Walia-LLM: Enhancing Amharic-LLaMA by Integrating Task-Specific and Generative Datasets

Israel Abebe Azime [1], Atnafu Lambebo Tonja [2,3], Tadesse Destaw Belay [2], Mitiku Yohannes Fuge [4], Aman Kassahun Wassie[4], Eyasu Shiferaw Jada, Yonas Chanie[5], Walelign Tewabe Sewunetie [6], Seid Muhie Yimam [7]

[∀] Masakhane NLP, [∀] Ethio NLP, [1] Saarland University, Germany, , [2] Instituto Politécnico Nacional, Mexico,[3] Lelapa AI,[4] AIMS
[5] Carnegie Mellon University, [6] Debre Markos University, [7]Universität Hamburg, Germany

Abstract

Large language models (LLMs) have received a lot of attention in natural language processing (NLP) research because of their exceptional performance in understanding and generating human languages. However, low-resource languages are left behind due to the unavailability of resources. In this work, we focus on enhancing the LLAMA-2-Amharic model by integrating task-specific and generative datasets to improve language model performance for Amharic. We compile an Amharic instruction fine-tuning dataset and fine-tuned LLAMA-2-Amharic model. The fine-tuned model shows promising results in different NLP tasks. We also explore the effectiveness of translated instruction datasets compared to the dataset we created. Our dataset creation pipeline, along with instruction datasets, trained models, and evaluation outputs, is made publicly available to encourage research in language-specific models. [1]

1 Introduction

Large language models (LLMs) such as GPT series (Brown et al., 2020), LLAMA-2 (Touvron et al., 2023), Phi2 (Javaheripi et al., 2023), Mistral (Jiang et al., 2023), Mixtral (Jiang et al., 2024), PaLM (Chowdhery et al., 2023), Gemini (Team et al., 2023), BLOOM (Workshop et al., 2022), have exhibited exceptional performance in understanding and generating human language, showcasing a range of capabilities from basic linguistic comprehension to complex text generation.

LLAMA-2 (Touvron et al., 2023), a family of pre-trained and fine-tuned large language models (LLMs), demonstrated impressive performance across multiple tasks, particularly in dialogue-based interactions. Regardless of these achievements, LLAMA-2 pre-training supports a small number of languages, which does not include low-resource languages like Amharic. This makes adapting LLMs to low-resource languages that are not included a significant challenge.

Adopting these LLMs to local languages requires the preparation of a quality instruction dataset. Amharic is one of the Semitic languages under the Afroasiatic language family spoken in Ethiopia with more than 57M speakers. There are numerous task-specific datasets for Amharic (Tonja et al., 2023) compared to other Ethiopian languages. This paper focuses on enhancing the LLAMA-2-Amharic (Andersland, 2024) model with quality datasets that are created by converting existing datasets in English into instruction-based Amharic datasets. Furthermore, we create new instruction datasets following the approach by Wei et al. (2022).

LLAMA-2-Amharic model by Andersland (2024) was created by pre-training LLAMA-2 7B model using open-source Amharic and translated corpus. After performing vocabulary expansion and pre-training, Andersland (2024) fine-tuned the created model by translating English instruction datasets into Amharic using commercial translation tools. In our research, we aim to improve the performance of the Amharic LLAMA model by integrating task-specific and generative datasets, as shown in Table 1. The contributions of this paper are as follows:

- Creating Amharic instruction fine-tuning data from existing NLP task-specific and generative datasets.
- Evaluating new and existing models' performance.
- Exploring the effect of carefully curated datasets by combining them with machine-translated instruction datasets.

[1]For data generation pipeline, see https://github.com/EthioNLP/afri-sft-data. For models and datasets, refer to https://huggingface.co/EthioNLP.

- Exploring the effect of instructions on the model's performance by introducing code-mixing instructions.

- Open-sourcing our dataset creation pipeline, instruction datasets, trained models, and evaluation outputs to promote language-specific studies on these models.

2 Related Work

The introduction of open-source LLMs like LLAMA-2 (Touvron et al., 2023) enabled the creation of several language models that focus on specific applications. This application gives more capabilities for these LLMs by teaching them to use tools (Schick et al., 2023), write code (Roziere et al., 2023), understand videos (Zhang et al., 2023a), or work for different languages (Cui et al., 2023). To achieve remarkable understanding and generation abilities, LLMs require large training data and huge compute resources (Hoffmann et al., 2022).

The work by Dong et al. (2023) explores how LLMs' generation, natural language understanding, and problem-solving abilities relate to the data they are trained on and its composition. This work suggests that the amount of composition data is more important for these abilities to show in a low-resource scenario.

Using self-instructed fine-tuning, the work by Wei et al. (2022); Taori et al. (2023); Cui et al. (2023) showed a new approach to align the generation outputs of the generative models through the application of NLP tasks. These tasks are structured around natural language instruction templates, providing a novel means to guide the model's generation process toward better adherence to task-specific requirements. LLAMA-Adapter (Zhang et al., 2023b) also shows that it is possible to reduce the fine-tuning time of LLAMA-7B by introducing lightweight adapters on top of the model.

Acquiring and preparing a dataset for instruction fine-tuning presents a significant challenge due to the extensive labor and resources required. There are several ways of acquiring instruction data, including manual dataset creation, using generative models (Wang et al., 2022; Taori et al., 2023), or using machine translation instruction data for training LLMs for specific languages (Cui et al., 2023).

Fine-tuning LLMs such as LLAMA-2 for specific tasks is an area of exploration as well. Advanced language model-based translator (ALMA) (Xu et al., 2023) outperformed state-of-the-art (SOTA) no language left behind (NLLB) (NLLB Team et al., 2022) model MT task. They worked on fine-tuning monolingual data and subsequent fine-tuning with parallel data. Apart from LLAMA-2, (Moslem et al., 2023) worked on Mistral 7B fine-tuning for medical domain machine translation, where they showed improvement in Spanish to English translation from baseline performance.

