# LGSE: Lexically Grounded Subword Embedding Initialization

**Anonymous ACL submission**


## Abstract

Adapting pretrained language models to low-resource, morphologically rich languages remains a significant challenge. Existing vocabulary expansion methods typically rely on arbitrarily segmented subword units, resulting in fragmented lexical representations and loss of critical morphological information. To address this limitation, we propose the Lexically Grounded Subword Embedding Initialization (LGSE) framework, which introduces morphologically informed segmentation for initializing embeddings of novel tokens. Instead of using random vectors or arbitrary subwords, LGSE decomposes words into their constituent morphemes and constructs semantically coherent embeddings by averaging pretrained subword or FastText-based morpheme representations. When a token cannot be segmented into meaningful morphemes, its embedding is constructed using character n-gram representations to capture structural information. During Language-Adaptive Pretraining, we apply a regularization term that penalizes large deviations of newly introduced embeddings from their initialized values, preserving alignment with the original pretrained embedding space while enabling adaptation to the target language. To isolate the effect of initialization, we retain the original XLM-R vocabulary and tokenizer and update only the new embeddings during adaptation. We evaluate LGSE on three NLP tasks: Question Answering, Named Entity Recognition, and Text Classification, in two morphologically rich, low-resource languages: Amharic and Tigrinya. Experimental results show that LGSE consistently outperforms baseline methods across all tasks, demonstrating the effectiveness of morphologically grounded embedding initialization for improving representation quality in underrepresented languages.


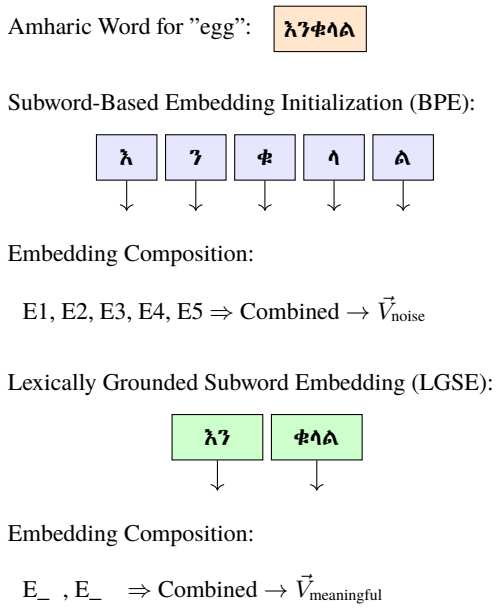


Figure 1: Comparison of embedding initialization strategies: standard BPE subword splits vs. linguistically grounded morphemes.

## 1 Introduction

Pretrained multilingual language models (PLMs) have become foundational in modern natural language processing (NLP), leveraging token sequences generated from word or subword-level units (Liu et al., 2024). A representative example is XLM-R (Conneau et al., 2020), a transformer-based model trained on over 100 languages using the SentencePiece algorithm for subword segmentation. Although XLM-R employs a shared vocabulary of 250K subword units, the effective average token coverage per language is relatively limited, approximately 2.5K subwords compared to monolingual models such as GPT, which typically utilize vocabularies in the range of 40K tokens (Wang et al., 2019).

Despite their wide language coverage, PLMs tend to favor high-resource languages, especially

those that are typologically or orthographically closer to English (e.g., French, Spanish). In contrast, morphologically rich languages such as German face heightened out-of-vocabulary (OOV) challenges due to their complex inflectional and derivational systems (Ataman and Federico, 2018; Lample et al., 2018; Wang et al., 2019). These issues are significantly exacerbated for low-resource languages, particularly those written in non-Latin scripts. Languages based on the Ge'ez script, such as Amharic and Tigrinya, suffer from poor lexical coverage and unreliable token representations due to a combination of script-specific orthographic complexity and minimal training data.

To mitigate out-of-vocabulary (OOV) issues, subword tokenization techniques such as Byte Pair Encoding (BPE), introduced by Sennrich et al. (2016), have become foundational in neural machine translation (NMT) and broader NLP pipelines (Hiraoka et al., 2019; Bostrom and Durrett, 2020). However, BPE operates purely on character co-occurrence frequency and disregards linguistic structure, often fragmenting morphologically rich words into arbitrary subword units. This segmentation undermines semantic coherence, particularly in agglutinative or templatic languages. The issue is especially pronounced in underrepresented languages using the Ge'ez script, such as Amharic and Tigrinya, where BPE frequently breaks full lexical units, including complete nouns, into semantically meaningless fragments.

