Story Generation with Large Language Models for African Languages

Catherine Nana Nyaah Essuman
African Institute for Mathematical Sciences
University of Cape Town
catherine@aims.ac.za

Jan Buys
University of Cape Town
jan.buys@uct.ac.za

Abstract

The development of Large Language Models (LLMs) for African languages has been hindered by the lack of large-scale textual data. Previous research has shown that relatively small language models, when trained on synthetic data generated by larger models, can produce fluent, short English stories, providing a data-efficient alternative to large-scale pretraining. In this paper, we apply a similar approach to develop and evaluate small language models for generating children's stories in isiZulu and Yoruba, using synthetic datasets created through translation and multilingual prompting. We train six language-specific models varying in dataset size and source, and based on the GPT-2 architecture. Our results show that models trained on synthetic low-resource data are capable of producing coherent and fluent short stories in isiZulu and Yoruba. Models trained on larger synthetic datasets generally perform better in terms of coherence and grammar, and also tend to generalize better, as seen by their lower evaluation perplexities. Models trained on datasets generated through prompting instead of translation generate similar or more coherent stories and display more creativity, but perform worse in terms of generalization to unseen data. In addition to the potential educational applications of the automated story generation, our approach has the potential to be used as the foundation for more data-efficient low-resource language models.

1 Introduction

In recent years, pretrained transformer language models have been used as the foundation of NLP systems for text generation, understanding and summarizing, and information extraction (Razumovskaia et al., 2024). However, most of the advancements have been concentrated on high-resource languages (HRLs) such as English and French, leaving low-resource languages (LRLs) and African languages in particular underrepresented in advancements in Language Models (LMs). Some of these languages, despite having millions of speakers, lack sufficient data online to train robust LMs or develop and deploy systems that can cater for their speakers. While many efforts have been made to create LMs, the lack of suitable datasets remains a significant challenge. In response, recent research has focused on creating datasets for African languages, either through manual annotations or through synthetic data generation (Adelani et al., 2023; Tonja et al., 2024; Adelani et al., 2025).

The use of synthetic data has proven to be essential for training LMs in low-resource settings. Liu et al. (2024) argues that synthetic data addresses data scarcity, allowing models to generalize better while high-quality synthetic data helps to avoid biases. Gunasekar et al. (2023) also demonstrates that synthetic generated datasets with high quality can enhance model learning. Our work is motivated by TinyStories (Eldan and Li, 2023) which uses curated synthetic data consisting of short stories using simple language to train small language models. That work shows that high-quality synthetic data can enable small models to match the performance of larger models by focusing on coherent and diverse content.

The aim of this paper is to investigate whether a similar approach can be applied to generate high-quality synthetic stories in low-resource languages, which can then be used to train small but capable language models. We train six models based on the GPT-2 architecture from scratch for isiZulu and Yoruba, using synthetic datasets of different sizes and approaches to generate children's stories. We evaluate the performance of these models using both qualitative and quantitative analysis in order to investigate whether LMs trained on synthetic LRL data can produce coherent and fluent stories. We compare the performance of models trained on

Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025), pages 115–125
July 31, 2025 ©2025 Association for Computational Linguistics

the translated stories to that of models trained on stories generated by prompting from a multilingual model.

2 Related Work

Generating stories with LLMs has proven to be a promising approach to generating texts that are coherent and appealing. TinyStories (Eldan and Li, 2023) showed that even small-scale models can generate fluent short stories, offering a promising approach to train effective story generation models with less computational resources. This result offers a promising path to develop models for low-resource languages (LRLs) in settings that also frequently lack the infrastructure required for large-scale pretraining.

Razumovskaia et al. (2024) investigated cross-lingual story generation by generating stories in multiple languages from a single plan in English. This work complements the findings of Eldan and Li (2023), drawing attention to the versatility of LLMs across languages, which is critical for African languages lacking considerable data. The two studies stress the importance of building LLMs that are capable of generating coherent stories in resource constrained environments, which is a major challenge for African languages. In a different approach, the GROVE framework (Wen et al., 2023) uses Retrieval-Augmented Generation (RAGs) to enhance the coherence and complexity of stories. This approach further underscores the importance of extra information (whether through cross-lingual plans or the retrieval of evidence) to improve the quality of stories. Both methods show that making use of external information can improve the capability of LLMs generations.

In our research, we build upon these ideas by using two methods for synthetic data generation: translation of existing stories into isiZulu and Yoruba using a multilingual translation model, and directly prompting a multilingual model to generate stories in both target languages, in order to create LMs which can generate stories in isiZulu and Yoruba.

3 Methodology

3.1 Dataset Generation

We follow two approaches to generate synthetic datasets: machine translation of English stories to the target languages, and prompting a multilingual language model to generate stories in the target languages.

We use the `TinyStories`[1] (Eldan and Li, 2023) dataset as the source of English stories to be translated. This dataset consists of stories generated by prompting GPT-3.5 and GPT-4. The prompts selected random keywords from a set of 3,000 nouns, verbs, and adjectives to generate stories aimed at children aged 3 to 5 years. The dataset contains approximately 2 million unique stories, but for the purpose of this study, an eighth (250,000), of these stories were used. The stories were translated from English into both isiZulu and Yoruba using the state-of-the-art Seamless Massively Multilingual and Multimodal Machine Translation (Seamless M4T-V2) model version 2 (Communication et al., 2023).

For the second data generation approach we utilized **`AfroLlama`**[2], a multilingual text generation model developed by Jacaranda Health, which was fine-tuned from Meta AI's Llama 3, to generate synthetic stories directly in the isiZulu and Yoruba. To generate the stories, we created prompts varying in content but with a consistent structure in the target languages aimed at guiding the model to produce children stories. Example prompts are shown in Table 1. We generated 10,000 unique short stories about different characters, with a clear beginning, middle and end.

All together we created six synthetic datasets:

- **isiZulu and Yoruba Large: 250,000** stories from the TinyStories dataset, translated into isiZulu and Yoruba.
- **isiZulu and Yoruba Mini: 10,000** stories sampled from the initial set of 250,000 TinyStories, translated into isiZulu and Yoruba.
- **isiZulu and Yoruba Prompt: 10,000** stories generated by prompting Afro Llama to generate stories in isiZulu and Yoruba.

We split each of the datasets into training (70%), validation (20%) and evaluation (10%) sets.

3.2 Model Initialization & Pretraining

For the models trained in this study, we initialized the weights randomly, meaning that no pretrained model weights were used during the training process. We trained the models entirely on the synthetic datasets generated from our corpus, with no

[1] https://huggingface.co/datasets/roneneldan/TinyStories
[2] https://huggingface.co/Jacaranda/AfroLlama_V1

	Prompt
1	Ko̱ itan awo̱n o̱mo̱de ni Yoruba nipa Lily ati Max ti o gba e̱bun airote̱le̱, o ni ipari ti o dara.
2	So̱ itan awo̱n o̱mo̱de ni Yoruba nibiti Emma nilo lati gafara fun o̱re̱ re̱ Thabo, o ni opin irora.
3	Bhala indaba emfushane yezingane ngesiZulu lapho uZandile no-Oliver behlangana nesilwane esikhulumayo, inesiphetho esihle kakhulu.
4	Xoxa izindaba zezingane ngesiZulu ngoNomsa owafunda izifundo ezibalulekile ngokuhlanganyela.

Table 1: Prompts for story generation in Yobura (1 & 2) and Zulu (3 & 4)

use of external corpora or multilingual pretraining. While this approach allows for an investigation of model performance based purely on the synthetic data, the lack of real-world language exposure may limit the models' ability to generalize effectively to unseen data. Training from scratch on synthetic data could result in biases that differ from those seen in models pre-trained on real-world data. Our motivation was to isolate the effects of our synthetic dataset and avoid potential transfer effects from external corpora.

3.3 Pre-processing & Model Training

We trained six language-specific models, one for each of the generated synthetic datasets. The text was tokenized with Byte-Level Byte-Pair Encoding (BPE) (Wang et al., 2020) for isiZulu and SentencePiece BPE (Kudo and Richardson, 2018) for Yoruba. Table 2 shows the number of tokens in each of the datasets.

We train story generation models using the GPT-2 (Generative Pre-trained Transformer 2) architecture, which is a transformer-based autoregressive language model (Radford et al., 2019). At its core lies the transformer decoder block introduced by (Vaswani et al., 2017), which uses self-attention mechanisms to process sequential data. Our implementation is based on Andrej Kaparthy's nanoGPT model [3]. Our models are smaller than the "small" variant of GPT-2, with the specifications given in Table 3. The model size is 30.59M parameters for the isiZulu models and 29.20M parameters for the Yoruba models. The model sizes were chosen in proportion to the size of the available training data, while allowing for computational feasibility in a low-resource setting. The aim is to show that with an even smaller model, fluent, coherent stories can still be generated in a low-resource language.

3.4 Generating Stories from the Trained Models

We use the models trained on our isiZulu and Yoruba datasets to generate new stories. We evaluate the models by evaluating the quality of the generated stories. Some evaluations also use the held-out evaluation datasets from the original datasets generated by translation or prompting. We generate 1,000 stories from each of the models to ensure there is enough data to evaluate the performance of the models based on the chosen evaluation metrics. To generate a story we prompt the model with the start token and sample stories using the hyperparameter values given in Table 4. We set these hyperparameters to ensure a balance between diversity and coherence in the generated stories. The maximum number of tokens of 512 is equal to the model context length during training. A temperature of 0.7 is used to ensure diversity in the model generations, while top-k sampling with k=50 limits the number of possible next words from which the model can sample to maintain coherence.