After the LLAMA-2 was released, researchers successfully adapted the model for other languages. The work by Cui et al. (2023) involved creating a unique tokenizer for Chinese, extending the pre-training phase, and then fine-tuning the model. This work incorporates secondary pre-training using Chinese data and fine-tunes the model with Chinese instruction datasets. The result shows a significant enhancement of the model's ability to comprehend and execute instructions.

To the same approach of the work by Cui et al. (2023), LLAMA-2 was also adopted for Amharic (Andersland, 2024). During pre-training, Andersland (2024) used an open-source Amharic corpus with some translated corpus from English, and for fine-tuning, available English instruction datasets were translated to Amharic using the Google Translate API and SeamlessM4T. Following the increase of the LLAMA vocabulary size from 32k to 51k and subsequent pre-training with a large Amharic text corpus, they conducted supervised instruction fine-tuning using machine-translated datasets. Then, they evaluated their model using the MMLU (Hendrycks et al., 2020) multiple-choice English dataset by translating it into Amharic. The model is available without original Amharic evaluations because no instruction-based datasets exist for Amharic.

3 Dataset preparation

In this work, we have converted existing NLP task-specific datasets, like sentiment analysis and machine translation, into instruction datasets. We created an instruction template for each task and developed a data creation pipeline (Figure 1) that merges each template instruction with appropriate data from a pre-existing dataset. This pipeline helps us to create instruction datasets from pre-existing NLP task datasets. For the new NLP task, we focused on collecting a new dataset that can be converted into instruction data. We also created

Data Source	Source Data			Is new	# Templates	Generated Data		
	train	val	test			train	val	test
Amharic QA	1723	595	299	NO	14	10000	595	299
MasakhaNews	11522	188	376	NO	11	7866	205	376
MT (amh-eng)	497739	1012	1012	NO	10	10000	997	1012
MT (eng-amh)	497739	1012	1012	NO	10	10000	997	1012
Summarization	5761	719	719	NO	9	10000	719	719
Text Expansion	5761	719	719	NO	9	10000	719	719
Sentiment Analysis (AfriSenti)	5984	1497	1999	NO	7	10000	1728	1999
NER	1750	500	250	NO	9	10000	500	250
News Title Generation	-	-	-	Yes	10	10000	5078	5078
Poem Generation	-	-	-	Yes	3	3885	69	70
Poem Completion	-	-	-	Yes	7	3885	69	70
Religious Lyrics Generation	-	-	-	Yes	3	4929	188	206
Religious Lyrics Completion	-	-	-	Yes	4	10000	1497	1728
Story generation	-	-	-	Yes	10	1665	24	25
Spelling Correction	-	-	-	Yes (modified)	9	10000	1438	1438
Non religious music Lyrics Generation	-	-	-	Yes	4	148	5	5
Non religious music Lyrics Completion	-	-	-	Yes	7	259	5	5
Total						122,637	14,911	15,011

Table 1: **Dataset** used for preparing instruction fine-tuning data. **Is new** = new custom dataset. Details of each data source are explained in Section 3. Figure 1 shows how a dataset is converted to instruction data using our Data processing Pipeline.

new datasets by tweaking existing datasets. Finally, we included an instruction-tuning dataset converted into Amharic language using machine translation systems. Table 1 shows a detailed distribution of instruction task data.

3.1 Instruction dataset from existing datasets

We have used several existing datasets to create an instruction dataset from an existing one. The production of these datasets includes web scraping, human labeling, and verification. By collecting and using this dataset for instruction, we ensure the quality of our instruction dataset. The other benefit of working with these datasets is that we ensure similar prompts across all our models for testing, which eliminates prompt-related performance variance that is usually reported while evaluating the performance of this dataset in generative LLMs.

For sentiment analysis data, we used *AfriSenti* (Muhammad et al., 2023), a sentiment analysis benchmark dataset for 14 African languages where Amharic is among the ones. The dataset is labeled with three sentiment classes: positive, negative, and neutral. The number of train, test, and val sets are shown in Table 1. We used the Amharic version of the classes for the test cases, and tests were done to check if the model gives one of the sentiment classes during generation.

We also worked with *MasakhaNews* (Adelani et al., 2023), a benchmark dataset for news topic classification covering 16 languages widely spoken in Africa. It provides an evaluation of baseline models from classical machine learning models and fine-tunes several language models.

To test if our model has the ability to identify names from sentences, we modified *MasakhaNER* (Adelani et al., 2021), which is a dataset for named entity recognition (NER) in ten African languages. We created questions to extract only personal names for this work, and we plan to include more in our future works.

AmharicQA (Abedissa et al., 2023) is a publicly available Amharic open-ended question-answering dataset. It is a crowdsourced 2,628 question-answer pairs over 378 Wikipedia articles. These question-answer pairs are supplemented with context that the language model can use to answer the questions. We have also used this dataset to evaluate our models by converting it into an instruction dataset.

For tasks like Amharic text summarization, we used *XL-Sum* (Hasan et al., 2021), a comprehensive and diverse dataset comprising 1M annotated article-summary pairs. The dataset covers 44 languages, ranging from low to high-resource ones. We utilized the Amharic portion of the dataset in two ways. First, we took the text and prepared an instructional dataset to test our model's ability to summarize the text. We also created a text expansion task where our model takes the shorter sentence and produces a detailed explanation of the text, the inverse of the text summarization task.

Finally, we used the dataset by Barrault et al. (2019); NLLB Team et al. (2022) to prepare training, validation, and testing for machine translation. Our training dataset is from WMT19 (Barrault et al., 2019), and validation and testing are from (NLLB Team et al., 2022).