As illustrated in Figure 1, the Amharic word **እንቁላል** ('Iənqulāl', meaning "egg") is decomposed into a series of subwords that fail to preserve its morphemic integrity. This fragmentation negatively affects subword embedding initialization by associating these noisy segments with ill-grounded or diluted vector representations, yielding embeddings that poorly capture the word's meaning. Consequently, morphologically aware tokenization is essential not merely for better segmentation but as a prerequisite for reliable and linguistically grounded embedding initialization in morphologically complex, low-resource languages.

Multilingual models like mBERT and XLM-R, which rely on shared vocabularies and embedding spaces across languages, often fail to encode the morphosyntactic nuances of such languages (Ahia et al., 2023; Wang et al., 2019). While language-adaptive pretraining (LAPT) (Chau et al., 2020) and similar transfer learning techniques aim to bridge this representational gap, they still struggle when applied to typologically distinct scripts. In particular, expanding the vocabulary or retraining the embedding matrix with newly introduced tokens disrupts alignment with the pretrained distribution, complicating the integration of linguistically informed tokenizers (Dobler and de Melo, 2023; de Vries and Nissim, 2021).

Although methods like vocabulary expansion and random or averaged embedding initialization provide partial relief, they fail to restore the structural grounding that morpheme-level units provide. This is especially critical for morphologically rich languages, where tokenization and embedding decisions are tightly coupled. As highlighted by Mofijul Islam et al. (2022) and Limisiewicz et al. (2023), subword-based models often produce semantically fragmented and unstable representations across languages, particularly in low-resource settings.

To address these limitations, we advocate for embedding strategies that align with morphologically aware tokenization. By respecting linguistic structure during both tokenization and embedding initialization, such methods promise not only improved representation quality but also fairer and more effective inclusion of underrepresented languages in NLP systems (Hangya et al., 2023). Our contributions are:

**(1)** We reveal that subword-based embeddings used in current multilingual pretrained models fail to capture the morphological structure of low-resource, morphologically rich languages, leading to fragmented and semantically weak representations;

**(2)** We propose Lexically Grounded Subword Embedding Initialization (LGSE), a novel strategy that respects linguistic boundaries by leveraging morpheme-aware segmentation for embedding initialization. Unlike conventional methods that rely on arbitrary subword fragments, LGSE creates semantically coherent representations, enabling more accurate and robust representation learning for underrepresented, morphologically rich languages such as Amharic and Tigrinya.

**(3)** We introduce the first human-annotated benchmark dataset for evaluating downstream NLP tasks and assessing model performance in identifying high-quality educational content for two morphologically rich, underrepresented languages Amharic and Tigrinya. This resource fills a critical gap for low-resource languages and provides a foundation for future research on cross-lingual

transfer, morphological modeling, and educational AI;

**(4)** We rigorously evaluate LGSE on three downstream NLP tasks Question Answering, Named Entity Recognition, and Text Classification using two morphologically complex, low-resource languages: Amharic and Tigrinya. Compared to strong multilingual baselines, LGSE achieves substantial and consistent improvements over conventional subword-based embedding methods, demonstrating the effectiveness of linguistically grounded initialization in challenging language settings.

## 2 Related Work

### 2.1 Over-Segmentation in Low-Resource Languages

Subword tokenization methods such as Byte-Pair Encoding (BPE) and SentencePiece are widely used in multilingual pretrained language models. However, these approaches often cause excessive fragmentation when applied to morphologically rich and low-resource languages (Rust et al., 2021; Muller et al., 2021). This over-segmentation leads to longer token sequences, which increase inference time (Hofmann et al., 2022; Sun et al., 2023), raise API costs (Ahia et al., 2023; Petrov et al., 2023), and degrade downstream task performance (Bostrom and Durrett, 2020; Fujii et al., 2023). Tokenizers trained on high-resource languages often produce segmentation mismatches in low-resource languages due to their lack of morphological awareness (Sun et al., 2023).

### 2.2 Vocabulary Expansion and Embedding Initialization

Vocabulary expansion is a common strategy for adapting pretrained models to underrepresented languages, particularly when the base vocabulary lacks coverage for non-Latin scripts or language-specific structures (Conneau et al., 2020; Downey et al., 2024; Pfeiffer et al., 2020). Pretrained models typically use a fixed vocabulary of approximately 50K tokens (Ushio et al., 2023), which often fails to adequately represent morphologically rich or low-resource languages.

To address out-of-vocabulary (OOV) issues, embedding initialization methods aim to leverage pretrained representations. Liu et al. (2021) propose synthesizing OOV embeddings using subword and hyperword information. UniBridge aligns non-overlapping tokens across languages via syntactic and semantic embeddings to enhance cross-lingual transfer (Pham et al., 2024). EVALM mitigates overfitting by initializing new tokens with high-resource language translations (InitHRL) and applying regularization during fine-tuning (Nag et al., 2023).

Recent work has also explored vocabulary expansion in decoder-only models such as LLaMA 2 and 3 to improve generative performance in low-resource settings (Balachandran, 2023; Larcher et al., 2023; Lin et al., 2024; Cui et al., 2023; Fujii et al., 2024; Choi et al., 2024b; Nag et al., 2025). These approaches typically add subwords based on frequency and continue pretraining or fine-tune with instruction data. While effective in reducing token overhead and improving fluency, they often rely on naive subword addition and initialization methods that do not account for linguistic structure. For example, Balachandran (2023) and Cui et al. (2023) introduce additional tokens for Tamil and Chinese using simple initialization, while Fujii et al. (2024) adapts LLaMA 2 for Japanese through cross-lingual pretraining. Similarly, Choi et al. (2024a) and Nguyen et al. (2024) extend coverage for Korean and Southeast Asian languages. However, these methods do not explicitly address fragmentation or incorporate morphological alignment in their tokenization strategies.

### 2.3 Linguistically Informed Embedding Alignment

Several studies have explored embedding reinitialization and alignment strategies that incorporate cross-lingual semantics. WECHSEL (Minixhofer et al., 2022) maps new subword embeddings to semantically similar words using multilingual vector alignment. Although it improves zero-shot transfer, it treats subwords as atomic units and overlooks morphological structure. OFA (Liu et al., 2024) introduces matrix factorization to compress the embedding space for scalable adaptation, yet it also ignores language-internal patterns.

Language-specific vocabulary augmentation has been shown to improve syntactic tasks in low-resource languages (Chau et al., 2020), and Mundra et al. (2024) provide a comparative analysis of embedding initialization methods. Nonetheless, existing approaches largely neglect morpheme-based segmentation and do not exploit morphological composition for embedding initialization.

## 3 Problem Statement

Multilingual pretrained models such as mBERT and XLM-R employ a shared subword vocabulary $\mathcal{V}$ across multiple languages $\mathcal{L} = \{L_1, L_2, \ldots, L_m\}$. For a word $w \in L_i$, a tokenizer $T$ segments it into subwords $T(w) = [s_1, s_2, \ldots, s_n]$, where each subword $s_i \in \mathcal{V}$ is associated with a pretrained embedding $\mathbf{e}_i \in \mathbb{R}^d$. However, this subword segmentation frequently fails to align with the word's true morphemic structure $M(w) = [m_1, m_2, \ldots, m_k]$, where each $m_j$ represents a linguistically meaningful morpheme.

This misalignment is particularly problematic for morphologically rich and low-resource languages, leading to suboptimal semantic representations and poorer generalization on unseen or infrequent tokens.

## 4 Vocabulary expansion and Initialization

This section introduces two key approaches to enhancing embedding initialization for morphologically rich and low-resource languages. Section 4.1 presents Lexically Grounded Subword Embedding Initialization Framework (LGSE) which leverages subword-level semantic representations from FastText (Bojanowski et al., 2017) to initialize embeddings for morpheme-aligned tokens. This method ensures that the initialized vectors capture meaningful morphological and semantic patterns, aligning with the language's internal structure. In Section 4.2, we describe embedding initialization for new morphologically grounded tokens, which addresses out-of-vocabulary (OOV) scenarios by generating embeddings for novel morpheme-based units using composition strategies informed by morphological structure and distributional semantics. Together, these strategies aim to improve vocabulary coverage, semantic coherence, and representation quality in low-resource, morphologically complex languages.