3.5 Evaluation Metrics

In order to assess the quality of our generated sample stories and the performance of our models, we employ a number of evaluation metrics:

1. **Perplexity** is a normalized measure of the probability of text scored by the model:

$$\text{PPL} = e^{-\frac{1}{N}\Sigma_{n=1}^{N} logP(w_n|w_1,w_2,\cdots,w_{n-1})}, \quad (1)$$

   which exponentiates the average Negative Log Likelihood, where N is the number of tokens in the evaluation set and $w_1, w_2, \cdots, w_N$ are the tokens.

2. **Diversity** We consider two metrics. Lexical Diversity, also known as Type-Token Ratio (TTR), measures the variety of vocabulary

[3] https://github.com/karpathy/nanoGPT

Dataset	ZuluLarge	YorubaLarge	ZuluMini	YorubaMini	ZuluPrompt	YorubaPrompt
Train	28,809,839	41,056,131	1,193,167	1,694,424	1,035,290	1,208,521
Validation	7,200,689	10,264,075	297,977	424,016	259,003	302,424
Evaluation	3,928,142	5,658,137	161,736	235,508	141,535	164,736

Table 2: Dataset sizes (number of tokens) for the generated synthetic datasets

Hyperparameter	Value
Layers	6
Attention Heads	6
Embedding Dimension	384
Dropout Rate	0.2

Table 3: Transformer architecture hyperparameters

Hyperparameter	Value
Maximum new Tokens	512
Temperature	0.7
Top-k Sampling	50

Table 4: Story Generation Hyperparameters

used in the generated stories:

$$\text{TTR} = \frac{\text{Number of Unique Words}}{\text{Total Number of Words}} \tag{2}$$

Semantic Similarity measures how different the generated stories are from each other in terms of meaning, which helps us to understand if our model is creative in generating unique stories:

$$\text{Cosine Similarity} = \frac{\mathbf{A} \cdot \mathbf{B}}{||\mathbf{A}|| \times ||\mathbf{B}||} \tag{3}$$

where $\mathbf{A}$ and $\mathbf{B}$ are the sentence embeddings of two different stories, and $||\mathbf{A}||$ and $||\mathbf{B}||$ are their respective magnitudes. We compute the semantic similarity score by performing pairwise comparisons between all stories within each generated set and evaluation set, averaging the cosine similarity scores.

3. **Quality evaluation using Gemini** We follow the methodology of Eldan and Li (2023), which prompted GPT-4 to score the generated stories. We prompt Gemini 1.5 Pro, an LLM developed by Google, to score the generated and reference evaluation set stories from each of the six models based on Grammar, Coherence, Plot, and Creativity. Each of the categories are scored out of 10 and an overall score is also given. The prompt used for this evaluation is shown in Table 5.

Prompt
Grade these isiZulu and Yoruba stories out of 40 based on:
1. Grammar (10)
2. Coherence (10)
3. Plot (10)
4. Creativity (10)
Provide short comments (1-2 sentences) for each category in the format:
- Grammar: [score], [comment]
- Coherence: [score], [comment]
- Plot: [score], [comment]
- Creativity: [score], [comment]
Overall Score: [score]

Table 5: Prompt for Story Evaluation

Model	Train Loss	Val Loss
isiZuluPlus	2.566	2.687
YorubaPlus	1.906	1.948
isiZuluLite	0.442	5.424
YorubaLite	1.094	2.547
isiZuluGuide	0.417	4.035
YorubaGuide	0.758	2.660

Table 6: Train and Validation Losses for the Models

4 Results and Discussion

We refer to the six trained models as follows:

- **isiZulu and Yoruba Plus Models:** This refers to the models trained on the isiZulu Large and Yoruba Large datasets.
- **isiZulu and Yoruba Lite Models:** This refers to the models trained on the isiZulu Mini and Yoruba Mini datasets.
- **isiZulu and Yoruba Guide Models:** This refers to the models trained on the isiZulu Prompt and Yoruba Prompt datasets.

Table 6 shows the training and validation losses for each of the six models after training for 20 epochs.

4.1 Model Evaluation

The results for the model evaluation, using Perplexity and the Diversity Scores (Token-Type Ratio and

Model	Perplexity of ↓ Generated Stories	Perplexity of ↓ Evaluation Sets	TTR for ↑ Generated Stories	TTR for ↑ Evaluation Sets	Semantic Similarity of ↑ Generated Stories	Semantic Similarity of ↑ Evaluation Sets
isiZuluPlus	34.37	15.82	0.0653	0.1210	0.6549	**0.7603**
YorubaPlus	**5.92**	7.47	0.0158	0.0284	0.7505	0.7474
isiZuluLite	40.83	14.33	**0.0743**	0.1202	0.7044	0.7657
YorubaLite	7.43	**7.13**	0.0181	0.0273	**0.7567**	0.7479
isiZuluGuide	15.98	154.05	0.0545	**0.0816**	0.7453	0.7530
YorubaGuide	14.42	275.62	0.0208	0.0316	0.7451	0.7449

Table 7: Perplexity, Type-Token Ratio, and Semantic Similarity for Generated Stories & Evaluation Datasets

Semantic Similarity) are given in Table 7.

We calculate the **perplexity** of each set of generated stories with the model used to generate the respective set of stories. The comparison reveals several key insights. YorubaPlus has the lowest perplexity of 5.92 for generating stories, indicating that it is more confident and accurate in generating coherent stories compared to the other models. This is in contrast to isiZuluPlus, which has a higher perplexity of 34.37, suggesting isiZuluPlus struggles more to generate coherent, accurate stories.

Additionally we calculate the perplexity of the Large evaluation sets for each language across each of the models. This allows comparing these perplexity results across models; lower perplexity indicates better generalization. On the evaluation sets, YorubaPlus still performs well with a perplexity of 7.47, whereas isiZuluPlus has a perplexity of 15.82, which is better but still higher than YorubaPlus, showing that the Yoruba model generalizes more effectively.

Similarly, isiZuluLite has a higher perplexity of 40.83 for generating stories, indicating that it is less confident in generating coherent text compared to YorubaLite, which has a perplexity of 7.43. YorubaLite performs significantly better in both generation and evaluation, with perplexities of 7.43 and 7.13, respectively, suggesting better generalization and more accurate generation.

When analyzing models trained with datasets generated through prompting, isiZuluGuide has a perplexity of 15.98 for generating stories, which is lower than that of isiZuluLite but still relatively high. However, isiZuluGuide displays a much higher perplexity of 154.05 on the evaluation sets, indicating that although it generates relatively good stories consistent with the training data, it struggles to generalize to unseen data. Note that this mismatch is due to the different training data source (which the Lite and Plus models have the same training data source, just using different data sizes).

For YorubaGuide, the perplexity for generating stories is 14.42, which is higher than YorubaLite (7.43) but lower than isiZuluGuide. However, the evaluation perplexity for YorubaGuide is 275.62, which is much higher than YorubaLite's 7.13, suggesting that YorubaGuide has significant challenges in generalization.

This **TTR (Type-Token Ratio)** is in the range of zero and one, where a higher TTR value indicates more diverse vocabulary usage. The comparison of TTR scores highlights several trends based on training data size and dataset type. isiZuluLite has a higher TTR of 0.0743 compared to isiZuluPlus' 0.0653, suggesting that models trained on smaller datasets may exhibit more lexical diversity, with this effect being more noticeable in isiZulu than in Yoruba.

When comparing models trained on prompted versus translated datasets, isiZuluGuide shows a lower TTR of 0.0545, indicating less vocabulary diversity than isiZuluLite (0.0743) and isiZuluPlus (0.0653). Conversely, models trained on Yoruba datasets generated through prompting (YorubaGuide: 0.0208) show more lexical diversity than the models trained on translated datasets (YorubaLite: 0.0181, YorubaPlus: 0.0158). Furthermore, evaluation sets consistently exhibit higher TTR scores than generated stories, indicating that evaluation datasets have richer vocabulary. For example, the TTR of the isiZuluLite evaluation set is 0.1202 compared to the generated stories' TTR of 0.0743. This suggests that while the models capture some token variety, the generated stories still lack the vocabulary richness seen in the evaluation sets, highlighting limitations in vocabulary diversity during story generation.

The **semantic similarity** scores are in the range of zero to one, with a score close to zero indicating no similarity between stories, and a score close to one indicating high similarity. The comparison of semantic similarity scores highlights the impact of training data size and dataset type. Models trained on larger datasets (250,000 stories) tend to have lower semantic similarity scores compared to those trained on smaller datasets (10,000 stories) for generated stories. For example, isiZuluPlus scores 0.6549, while isiZuluLite scores 0.7044, suggesting that larger datasets lead to slightly less similar stories in isiZulu. However, this trend is not as pronounced in Yoruba models, where YorubaLite scores 0.7567 and YorubaPlus 0.7505, indicating that dataset size has less impact on generated story similarity for Yoruba.

Models trained on prompt-generated datasets (isiZuluGuide: 0.7453, YorubaGuide: 0.7451) show more consistent semantic similarity scores compared to those trained on translated datasets, like isiZuluLite (0.7044), suggesting that models trained on datasets generated through prompting leads to more stable story generation. When comparing the generated stories to the evaluation sets, the evaluation datasets consistently show higher similarity scores. For example, isiZuluPlus' generated stories score 0.6549, while the evaluation set score 0.7603. This pattern is seen across all models, with the isiZulu models showing larger gaps between generated stories and the evaluation set, indicating more diversity in the generated stories compared to the evaluation dataset. Note that here we use the evaluation sets corresponding to each of the models, which explains why the Plus and Lite model results are very close to each other, with the Guide results diverging.

For the **Quality Evaluation**, we score a subset of 200 stories generated from each of the models and 200 stories from our evaluation datasets (which were generated through translation or prompting). Gemini 1.5 Pro is prompted to give a score of out 10 for each of the following categories: Grammar, Coherence, Plot and Creativity. Table 8 presents the average scores for each of the categories.

When comparing models trained on 250,000 stories (isiZuluPlus and YorubaPlus) to those trained on 10,000 stories (isiZuluLite and YorubaLite), the impact of dataset size is evident. Larger datasets result in better performance in grammar and coherence, as seen with YorubaPlus scoring 7.196 in grammar compared to YorubaLite's 6.865, and isiZuluPlus scoring 5.120 in coherence compared to isiZuluLite's 3.475. However, no significant differences are observed in creativity and plot scores, suggesting that these aspects depend more on the nature of the story than the dataset size. Models trained on datasets generated through prompting, such as isiZuluGuide and YorubaGuide, outperform their translation-based counterparts (isiZuluLite and YorubaLite) in grammar and creativity, with isiZuluGuide scoring 6.830 in grammar and 5.490 in creativity compared to isiZuluLite's 4.615 and 3.955. Similarly, YorubaGuide improves creativity with a score of 5.890 compared to 4.910 for YorubaLite, indicating that prompting can enhance diversity in training data. When comparing the generated stories to the evaluation datasets, the evaluation sets consistently score higher across all categories, demonstrating that while the models capture certain quality aspects, they fall short in fully replicating the complexities of the original stories generated through translation and prompting. For example, the isiZuluPlus evaluation set scores 8.650 in grammar, higher than the generated story score of 6.440, and YorubaLite's evaluation set scores 5.775 for creativity compared to 3.955 for the generated stories.

4.2 Example Generations

Figure 1 shows two examples of generated stories, one from the YorubaPlus model and one from the isiZuluGuide model, along with their English translations.

4.3 Discussion

Overall, across the six models, we see differences in model performance between the isiZulu and Yoruba models, as well as between models trained on datasets generated through prompting versus translation.

YorubaPlus consistently shows the lowest perplexity scores, indicating better coherence and generalization, both for generating stories and for the evaluation sets. In contrast, isiZuluPlus and models trained with prompting (isiZuluGuide, YorubaGuide) show higher perplexities, especially during evaluation, indicating they struggle with generalization to unseen data. This demonstrates that while datasets generated through prompting may help with generating more coherent stories during training, it does not necessarily improve the model's ability to generalize across unseen data. However, the smaller size of the prompt-generated training sets is a possible confounding factor here. Models trained on larger datasets (YorubaPlus and isiZuluPlus), tend to generalize better, as seen by their lower evaluation perplexities.

Models trained on smaller datasets tend to have a higher lexical diversity than those trained on larger

Generated Story - YorubaPlus

Ìgbà kan wà tí e̱ye̱ kékeré kan wà tí orúko̱ rè̱ ń jé̱ Tweety. Tweety fé̱ràn láti máa fò lo̱ sókè lójú ò̱run, kó sì máa bá àwo̱n ò̱ré̱ ré̱ ṣeré. Ló̱jó̱ kan, nígbà tí Tweety ń fò lo̱, ó rí adágún omi kan tó lé̱wà ní òdìkejì adágún náà. Ó fé̱ sún mó̱ o̱n, àmó̱ ó ga jù fún un. Tweety so̱ pé: "Mo lè ràn é̱ló̱wó̱". Tweety àti Tweety jo̱ gbé e̱ye̱ kékeré náà kúrò nínú adágún náà. Lé̱yìn tí e̱ye̱ kékeré náà ti fò lo̱, Tweety so̱ pé: "O ṣeun fún ìrànló̱wó̱ re̱, Tweety kékeré. Inú mi dùn pé o ràn mí ló̱wó̱". Láti o̱jó̱ ye̱n ni Tweety àti Tweety ti di ò̱ré̱ tímó̱tímó̱, tí wó̱n sì máa ń jo̱ máa ṣeré lójoojúmó̱.

English Translation - YorubaPlus

There was a little bird that was called Tweety. Tweety liked to fly in the sky, and one day when he was flying, he found a beautiful pond on the other side of the lake. He wanted to get closer, but it was too high for him. Tweety said, "I can help you." Tweety and the little bird together helped the little bird out of the pond. After the little bird flew away, Tweety said, "Thank you for your help, little Tweety. I'm happy you helped me." From that day on, Tweety and the little bird became close friends and played together every day.

Generated Story - isiZuluGuide

Kwesinye isikhathi, kwakukhona umfana ogama lakhe lingu Liam owayehlala edolobhaneni elincane. Wayenobungani obukhulu nomndeni wakhe nabangane bakhe, kodwa wayezizwa eyedwa kancane. Ngelinye ilanga, uLiam wahlangana nentombazane egama layo lingu Zinhle. Njengoba bengazi omunye nomunye, baqala ukuba ngabangane bokuphila okujabulisayo ndawonye. Bahamba baya ohambweni ndawonye, babelana ngamabali nezindaba, futhi benza izinkumbulo eziningi ezimnandi. Ekugcineni, ubungani babo babuthuthukiswa ngaphezu kokuba ngabangani, futhi bobabili babazi ukuthi bazohlale bekhona komunye nomunye.

