VEXMLM: Vocabulary Expansion for Multilingual Models to Address Tokenization and OOV Challenges in Underrepresented Languages

Anonymous ACL 2025 submission

Abstract

Multilingual models have shown effectiveness in natural language processing (NLP) tasks, but their performance often declines for low-resource languages due to a predominant focus on high-resource languages during training. This leads to challenges such as out-of-vocabulary (OOV) and over-segmentation, mainly resulting from English-centric tokenization methods. Vocabulary expansion using target language tokens is a common strategy to address these problems. However, existing research mainly focuses on high-resource settings and overlooks the potential of vocabulary expansion to address OOV and over-segmentation in low-resource languages. To fill this gap, we introduce VEXMLM, an enhanced version of XLM-R optimized for low-resource languages through effective vocabulary expansion. Our approach involves creating a human-annotated benchmark dataset and training a language-specific tokenizer by maintaining semantic coherence morphological insights to build comprehensive vocabularies and integrating these tokens into the model via embedding initialization. VEXMLM is evaluated on 19 African languages with varying scripts and resource availability across four tasks: Question Answering, Named Entity Recognition, Sentiment Analysis, and Educational Quality Classification. Comparative experiments demonstrate that VEXMLM significantly outperforms baseline models, XLM-R and Glot500, on low-resource languages while improving performance for high-resource languages. The model, code, and dataset will be publicly available for research.

1 Introduction

Multilingual models often use auto-tokenizers that map unrecognized words to a single <UNK> token (Liu et al., 2024). The models struggle to distinguish between different scripts and lack proper encoding for scripts like Geez, causing unrecognized scripts to <UNK>. These models typically assign a generic <UNK> token for completely unrecognized scripts or characters, as they cannot decompose these elements further (Xue et al., 2022). Also, the vocabulary size of multilingual models is generally small, especially for low-resource languages (Wang et al., 2019a).

(Sennrich et al., 2016) proposes a method to tackle the open-vocabulary issue in neural machine translation (NMT) by encoding rare and unknown words as subword sequences. Using techniques like byte pair encoding, this approach breaks down words into smaller units such as characters or subword segments, improving NMT performance over traditional dictionary-based methods. For example, a subword tokenizer might split "doghouse" into "dog" and "house," even if "doghouse" is not in the vocabulary. This flexibility has made subword tokenizers the standard for text tokenization in recent years (Hiraoka et al., 2019; Bostrom and Durrett, 2020).

Over-segmentation, typos, variants in spelling and capitalization, and morphological changes can all cause the token representation of a word or phrase to change completely, which can result in mispredictions (Xue et al., 2022; Ahia et al., 2024). Furthermore, unknown characters (e.g., from a language that was not seen when the subword vocabulary was built) are typically OOV for a subword model (Xue et al., 2022).

Prior research has focused on tokenization algorithms and optimal vocabulary sizes for machine translation in English (Ahia et al., 2023b). However, low-resource languages often have smaller datasets, causing subword tokenizers trained in multiple languages to over-segment tokens in these languages (Ahia et al., 2023a). Likewise, a challenge in pre-trained multilingual models is limited vocabulary coverage or the exclusion of languages during training, resulting in poor representation of low-resource languages. (Wang et al., 2019a).

Many pre-trained models employ a subword vocabulary, which greatly reduces the issue of out-of-vocabulary tokens. However, it can still result in performance decline if domain-specific terms are overly fragmented or insufficiently represented due to limited training data (Ebrahimi and Kann, 2021a).

Out-of-vocabulary (OOV) words and over-segmentation are challenges in pre-trained multilingual models, especially for low-resource languages. These challenges underscore the need for efficient strategies to improve these models.

We investigate two main approaches for adapting a pre-trained language model to new target languages. The first approach involves fully adapting the model by replacing the tokenizer and focusing on the performance of the new target language (Minixhofer et al., 2022; Mundra et al., 2024b). The second approach extends the model's language support by adding new target tokens while preserving the performance of the original language (Garcia et al., 2021; Liu et al., 2024). While we explore both approaches, we focus primarily on the second approach, which involves extending the tokenizer and initializing the model's embedding layer and LM head for the newly added tokens.

This paper aims to (i) evaluate whether the expansion approach effectively addresses OOV words and over-segmentation, while ensuring balanced language representation and optimized performance across diverse linguistic contexts, particularly in low-resource language, and (ii) identify the best approaches for vocabulary expansion and initialization to support language adaptation in low-resource environments, maintaining performance comparable to source models. The key contributions are as follows:

- We compiled a human-annotated dataset for educational quality classification in Tigrinya and Amharic, following the approach outlined in (Lozhkov et al., 2024). Annotators rated each context on a 1 to 6 scale, based on a detailed guideline. We calculated the inter-annotator agreement to ensure annotation reliability, assessing label consistency and quality. This dataset will serve as a ground truth for evaluating educational content in these languages. Using the SentencePiece algorithm, we also trained a language-specific tokenizer, ensuring compatibility with the source model's tokenizer. Our evaluation of various multilingual subword-based models shows that our tokenizer performs particularly well for languages using the Geez script.
- We introduce VEXMLM, an extended XLM-R model that expands vocabulary to address OOV issues, over-segmentation, and poor representation of low-resource languages. Our results show that average vocabulary initialization, using the mean and standard deviation of token embeddings from the original model, is more effective. New script token embeddings are initialized by sampling from a normal distribution based on the source model's script parameters. We fine-tuned VEXMLM on four tasks and evaluated it across 19 low-resource languages with diverse scripts, and VEXMLM outperformed baseline models in several tasks.
- VEXMLM exceeds baseline models in downstream tasks for low-resource languages, such as Question Answering (QA), Sentiment Analysis, and Educational Value Classification. It also enhances performance for high-resource languages, demonstrating its considerable vocabulary sharing and generalization capabilities across diverse linguistic contexts.

2 Related Work

Multilingual language models often use subword-based tokenizers, which can cause unwanted behaviors in low-resource languages, such as over-segmentation and the vocabulary bottleneck limiting the representational capabilities of multilingual models (Limisiewicz et al., 2023; Beinborn and Pinter, 2023).

(Liang et al., 2023) proposes assigning vocabulary capacity to achieve sufficient coverage for each language and using a semantically meaningful tokenizer to overcome the issues in multilingual language models. This approach is computationally expensive. (Mundra et al., 2024a) introduces Constrained Word2Vec (CW2V) for expanding language model vocabularies without needing cross-lingual embeddings.

(Reimers and Gurevych, 2020) extend existing sentence embedding models to new languages. This allows the creation of multilingual versions from previously monolingual models.

The NLP community has mostly focused on vertical scaling, optimizing models for a few well-resourced languages, often neglecting horizontal

scaling to include low-resource languages. Vertical scaling enhances performance for a limited set of languages but lacks broader multilingual support (ImaniGooghari et al., 2023). In contrast, horizontal scaling extends model performance to a wider range of languages, including those with limited resources. (ImaniGooghari et al., 2023) addressed this by creating Glot500-m, a model that continues pretraining across 511 low-resource languages, marking a significant step forward for low-resource language representation. Thus OOV words present significant challenges across various NLP tasks (Chen et al., 2022; Wang et al., 2019b; Garcia-Bordils et al., 2022; Lochter et al., 2020; Zhuang et al., 2023; Shiao et al., 2024; VH and Chacko, 2024; Wang et al., 2019a)

Prior work on addressing OOV issues in multilingual settings, such as BERT, uses subword tokenization (e.g., WordPiece) instead of full word tokenization (Wang et al., 2019a; Jaffe, 2017; Platanios et al., 2018). This breaks unknown words into known subword units, like "autonomic" being tokenized as "auto" + "##nom" + "##ic," ensuring coverage even if the full word is not in the vocabulary. However, subword tokenization still struggles with the rich inflectional and derivational processes of morphologically complex languages (Chai et al., 2024).

Vocabulary-free models like ByT5 (Xue et al., 2022) and CANINE (Clark et al., 2022) offer competitive performance compared to subword-based models but face slower training and inference speeds. ByT5 is slower than mT5 (Xue et al., 2021), and CANINE, despite optimizations, still lags behind BERT models (Liang et al., 2023). BBPE improves translation quality by maximizing vocabulary sharing across languages (Wang et al., 2020). Recently, (Pagnoni et al., 2024) introduced the Byte Latent Transformer (BLT), which processes raw byte sequences directly, bypassing the need for tokenization and offering a scalable solution without fixed vocabulary limitations. Recently, (Pagnoni et al., 2024) introduced the Byte Latent Transformer (BLT), an innovative architecture that eliminates the need for tokenization. Instead, it processes raw byte sequences directly. This approach demonstrates the potential for scaling models trained on raw bytes, bypassing the limitations of a fixed vocabulary.

2.1 Language Adaptive Pretraining

Language Adaptive Pretraining (LAPT) is a promising method for adapting multilingual models to multiple languages simultaneously (Dobler and de Melo, 2023). (Alabi et al., 2022) demonstrated its use with XLM-R for 20 African languages, while (Ebrahimi and Kann, 2021b) and (Wang et al., 2022a) leveraged resources like the Bible and lexicons for model adaptation. (Muller et al., 2021) improved performance by transliterating unseen languages into Latin script. (Pfeiffer et al., 2020) introduced adapter modules to preserve pre-trained weights, though this incurs computational costs. However, none of these approaches address out-of-vocabulary (OOV) issues or over-segmentation in low-resource languages.

3 Proposed Method

We use XLM-R as the pre-trained model. First, we outline its pre-training procedure (Section 3.1). We then explore methods to address out-of-vocabulary (OOV) issues and improve the representation of low-resource languages by expanding the vocabulary (Section 3.2). Finally, we present our benchmark dataset (Section 4.3), expansion initialization (Section 3.3), and model initialization (Section 3.4). These methods apply to similar models and languages.

3.1 Multilingual Models

XLM-R is a transformer-based (Vaswani et al., 2017) multilingual masked language model with state-of-the-art performance on cross-lingual understanding tasks. It is pre-trained on text in 100 languages, making it a highly scalable model (Conneau et al., 2020). This large-scale pre-training allows XLM-R to learn a rich representation of language that can be fine-tuned for specific downstream tasks. The model's architecture is based on the RoBERTa (Liu et al., 2020) architecture, which is a variant of the BERT (Devlin et al., 2019) model. This architecture is highly effective for language modeling tasks, and the addition of multilingual pre-training data has further improved its performance (Conneau et al., 2020).

Recent research has focused on adapting large language models (LLMs) to support low-resource languages, noting that multilingual capabilities improve with adding more languages. (Wang et al., 2019a) explore vocabulary expansion, while (Ebrahimi and Kann, 2021a) adapt pre-trained mul-

tilingual models to nearly 1600 languages using the New Testament, showing that continued pretraining yields the best results despite limited resources. (Alabi et al., 2022) adapt XLM-R to 17 African languages and three widely spoken African languages to enhance cross-lingual transfer learning, and (Wang et al., 2022b) extend LLMs to low-resource languages using bilingual lexicons. Additionally, (ImaniGooghari et al., 2023) propose Glot500-m, trained on a 600GB corpus covering over 500 languages, demonstrating that expanding LLMs improves knowledge transfer from high-resource to low-resource languages. While these methods expand language coverage, challenges such as noise, corpus size for underrepresented languages, over-segmentation, script encoding, and tokenization still need attention.

3.2 Vocabulary Expansion and Adaptation

Vocabulary expansion often requires adapting a pre-trained model to a new task, which can be challenging due to differences in vocabulary, syntax, and semantics between the original training data and the target languages (Conneau et al., 2020). Typically, language models (LMs) are trained with a fixed vocabulary, often consisting of around 50,000 tokens (Ushio et al., 2023). A major limitation to adapting models to new languages is the vocabulary, which often fails to cover unseen scripts (Downey et al., 2024; Pfeiffer et al., 2020) or tokenizes target text inefficiently (Ahia et al., 2023b). (Muller et al., 2021) demonstrate that script is a critical factor in predicting transfer success.

Various methods have been suggested to address this problem, including adapting by replacing the tokenizer (Minixhofer et al., 2022; Mundra et al., 2024b), adapting by adding the new target tokens while maintaining the original language's performance (Garcia et al., 2021; Liu et al., 2024).

To address OOV issues and tokenization challenges in multilingual models for low-resource languages, we implement vocabulary expansion strategies. This involves integrating separately trained vocabularies from the Geez script and initializing the model's embedding layer for the new tokens. Since new tokens lack pre-trained embeddings, the embedding matrix is resized to include them, while existing token embeddings remain unchanged. This helps the model better manage vocabulary expansion and enhances its ability to generalize across different target languages by fine-tuning VEXMLM for those languages.

3.3 Average Initialization

We adopt the same vocabulary expansion problem formulation as (Mundra et al., 2024b; Hewitt, 2021). We conclude the average of the existing embeddings as the default initialization for new word embeddings for pre-trained language models is effective for better performance. Let θ be the parameters of a pre-trained neural source language model LM^s_θ, and let $\mathcal{V}^s = \{v^s_1, v^s_2, \ldots, v^s_n\}$ be the vocabulary of LM^s_θ. We will refer to $\mathcal{V}^s$ as the source vocabulary. Let $e^s_i \in \mathbb{R}^d$ be the sub-word embedding for word $i \in \mathcal{V}^s$. Let $\mathcal{E}^s$ denote the language modeling head's (henceforth LM head) embedding matrix of LM^s_θ and this is our source embedding matrix. The probability of occurrence of the next word w_i given the previous word sequence $w_1 : i-1$, $p_\theta(w_i|w_1 : i-1)$, is given by

$$p_\theta(w_i|w_1 : i-1) = \frac{\exp(h^\top_{i-1} e^s_{w_i})}{\sum_{j \in \mathcal{V}^s} \exp(h^\top_{i-1} e^s_j)}, \quad (1)$$

where $h_{i-1} = \phi(w_1 : i-1; LM^s_\theta) \in \mathbb{R}^d$ is the neural representation of the prefix using LM^s_θ.