### 4.1 Lexically Grounded Subword Embedding Initialization (LGSE)

Given access to a morphologically-aware tokenizer and pretrained FastText embeddings, we represent a new token $t$ segmented into morphemes $M(t) = [m_1, m_2, \ldots, m_k]$. Each morpheme $m_j$ is further represented by a set of character $n$-grams

$$G_j = \{g_{j1}, g_{j2}, \ldots, g_{jn_j}\}.$$

Each $n$-gram $g$ has an associated FastText embedding $\mathbf{f}_g \in \mathbb{R}^d$. The embedding for morpheme $m_j$ is computed as the average of its constituent $n$-gram embeddings

$$\mathbf{m}_j = \frac{1}{|G_j|} \sum_{g \in G_j} \mathbf{f}_g$$

while the initial token embedding $\mathbf{e}_t$ is obtained by averaging over all morpheme embeddings, i.e.,

$$\mathbf{e}_t = \frac{1}{k} \sum_{j=1}^{k} \mathbf{m}_j = \frac{1}{k} \sum_{j=1}^{k} \left( \frac{1}{|G_j|} \sum_{g \in G_j} \mathbf{f}_g \right).$$

To align the FastText embedding space with the pretrained model embedding space, a learned linear projection $\mathbf{W} \in \mathbb{R}^{d \times d}$ is applied, i.e.,

$$\mathbf{e}_t^{\text{aligned}} = \mathbf{W}\mathbf{e}_t.$$

### 4.2 Embedding Initialization for New Lexically Grounded Subword Tokens

We initialize the embedding for a new token as the average of its morpheme embeddings computed via FastText-based pooling:

$$\mathbf{e}_{\text{new}} = \frac{1}{k} \sum_{j=1}^{k} \mathbf{m}_j^{\text{aligned}},$$

where $\mathbf{m}_j^{\text{aligned}}$ are morpheme embeddings after projection. For tokens without known morpheme segmentation or embeddings, we initialize by sampling from a multivariate normal distribution estimated from existing pretrained embeddings:

$$\mathbf{e}_{\text{new}} \sim \mathcal{N}(\mu, \Sigma),$$

where $\mu$ and $\Sigma$ are the mean and covariance matrix of pretrained embeddings. To prevent excessive deviation of new embeddings from their initialization during continual pretraining or fine-tuning, we apply the regularization loss

$$\mathcal{L}_{\text{reg}} = \lambda \left\| \mathbf{e}_{\text{new}} - \boldsymbol{\mu} \right\|^2,$$

where $\boldsymbol{\mu}$ is the initial embedding vector (e.g., from FastText projection), and $\lambda$ controls the regularization strength, balancing stability and adaptability.

## 5 Experimental Setup

Our experiments are conducted using the multilingual encoder-based model XLM-R (Conneau et al., 2020) as the foundational architecture. XLM-R is selected for its proven cross-lingual transfer performance and extensive use in multilingual NLP research. Its decoupled SentencePiece tokenizer enables straightforward integration of morpheme-level tokens without modifying the underlying model architecture. To ensure fair comparison and reproducibility, all experiments utilize the base version of XLM-R and maintain consistent hyperparameters across both baseline and LGSE-enhanced models.

The model contains approximately 125 million parameters. Training was performed on a single GPU with 4 CPU cores and 46 GB RAM, with each run allocated up to 24 GPU hours on an Ampere architecture GPU. The computational environment was managed using Anaconda to ensure consistency and reproducibility.

### 5.1 Morphology-Aware Tokenization

We adopt a **morphologically informed tokenization strategy** that segments words into lexically grounded morphemes using supervised morphological analysis applied to monolingual corpora $P_{\text{am}}$ (Amharic) and $P_{\text{ti}}$ (Tigrinya). Unlike conventional tokenizers that rely solely on frequency-based subword segmentation, our approach respects linguistic boundaries to preserve morphological integrity. To construct a vocabulary that is both *linguistically meaningful* and *computationally efficient*, we combine high-frequency morpheme tokens with subword units learned via **Byte-Pair Encoding (BPE)**. A hyperparameter $r \in [0, 1]$ controls the ratio of morpheme tokens, yielding a hybrid vocabulary:

$$\begin{aligned} V &= V_{\text{BPE}_{\text{small}}} \cup V_{\text{morph}}, \\ |V_{\text{BPE}_{\text{small}}}| &= s(1-r), \\ |V_{\text{morph}}| &= sr. \end{aligned} \tag{1}$$

Tokenization proceeds in two stages: $(i)$ words are first segmented into morphemes; $(ii)$ BPE is then applied **within each morpheme**, preventing merges across morpheme boundaries.

Formally, for a word $w = m_1 m_2 \cdots m_k$, the tokenizer output is:

$$\text{Tokenizer}(w) = \bigcup_{i=1}^{k} \text{BPE}_{\text{small}}(m_i).$$

This **morphology-aware BPE** forms the foundation of our **Lexically Grounded Subword Embedding Initialization (LGSE)** framework. By aligning embeddings with linguistically interpretable morphemes and subwords, LGSE mitigates semantic fragmentation and noise introduced by arbitrary subword splits, thereby producing representations that better capture the morphological richness of underrepresented languages. For practical efficiency, we train the tokenizer to generate a vocabulary of 50,000 tokens per language using parallel and monolingual corpora from the **No Language Left Behind (NLLB)** project (Fan et al., 2021) for both Amharic and Tigrinya.

### 5.2 Baselines

To evaluate the effectiveness of our proposed LGSE approach, we compare it against several strong baselines. In all cases, the original XLM-R encoder layers remain frozen during initialization to isolate the effect of embedding strategies. All models subsequently undergo Language Adaptive Pretraining (LAPT) under identical settings for fairness.

- **XLM-R Off-the-Shelf:** The unmodified XLM-R model is used in a zero-shot setting without any additional training. This baseline provides a reference point for assessing the inherent transfer capabilities of the pretrained model in our target languages.

- **XLM-R + LAPT:** The original XLM-R vocabulary and embeddings are preserved, and the model is further adapted using Language Adaptive Pretraining on monolingual target language data. This measures the gains from LAPT alone without modifying the tokenizer or embeddings.

- **Random Initialization for Newly Added Tokens + LAPT:** When expanding the vocabulary with morphologically grounded tokens, only the embeddings for these new tokens are randomly initialized, while the pretrained embeddings and encoder parameters remain unchanged. Each new embedding vector is sampled from a Gaussian distribution estimated from the original embedding matrix:

  $$\mathbf{e}_t \sim \mathcal{N}(\mu, \Sigma),$$

  where $\mu$ and $\Sigma$ are the empirical mean and covariance of the original embeddings. This

baseline isolates the contribution of linguistically informed initialization by comparing against a purely random, statistically coherent initialization strategy.

- **Subword-Based Initialization (FOCUS) + LAPT:** We adopt FOCUS (Dobler and de Melo, 2023), a subword-level embedding refinement method that computes weighted combinations of overlapping pretrained subword tokens using Sparsemax. This improves representations for rare or unseen tokens without modifying the original tokenizer or vocabulary.

- **Lexically Grounded Subword Embedding Initialization (LGSE) + LAPT:** Our proposed approach combines morphology-aware tokenization with embedding initialization based on FastText-derived morpheme embeddings, aligned via a learned projection layer. This linguistically informed strategy mitigates over-fragmentation and enhances coverage of morphologically rich words, improving representation quality for low-resource languages.

## 6 Language-Adaptive Pretraining (LAPT)

To enhance the XLM-R model's performance on morphologically complex, low-resource languages such as Tigrinya and Amharic, we move away from subword-based approaches that utilize BPE vocabularies, such as FOCUS (Dobler and de Melo, 2023). Instead, we initialize the embedding layer with lexically grounded representations derived from a morphology-aware tokenizer trained on linguistically annotated corpora. This tokenizer segments text into morphemes, preserving the language's meaningful lexical and grammatical structures, unlike arbitrary subword units.

We employ *Language-Adaptive Pretraining (LAPT)* with a morpheme-level Masked Language Modeling (MLM) objective, initializing the embedding layer with morpheme-aware representations from annotated corpora to maintain the linguistic integrity of the target languages. For Amharic, we utilize the CC100 corpus (133M tokens), previously used in XLM-R pretraining (Conneau et al., 2020), while for Tigrinya, we rely on data from (Gaim et al., 2021), totaling approximately 0.5GB. