a set of instructions was provided to annotators, including (i) "the translation should be manually generated without using any translation tool" and (ii) "the on-screen keyboard or any other keyboard that supports the Hausa special characters should be used." These simple and easy-to-remember instructions were adopted to ensure data authenticity and the quality of the dataset. Importantly, the picture has always been presented to the translators.

4.2 Data Validation

After the annotation, each question and the answer were validated to ensure the quality and consistency of the translations, including the basic check that the translations were done using the correct alphabet and special characters. We validated the translations by relying on 7 Hausa language experts. A separate interface was created for the validators to be able to see the images, the original English Question-Answer pair, and the translations of all the pairs that were generated by the annotators in the first phase at once.

[4] https://www.ethnologue.com/language/hau
[5] https://www.bbc.com/hausa
[6] https://www.voahausa.com/
[7] https://www.dw.com/ha/labarai/s-11605
[8] https://visualgenome.org/static/data/dataset/question_answers.json.zip

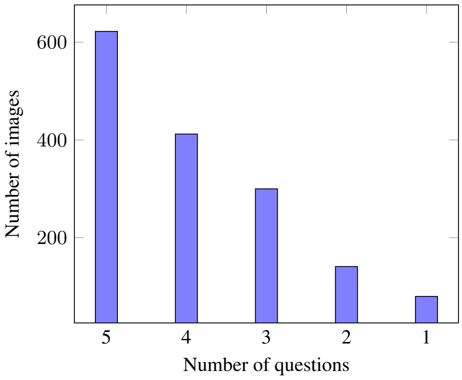


Figure 2: Number of questions per image

Item	Count
Number of Images	1,555
Number of Questions	6,020
Number of Answers	6,020
Number of Counting Questions	616

Table 1: HaVQA Dataset Statistics

The common problem was that the annotators mixed up the choice of words when translating objects that did not have a clear masculine or feminine grammatical gender. Examples of such cases are: "Ina ƴar tsana ruwan hoda?" (**gloss:** Where is the pink teddy bear?), where "ƴar" (feminine) was used, but "yake" (masculine) was used in "Me teddy bear ɗin yake sanye dashi?" (**gloss:** What is the teddy bear wearing?).

Another problem was that the annotators were still using b, d, k and y instead of the special characters ɓ, ɗ, ƙ and ƴ. Using plain ASCII instead of accented symbols introduces ambiguities that can be sometimes resolved only by consulting the original English questions or answers or the associated image. An example is the different meanings of the words "kare" (dog) and "ƙare" (finish). Some annotations included 'y for ƴ (e.g., "Ina 'yar tsanar dabbar?", instead of "Ina ƴar tsanar dabbar?").

4.3 Data Analysis

The HaVQA consists of questions and their corresponding answers. For each image, at least one and at most five questions were asked and answered; see the distribution in Figure 2. We used the punkt[9] tokenizer in the NLTK toolkit (Bird et al., 2009) for tokenization. Some relevant statistics in the created HaVQA dataset are shown in Table 1.

[9] https://www.nltk.org/api/nltk.tokenize.punkt.html

Question Type		%
Hausa	**Gloss**	
"Mene ne/Mene/Me/Waɗanne"	What	56.3
"Me/Mai yasa"	Why	2.9
"Yaya (Nawa/Guda nawa)"	How (much/many)	12.9
"Yaushe"	When	5.6
"A ina/Ina"	Where	16.4
"Waye/(wace/wacce/wace ce)"	Who/(whose)	5.9

Table 2: HaVQA Question Types Statistics

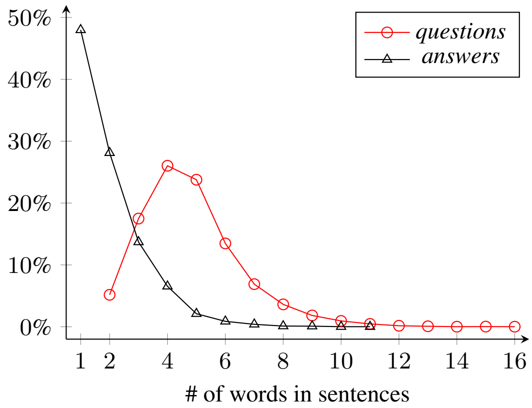


Figure 3: Percentage of Hausa questions and answers with different word counts in HaVQA.

4.3.1 Questions

A variety of question types were included in the original English data; they start with the question words: what, why, how, when, where, and who. In the created HaVQA, these words were translated based on the context in which they appeared. In the Hausa language, these question types vary according to usage, gender, and dialect. For example, the word "who" is translated as "**wanne**" if it is associated with the male gender, or "**wacce**" or "**wace**" for female. The statistics of the different question types (based on the words that start the question) are shown in Table 2. The Hausa questions vary, from as short as 2 to as long as 16 words. The length distribution is shown in Figure 3.

In Figure 4, we show the distribution of Hausa words used when asking a question based on the English question tags of the original dataset. While most of the question tags are used at the beginning of the sentence, as in English, "nawa/nawane" (how much) is mostly used when the subjects that need counting are mentioned, e.g., 2^{nd} in the example "Jirgin leda **nawa** ne a jikin wannan hoton?" (How many kites are there in this picture?) and 3^{rd} in the example "Maza **nawa** ne akan dusar ƙanƙarar?" (How many men are there in the snow?).

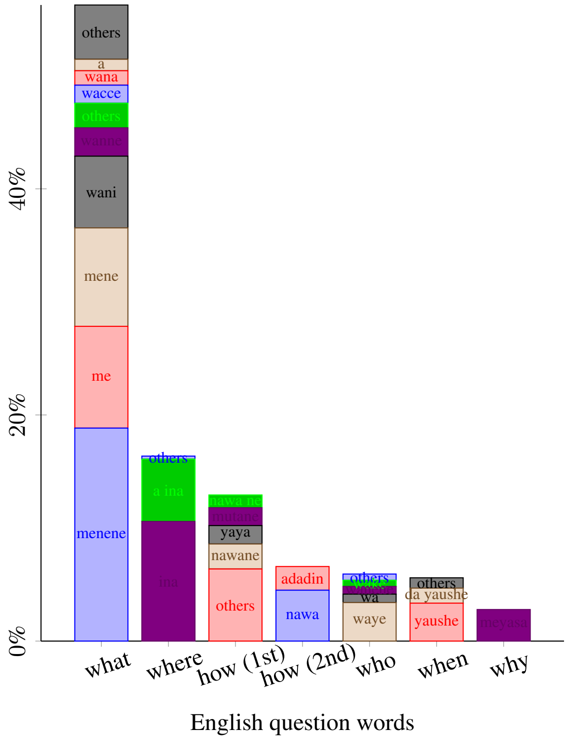


Figure 4: Distribution of Hausa words used in asking questions in relation to their English counterparts. We provide the gloss of these words in Table 8.

4.3.2 Answers

Based on the questions, various answers are included, ranging from a single word (or number) to a short description. The distribution of answer lengths is shown in Figure 3. The distribution of answers for the question types is shown in Figure 10.

5 Sample Applications of HaVQA

We tested the HaVQA data by experimenting with the following NLP tasks: *i)* Questions, Answers, and Images for visual question answering, multimodal machine translation, *ii)* Questions, and Images for visual question elicitation, and *iii)* Questions and Answers for text-only machine translation.

5.1 Visual Question Answering

We used multimodal Transformer-based architecture for VQA consisting of three modules: *i)* feature extraction module—which extracts features from the image and question, *ii)* fusion module—which combines both textual and image features, and *iii)* classification module—which obtains the answer (Siebert et al., 2022).

Similarly, we used Visual Transformer (ViT) for image feature extraction (Dosovitskiy et al., 2020) and multilingual BERT for Hausa, i.e., Hausa BERT[10] for extracting features from the Hausa questions. The classifier which obtains the answer is a fully connected network with output having dimensions equal to the answer space. This architecture is illustrated in Figure 5. We used the Wu and Palmer metric for VQA evaluation (Wu and Palmer, 1994b).

Set	Q/A pairs	Tokens (En)	Tokens (Ha)
Train	4816	35,634	32,142
Dev	602	4,508	4,112
Test	602	4,554	4,084
Total	6,020	44,696	40338

Table 3: Statistics of our data (questions) used in the ***Visual Question Answering*** task: the number of sentences and tokens.

5.2 Machine Translation

We performed text-only and multi-modal translation using the HaVQA dataset. We partitioned the dataset into train/dev/test sets in the ratio of 80:10:10 as shown in Table 4.

5.2.1 Text-Only Translation

We used the questions and answers in English and Hausa for text-only translation. We utilized two approaches for training the Transformer model (Vaswani et al., 2018): training from scratch and fine-tuning a pre-trained multi-lingual model. We evaluated the models' performance using SacreBLEU (Post, 2018) for the dev and test set.

Transformer Trained from Scratch We used the Transformer model as implemented in OpenNMT-py (Klein et al., 2017).[11] Subword units were constructed using the word pieces algorithm (Johnson et al., 2017b). Tokenization is handled automatically as part of the pre-processing pipeline of word pieces.

We jointly generated a vocabulary of 32k subword types for both the source and target languages, sharing it between the encoder and decoder. We used the Transformer base model (Vaswani et al., 2018). We trained the model on the Google Cloud Platform (8 vCPUs, 30 GB RAM) and followed

[10]https://huggingface.co/Davlan/bert-base-multilingual-cased-finetuned-hausa

[11]http://opennmt.net/OpenNMT-py/quickstart.html

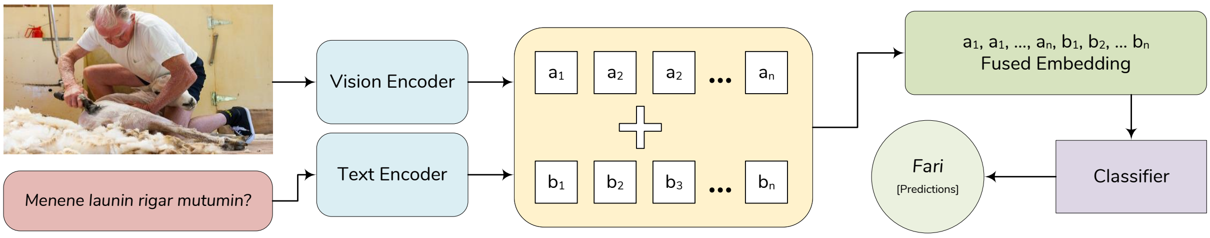


Figure 5: Visual Question Answering. The question is encoded by Hausa BERT, and the context [image] is encoded using a Vision Transformer. The created fused embedding is passed through a classifier to yield the best possible answer. **Gloss:** Input–What color is the man's shirt?; Prediction–White

Set	Sentences	Tokens	
		English	Hausa
Train	9,632	35,634	32,142
Dev	1,204	4,508	4,112
Test	1,204	4,554	4,084
Total	12,040	44,696	40,338

Table 4: Statistics of our data used in the English↔Hausa text-only and multimodal translation.

the standard "Noam" learning rate decay,[12] see Vaswani et al. (2017) or Popel and Bojar (2018) for more details. Our starting learning rate was 0.2, and we used 8000 warm-up steps. The model was trained using 200k training steps and 3k validation steps, and the checkpoints were saved at 3k steps.

Fine-tuning We also employed fine-tuning a large pre-trained model on our domain. This approach has been shown to leverage monolingual data and multilingualism to build a better translation model (Adelani et al., 2022a). We used the M2M-100 pretrained encoder-decoder model (Fan et al., 2022). The model was built to translate between 100 language pairs, including Hausa and 16 other African languages. Specifically, we fine-tuned the 418 million parameter version of M2M,[13] for three epochs. We used a maximum of 128 tokens for both the target and source sentences and a beam size of 5 during decoding. We trained the model on Google Colab (1 GPU, Tesla T4).

5.2.2 Multi-Modal Translation

We used the QA pairs of English and Hausa and the associated images for multimodal machine translation (MMT). Multimodal translation involves utilizing the image modality and the English text for translation to Hausa. It extracts automatically learned features from the image to improve translation quality. We take the MMT approach using object tags derived from the image (Parida et al., 2021).

We first extract the list of English object tags for a given image using the pre-trained Faster RCNN (Ren et al., 2015c) with ResNet101 (He et al., 2016) backbone. We consider up to top 10 object tags for each image based on their confidence scores. The object tags are concatenated to the English sentence, which needs to be translated into Hausa. The concatenation uses the special token '##' as the delimiter, followed by comma-separated object tags. Adding object labels enables the otherwise text-only model to utilize visual concepts which may not be readily available in the original sentence, and to supply context information for easier disambiguation. The English sentences and object tags are fed to the encoder of a text-to-text Transformer model, as shown in Figure 6.

We used the pre-trained M2M-100 Transformer. We trained the model on the Google Cloud Platform (1 GPU, NVIDIA T4). For comparison, we keep the dataset division the same as the text-only translation, as shown in Table 4.

5.3 Visual Question Elicitation

Similar to image captioning, we used the images and associated questions to train an automatic visual question elicitation (VQE) model. We extracted visual features using the images and fed them to an LSTM decoder. The decoder generates the tokens of the caption autoregressively using a greedy search approach (Soh, 2016). Trained to minimize the cross-entropy loss on the questions from the training data (Yu et al., 2019a) was minimized. The architecture is illustrated in Figure 7.

Image encoder All the images were resized to 224×224 pixels, and features from the whole image were extracted to train the model. The feature

[12] https://nvidia.github.io/OpenSeq2Seq/html/api-docs/optimizers.html
[13] https://huggingface.co/facebook/m2m100_418M

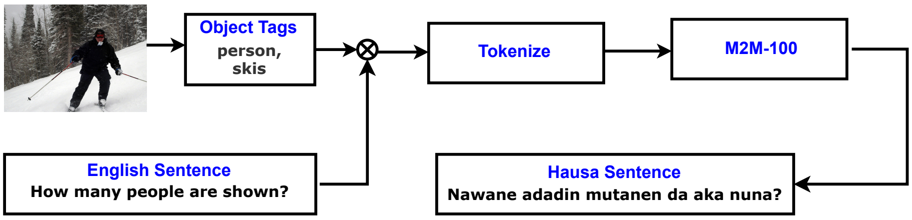


Figure 6: Multimodal machine translation. The object tags are extracted from images and the English source text input to the M2M-100 to generate the Hausa translation output.

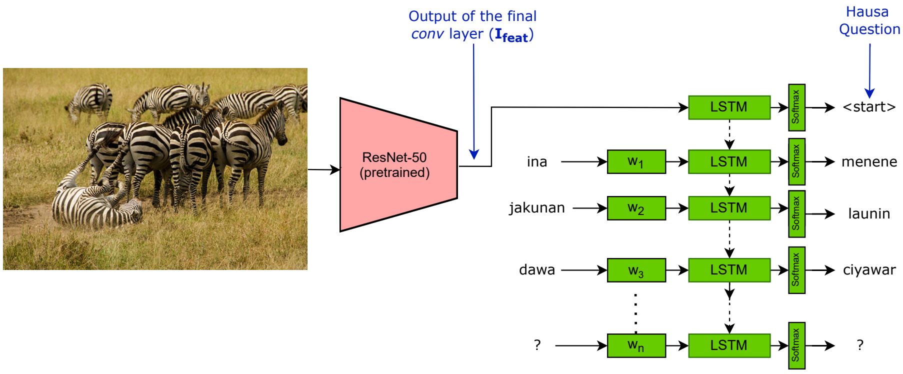


Figure 7: Architecture of Visual Question Elicitation using ResNet-50 and LSTM. The training question was "**ina jakunan dawa?**" (**gloss:** where are the zebras?). When the image was passed during inferencing, the question elicited was "**menene launin ciyawar?**" (**gloss:** what color is the grass?)

Set	**Sentences**	**Tokens**	
		English	Hausa
Train	4816	27,091	23,148
Dev	602	3,306	2,876
Test	602	3,320	2,798
Total	6,020	33,717	28,822

Table 5: Statistics of our data (questions) used in the ***Visual Question Elicitation*** task: the number of sentences and tokens.

vector is the output of the final convolutional layer of ResNet-50. It is a 2048-dimensional feature representation of the image. The encoder module is a fixed feature extractor and, thus, non-trainable.

LSTM decoder A single-layer LSTM, with a hidden size of 256, was used as a decoder. The dropout is set to 0.3. During training, for the LSTM decoder, the cross-entropy loss is minimized and computed using the output logits and the tokens in the gold caption. Weights are optimized using the Adam optimizer (Kingma and Ba, 2014) with an initial learning rate of 0.001. Training is halted when the validation loss does not improve for ten epochs. We trained the model for 100 epochs.

VQE Dataset We used the Hausa Visual Genome (Abdulmumin et al., 2022) and HaVQA datasets to build our Hausa vocabulary, resulting in 7679 Hausa word types for question generation. The question elicitation experiment was carried out using the 1,555 images present in the HaVQA dataset. For training and evaluation of the visual question elicitation, we have considered images and questions as shown in Table 5. As in VQE, we only considered images and their associated questions ignoring answers which are not necessary for the question generation, the statistics of the dataset differ from the multimodal translation tasks.

6 Results and Discussion

This section presents the results obtained after implementing the experiments described in Section 5.

6.1 Visual Question Answering

We employed state-of-the-art language and vision models for Visual Question Answering to report a viable baseline. Table 6 presents the different image encoders we use to obtain our experiment results in combination with the Hausa BERT-based text encoder. The text encoder remains the same across all these experiments. The *WuPalmer* (Wu and Palmer, 1994a) score was chosen as the metric

Image Encoder	Text Encoder	WuPalmer Score
BEiT-large-P-224	Bert-base-Hausa	27.76
ViT-base-P-224	Bert-base-Hausa	28.91
ViT-large-P-224	Bert-base-Hausa	29.67
DeiT-base-P-224	**Bert-base-Hausa**	**30.86**

Table 6: Results of the proposed baseline for Visual Question Answering on our HaVQA dataset. It uses the Multimodal Transformer-based architecture.

to evaluate the baselines. The WuPalmer metric measures semantic similarity between words based on their depth in a lexical hierarchy and the depth of their common ancestor. The metric ranges from 0 to 1, with higher values indicating greater similarity. It is widely employed in tasks such as word sense disambiguation, information retrieval, and semantic relatedness estimation.

From the results reported, the Data-Efficient Image Transformers (DeiT, Touvron et al., 2021 model proposed by Facebook yielded the best results in our architecture. It reached a score of 30.85 and became our best-performing baseline. ViT-base and BEiT Large yielded scores of 28.90 and 27.75, respectively. ViT Large reported a WuPalmer score of 29.67. The DeiT models utilize a distillation token to transfer knowledge from a teacher to a student model through backpropagation. This transfer occurs via the self-attention layer, involving the class token (representing the global image representation) and patch tokens (representing local image features). The distillation token interacts with these tokens, assimilating important information from the teacher model and effectively transferring its knowledge. As a result, the student model trained with the distillation token demonstrates improved performance compared to models trained solely with supervised learning.

In our study, we conducted manual validation of the results generated by the Visual Question Answering (VQA) model. Our analysis revealed that the model exhibited higher performance when tasked with answering questions that required one-word answers. In these cases, the model consistently provided precise answers for the majority of questions and achieved a very good for the remaining ones.

The training dataset used for training the VQA model consisted of 5500 instances, while the test dataset comprised 520 instances. To provide further insight into the distribution of answers, we

Figure 8: Example of a prediction by the VQA model. The question was "**ina cin abincin nan ke faruwa?**" (**gloss:** where is this meal taking place?). The ground truth was "**a restaurant**" and the predicted answer was "**kan tebur**" (**gloss:** on the table)

.

presented Figure 3, which plots the distribution of word counts in the answers.

By focusing on questions that necessitate one-word answers, our study aimed to explore the extent to which the VQA model can excel in a more restricted task akin to classification. The choice to emphasize single-word answers allowed us to investigate the model's capabilities within a specific context and assess the potential impact of this narrowed scope on its performance. The observed errors were mainly associated with cases where there is a dominant object in the picture. The dominant object is returned as the answer regardless the question, see the example prediction in Figure 8. More sample VQA outputs are provided in Figure 11 in the Appendix. Example 4 in Figure 11 illustrates the same problem when answering the question "Wacce **dabba** ce ta fito?" (**gloss:** What **animal** is shown?). Some systems respond with the word "**ciyawa**" (grass), because it is the dominant element in the picture.

6.2 Machine Translation

The text-to-text and multi-modal translation model results are shown in Table 7. For the text-only translation, fine-tuning the Facebook M2M-100 model on the questions and answers for English→Hausa

Method	English→Hausa	Hausa→English
Text-Only		
Transformer	27.1	47.1
M2M-100	35.5	58.7
MultiModal		
M2M-100	26.3	-

Table 7: Results of text-only and multimodal translation on the HaVQA test set.

translation delivers a score by about +8.4 BLEU points better than training a Transformer model, and +11.6 for Hausa→English. The multimodal translation model achieved a decent performance comparable to the text-only translation (−0.8 BLEU points).

The better performance by the text-only translation model is expected because, unlike in the Visual Genome dataset (Abdulmumin et al., 2022), the sentences in HaVQA are mostly unambiguous and, hence, do not require the context that was provided by the images. Also, it is possible that the text captions extracted from the image brought different synonyms than what the single reference translation in Hausa expects. This situation would lead to a comparably good translation quality when assessed by humans but a decreased BLEU score.

6.3 Visual Question Elicitation

Because it is difficult to measure the quality of the generated questions using automatic evaluation metrics, we manually evaluated the sample-generated questions, relying on a native Hausa speaker. We sampled about 10% of the elicited questions and subjected them to manual evaluations. We categorized each of the sampled questions as either "**Exact**", "**Correct**", "**Nearly Correct**" and "**Wrong**". We present the distribution of these classes in Figure 9, and provide some samples in Figure 12 in the Appendix.

All the generated predictions were valid (reasonable) questions, with all but 3 (99.5%) having the question mark ('?') appended at the end of the question. The distribution of the question types are: "menene/me/mene" (what)–64.1%, "ina/a ina" (where)–26.4%, "yaushe/da yaushe" (when)–2.8%, "waye/wacce/wana/wanne" (who)–3.7%, "meyasa" (why)–1.5%, "nawane/nawa" (how much)–1.32% and "yaya" (how)–0.2%.

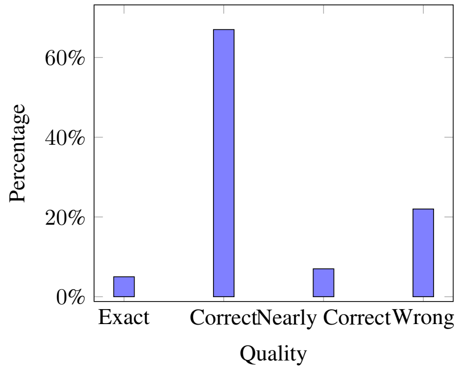


Figure 9: Quality distribution of automatically generated questions.

7 Conclusion and Future work

We present HaVQA, a multimodal dataset suitable for many NLP tasks for the Hausa language, including visual question answering, visual question elicitation, text and multimodal machine translation, and other multimodal research.

The dataset is freely available for research and non-commercial usage under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 License at: http://hdl.handle.net/11234/1-5146. We released our experimental code through Github.[14]

Our planned future work includes: *i)* extending the dataset with more images and QA pairs, *ii)* providing ground truth for all images for image captioning experiments, and *iii)* organizing a shared task using HaVQA.

Ethics Statement

We do not envisage any ethical concerns. The dataset does not contain any personal, or personally identifiable, information, the source data is already open source, and there are no risks or harm associated with its usage.

Limitations

The most important limitation of our work lies in the size of the HaVQA dataset. However, substantial further funding would be needed to resolve this. For the baseline multimodal experiments, we did not use the image directly but resorted to extracting textual tags and including them in the text-only translation input. A tighter fusion technique may give better performance.

[14] https://github.com/shantipriyap/HausaVQA/tree/main

Acknowledgements

The HaVQA dataset was created using funding from the HausaNLP research group. This work is supported by Silo AI, Helsinki, Finland. This work has received funding from the grant 19-26934X (NEUREM3) of the Czech Science Foundation and has also been supported by the Ministry of Education, Youth, and Sports of the Czech Republic, Project No. LM2018101 LINDAT/CLARIAH-CZ.

References

Idris Abdulmumin, Satya Ranjan Dash, Musa Abdullahi Dawud, Shantipriya Parida, Shamsuddeen Muhammad, Ibrahim Sa'id Ahmad, Subhadarshi Panda, Ondřej Bojar, Bashir Shehu Galadanci, and Bello Shehu Bello. 2022. Hausa visual genome: A dataset for multi-modal English to Hausa machine translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6471–6479, Marseille, France. European Language Resources Association.

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. 2022a. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070, Seattle, United States. Association for Computational Linguistics.

David Ifeoluwa Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba O. Alabi, Shamsuddeen H. Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire M. Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Elvis Mboning, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo L. Mokono, Ignatius Ezeani, Chiamaka Chukwuneke, Mofetoluwa Adeyemi, Gilles Q. Hacheme, Idris Abdulmumin, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu Ngoli, and Dietrich Klakow. 2022b. Masakhaner 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Benjamin Akera, Jonathan Mukiibi, Lydia Sanyu Naggayi, Claire Babirye, Isaac Owomugisha, Solomon Nsumba, Joyce Nakatumba-Nabende, Engineer Bainomugisha, Ernest Mwebaze, and John Quinn. 2022. Machine translation for african languages: Community creation of datasets and models in uganda. In *3rd Workshop on African Natural Language Processing*.

Saminu Mohammad Aliyu, Gregory Maksha Wajiga, Muhammad Murtala, Shamsuddeen Hassan Muhammad, Idris Abdulmumin, and Ibrahim Said Ahmad. 2022. HERDPhobia: A Dataset for Hate Speech against Fulani in Nigeria. In *Proceedings of the The Sixth Widening NLP Workshop (WiNLP)*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433.

Dorcas Awino, Lawrence Muchemi, Lilian D.A. Wanzare, Edward Ombui, Barack Wanjawa, Owen McOnyango, and Florence Indede. 2022. KenSpeech: Swahili Speech Transcriptions.

Asma Ben Abacha, Mourad Sarrouti, Dina Demner-Fushman, Sadid A. Hasan, and Henning Müller. 2021. Overview of the vqa-med task at imageclef 2021: Visual question answering and generation in the medical domain. In *CLEF 2021 Working Notes*, CEUR Workshop Proceedings, Bucharest, Romania. CEUR-WS.org.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. 2016. Human attention in visual question answering: Do humans and deep networks look at the same regions? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 932–937, Austin, Texas. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv e-prints*, pages arXiv–2010.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2022. Beyond english-centric multilingual machine translation. *Journal of Maching Learning Research*, 22(1).

Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. 2015. Are you talking to a machine? dataset and methods for multilingual image question answering. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, page 2296–2304, Cambridge, MA, USA. MIT Press.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Deepak Gupta, Pabitra Lenka, Asif Ekbal, and Pushpak Bhattacharyya. 2020. A unified framework for multilingual and code-mixed visual question answering. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 900–913.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Xuehai He, Zhuo Cai, Wenlan Wei, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2021. Towards visual question answering on pathology images. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 708–718, Online. Association for Computational Linguistics.

Florence Indede, Owen McOnyango, Lilian D.A. Wanzare, Barack Wanjawa, Edward Ombui, and Lawrence Muchemi. 2022. KenPos: Kenyan Languages Part of Speech Tagged dataset.

P.J. Jaggar. 2001. *Hausa*. London Oriental and African language library. John Benjamins Publishing Company.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017a. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017b. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Kushal Kafle and Christopher Kanan. 2017. Visual question answering: Datasets, algorithms, and future challenges. *Computer Vision and Image Understanding*, 163:3–20.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. Cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proc. ACL*.

Venkat Kodali and Daniel Berleant. 2022. Recent, rapid advancement in visual question answering: a review. In *2022 IEEE International Conference on Electro Information Technology (eIT)*, pages 139–146. IEEE.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Gokul Karthik Kumar, Abhishek Gehlot, Sahal Shaji Mullappilly, and Karthik Nandakumar. 2022. Mucot: Multilingual contrastive training for question-answering in low-resource languages. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 15–24.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5(1).

Hui Li, Peng Wang, Chunhua Shen, and Anton van den Hengel. 2019. Visual question answering as reading comprehension. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6312–6321, Los Alamitos, CA, USA. IEEE Computer Society.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

Mateusz Malinowski and Mario Fritz. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'14, page 1682–1690, Cambridge, MA, USA. MIT Press.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Sa'id Ahmad, Meriem Beloucif, Saif M. Mohammad, Sebastian Ruder, Oumaima Hourrane, Pavel Brazdil, Felermino Dário Mário António Ali, Davis David, Salomey Osei, Bello Shehu Bello, Falalu Ibrahim, Tajuddeen Gwadabe, Samuel Rutunda, Tadesse Belay, Wendimu Baye Messelle, Hailu Beshada Balcha, Sisay Adugna Chala, Hagos Tesfahun Gebremichael, Bernard Opoku, and Steven Arthur. 2023a. AfriSenti: A Twitter Sentiment Analysis Benchmark for African Languages.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Seid Muhie Yimam, David Ifeoluwa Adelani, Ibrahim Sa'id Ahmad, Nedjma Ousidhoum, Abinew Ali Ayele, Saif M. Mohammad, Meriem Beloucif, and Sebastian Ruder. 2023b. SemEval-2023 Task 12: Sentiment Analysis for African Languages (AfriSenti-SemEval). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, Toronto, Canada. Association for Computational Linguistics.

Shamsuddeen Hassan Muhammad, David Ifeoluwa Adelani, Sebastian Ruder, Ibrahim Sa'id Ahmad, Idris Abdulmumin, Bello Shehu Bello, Monojit Choudhury, Chris Chinenye Emezue, Saheed Salahudeen Abdullahi, Anuoluwapo Aremu, Alípio Jorge, and Pavel Brazdil. 2022. NaijaSenti: A nigerian Twitter sentiment corpus for multilingual sentiment analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 590–602, Marseille, France. European Language Resources Association.

Shantipriya Parida, Subhadarshi Panda, Satya Prakash Biswal, Ketan Kotwal, Arghyadeep Sen, Satya Ranjan Dash, and Petr Motlicek. 2021. Multimodal neural machine translation system for English to Bengali. In *Proceedings of the First Workshop on Multimodal Machine Translation for Low Resource Languages (MMTLRL 2021)*, pages 31–39, Online (Virtual Mode). INCOMA Ltd.

Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin O Steitz, Stefan Roth, Ivan Vulić, and Iryna Gurevych. 2021. xgqa: Cross-lingual visual question answering. *arXiv preprint arXiv:2109.06082*.

Martin Popel and Ondřej Bojar. 2018. Training tips for the transformer model. *The Prague Bulletin of Mathematical Linguistics*, 110(1):43–70.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Mengye Ren, Ryan Kiros, and Richard Zemel. 2015a. Exploring models and data for image question answering. *Advances in neural information processing systems*, 28.

Mengye Ren, Ryan Kiros, and Richard S. Zemel. 2015b. Exploring models and data for image question answering. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, page 2953–2961, Cambridge, MA, USA. MIT Press.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015c. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 91–99, Cambridge, MA, USA. MIT Press.

Himanshu Sharma and Anand Singh Jalal. 2022. Comparison of visual question answering datasets for improving their reading capabilities. In *International*

Conference on Artificial Intelligence and Sustainable Engineering, pages 525–534, Singapore. Springer Nature Singapore.

Nobuyuki Shimizu, Na Rong, and Takashi Miyazaki. 2018. Visual question answering dataset for bilingual image understanding: A study of cross-lingual transfer using attention maps. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1918–1928, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Tim Siebert, Kai Norman Clasen, Mahdyar Ravanbakhsh, and Begüm Demir. 2022. Multi-modal fusion transformer for visual question answering in remote sensing. In *Image and Signal Processing for Remote Sensing XXVIII*, volume 12267, pages 162–170. SPIE.

Moses Soh. 2016. Learning cnn-lstm architectures for image caption generation.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2021. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR.

Aisha Urooj, Amir Mazaheri, Mubarak Shah, et al. 2020. Mmft-bert: Multimodal fusion transformer with bert encodings for visual question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4648–4660.

Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2tensor for neural machine translation. In *Proc. of AMTA (Volume 1: Research Papers)*, pages 193–199.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Barack Wanjawa, Lilian Wanzare, Florence Indede, Owen McOnyango, Edward Ombui, and Lawrence Muchemi. 2022. Kencorpus: A kenyan language corpus of swahili, dholuo and luhya for natural language processing tasks.

Zhibiao Wu and Martha Palmer. 1994a. Verb semantics and lexical selection. *arXiv preprint cmp-lg/9406033*.

Zhibiao Wu and Martha Palmer. 1994b. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138.

Jun Yu, Jing Li, Zhou Yu, and Qingming Huang. 2019a. Multimodal transformer with multi-view visual representation for image captioning. *IEEE transactions on circuits and systems for video technology*, 30(12):4467–4480.

Licheng Yu, Eunbyung Park, Alexander C. Berg, and Tamara L. Berg. 2015. Visual madlibs: Fill in the blank description generation and question answering. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, page 2461–2469, USA. IEEE Computer Society.

Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019b. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6281–6290.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4995–5004.

Appendix

A Annotators and Validators Recruitment

We recruited Hausa natives from the team of experienced translators at the HausaNLP research group as annotators and validators. The team of annotators consisted of 4 females and 3 males, while the validators included 3 females and 4 males. Each member of the annotation/validation team have at least an undergraduate degree. They all reside in different parts of Northern Nigeria.

B Annotation Guidelines

The following guidelines were provided to the native Hausa annotators and validators:

1. Remember to read the Hausa typing rules. Before starting annotation, test it once and report for any issues.

2. The annotator must be a native speaker of the Hausa language.

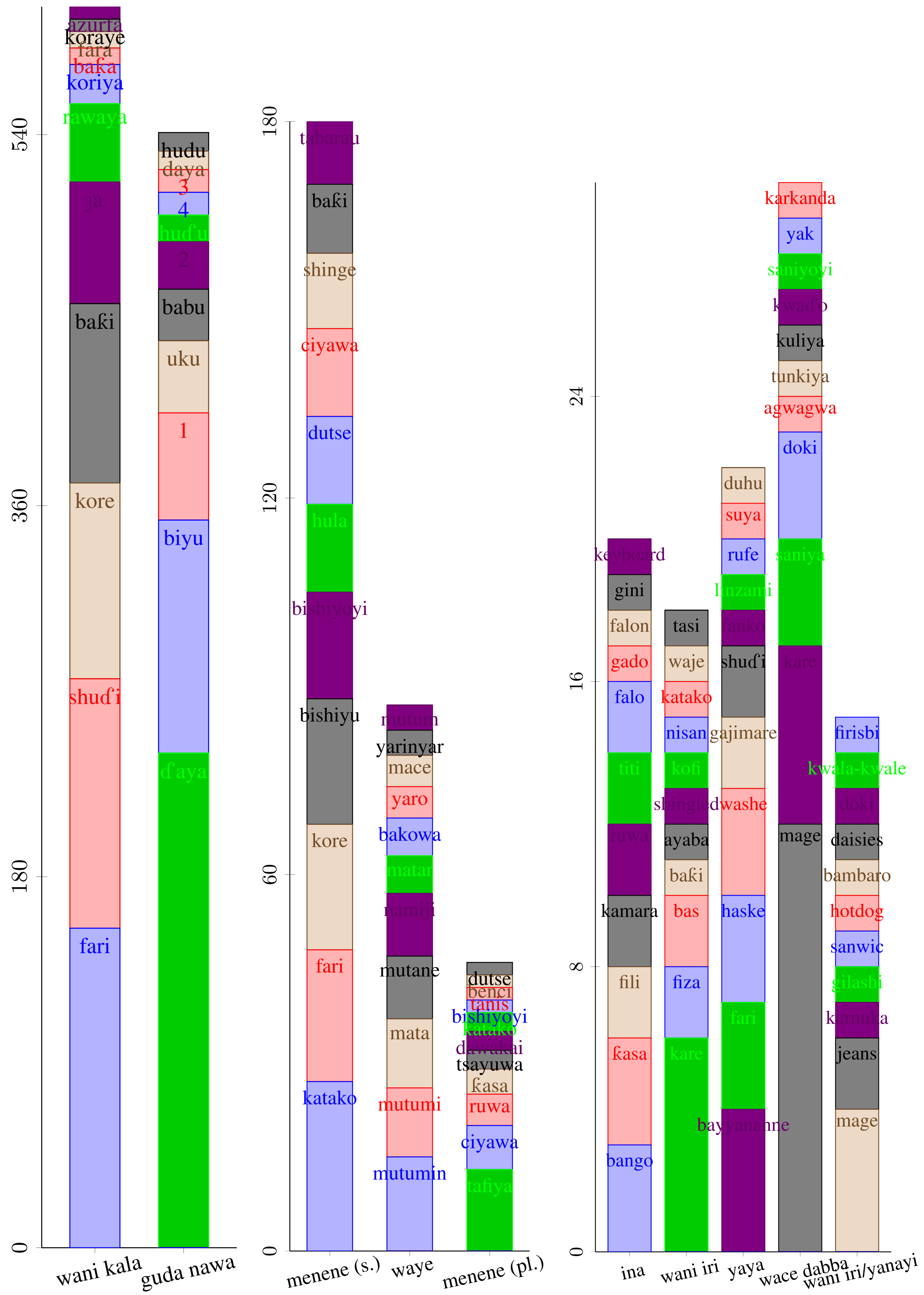
3. Look at the image before annotating.

4. Try to understand the task, i.e., translate the questions and answers into the Hausa language.

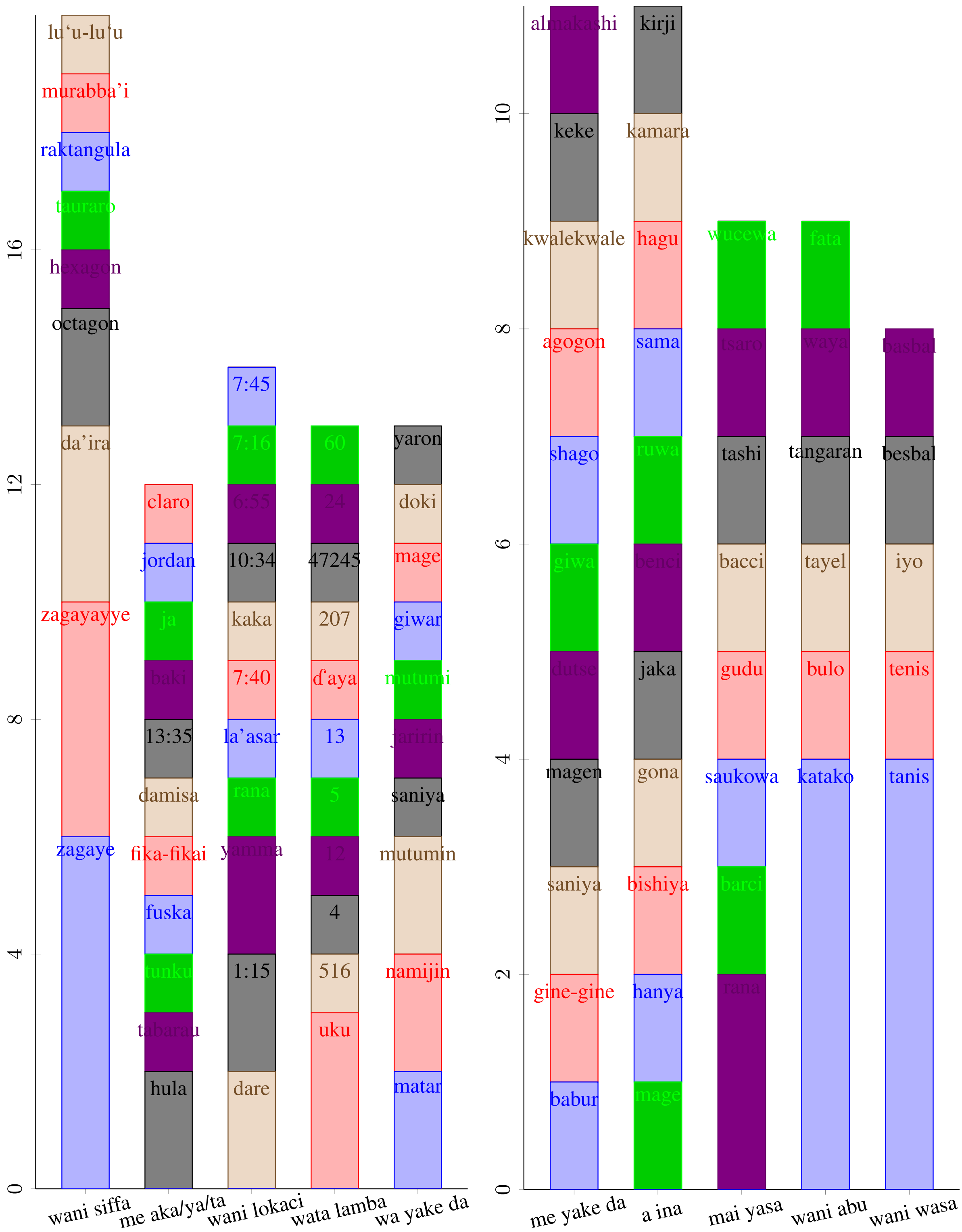
5. Do not use any Machine translation system for annotation.

6. Do not enter dummy entries for testing the interface.
7. Data will be saved at the backend.
8. Press the Shift Key on the virtual keyboard for complex consonants.
9. Contact the coordinator for any clarification/support

C Question-Type Answer Distribution



A. **B.** **C.**



D. **E.**

Figure 10: One-word answer distribution for the Hausa question types

D Visual Question Answer Sample Predictions

S/N	Example	Model	Instances		Prediction	Gloss.
			Test	*Train*		
1.		microsoft/beit-large	520	5500	*rana*	daytime
		google/vit-base	520	5500	*rana*	daytime
		facebook/deit-base	520	5500	*rana*	daytime
		google/vit-large	520	5500	*rana*	daytime
	Question: *Yaushe aka ɗauki wannan hoton?* **gloss.** When was this photo taken?		**Answer:** *rana* **gloss.** Daytime			
2.		microsoft/beit-large	520	5500	*Biyu*	Two
		google/vit-base	520	5500	*Biyu*	Two
		facebook/deit-base	520	5500	*Biyu*	Two
		google/vit-large	520	5500	*Biyu*	Two
	Question: *Hawa nawa ne a gefen hagun ginin?* **gloss.** How many floors is the left of the building?		**Answer:** *uku* **gloss.** three			
3.		microsoft/beit-large	520	5500	*Kore*	Green
		google/vit-base	520	5500	*kore*	green
		facebook/deit-base	520	5500	*kore*	green
		google/vit-large	520	5500	fari	white
	Question: *Menene launin rigar sa?* **gloss.** What color is his shirt?		**Answer:** *fari* **gloss.** white			
4.		microsoft/beit-large	520	5500	*ciyawa*	grass
		google/vit-base	520	5500	*ciyawa*	grass
		facebook/deit-base	520	5500	*dawa*	wild
		google/vit-large	520	5500	*dawa*	wild
	Question: *Wacce dabba ce ta fito?* **gloss.** What animal is shown?		**Answer:** *giwa* **gloss.** elephant			

Figure 11: Examples of answers generated by the VQA model. All the models were trained using a batch size of 16.

E Visual Question Elicitation Sample Predictions

Example 1

Exact

Ref. Question.: *Mene launin ciyawar?*
Gloss: What is the color of the grass?
Pred. Question.: *menene launin ciyawar?*
Gloss: what is the color of the grass?

Example 2

Correct

Ref. Question.: *A ina aka ɗauki hoton?*
Gloss: Where was the picture taken?
Pred. Question.: *me raƙumin dawan yakeyi?*
Gloss: what is the giraffe doing?

Example 3

Nearly Correct

Ref. Question.: *Ina agogon hasumiya?*
Gloss: Where is the tower clock?
Pred. Question.: *ina fitilolin mota?*
Gloss: where are the car lights?

Example 4

Wrong

Ref. Question.: *Akan me karen yake?*
Gloss: Where is the dog lying?
Pred. Question.: *menene launin idon magen?*
Gloss: what is the color of the cat's eye?

Figure 12: Examples of Questions elicited by the VQE model.

F Glossaries

Hausa	Gloss	Hausa	Gloss	Hausa	Gloss
a	in	*da yaushe*	when	*kore*	green (s.)
a benci	on bench	*daya*	one	*koriya*	green
a bishiya	on tree	*ɗaya*	one	*kuliya*	cat
a gona	on farm	*doki*	horse	*kwaɗo*	frog
a hagu	on left	*duhu*	dark	*kwala-kwale*	canoe
a hanya	on way (road)	*dutse*	stone	*kwalekwale*	canoe
a ina	where (is/are)	*falo*	parlour	*la'asar*	evening
a jaka	in bag	*falon*	the parlour	*linzami*	bridle
akan (me)	on what	*fanko*	empty	*lu'u-lu'u*	diamond
a kamara	on camera	*fara*	white (she)	*mace*	woman
a kirji	on chest	*fari*	white (he)	*mage*	cat
a ruwa	in water	*fata*	skin	*magen*	the cat
a sama	in air	*fika-fikai*	wings	*mai yasa*	why is
a yaushe	when	*fili*	field	*mata*	woman
adadin	the quantity	*firisbi*	frisbee	*matar*	the woman
almakashi	scizzors	*fiza*	pizza	*me*	what
agogon	the clock	*fuska*	face	*me aka/ya/ta*	what does
agwagwa	duck	*gado*	bed	*me yake da*	what has
ayaba	banana	*gajimare*	cloud	*me yasa*	why
azurfa	silver	*gilashi*	glass	*mene*	what
babu	nothing	*gine-gine*	buildings	*menene* (s.)	what is [it]
babur	motorcycle	*gini*	building	*menene* (pl.)	what are
bacci	sleep	*giwa*	elephant	*meyasa*	why [did]
baƙa	black (she)	*giwar*	the elephant	*meye*	what
baƙi	black (he)	*guda nawa*	how many	*mutane*	people
bakowa	nobody	*gudu*	run	*murabba'i*	square/quarter
bambaro	straw	*haske*	light	*mutum*	person
bango	wall	*hoton*	the image	*mutumi*	person (m.)
barci	sleep	*hudu*	four	*mutumin*	the man
bas	bus	*huɗu*	four	*namiji*	male
basbal	baseball	*hula*	cap	*namijin*	the man
bayyananne	clear	*ina*	where (is)	*nawa*	how much/mine
benci	bench	*inane*	where is	*nawane*	how much [is]/it's mine
bishiyoyi	trees	*iyo*	swimming	*na wane*	how much [is]/it's mine
bishiyu	trees	*ja*	red	*raktangula*	rectangle
biyu	two	*jaririn*	the baby	*rana*	day/sun
bulo	block/brick	*kaka*	autumn	*rawaya*	yellow
ciyawa	grass	*kamara*	camera	*rufe*	close [it]
daga (me)	from what	*kare*	dog	*ruwa*	water
daga ina	from where	*karkanda*	rhinoceros	*saniya*	cow
da'ira	round	*karnuka*	dogs	*saniyoyi*	cows
da me	with what	*ƙasa*	sand	*saukowa*	coming down
dame	with what	*katako*	wood	*sanwic*	sandwich
damisa	tiger	*keke*	bicycle	*shago*	store
dare	night	*kofi*	cup	*shinge*	wall
dawakai	horses	*koraye*	green (pl.)	*shuɗi*	blue

Continued on Next Page…

Table 8 – Continued

Hausa	Gloss	Hausa	Gloss	Hausa	Gloss
su waye	who are they	*wa yake da*	who has	*wata lamba*	what number
suwaye	who are they	*wacce*	what/which is (she)	*waya*	phone/who did ...
suya	frying	*wace*	what/which (she)	*wayake*	who is ... (he)
ta ina	where	*wace dabba*	what animal	*waye*	who is it (he)
ta yaya	how	*waje*	outside	*wucewa*	walking past
tabarau	glass	*wake*	who is ... (he/she)	*ya*	how/sister
tafiya	walking	*wana*	what/which is (it)	*yaya*	how/sister
tangaran	ceramic/china	*wane*	what/which/who is (he)	*yamma*	evening
tanis	tennis	*wanene*	who is it (he)	*yarinyar*	the girl
tashi	wake/stand	*wani*	what/which (he)	*yaro*	boy
tasi	taxi	*wani abu*	what material	*yaron*	the boy
tauraro	star	*wani iri*	what type	*yaushe*	when
tayel	tie/tile	*wani iri/yanayi*	what kind	*yaushene*	when is it
titi	road	*wani kala*	what color	*zagayayye*	round
tsaro	security	*wani lokaci*	what time	*zagaye*	round
tsayuwa	standing	*wani siffa*	what shape		
tunkiya	sheep	*wani wasa*	what sport		
tunku	bear	*wanne*	what/which is (he)		
uku	three	*washe*	clear		
wa	who	*wasu*	what/which (plural)		

Table 8: Gloss of Hausa words used in Figures 4 and 10.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 7.2

☒ A2. Did you discuss any potential risks of your work?
We couldn't find any potential risk in our work.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
No response.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☑ Did you run computational experiments?

Section 5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 5

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 5

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Sections 5 and 6

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Sections 4 and 5

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Section 4, Appendix E

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix F

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Appendix E

☒ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No personal data was used in this work

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
We annotated an open source data and used it in this work

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Appendix E