In vocabulary expansion, we add n' new subwords $\notin \mathcal{V}^s$ forming the target vocabulary $\mathcal{V}^t = \{v^t_1, v^t_2, \ldots, v^{\prime t}_n\}$. This implies we need a new word embedding e^t_j for each $j \in \mathcal{V}^t$ comprising in E^t. The new language model $LM^t_{\theta'}$ has parameters $\theta' = \theta \cup \{e^t_j; j \in \mathcal{V}^t\}$. The output distribution of $LM^t_{\theta'}$ given by $p_{\theta'}(w_i|w_1 : i-1)$ is defined similarly as $p_\theta(w_i|w_1 : i-1)$ but with the normalization factor involving $\mathcal{V}^s \cup \mathcal{V}^t$.

Our goal is to find initializations for E^t such that the extended model not only retains its previous behavior but also can lead to good downstream performance for the languages corresponding to the new vocabulary with minimal continual pretraining. Note that in our notations so far we have only mentioned the LM head, but just as the LM head has an expansion (E^t_{lmhead}), the input embedding matrix also has an expansion (E^t_{input}). This is trivial if both matrices are shared but in case they are not, we also need to find initializations for the latter. Following (Hewitt, 2021), we can use the same approach to initialize E^t_{input} as we do for E^t_{lmhead}.

3.4 Model Initialization

To adapt XLM-R for 19 low-resource languages with diverse scripts including Geez-script languages (e.g., Amharic and Tigrinya), we first classify the model's tokens by their Unicode block to

map them to their respective scripts. For each script, we calculate the mean and standard deviation of its token embeddings in the original embedding space. New embeddings for Geez-script tokens are then generated by sampling from a multivariate Gaussian distribution parameterized by these statistics. The new embeddings are integrated with the original embedding matrix, and the model is resized to accommodate the extended vocabulary, ensuring continuity of prior capabilities while enabling improved performance on tasks involving the new script.

4 Experimental Setup

Instead of creating an entirely new vocabulary, our experiment uses a simple and effective approach to expand the source model's token set by incorporating 30k new tokens from the Geez Script languages. The model's embedding layer and LM head are initialized for these additional tokens, followed by continuous pretraining. This process produces the VEXMLM model, which is then fine-tuned to enhance performance across 19 target languages in four downstream tasks. you can see from Figure 2 for the extended vocabulary.

4.1 Experimental Details

Through vocabulary expansion, XLM-R has been extended to create VEXMLM, a novel variant specifically designed to enhance support for low-resource languages and resolve challenges associated with out-of-vocabulary (OOV) tokens and tokenization errors.

VEXMLM maintains the core architecture of the XLM-RoBERTa-base, featuring 12 layers, a hidden size of 768, and the ability to handle sequences up to 514 tokens long.

Key aspects of VEXMLM include: ELU activation function, Layer normalization epsilon of 1e-05, Special token IDs: BOS (0), EOS (2), PAD (1), attention heads in 12 hidden layers. The model underwent a three-epoch extension while maintaining its "XLM-RoBERTa" classification to enhance cross-lingual performance, especially for low-resource languages. Afterward, VEXMLM was fine-tuned for tasks such as sentiment analysis, question answering, multilingual named entity recognition, and educational value classification. Its performance was benchmarked against models like XLM-R and Glot500, which also include low-resource languages. This approach aims to leverage VEXMLM's expanded linguistic capabilities and improve performance on cross-lingual tasks, particularly for underrepresented languages in Geez Script.

4.2 Tokenizer and Languages

Tokenizer

In multilingual settings, subword tokenization can be inefficient due to segmentation mismatches and over-segmentation in low-resource languages (Sun et al., 2023a). Factors such as pretraining data size and writing systems contribute to inconsistent tokenization across languages (Ahia et al., 2023a). Over-tokenization increases sentence length and leads to unknown (UNK) tokens (Zhang et al., 2022). Geez Script languages, with their distinct morphological structures, are especially affected by over-segmentation (Ahia et al., 2023a). To improve performance for Geez Script languages, we aim to reduce token usage using a SentencePiece tokenizer (Kudo and Richardson, 2018) and analyze the morphological structure of out-of-vocabulary (OOV) words specific to their script.

Languages

We start by training a tokenizer tailored for Geez Script languages to address over-segmentation in multilingual tokenization. Then, our extended model, VEXMLM, is fine-tuned on 19 low-resource languages with diverse scripts across four downstream tasks.

4.3 Datasets and Tasks

Finally, we evaluate the VEXMLM model on 19 low-resource languages and four tasks.
(1) Text classification task: that aims to assign predefined labels (e.g., sentiment, score value, etc) to a given text (Sun et al., 2023b).
(A). Educational quality classification Dataset : We created a benchmark dataset of 1,500 human-annotated samples for Educational Quality classification in Amharic and Tigrinya, in collaboration with the community. The dataset includes structured data from public educational blogs, collected in two stages: first, diverse online text snippets were gathered from manuals and blogs, and second, they were annotated on a scale of 1 to 6 based on educational quality. Details and samples are provided in B.

(B) Multilingual sentiment analyses: We used AfriSenti,a sentiment analysis dataset for low-

resource African languages, containing over 110K annotated tweets in 14 languages, for detail of the languages (Muhammad et al., 2023).

(2) Multilingual NER: We used the MasakhaNER dataset (Adelani et al., 2021) and Tigriyna (Yohannes and Amagasa, 2022)dataset both, offer NER tasks for 11 low-resourced African languages.

(3)Question and Answer: TIGQA dataset (Teklehaymanot et al., 2024) which offers a question-answering expertly annotated dataset. Additionally, used the Amharic(AmQA) dataset (Taffa et al., 2024).

Model Training

To train VEXMLM, we first implement vocabulary expansion refer to the section 3.2. Next, we finetune XLM-R, Glot500, and the extended model VEXMLM. We evaluate VEXMLM on four subdomain tasks, which include Named Entity Recognition(NER), Question Answering(QA), Educational quality classification(EQC), and Sentiment Analysis(SA), see Table 3 with results.

Baselines

In addition to VEXMLM, we establish the following baseline models for comparison:

- XLM-R (Conneau et al., 2020)
- Glot500 (ImaniGooghari et al., 2023)

5 Results and Evaluation

We utilized a comprehensive evaluation strategy for both intrinsic and extrinsic metrics. The intrinsic evaluation encompassed Parity (fairness across languages) (Petrov et al., 2024), Fertility (token efficiency) (Rust et al., 2021), and Compression (text reduction) (Goldman et al., 2024). Additionally, we analyzed language and script coverage, with a focus on low-resource languages, as well as Out-Of-Vocabulary (OOV) handling and over-tokenization. For the extrinsic evaluation, we compared our model, VEXMLM, with seven multilingual subword-based tokenizer models using Geez Script languages such as Amharic, Tigrinya, and Tigre, demonstrating that our model tokenized with fewer tokens. Furthermore, we fine-tuned the model on 19 low-resource languages and evaluated its performance across four downstream tasks.

Figure 4 compares the parameter count and vocabulary size of XLM-R, VEXMLM, and Glot500. VEXMLM, an enhanced version of XLM-R, boasts 301 million parameters and 280k tokens, which enhances representation for languages that use the Geẹz script. XLM-R contains 279 million parameters and 250k tokens, while Glot500 provides broader linguistic coverage with 350 million parameters and 400k tokens. VEXMLM maintains a balance between model size and linguistic efficiency.

Parity is a metric that assesses the fairness of a tokenizer in processing equivalent sentences across different languages. Table 1 illustrates the comparative performance of models in low-resource languages versus English at the sentence level. VEXMLM consistently shows the best overall parity across most languages, indicating its superior ability to perform in these contexts relative to the English baseline. While Glot500 excels in Tigrinya and Amharic, its performance is less consistent across other languages, and XLM-R exhibits lower parity scores, signifying greater difficulty in handling low-resource languages.

VEXMLM, provides the most balanced performance by minimizing the disparity between English and low-resource languages, it also offers more compact encoding for Tigrinya and Amharic. Overall, VEXMLM achieves near-parity with English in Afar, Oromo, Afrikaans, German, and Arabic, although Glot500 surpasses it in Amharic.

Language	Models		
	XLM-R	Glot500	VEXMLM
Tigrinya	1.36	**1.20**	0.27
Amharic	0.82	**0.90**	0.39
Ge'ez	0.45	0.31	**0.73**
Tigre	0.82	0.88	**0.89**
Harari	0.41	0.14	**0.88**
Gurage	0.62	0.12	**0.89**
Afar	0.54	0.30	**0.90**
Oromo	1.27	0.90	**0.93**
Afrikaans	1.18	1.30	**0.96**
German	1.09	1.10	**0.98**
Arabic	1.27	1.30	**1.16**

Table 1: Comparative Performance of Models in Parity Across Different Languages Versus English at the Sentence Level.

Fertility (Rust et al., 2021) is a metric that assesses the number of tokens produced relative to the original text, helping evaluate the granularity and

efficiency of tokenization. In Figure 1 VEXMLM shows strong specialization in less-resourced languages like Amharic and Tigrinya, making it highly effective for tasks in those languages. Glot500 demonstrates a well-rounded ability across all three languages, especially excelling in English, while XLM-R shows potential but struggles with the linguistic diversity in Amharic and Tigrinya.

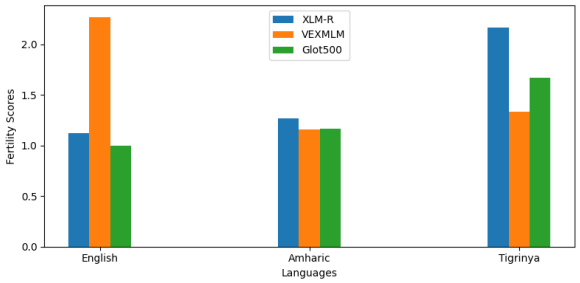

Figure 1: Comparison of Fertility Score Across Three Languages.

Compression (Goldman et al., 2024) Compression is a reliable intrinsic indicator of tokenization quality, we evaluate how efficiently a tokenizer reduces the text length while maintaining information quality and effectiveness. Figure 3 assesses the trade-off between reducing the token count and preserving the meaningful content of the text. Moreover, in Table 1 the 0.27 score suggests that VEXMLM employs a more compact tokenization strategy for Tigrinya compared to English. The lower token count implies that the model can encode the same information using fewer tokens, potentially indicating a more effective representation of the language's structure and patterns.

Language and Script Analysis: In this analysis, we examine languages, more spoken countries, and scripts in Low-Resource African Languages (see Table 5 in B for details). In out-of-vocabulary analyses at the word level,

VEXMLM addressed XLM-R's limitations for Ge'ez script languages while maintaining or improving performance for other scripts.

OOV Word Analysis: We evaluate our model against a baseline at the word level, focusing on how out-of-vocabulary (OOV) words are handled across scripts and languages. Addressing the OOV issue is crucial in multilingual contexts. For example, in Named Entity Recognition (NER) tasks for 11 low-resource African languages, most errors stem from OOV words (see Table 2). The

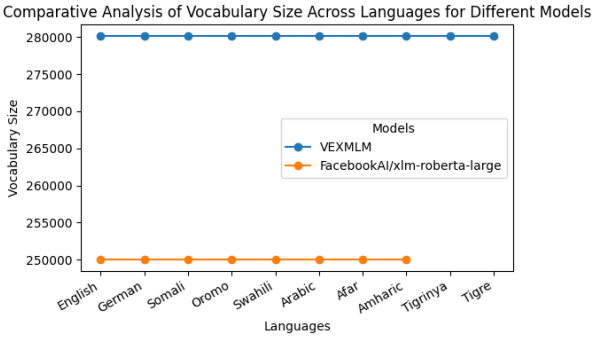

Figure 2: Comparative Analysis of Sample Vocabulary Size Across Different Scripts and Languages Using XLM-R and VEXMLM.

original XLM-R model shows significantly lower accuracy for OOV words compared to non-OOV words in these languages. In contrast, the expanded VEXMLM model shows notable improvement in OOV word accuracy, especially for Amharic and Tigrinya.

Language	**XLM-R**		**VEXMLM**	
	Non-OOV	**OOV**	**Non-OOV**	**OOV**
amh	98.1	91.1	98.1	**92.2**
tig	78.1	96.1	89.5	**98.2**
hau	97.0	90.2	97.2	95.6
ibo	97.8	91.9	97.7	94.5
kin	98.8	84.9	99.0	93.4
lug	98.8	91.3	96.4	94.8
luo	97.8	91.4	98.6	95.2
pcm	98.6	89.6	97.5	97.0
swa	98.2	80.2	93.9	95.2
wol	92.5	89.1	98.7	93.0
yor	91.6	81.8	92.8	89.1
Average	**95.2**	**81.43**	**96.3**	**94.3**

Table 2: Evaluation of OOV word accuracy in NER tasks for 11 low-resource African languages, comparing the performance of the original XLM-R model and the expanded VEXMLM model.

Over-tokenization: The distribution of languages in the corpus impacts tokenization, with dominant languages remaining intact and underrepresented languages being excessively tokenized. This over-tokenization increases sentence length and results in unknown (UNK) tokens (Talat et al., 2022). Using fewer tokens to represent input data can improve inference speed, reduce costs, and enhance utility (Liang et al., 2023). This approach

also helps the model manage longer contexts and mitigates over-tokenization in low-resource languages (Rust et al., 2021).

Figure 3 displays the average token count generated by various multilingual models on our Geez Script(Amharic, Tigriyna) benchmark dataset. As shown, VEXMLM outperforms other subword-based multilingual tokenizers, significantly reducing the average input sequence length.

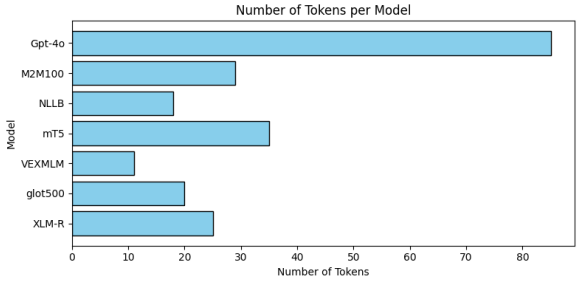

Figure 3: We compare our model with other multilingual tokenizers based on the number of tokens generated. Fewer tokens indicate higher efficiency, as they represent the text more compactly.

Downstream performance: Table 3 compares the performance of XLM-R, VEXMLM, and Glot500 models on four different downstream NLP tasks: Educational quality classification (EQC), Sentiment Analysis (SA), Named Entity Recognition (NER), and Question Answering (QA). The results are presented in terms of accuracy (Acc), Exact match(EM), and F1 score (F1), reflecting cross-lingual transfer after fine-tuning. VEXMLM's superior generalization capabilities on the dataset in section 4.3.

Tasks	XLM-R	VEXMLM	Glot500
SA (Acc)	0.77	**0.80**	0.46
EQC(Acc)	0.96	**0.98**	0.70
NER (Acc)	0.75	0.78	**0.92**
QA (EM)	0.66	**0.87**	0.74
QA (F1)	0.78	**0.90**	0.78

Table 3: compares the performance of XLM-R, VEXMLM, and Glot500 on four NLP tasks: Sentiment Analysis (SA), Educational Quality Classification (EQC), Named Entity Recognition (NER), and Question Answering (QA). Evaluation metrics include accuracy for SA, EQC, and NER, and Exact Match (EM) and F1-score for QA.

5.1 Discussions

Addressing the OOV issue in multilingual settings is important. Using the NER task as an example, we find that most errors occur at OOV positions (Table 2). Both XLM-R and VEXMLM perform well on non-OOV words, but XLM-R shows much lower accuracy for OOV words. VEXMLM significantly improves OOV word accuracy, resulting in overall better performance by reducing OOV errors. Additionally, VEXMLM outperforms other multilingual tokenizers in efficiency, particularly for low-resource languages, reducing the token count by 12.94% (Figure 3). This reduction leads to shorter input sequences, easing computational load and memory usage. VEXMLM's efficient tokenization addresses over-tokenization in Geez Script languages like Amharic and Tigrinya, avoiding fragmentation and meaning loss. Lastly, Sections 3.3 and 3.4 discuss how initialization affects model performance. Using the average of existing embeddings for new word embeddings is more effective, improving expansion performance (see Section 3.3). Grammarly, ChatGPT, were utilized to improve language clarity and refine phrasing in our original content.

6 Conclusion

We introduce VEXMLM, a model that enriches the vocabulary of the XLM-R tokenizer by incorporating language-specific tokenization methods. This approach preserves semantic coherence and utilizes morphological insights to develop extended vocabularies from Geez script languages, we effectively initialize the model's embedding layer to support the new tokens. VEXMLM is evaluated on 19 African languages with diverse scripts and exceeds, particularly for Geez script languages like Amharic and Tigrinya. Also, languages that share similar characters or token structures alike benefit from this model, mainly given the variability in tokenization. For instance, the token "BPE" can be represented as either ("BP", "E") or ("B", "PE"), and this adaptability aids other low-resource languages with comparable scripts.

Achieving an average accuracy score of 94.3 across 11 low-resource languages, VEXMLM significantly reduces the number of tokens for Geez script languages. We assess token efficiency using metrics such as fertility and average token count. We believe this work will greatly contribute to NLP for low-resource languages.

Limitations:

Despite the extensiveness of our work, it faces the following limitations.

First, Our vocabulary expansion Method was based solely on data from Amharic and Tigrinya, which may have limited generalizability. While token overlap was observed, the unique characteristics of 19 low-resource languages highlight the need for further validation and potential adjustments in tokenization strategies.

Second, while the Ge'ez script is the basis for several other languages, Ge'ez itself lacks native speakers who can offer essential linguistic insights, unlike languages such as Amharic and Tigrinya. Moreover, no computational linguistic research on Ge'ez has been conducted to date. Consequently, there may be inaccuracies or improperly formulated sentences when performing sentence-level parity comparisons for this language.

Third, we did not report the perplexity metric, as our primary focus was to assess whether our approach addresses challenges related to out-of-vocabulary (OOV) tokens, over-segmentation, and the representation of new languages. We argue that metrics such as parity, fertility, and compression are sufficient to answer the research question, without the need to rely on perplexity. Finally, human-annotated resource data was only prepared for Amharic and Tigrinya. Developing similar resources for additional low-resource languages is crucial for further advancements, highlighting this as future work for other researchers.

References

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.

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Valentin Hoffman, Tomasz Limisiewicz, Yulia Tsvetkov, and Noah A Smith. 2024. Magnet: Improving the multilingual fairness of language models with adaptive gradient-based tokenization. *arXiv preprint arXiv:2407.08818*.

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David Mortensen, Noah Smith, and Yulia Tsvetkov. 2023a. Do all languages cost the same? tokenization in the era of commercial language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9904–9923, Singapore. Association for Computational Linguistics.

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David R Mortensen, Noah A Smith, and Yulia Tsvetkov. 2023b. Do all languages cost the same? tokenization in the era of commercial language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Lisa Beinborn and Yuval Pinter. 2023. Analyzing cognitive plausibility of subword tokenization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4478–4486, Singapore. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Kaj Bostrom and Greg Durrett. 2020. Byte pair encoding is suboptimal for language model pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4617–4624.

Yekun Chai, Yewei Fang, Qiwei Peng, and Xuhong Li. 2024. Tokenization falling short: On subword robustness in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1582–1599, Miami, Florida, USA. Association for Computational Linguistics.

Lihu Chen, Gael Varoquaux, and Fabian Suchanek. 2022. Imputing out-of-vocabulary embeddings with LOVE makes LanguageModels robust with little cost. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3488–3504, Dublin, Ireland. Association for Computational Linguistics.

Jonathan H Clark, Dan Garrette, Iulia Turc, and John Wieting. 2022. Canine: Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, 10:73–91.

Kartikay Conneau, Alexis Workshop Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Konstantin Dobler and Gerard de Melo. 2023. Focus: Effective embedding initialization for monolingual specialization of multilingual models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

C. M. Downey, Terra Blevins, Dhwani Serai, Dwija Parikh, and Shane Steinert-Threlkeld. 2024. Targeted multilingual adaptation for low-resource language families. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15647–15663, Miami, Florida, USA. Association for Computational Linguistics.

Abteen Ebrahimi and Katharina Kann. 2021a. How to adapt your pretrained multilingual model to 1600 languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4555–4567, Online. Association for Computational Linguistics.

Abteen Ebrahimi and Katharina Kann. 2021b. How to adapt your pretrained multilingual model to 1600 languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4555–4567.

Xavier Garcia, Noah Constant, Ankur Parikh, and Orhan Firat. 2021. Towards continual learning for multilingual machine translation via vocabulary substitution. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1184–1192, Online. Association for Computational Linguistics.

Sergi Garcia-Bordils, Andrés Mafla, Ali Furkan Biten, Oren Nuriel, Aviad Aberdam, Shai Mazor, Ron Litman, and Dimosthenis Karatzas. 2022. Out-of-vocabulary challenge report. In *European Conference on Computer Vision*, pages 359–375. Springer.

Omer Goldman, Avi Caciularu, Matan Eyal, Kris Cao, Idan Szpektor, and Reut Tsarfaty. 2024. Unpacking tokenization: Evaluating text compression and its correlation with model performance. *arXiv preprint arXiv:2403.06265*.

John Hewitt. 2021. Initializing new word embeddings for pretrained language models. *URL: https:/nlp. stanford. edu/˜ johnhew//vocab-expansion. html*.

Tatsuya Hiraoka, Hiroyuki Shindo, and Yuji Matsumoto. 2019. Stochastic tokenization with a language model for neural text classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1620–1629.

Ayyoob ImaniGooghari, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

Alan Jaffe. 2017. Generating image descriptions using multilingual data. In *Proceedings of the second conference on machine translation*, pages 458–464.

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.

Davis Liang, Hila Gonen, Yuning Mao, Rui Hou, Naman Goyal, Marjan Ghazvininejad, Luke Zettlemoyer, and Madian Khabsa. 2023. XLM-V: Overcoming the vocabulary bottleneck in multilingual masked language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13142–13152, Singapore. Association for Computational Linguistics.

Tomasz Limisiewicz, Jiří Balhar, and David Mareček. 2023. Tokenization impacts multilingual language

modeling: Assessing vocabulary allocation and overlap across languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5661–5681, Toronto, Canada. Association for Computational Linguistics.

Yihong Liu, Peiqin Lin, Mingyang Wang, and Hinrich Schuetze. 2024. OFA: A framework of initializing unseen subword embeddings for efficient large-scale multilingual continued pretraining. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1067–1097, Mexico City, Mexico. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Roberta: A robustly optimized bert pretraining approach. In *International Conference on Learning Representations*.

Johannes V Lochter, Renato M Silva, and Tiago A Almeida. 2020. Deep learning models for representing out-of-vocabulary words. In *Brazilian Conference on Intelligent Systems*, pages 418–434. Springer.

Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. 2024. Fineweb-edu: the finest collection of educational content.

Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. 2022. WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3992–4006, Seattle, United States. Association for Computational Linguistics.

Shamsuddeen Hassan Muhammad, Seid Yimam, Idris Abdulmumin, Ibrahim Sa'id Ahmad, Nedjma Ousidhoum, Abinew Ayele, David Adelani, Sebastian Ruder, Meriem Beloucif, Shehu Bello Bello, and Saif M. Mohammad. 2023. SemEval-2023 task 12: Sentiment analysis for african languages (AfriSenti-SemEval). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*.

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021. When being unseen from mbert is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462.

Nandini Mundra, Aditya Nanda Kishore Khandavally, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, and Mitesh M Khapra. 2024a. An empirical comparison of vocabulary expansion and initialization approaches for language models. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 84–104, Miami, FL, USA. Association for Computational Linguistics.

Nandini Mundra, Aditya Nanda Kishore, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, and Mitesh M Khapra. 2024b. An empirical comparison of vocabulary expansion and initialization approaches for language models. *arXiv preprint arXiv:2407.05841*.

Artidoro Pagnoni, Ram Pasunuru, Pedro Rodriguez, John Nguyen, Benjamin Muller, Margaret Li, Chunting Zhou, Lili Yu, Jason Weston, Luke Zettlemoyer, et al. 2024. Byte latent transformer: Patches scale better than tokens. *arXiv preprint arXiv:2412.09871*.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, Thomas Wolf, et al. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. *arXiv preprint arXiv:2406.17557*.

Aleksandar Petrov, Emanuele La Malfa, Philip Torr, and Adel Bibi. 2024. Language model tokenizers introduce unfairness between languages. *Advances in Neural Information Processing Systems*, 36.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. Mad-x: An adapter-based framework for multi-task cross-lingual transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673.

Emmanouil Antonios Platanios, Mrinmaya Sachan, Graham Neubig, and Tom Mitchell. 2018. Contextual parameter generation for universal neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 425–435.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

William Shiao, Mingxuan Ju, Zhichun Guo, Xin Chen, Evangelos Papalexakis, Tong Zhao, Neil Shah, and

Yozen Liu. 2024. Improving out-of-vocabulary handling in recommendation systems. *arXiv preprint arXiv:2403.18280*.

Jimin Sun, Patrick Fernandes, Xinyi Wang, and Graham Neubig. 2023a. A multi-dimensional evaluation of tokenizer-free multilingual pretrained models. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1725–1735, Dubrovnik, Croatia. Association for Computational Linguistics.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023b. Text classification via large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8990–9005, Singapore. Association for Computational Linguistics.

Tilahun Abedissa Taffa, Ricardo Usbeck, and Yaregal Assabie. 2024. Low resource question answering: An Amharic benchmarking dataset. In *Proceedings of the Fifth Workshop on Resources for African Indigenous Languages @ LREC-COLING 2024*, pages 124–132, Torino, Italia. ELRA and ICCL.

Zeerak Talat, Aurélie Névéol, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, Shanya Sharma, Arjun Subramonian, Jaesung Tae, Samson Tan, Deepak Tunuguntla, and Oskar Van Der Wal. 2022. You reap what you sow: On the challenges of bias evaluation under multilingual settings. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 26–41, virtual+Dublin. Association for Computational Linguistics.

Argilla Team. n.d. Argilla: An open-source framework for managing and labeling data. Accessed: [2023].

Hailay Kidu Teklehaymanot, Dren Fazlija, Niloy Ganguly, Gourab Kumar Patro, and Wolfgang Nejdl. 2024. TIGQA: An expert-annotated question-answering dataset in Tigrinya. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16142–16161, Torino, Italia. ELRA and ICCL.

Asahi Ushio, Yi Zhou, and Jose Camacho-Collados. 2023. Efficient multilingual language model compression through vocabulary trimming. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14725–14739, Singapore. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Akhila VH and Anu Mary Chacko. 2024. Cooperative embedding-a novel approach to tackle the out-of-vocabulary dilemma in bot classification. In *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*, pages 1479–1486.

Changhan Wang, Kyunghyun Cho, and Jiatao Gu. 2020. Neural machine translation with byte-level subwords. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34(05), pages 9154–9160.

Hai Wang, Dian Yu, Kai Sun, Jianshu Chen, and Dong Yu. 2019a. Hai wang, dian yu, kai sun, janshu chen, and 791 dong yu. 2019. improving pre-trained multilingual 792 models with vocabulary expansion. arxiv preprint 793 arxiv:1909.12440. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 316–327, Hong Kong, China. Association for Computational Linguistics.

Hai Wang, Dian Yu, Kai Sun, Jianshu Chen, and Dong Yu. 2019b. Improving pre-trained multilingual model with vocabulary expansion. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 316–327.

Xinyi Wang, Sebastian Ruder, and Graham Neubig. 2022a. Expanding pretrained models to thousands more languages via lexicon-based adaptation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Xinyi Wang, Sebastian Ruder, and Graham Neubig. 2022b. Expanding pretrained models to thousands more languages via lexicon-based adaptation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 863–877, Dublin, Ireland. Association for Computational Linguistics.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Hailemariam Mehari Yohannes and Toshiyuki Amagasa. 2022. A method of named entity recognition for tigrinya. *ACM SIGAPP Applied Computing Review*, 22(3):56–68.

Shiyue Zhang, Vishrav Chaudhary, Naman Goyal, James Cross, Guillaume Wenzek, Mohit Bansal, and Francisco Guzman. 2022. How robust is neural machine translation to language imbalance in multilingual tokenizer training? In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 97–116, Orlando, USA. Association for Machine Translation in the Americas.

Zhongyu Zhuang, Ziran Liang, Yanghui Rao, Haoran Xie, and Fu Lee Wang. 2023. Out-of-vocabulary word embedding learning based on reading comprehension mechanism. *Natural Language Processing Journal*, 5:100038.

A Educational Quality Classification Dataset

The scarcity of labeled data is a particularly notable issue in low-resource languages. With this in mind, we collect human-annotated educational content evaluation data for training and evaluating multilingual models in two low-resource languages, Amharic and Tigriyna, carried out from October 2024 up to December 2024. Community-driven annotator primarily from Ethiopia and is actively involved in data development. All the collected content is sourced from public educational blogs. Community-driven efforts achieve this by creating a human-annotated dataset. Annotators assessed the educational value of each context on a scale from 1 to 6 following a detailed annotation guideline. This dataset is intended to train and evaluate educational quality classifier models.

Language Selection

To compile the dataset, we employ a methodology akin to that used in the original FineWeb-Edu datasets (Penedo et al., 2024). We focus on two specific languages, Tigrinya and Amharic, to establish a ground truth dataset for assessing educational value in low-resource languages.

FineWeb-Edu comprises 1.3 trillion tokens, specifically optimized for educational content, and significantly surpasses all openly available web-based datasets in several reasoning- and knowledge-intensive benchmarks, including MMLU, ARC, and OpenBookQA (Penedo et al., 2024). Unlike FineWeb, which relies solely on web content scraped through Common Crawl and often includes unstructured, noisy, and low-quality material. Unlike the FineWeb-Edu datasets (Penedo et al., 2024), we enhance our dataset with structured data from online manuals and public educational blogs to improve quality and diversity. While the initial annotations were generated using a large language model (LLM), we refined and verified these annotations through human annotation to ensure accuracy and reliability.

As shown in Table 4, we compiled a dataset for educational quality classification using a variety of academic blogs and manual excerpts as context for Amharic and Tigrinya languages. This annotated dataset will serve as a reliable ground truth for researchers working with these low-resource languages, providing an invaluable resource for developing and evaluating models that distinguish educational content from other types of online material.

Statistic	Value
Total Contexts in Amharic	750
Total Contexts in Tigriyna	750
Total Contexts in Both Languages	1500
Total Words in Amharic	39,477
Total Words in Tigriyna	44,210
Total Words (Both Languages)	83,687
Total Unique Tokens	25,795

Table 4

Preprocessing

After collecting the row texts, we perform several preprocessing steps. First, we eliminate any remaining texts in other languages using a FastText-based language identifier (Bojanowski et al., 2017). Additionally, we filter out texts containing abusive language by applying straightforward rule-based heuristics. To maintain high-quality text content, we remove entries containing URLs or emojis. Finally, tokenized text was split into sentences and further into individual words.

Annotation

Each text entry in the dataset was annotated by five coders, with each coder selecting one or more labels from six category classes. The coders who participated in this task were volunteers contributing to a community engagement effort. The annotation process was carried out using the open-source tool Argilla (Team, n.d.) as you see in Fig

B Annotation Guidline

Guidelines for Rating Educational Value of the Content. It comprises six categories: None, Minimal, Basic, Good, Excellent Problematic Content Rate the content using the following criteria:

[1] **No** Educational Value:
Definition: No educational purpose whatsoever. Purely entertainment, advertisements, or personal

content with nothing to learn.
Examples: Social media conversations about daily life Online shopping product listings Advertisement pages Personal blog posts about someone's day Forum discussions about entertainment Comment sections Sports match reports.

[2] **Minimal** Educational Value:
Definition: Contains a few facts or pieces of information, but the content is mostly non-educational. Information is incidental or not the main focus. **Examples**: News article that mentions some historical facts A travel blog with basic information about a location Product review with some technical details Company website with brief industry information Recipe that briefly explains a cooking technique Entertainment article with occasional facts.

[3] **Basic** Educational Content:
Definition: Attempts to explain or teach something, though the information might be scattered or disorganized. Mixed with non-educational content. **Examples**: A basic how-to guide with ads Simple Wikipedia-style article Blog post explaining a concept but lacking depth Amateur tutorial video transcript Brief explanation of a scientific concept Quick overview of a historical event.

[4] **Good** Educational Content:
Definition: Has a clear teaching purpose and well-organized information. Suitable for learning but may have minor limitations. Examples: Detailed tutorial with clear steps Well-written educational blog post Comprehensive guide to a topic Clear explanation of a scientific process Structured learning material Educational website article with examples.

[5] **Excellent** Educational Content:
Definition: Outstanding teaching material with a clear structure and thorough explanations. Includes helpful examples and lacks distracting content. **Examples**: Professional educational resource Well-crafted learning module In-depth guide with clear examples Comprehensive educational article High-quality teaching material Expert explanation with practical applications.

[6] **Problematic Content**
Definition: Unreadable or corrupted text, inappropriate content, or machine-generated nonsense. **Examples**: Text in a different language than expected Garbled characters or formatting AI-generated spam content Inappropriate or offensive material Broken/partial webpage content Content that's too technical to evaluate.

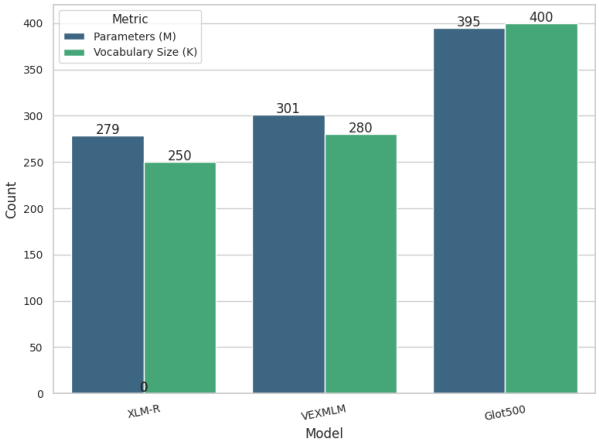

Figure 4: Vocabulary size of the models

Language	Country	Script
Algerian Arabic (arq)	Algeria	Arabic
Amharic (ama)	Ethiopia	Geez
Hausa (hau)	Nigeria, Niger, Ghana, Cameroon, Benin, Togo, Chad, Sudan	Arabic
Igbo (ibo)	Nigeria, Equatorial Guinea, Barbados, Cuba, Jamaica	Latin
Kinyarwanda (kin)	Rwanda	Latin
Moroccan Arabic/Darija (ary)	Morocco	Arabic
Mozambique Portuguese (pt-MZ)	Mozambique	Latin
Nigerian Pidgin (pcm)	Nigeria	Latin
Oromo (orm)	Ethiopia	Latin
Swahili (swa)	Kenya, Tanzania, Uganda, Democratic Republic of the Congo (DRC), Mozambique	Latin
Tigrinya (tir)	Eretiria, ethiopia and mainly in Australia and America	Ge'ez
Twi (twi)	Ghana	Latin
Xithonga (tso)	Mozambique	Latin
Yoruba (yor)	Nigeria, Benin, Togo	Latin
Kinyarwanda	Rwanda	Latin
Luganda	Uganda	Latin
Luo	Kenya, Tanzania	Luo Lakeside
Naija Pidgin	Nigeria (also spoken in some West African countries)	Latin
Wolof	Senegal, Gambia, Mauritania	Latin
Gurage	Ethiopia	Geez
Harari	Ethiopia	Ethiopic
Afar	Ethiopia	Ge'ez
Tigre	Eritrea	Ge'ez
Ge'ez	now primarily used as a liturgical language in the Ethiopian Orthodox Church	Ge'ez

Table 5: Languages, Countries, and Scripts in Low Resource African Languages

	Context	Label
1	ዘፈር ልምዓት ላዕለዎት ትምህርቲ ብሚኒስትሪ ትምህርቲ ዝመሓደር ኮይኑ ብኣርባዕተ ዋና ኣካየድቲ ስራሕን 17 ዴስክን ዝቐመ ዓውዲ እዩ። እቲ ዓውዲ ፅሬት ዘለዎ ትምህርቲ፣ መፅናዕትን ምህዞን ከምኡ እውን ላዕለዎት ትምህርቲ ብምስፋሕ ብማሕበረሰብ ብርኪ ግልጋሎት ብምሃብ ብቑዓትን ተወዳደርትን ዜጋታት ንምፍራይ ይሰርሕ።	2
2	ዘፈር ልምዓት ላዕለዎት ትምህርቲ ብሚኒስትሪ ትምህርቲ ዝመሓደር ኮይኑ ብኣርባዕተ ዋና ኣካየድቲ ስራሕን 17 ዴስክን ዝቐመ ዓውዲ እዩ። እቲ ዓውዲ ፅሬት ዘለዎ ትምህርቲ፣ መፅናዕትን ምህዞን ከምኡ እውን ላዕለዎት ትምህርቲ ብምስፋሕ ብማሕበረሰብ ብርኪ ግልጋሎት ብምሃብ ብቑዓትን ተወዳደርትን ዜጋታት ንምፍራይ ይሰርሕ።	4

Figure 5: EQC Dataset Samples for Tigriyna

	Context	Label
1	የሰው ልጅ ኑሮ በጉርብጥብጥና በሸለቆ በተሞላበት እውነታ ውስጥ ነፃነትና ሰንሰለቶች ጉራማይሌ ህልውና አላቸው። ዕውቀት፣ ጥበብ፣ ሥነ ምግባር፣ ርዕዮተ ዓለም፣ እምነት፣ ወዘተ ከኅብረተሰብ ወደ ግለሰብ፣ ከግለሰብ ወደ ኅብረተሰብና ከግለሰብ ወደ ግለሰብ ሲተላለፉ እንደኖሩ ሁሉ የሰንሰለቶች ቀለበቶችም እንደዚያ ይሸጋገራሉ። ከዚህ በፊት እንደተባለው ለህሊና ሰንሰለቶች መተላለፊያ ሠረገላቸው ብዙ ነው። መደበኛና ኢመደበኛ ትምህርት፣ ሥነ ቃል፣ ሙዚቃ፣ እምነት፣ ወዘተ ሁሉ ሠረገላ ሊሆኑ ይችላሉ። ዛሬ በፈጣሪና በመላዕክት እንደሚያለው የእኔ ጋይለ ሥላሴ ስምም መማያ የነበረበት ጊዜ ነበር። የጋይለ ሥላሴ ውዳሴ በእነዚህ በእነዚህ መንገዶች ናኘ ወይም ወደ ሕዝብ ሠረፀ ብሎ ለመዘርዘር ከመሞከር ይልቅ፣ ለዘመኑ ኅብረተሰብ በጡጦ መሰጠት የቀረው ምግብ ነበር ብሎ መናገር ይቀላል። ትችትን ያነወረ ውዳሴን መቅለብ፣ ምን ያህል በሰንሰለት መያዝ እንደሆነ ወይም ምን ያህል እንደሚገንዝና የለውጥ ተቀናቃኞች መጫወቻ ለመሆን እንደሚዳርግ ትናንትናም ታይቷል፣ ዛሬም እያየነው ነው።	4
2	ኢትዮጵያ አዲስ ምዕራፍ ላይ ትገኛለች ስንል፣ ከላይ ከጠቀስነው አካባቢያዊና ዓለም-ገብ ሁኔታ ጋር ከተገናኙ ብዙ አሮጌ ችግሮች ጋር ገና የሚያታግል ፈተና እንዳለብን ሳንዘነጋ ነው። የውጭና የውስጥ ኃይሎች በተቀናበሩበት ትግግዝ፣ ኢትዮጵያን በተነጣይ የጠመንጃ ትግሎች እየሸራረፉ የማሳነስ ሙከራ አንጋፋ ችግራችን ነው። በአንደኛው የዓለም ጦርነትነት ውጤትነት ከአስትሮ-ሃንጋሪና ከኦቶማን ኢምፓየሮች ማኅፀን አገሮች እንደተፈለፈሉ ሁሉ፣ ኢትዮጵያንም በጦርነት አሸመድምዶ ወደ ሰላም-አልባ ብጥስጣሽ አገሮች እንድትወርድ ተለፍቷል። ኢትዮጵያን በጦርነት ምርኮ ይዞ ዕጣዋን መወሰን የማይሆን፣ ከሆነም የእርስ በርስ ጭፍን ፍጅት አቀጣጥሎ በ‹ዘር ፍጅት› ጥቃት ዓለም የደገፈው መነጠልን ለማዋለድም ተሞክሯል። ደባዎቹ ገና አልደረቁም። ጥይት ተኳሽ የውስጥ በጥባጭና ተነጣይ አርብተውና ተጣብተው ኢትዮጵያን ለመሰባበር የሚመኙ የጥፋት ኃይሎች ዛሬም አገራችንን ዙሪያ እየዞሯት ናቸው። የውጭ ተቀፂላ በሆነ ጦርነት አማካይነት መነጠልን ማሳካት የሚሹ በውስጣችን አሉ። ከጦርነት በመለስ ጭቅጭቅ ያለባቸውን መሬቶች ሉዓላዊ ይዞታችን በሚል ብልጣ ብልጥነት አንቀው ዴሞክራሲ ሳይደላደል በፊት ባልፀዳ የሪፈረንደም ‹‹ጥቦብ›› ለማምለጥ ያደቡም ቡድኖች አሉ።	3

Figure 6: EQC Dataset Samples for Amharic