The *Amharic spell correction* dataset is designed

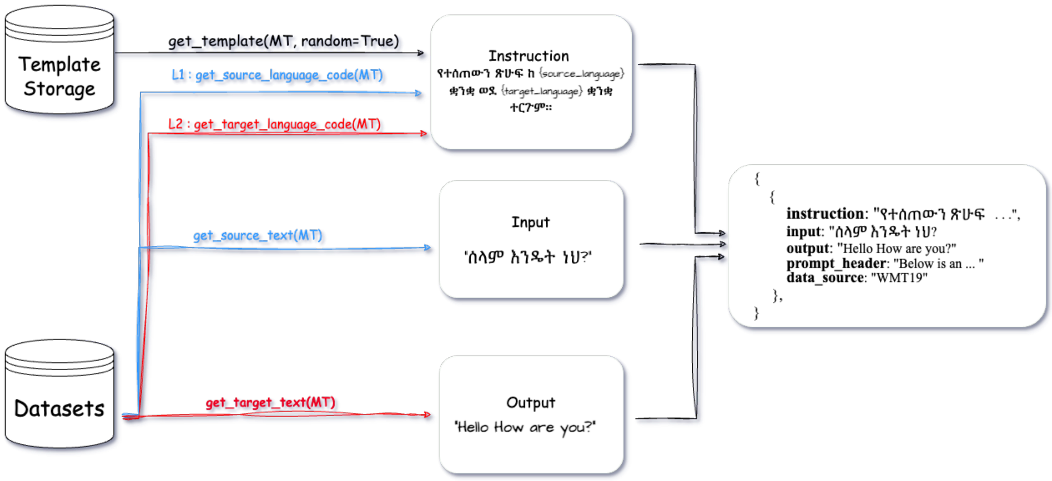


Figure 1: Data processing Pipeline. The pipeline creates instruction data from existing task datasets, and from generative datasets, we collected. All instructions, input, and output are in Amharic except for the MT case, as shown in the picture. The data source will not be used during training.

to assess the effectiveness of models in correcting Amharic spelling errors, covering common misspellings to advance NLP tools for the language. We leveraged Amharic BBC news texts from XL-Sum (Hasan et al., 2021) for this task. We also leveraged the text augmentation library nlpaug (Ma, 2019). We introduced some random character augmentations, including *insertion*, *substitution*, *swapping*, *deletion*, and *word cropping*. This augmentation is done randomly and applied to part of the dataset.

After preparing each dataset, we found that the machine translation dataset we have was significantly larger than the other tasks, so we set a maximum threshold of 10k instructions randomly for the training split of each dataset. For validation and testing, we only used one template per task, and we did not expand the data sizes. More dataset examples and explanations are found in the Appendix B.

3.2 New Custom Datasets

Most of the task datasets we prepared in Section 3.1 did not focus on generation tasks. Generation tasks are less explored for low-resource languages like Amharic, so we created original datasets collected from publicly available sources.

In Amharic, music, stories, and poems represent fascinating cultural artifacts. We have created three new datasets to facilitate the training and evaluation of models' capabilities in processing these tasks. The first track we considered is *religious music lyrics generation*. We included several types of music lyrics in this dataset. We collected the above 2k Amharic spiritual song lyrics from WikiMezmur[2]. Despite the popularity of non-religious music in Ethiopia, finding a freely available source to include this in our data was difficult; hence, our non-religious music data was smaller than the others. To expand this dataset, we split the lyrics into verses and created a new completion task where the input is the first verse and the output is the remaining whole verse.

To understand the story generation abilities of different models, we created a dataset for *Ethiopian folktales*: We collected several Ethiopian folktales from EthopianFolkTales[3]. These stories are collected from all Ethiopian regions. Given that the dataset comprises traditional Ethiopian stories, there is no copyright restriction on them, and our usage is only for research purposes. We also collected *Amharic poem* from several public telegram channels.

For *news title generation*, we collected 50k news title and body pairs from different Amharic news sources such as BBC News[4], Deutsche Welle (DW) news[5], Sheger FM[6], Addis Admass Newspaper[7], and VOA Amharic[8]. To save GPT-4 credits, we did our testing only on the first 1300 items of this data.

[2] https://wikimezmur.org/am
[3] https://www.ethiopianfolktales.com/am
[4] https://www.bbc.com/amharic
[5] https://www.dw.com/am
[6] https://www.shegerfm.com/
[7] https://www.addisadmassnews.com/
[8] https://amharic.voanews.com

3.3 Translated instruction fine-tuning dataset

During LLAMA model self-instructed fine-tuning (Touvron et al., 2023), instruction datasets like Alpaca (Taori et al., 2023) and dolly (Conover et al., 2023) have been widely used. In the work Andersland (2024), machine translation systems were used to translate these datasets into Amharic instruction fine-tuning data. This method is used by most papers that try to adopt LLAMA models for their language, like (Cui et al., 2023). For Amharic versions of alpaca and dolly datasets, we used datasets used by LLAMA-2-Amharic (Andersland, 2024) training. We explored the effect of training a model by using only our relatively clean and human-verified data alone and in combination with this machine translation data.

4 Experiments

We followed Chinese LLAMA (Cui et al., 2023) experiments to perform supervised fine-tuning (SFT) on our dataset using different types of the dataset we created. Figure 2 shows the full training pipeline that summarizes the overall experiment steps we followed. We used codes available on the Chinese-LLAMA-Alpaca[9] repository. We used 4 A100 GPUs with the default parameters in the repository. All training is done for three epochs. All models and evaluation codes will be available in our repository. For the MT task, we also worked on M2M100 (Fan et al., 2021) and NLLB (NLLB Team et al., 2022) models.

During the evaluation of the models, we used `gpt-4-0613` for GPT-4. For our LLAMA-based models, we used fixed generation parameters across the models. We also evaluated various models that purportedly support this language but excluded them due to their inability to perform the required tasks. This is further discussed in Appendix A.

Our main experiment includes:

- Evaluating existing models on our dataset.
- Fine-tuning the model using a dataset detailed in Table 1, referred to as **Walia (task data)**. Unlike the approach taken in the LLAMA-2-Amharic model (Andersland, 2024), this experiment did not incorporate machine-translated instructional data.
- Fine-tuning the model using **Walia (combined data)**, which consists of our prepared dataset along with the machine-translated instructional data previously utilized in the LLAMA-2-Amharic model (Andersland, 2024).
- Fine-tuning **Walia MT**, the MT model we trained to perform the machine translation task in our dataset. In this experiment, we used only MT datasets from Table 1 and scaled them to 200k data rather than 20k as shown in the table.
- Exploring the effect of prompts in existing and available models for Amharic tasks. Additionally, how code mixing affects the performance of models. This include experiments discussed in 4.3.

4.1 Datasets

The first set of experiments we conducted involved evaluating the base LLAMA-2-Amharic model (Wei et al., 2022) on our custom test set, which was created from different NLP task datasets. This will provide us with a baseline performance for Amharic tasks. The next set of experiments used different NLP task datasets that were converted into an instruction dataset by our data generation pipeline. We used LLAMA-2-Amharic model (Andersland, 2024), which is pre-trained using the LLAMA-2 model for the Amharic language and performed supervised instruction fine-tuning on the task datasets. This ensures our model only has access to quality datasets that were adopted from verified NLP tasks. Finally, we combined our instruction dataset with the machine-translated instruction datasets. In the different datasets above, we have capped our training dataset to a maximum of 10k data from individual tasks, as shown in Table 1. We kept fixed instruction and data frequency in our validation and test set to avoid any performance variation because of instruction differences. For machine translation experiments, we created additional data that contains 200k data points from Barrault et al. (2019) and NLLB Team et al. (2022).

4.2 Evaluation Metrics

For selected NLP tasks in this paper, we used different evaluation metrics. For *sentiment analysis* and *news classification* tasks, we have used the weighted f1 score. For these classification tasks, we also keep track of the number of times the model returns output that cannot be classified as one of the classes.

[9] `https://github.com/ymcui/Chinese-LLaMA-Alpaca`

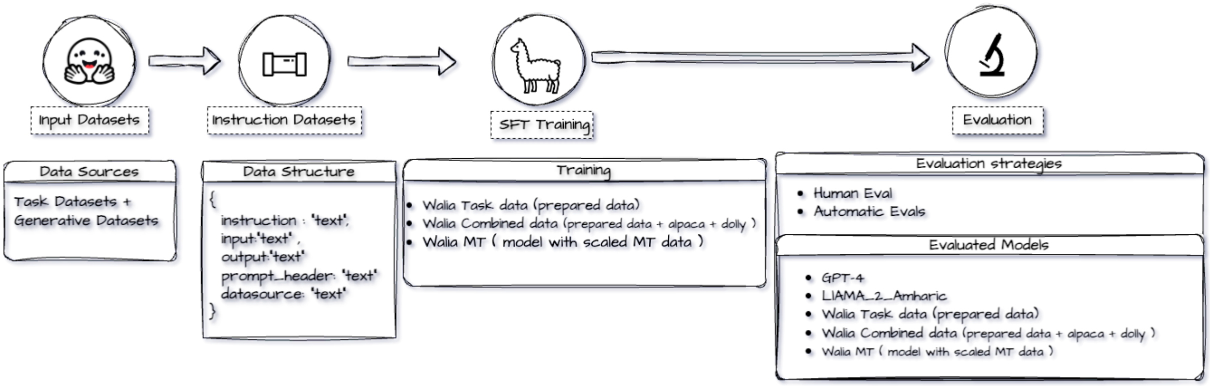


Figure 2: Full training pipeline that summarizes the work done.

For *generation tasks*, we used Rogue (Lin, 2004) scores. We used Rogue scores to evaluate *xlm-summarization*, *reverse summarization*, and *AmharicQA* tasks. We reported Rogue1, Rogue2, and RugueL metrics for generation tasks, but we heavily rely on RogueL for analysis since it focuses on the longest common subsequence rather than n-grams. We observed that most of our generation outputs do not share common n-grams when n is greater than 2, and the generations from systems like GPT-4 tend to be longer where the n-gram comparison methods express the results less. Additionally, we used word-based evaluation metrics, SacreBLEU (Post, 2018) and character-based evaluation metrics, chrF++ (Popović, 2017) automatic evaluation metrics for MT tasks. We added character-based metrics (chrf++) because, for low-resource languages with complex morphologies, chrF++ offers a more robust and adaptable metric compared to word-based metrics like SacreBLEU.

Finally, we performed human evaluation for generative tasks such as music, poetry, and story generation. We sampled 120 individual items and conducted blind reviews using three people for each question. We created a rating system from 1 to 5 with detailed instructions and reported the average rating per task and model.

We did several evaluations for some tasks that were hard to evaluate, e.g., we used accuracy and SacreBLEU scores as evaluation metrics for AmharicQA following the suggestion by Abedissa et al. (2023); Lee et al. (2021). For tasks that require specific text output, we performed character normalization and text cleaning on the outputs before evaluation to avoid analysis because of typos and formatting issues.

In addition to the evaluation methods mentioned above, we explored the possibility of using GPT-4 for evaluation purposes, following the work from the Chinese LLaMA (Cui et al., 2023). Our assessment covered various generation tasks, showing that GPT-4 performs well in most areas. However, it shows inconsistency in scoring due to differences in the rating scale it assigns during each run. In addition, it struggles with evaluating poem and music generation tasks, as it does not fully understand Amharic poetic structure. Additionally, it encounters some challenges in evaluating machine translation, often missing grammatical details in Amharic sentences. Despite these limitations, GPT-4 has the potential for evaluating tasks if it is coupled with manual checks to ensure consistency. We expect similar difficulties in other low-resource languages based on our preliminary findings. While we did not include GPT-4 scores in our current reports due to time and cost constraints, we plan to include them in future research.

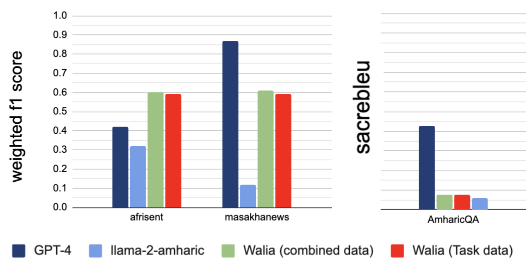


Figure 3: Generation scores: weighted f1 scores for AfriSenti and MasakhaNews (left) and SacreBLEU score for Amharic QA (right)

4.3 Prompt based experiments

Throughout our investigation, we observed using only one instruction for a task introduces a high dependency on the prompt, leading to prompt over-

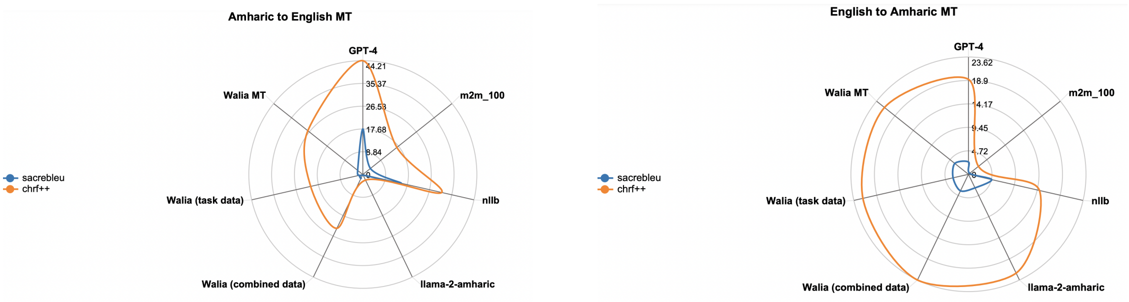


Figure 4: Scores for **machine translation**. Amharic to English translation scores (Right) and English to Amharic translation scores(left).

fitting. In this case, models fail to do tasks like sentiment classification when presented with different instruction prompts. To deal with this problem, we worked on manually produced templates for each task as shown in Table 1.

Additionally, we experimented with the prompt header part of the dataset shown in 1. The prompt header is additional English text stating, `"Below is an instruction that describes a task. Write a response that appropriately completes the request."`. Introducing this in models like GPT-4 yielded a significant reduction in classifiable outputs, which highlights the effectiveness of incorporating clear English instructions to steer the model toward the desired outcome.

Tasks	GPT-4	LLAMA-2-Am	Walia I	Walia II
Text summarization	3.34	0.62	1.13	0.80
Text expansion	3.10	3.22	2.05	2.82
Amharic QA	28.23	2.83	5.37	6.25

Table 2: I = Walia (task data), II = Walia (combined data). **ROUGEL** scores for text summarization, Text expansion and Amharic QA.

5 Results

Below, we discuss the performance of each model we tested by task type and evaluation strategy.

5.1 Classification Results

For classification tasks, we used two metrics. Our models improve LLAMA-2-Amharic scores as shown in AfriSenti, MasakhaNews, and QA tasks in Figure 3. The other metrics we reported measure how many times the model returns one of the categories. For the AfriSenti classification task, 759 and 52 out of 1999 test data are not in any of the three classes for LLAMA-2-Amharic and GPT-4, respectively. Our models reduce these unusable results and do not produce unusable outputs. For MasakhaNews 248, 136, 106, and 3, results are unusable for LLAMA-2-Amharic, Walia (task data), Walia (combined data), and GPT-4, respectively. In MasakhaNews case, GPT-4 tends to take the lead in producing reasonable outputs.

Tasks	GPT-4	LLAMA-2-Am	Walia I	Walia II
Story generation	2.93	1.00	3.60	1.73
Poem completion	2.53	1.46	1.73	2.26
Poem generation	2.13	1.00	2.46	2.00
Religious Lyrics Gen.	2.86	1.46	1.60	1.46
Religious Lyrics Compl.	3.60	1.40	2.13	1.93
Non religious Lyrics Gen.	3.53	1.00	1.60	2.06

Table 3: I = Walia (task data), II = Walia (combined data). Average blind **human evaluation** out of 5, for three people in each task. **(1)** empty or non Amharic text. **(2)** not written in task format. **(3)** written in task format but no consistent idea and spelling errors. **(4)** looks like that specific generation task but has spelling and grammar errors. **(5)** this looks like a perfect generation of the task. Underlined text indicates cases where we see improvement compared to LLAMA-2-Amharic.

5.2 Generation Results

As explained in Section 4.2, we focus on RogueL metrics for our analysis. Across text summarization and AmharicQA, GPT-4 takes the lead, showing the model's generation ability is very high. We were able to improve the LLAMA-2-Amharic model's ability for this task using our data, as shown in Table 2.

We conducted a human evaluation for the models that do not have fixed gold labels, as shown in Table 3. Table 3 result shows that the generation ability of LLLAMA-2-Amharic can be enhanced by adding generation-specific datasets. Walia lacks an understanding of the specific formatting of texts because of the limitations in our pre-processing. However, it shows significant improvement where the LLAMA-2-Amharic fails to understand the query.

5.3 Machine Translation (MT)

For the MT task we evaluated two open-source sequence-to-sequence models (M2M100 (Fan et al., 2021) and NLLB (NLLB Team et al., 2022)), GPT-4, LLAMA-2-Amharic, and our models. Figure 4 shows SacreBLEU and chrF++ results for the above MT models. As shown in the figure, from MT models, GPT-4 showed better results than the other models when using English as the target language. However, our models showed results comparable to the NLLB and m2m100 models and outperformed the LLAMA-2-Amharic model for the Amharic-English translation. For the English-Amharic translation, the NLLB model outperformed the others in the SacreBLEU score, while our models showed comparable results and outperformed GPT-4, LLAMA-2-Amharic and m2m100 models in this translation direction. In our MT evaluation, we noticed irregularities between the results of the two evaluation metrics. Since SacreBLEU is a word-based metric, the results show that the scores are too low. This shows that using only automatic evaluation metrics makes interpreting and generalizing the results hard. We will add metrics like human evaluation to evaluate MT results in the future.

6 Conclusion and Future Works

In this work, we created Amharic instruction fine-tuning dataset, evaluated the performance of existing and our fine-tuned models in the new dataset, and explored the effect of carefully curated datasets on the models' performance. We observed a possibility of reusing task-specific datasets to improve the generation and task performance of the existing LLAMA-2-Amharic model.

Our data generation pipeline that generates instruction datasets from task datasets can be used for the generation of similar datasets for other languages given template instructions. We are working on this kind of dataset for all languages included in MasakaNER (Adelani et al., 2021), MasakhaNews (Adelani et al., 2023), AfriSenti (Muhammad et al., 2023) and more to improve multilingual LLAMA models. We plan to open-source the instruction datasets with the generation code.

Moving forward, we aim to improve both the quality and volume of the data utilized. Task-specific dataset creations are meant to complement, not replace, language-specific instruction dataset creations, and we plan to work on creating quality instruction datasets in addition to using existing task datasets. We also plan to explore the relevance of using LLMs for evaluation in low-resource languages like Amharic and incorporate LLMs to evaluate our LLMs.

7 Limitations

One limitation we observed in our work is the lack of reliable generation metrics for our tasks. The models tend to generate wordy and explained outputs despite our attempts to design the instruction template specifically. As a solution, we used several metrics that can express one task's ability and reported the best-suited one.

In our current evaluation of all the models, we observed significant limitations while preforming the spell correction and NER tasks. For Amharic spell correction all four generation models, including GPT-4, try to generate other things related to the text, and the word error rate for all of them is close to 99%.

We have yet to explore the effect of using machine-translated instruction datasets for building language-specific LLMs with regard to introducing cultural bias.

References

Tilahun Abedissa, Ricardo Usbeck, and Yaregal Assabie. 2023. Amqa: Amharic question answering dataset. *arXiv preprint arXiv:2303.03290*.

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, et al. 2021. Masakhaner: Named entity recognition for african languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

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, Abraham Owodunni, Nnaemeka Obiefuna, Muhidin Mohamed, Shamsuddeen Hassan Muhammad, Teshome Mulugeta Ababu, Saheed Abdullahi Salahudeen, Mesay Gemeda Yigezu, Tajuddeen Gwadabe, Idris Abdulmumin, Mahlet Taye, Oluwabusayo Awoyomi, Iyanuoluwa Shode, Tolulope Adelani, Habiba Abdulganiyu, Abdul-Hakeem Omotayo, Adetola Adeeko, Abeeb Afolabi, Anuoluwapo Aremu, Olanrewaju Samuel, Clemencia

Siro, Wangari Kimotho, Onyekachi Ogbu, Chinedu Mbonu, Chiamaka Chukwuneke, Samuel Fanijo, Jessica Ojo, Oyinkansola Awosan, Tadesse Kebede, Toadoum Sari Sakayo, Pamela Nyatsine, Freedmore Sidume, Oreen Yousuf, Mardiyyah Oduwole, Kanda Tshinu, Ussen Kimanuka, Thina Diko, Siyanda Nxakama, Sinodos Nigusse, Abdulmejid Johar, Shafie Mohamed, Fuad Mire Hassan, Moges Ahmed Mehamed, Evrard Ngabire, Jules Jules, Ivan Ssenkungu, and Pontus Stenetorp. 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.

Michael Andersland. 2024. Amharic llama and llava: Multimodal llms for low resource languages. *Preprint*, arXiv:2403.06354.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Mike Conover, Matt Hayes, Ankit Mathur, Xiangrui Meng, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Xin Reynold. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

Guanting Dong, Hongyi Yuan, Keming Lu, Chengpeng Li, Mingfeng Xue, Dayiheng Liu, Wei Wang, Zheng Yuan, Chang Zhou, and Jingren Zhou. 2023. How abilities in large language models are affected by supervised fine-tuning data composition. *arXiv preprint arXiv:2310.05492*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Tahmid Hasan, Abhik Bhattacharjee, Md Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M Sohel Rahman, and Rifat Shahriyar. 2021. Xl-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Mojan Javaheripi, Sébastien Bubeck, Marah Abdin, Jyoti Aneja, Sebastien Bubeck, Caio César Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, et al. 2023. Phi-2: The surprising power of small language models.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Joongbo Shin, and Kyomin Jung. 2021. KPQA: A metric for generative question answering using keyphrase weights. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2105–2115, Online. Association for Computational Linguistics.

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

Peiqin Lin, Shaoxiong Ji, Jörg Tiedemann, André F. T. Martins, and Hinrich Schütze. 2024. Mala-500: Massive language adaptation of large language models. *Preprint*, arXiv:2401.13303.

Edward Ma. 2019. Nlp augmentation. https://github.com/makcedward/nlpaug.

Yasmin Moslem, Rejwanul Haque, and Andy Way. 2023. Fine-tuning large language models for adaptive machine translation. *arXiv preprint arXiv:2312.12740*.

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 Destaw Belay, Wendimu Baye Messelle, Hailu Beshada Balcha, Sisay Adugna Chala, Hagos Tesfahun Gebremichael, Bernard Opoku, and Steven Arthur. 2023. AfriSenti: A Twitter sentiment analysis benchmark for African languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13968–13981, Singapore. Association for Computational Linguistics.

Marta R. Costa-jussà NLLB Team, 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, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

Matt Post. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, et al. 2023. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Atnafu Lambebo Tonja, Tadesse Destaw Belay, Israel Abebe Azime, Abinew Ali Ayele, Moges Ahmed Mehamed, Olga Kolesnikova, and Seid Muhie Yimam. 2023. Natural language processing in Ethiopian languages: Current state, challenges, and opportunities. In *Proceedings of the Fourth workshop on Resources for African Indigenous Languages (RAIL 2023)*, pages 126–139, Dubrovnik, Croatia. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*.

Jason Wei, Maarten Paul Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew Mingbo Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. A paradigm shift in machine translation: Boosting translation performance of large language models. *arXiv preprint arXiv:2309.11674*.

Hang Zhang, Xin Li, and Lidong Bing. 2023a. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. 2023b. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*.

A Experimental details

We adopted our LLAMA-2 instruction tuning experimental code from Chinese-LLAMA-Alpaca[10] repository by (Cui et al., 2023). we performed our instruction finetuning for three epochs on 4 A100 GPUs using the parameters in Table 4. In the generation phase, we used parameters in Table 5 for all models except GPT-4.

Parameter	Value
epoch	3
lr	1e-4
lora_rank	8
lora_alpha	32
lora_dropout	0.05
per_device_train_batch_size	1
per_device_eval_batch_size	1
gradient_accumulation_steps	8

Table 4: Training Parameters

Parameter	Value
seed	42
do_sample	True
min_length	None
top_p	1.0
temperature	1.0
top_k	5
repetition_penalty	5.0
length_penalty	1

Table 5: Configuration settings for token generation.

B Dataset details

Figure 5 shows how we repurposed existing sentiment analysis data to convert it into an instruction dataset. We utilized a task template selected at random from our collection. The number of templates collected for each task is shown in the table. The reason for keeping the prompt header in the English language is discussed in Section 4.3.

To avoid instruction overfitting, as discussed in Section 4.3, we collected a variety of instructions, like the example shown in Figure 6. The example displays different instructions that can be paired with a machine translation dataset to create a machine translation instruction dataset. This step is repeated for all tasks, including tasks like poem generation, which we created by collecting from different websites.

Due to the difficulty we faced in evaluating generation tasks, we employed human evaluation. Figure 7 shows how evaluators scored each generation output for the case of the story generation task.

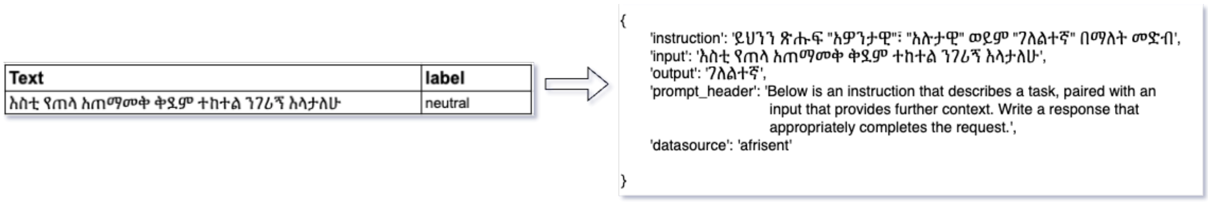


Figure 5: Example data output from our dataset creation pipeline. 1

C Result details

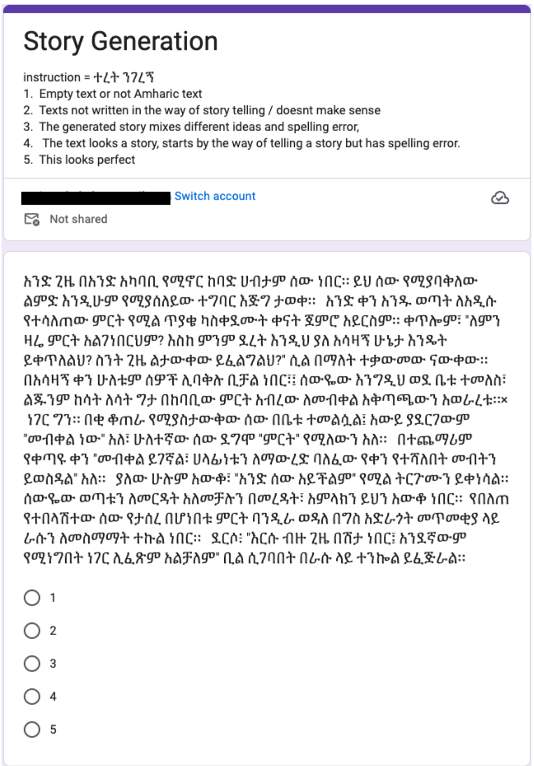
In Table 6, we present detailed scores for text summarizing, text expansion, and Amharic QA. The choice of the right parameter needs further exploration but RogueL is a good metrics to show improvement because it doesn't depend on specific n-gram similarities.

In Figure 8 we demonstrate that even if the model's output is not an exact match, we have created a pipeline to verify and identify the correct output from the generated sequence. We limit our-self to GPT-4 as an external model because it's not explored in other well-known models. Additionally, we reveal that the mala-500 (Lin et al., 2024) model produces unrelated outputs, which merits deeper examination.

[10] https://github.com/ymcui/Chinese-LLaMA-Alpaca

ID	Instruction templates
1	የተሰጠውን ጽሁፍ ከ {source_language} ቋንቋ ወደ {target_language} ቋንቋ ተርጉም:: *(Translate the writing given in {source_language} into {target_language} text)*
2	የተሰጠውን የ{source_language} አረፍተ ነገር ወደ {target_language} ተርጉም *(Translate {source_language} sentence to {target_language} sentence)*
3	ይህ የ {source_language} ፅሁፍ በ{target_language} እንዴት ሊፃፍ ይችላል (This is {source_language} sentence , How can it be written in {target_language})

Figure 6: Example templates for machine translation task with English translation. By using random instructions for tasks, we ensure that the model does not fit the specific instructions for tasks.


Story Generation

instruction = ተረት ንገረኝ
1. Empty text or not Amharic text
2. Texts not written in the way of story telling / doesnt make sense
3. The generated story mixes different ideas and spelling error,
4. The text looks a story, starts by the way of telling a story but has spelling error.
5. This looks perfect

[redacted] Switch account
Not shared

አንድ ጊዜ በአንድ አካባቢ የሚኖር ከባድ ሀብታም ሰው ነበር። ይህ ሰው የሚያባቅለው ልምድ እንዲሁም የሚያሰለይው ተግባር እጅግ ታወቀ። አንድ ቀን አንዱ ወጣት ለአዲሱ የተሳለጠው ምርት የሚል ጥያቄ ካስቀደሙት ቀናት ጀምሮ አይርስም። ቀጥሎም፣ "ለምን ዛሬ ምርት አልገነበርህም? እስከ ምንም ደረት እንዲህ ያለ አሳዛኝ ሁኔታ እንዴት ይቀጥለልህ? ስንት ጊዜ ልታውቀው ይፈልግልህ?" ሲል በማለት ተቃውመው ናውቀው። በአሳዛኝ ቀን ሁለቱም ሰዎች ሊባቅሉ ቢቻል ነበር፡፡ ሰውዬው እንግዲህ ወደ ቤቱ ተመለስ፣ ልጁንም ከሳት ለሳት ግታ በከባቢው ምርት አብረው ለመብቀል አቅጣጫውን አወራረቱ።× ነገር ግን። በቂ ቅጠራ የሚያስታውቅው ሰው በቤቱ ተመልሷል፤ አውይ ያደርገውም "መብቀል ነው" አለ፣ ሁለተኛው ሰው ደግሞ "ምርት" የሚለውን አለ። በተጨማሪም የቀጣዩ ቀን "መብቀል ይገኛል፣ ህላፊነቱን ለማውረድ ባለፈው የቀን የተሻለበት መብትን ይወስዳል" አለ። ያለው ሁሉም አውቆ፣ "አንድ ሰው አይችልም" የሚል ትርጉሙን ይቀነሳል። ሰውዬው ወጣቱን ለመርዳት አለመቻሉን በመረዳት፣ አምላክን ይህን አውቆ ነበር። የበለጠ የተበላሸተው ሰው የታሰረ በሆነበቱ ምርት ባንዲራ ወዳለ በግስ አድራጎት መጥመቂያ ላይ ራሱን ለመስማማት ተኩል ነበር። ደርሶ፣ "እርሱ ብዙ ጊዜ በሽታ ነበር፤ አንደኛውም የሚነግበት ነገር ሊፈጽም አልቻለም" ቢል ሲገባበት በራሱ ላይ ተንኰል ይፈጅራል።

1
2
3
4
5

Figure 7: Form used for human annotation with labeling instruction. We see in the figure how one question of the sample story generation task is being validated.

Tasks	GPT-4	LLaMA-2-Amharic	Walia (Task data)	Walia (combined data)
Text summarization	3.41/0.11/3.34	0.61/0.00/0.62	1.12/0.00/1.13	0.78/0.00/0.80
Text expansion	3.11/0.11/3.10	3.35/0.02/3.22	2.14/0.02/2.05	2.89/0.10/2.82
Amharic QA	28.22/8.00/28.23	2.83/0.66/2.83	5.36/0.67/5.37	6.34/1.56/6.25

Table 6: **Rogue1/Rogue2/RogueL** scores for text summarization, Text expansion and AmharicQA

		output analysis
instructions	የተሰጠው ጽሑፍ አስተያየት ምን አይነት ነው? "አዎንታዊ"፣ "አሉታዊ" ወይም "ገለልተኛ" የምል ምላሽ ስጥ።	
input	@user ክብር እና ምስጋና ለዓለማት ፈጣሪ ይሁን'	
output	አዎንታዊ	
prompt_header	Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.	
GPT-4 output	ይህ የተሰጠው ጽሑፍ ዐዎንታዊ አስተያየት ነው።	correct
llama-2-amharic output	ጹሑፉ አዎንታዊ መልእክት ያስተላልፋል።	correct
Walia (combined data) output	አዎንታዊ	correct
Walia (task data) output	አዎንታዊ	correct
mala-500 output	(Ammohay) \'amha-r\', ammoroy\': «assistent» - used for both males & females; e., "\'amma roo(/ father)-asstant or helper... more » Less • new window -- edit translation-- disclaimers * This dictionary does not contain any content that falls into these categoriesmore about this | how do I know which category my entry belongsless About · Contact Us All'}	incorrect

		output analysis
instructions	የተሰጠው ጽሑፍ አስተያየት ምን አይነት ነው? "አዎንታዊ"፣ "አሉታዊ" ወይም "ገለልተኛ" የምል ምላሽ ስጥ።	
input	@user @user አዎ አረ ቤቱ ይቁጠረው'	
output	አሉታዊ	
prompt_header	Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.	
GPT-4 output	አዎንታዊ	incorrect
llama-2-amharic output	ቤቱን ማንቀሳቀስ!	incorrect
Walia (combined data) output	ገለልተኛ	incorrect
Walia (task data) output	ገለልተኛ	incorrect
mala-500 output	\'[Ancient Greek word for slave][1]] (or servant) or helper; used when there are many servants who assist their lord/lady of higher rank.[]. For example Assistent Ministers can be appointed by minister-of more senior ranks but not yet Prime minster himself They could also work at lower level positions like Manager /ator etc... In this context its meaning has evolved into one that means helping people out especially those less fortunate than'	incorrect

Figure 8: Example of analysis we did on the model outputs.