English Translation - isiZuluGuide

Once upon a time, there was a boy named Liam who lived in a small town. He had a great friendship with his family and friends, but he felt a little lonely. One day, Liam met a girl named Zinhle. As they got to know each other, they began to be friends and share exciting adventures together. They went on journeys, shared stories, and created many wonderful memories. Eventually, their friendship grew beyond just being friends, and they both knew they would always be there for each other.

Figure 1: Example generated stories and their English translations from YorubaPlus and isiZuluGuide

Model	Grammar		Coherence		Plot		Creativity	
	Gen	Eval	Gen	Eval	Gen	Eval	Gen	Eval
isiZuluPlus	6.440	8.650	5.120	8.650	3.860	5.725	4.475	5.735
YorubaPlus	7.196	8.205	5.412	8.8805	4.185	5.675	5.155	5.620
isiZuluLite	4.615	8.635	3.475	8.885	2.495	5.815	3.955	5.775
YorubaLite	6.865	8.340	5.195	8.655	4.070	5.675	4.910	5.545
isiZuluGuide	6.830	8.545	4.925	8.295	4.020	5.365	5.490	5.780
YorubaGuide	7.475	8.270	5.340	8.065	4.170	5.160	5.890	5.535

Table 8: Average Scores for isiZulu and Yoruba Models and Evaluation Datasets

datasets, as is shown by the high TTR and semantic similarity scores. Models trained on small datasets may produce stories with more varied vocabulary, but will lead to generated stories with high similarity among them. We see this more in the isiZulu models as compared to that of the Yoruba models, which suggests that the size of the dataset has an impact on features across these languages.

Models trained on datasets generated through prompting tend to produce semantic similarity and TTR scores that are comparable to those trained on datasets generated through translations. The isiZulu models produce stable outputs from the prompting-based datasets as compared to the translation-based datasets.

In terms of quality evaluation, the results suggest that models trained on datasets generated through prompting generally perform better in creativity compared to the models trained on translated datasets. This reinforces the idea that models which are trained on stories generated through prompting may be better at capturing imaginative elements in the story generation procedure.

Models trained on large datasets tend to perform better in terms of grammar and coherence of the generated stories. This implies models may need to be trained on larger datasets to be able to capture the linguistic features of African languages. However, scores in the creativity and plot categories are not highly sensitive to the data size, indicating that training on a large synthetic dataset may not be enough to enhance creativity and plot of the generation process. The Gemini quality evaluation confirms that while our models can generate stories that perform well grammatically and with coherence, they struggle in producing creative stories with a consistent plot.

Histograms of the distribution of the Gemini scores per category over each of the model's generated stories and the evaluation dataset stories are given in Appendix A.

5 Conclusion

This paper investigated the feasibility of training models for story generation in low-resource African languages using synthetic data. The results show that it is possible to train models that can generate grammatical and coherent stories, which is promising in particular considering the relatively small training data sizes. Models trained on stories generated through prompting an existing large multilingual model showed particular strength in terms of the quality of the generated outputs, but displayed less generalization than models trained on translated stories, which exhibit more diversity. Overall, in addition to providing new datasets of children stories in isiZulu and Yoruba, which might be of practical usage, e.g. in reading tutoring applications, our results suggest that pretraining on controlled synthetic datasets might be a promising avenue for future investigation of pretraining general-purpose low-resource language models.

Limitations

Our approach relies on the availability of sufficiently high-quality translation models or multilingual LLMs for the target languages, which are not always available for low-resource African languages. However, translation models generally require less training data than general-purpose multilingual language modeling training. Adding generation constraints or quality filters could help to improve synthetic data quality in lower-resource settings. Larger synthetic training datasets would likely have led to higher-quality models, however the study was performed within limited available computational resources. Pretraining models on a combination of real and synthetic data is likely to lead to better models. Fine-tuning and evaluating the models on instruction tuning datasets will enable better evaluation of the potential of this approach to scale beyond story generation.

Acknowledgements

Catherine Essuman was financially supported by Google DeepMind African Institute for Mathematical Sciences (AIMS) AI for Science Master's program. This work is based on the research supported in part by the National Research Foundation of South Africa (Grant Number 151601).

References

David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure F. P. Dossou, Akintunde Oladipo, Doreen Nixdorf, Chris Chinenye Emezue, Sana Al-azzawi, Blessing Sibanda, Davis David, Lolwethu Ndolela, Jonathan Mukiibi, Tunde Ajayi, Tatiana Moteu, Brian Odhiambo, and 46 others. 2023. MasakhaNEWS: News topic classification for African languages. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 144–159, Nusa Dua, Bali. Association for Computational Linguistics.

David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba Oluwadara Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, Chiamaka Ijeoma Chukwuneke, Happy Buzaaba, Blessing Kudzaishe Sibanda, Godson Koffi Kalipe, Jonathan Mukiibi, Salomon Kabongo Kabenamualu, Foutse Yuehgoh, Mmasibidi Setaka, Lolwethu Ndolela, and 8 others. 2025. IrokoBench: A new benchmark for African languages in the age of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2732–2757, Albuquerque, New Mexico. Association for Computational Linguistics.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, and 49 others. 2023. Seamlessm4t: Massively multilingual & multimodal machine translation. *Preprint*, arXiv:2308.11596.

Ronen Eldan and Yuanzhi Li. 2023. Tinystories: How small can language models be and still speak coherent english? *Preprint*, arXiv:2305.07759.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks are all you need. *Preprint*, arXiv:2306.11644.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, and Andrew M. Dai. 2024. Best practices and lessons learned on synthetic data. In *First Conference on Language Modeling*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Evgeniia Razumovskaia, Joshua Maynez, Annie Louis, Mirella Lapata, and Shashi Narayan. 2024. Little red riding hood goes around the globe: Crosslingual story planning and generation with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10616–10631, Torino, Italia. ELRA and ICCL.

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

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 *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 6000–6010.

Changhan Wang, Kyunghyun Cho, and Jiatao Gu. 2020. Neural machine translation with byte-level subwords. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9154–9160.

Zhihua Wen, Zhiliang Tian, Wei Wu, Yuxin Yang, Yanqi Shi, Zhen Huang, and Dongsheng Li. 2023. GROVE: A retrieval-augmented complex story generation framework with a forest of evidence. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3980–3998, Singapore. Association for Computational Linguistics.

A Appendix

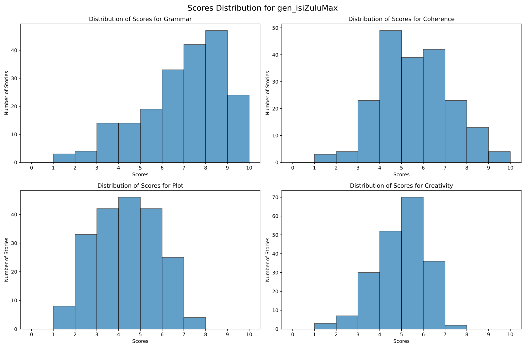

Figure 2: Gemini Score Distribution for isiZulu-Plus Model Generations

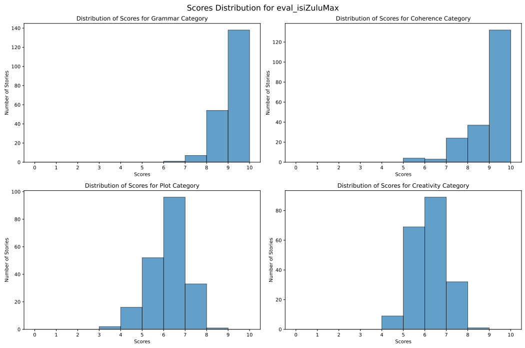

Figure 3: Gemini Score Distribution for isiZulu-Plus Evaluation Set

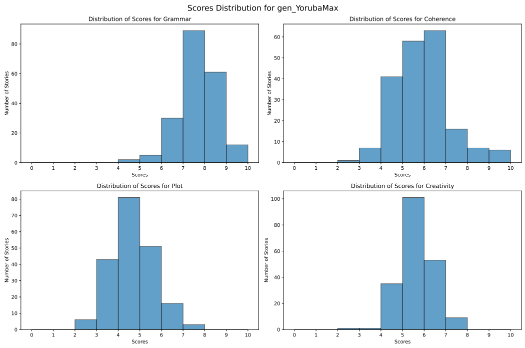

Figure 4: Gemini Score Distribution for YorubaPlus Model Generations

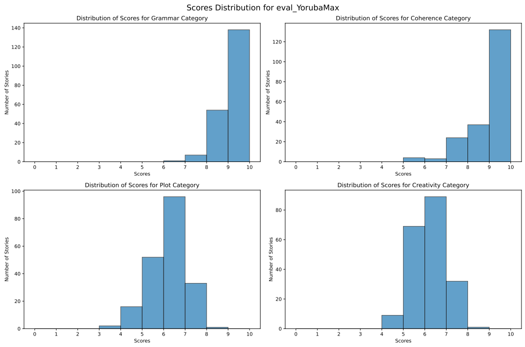

Figure 5: Gemini Score Distribution for YorubaPlus Evaluation Set

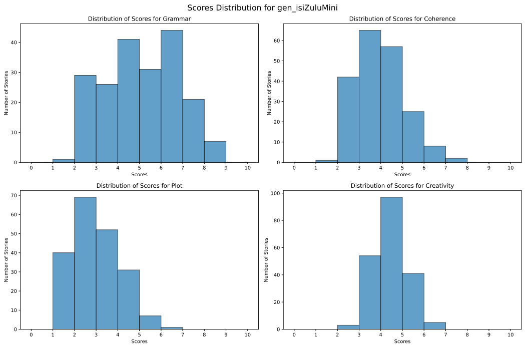

Figure 6: Gemini Score Distribution for isiZul-uLite Model Generations

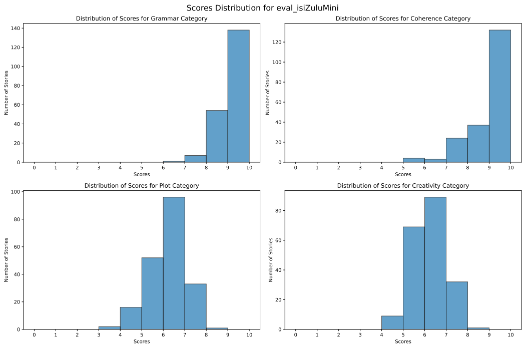

Figure 7: Gemini Score Distribution for isiZul-uLite Evaluation Set

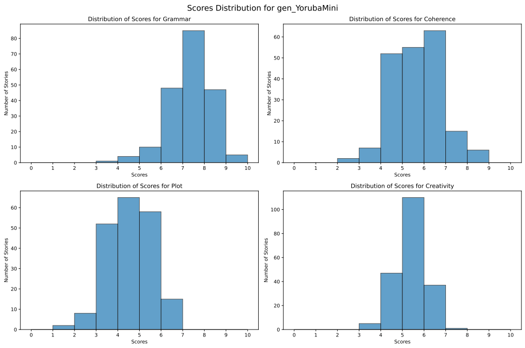

Figure 8: Gemini Score Distribution for YorubaLite Model Generations

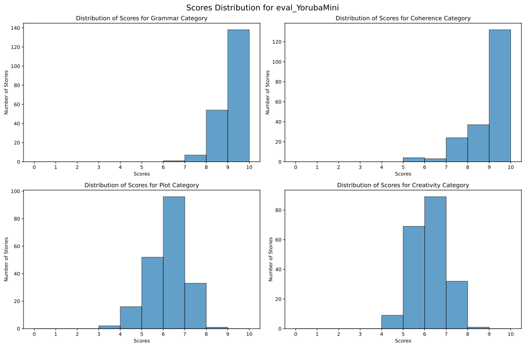

Figure 9: Gemini Score Distribution for YorubaLite Evaluation Set

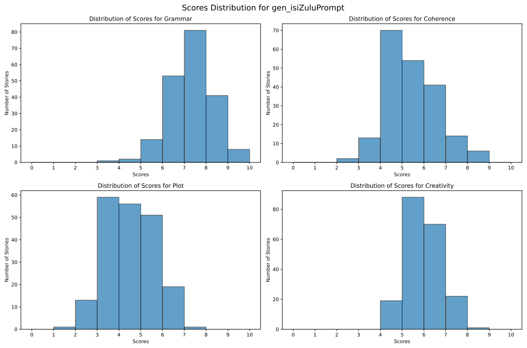

Figure 10: Gemini Score Distribution for isiZuluGuide Model Generations

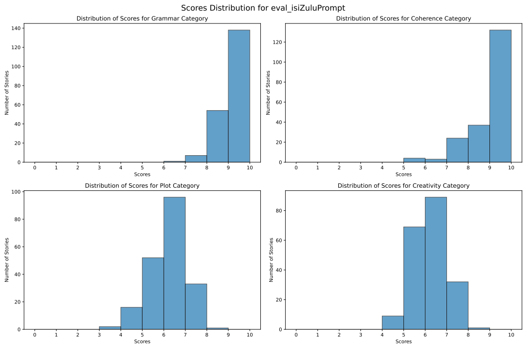

Figure 11: Gemini Score Distribution for isiZuluGuide Evaluation Set

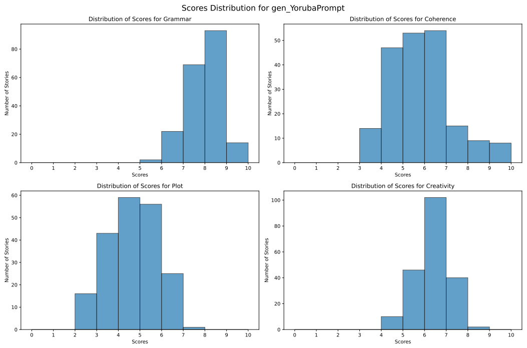

Figure 12: Gemini Score Distribution for YorubaGuide Model Generations

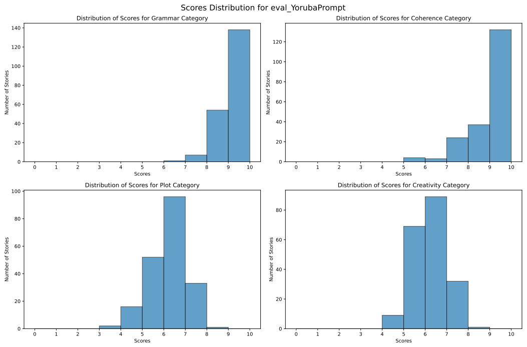

Figure 13: Gemini Score Distribution for YorubaGuide Evaluation Set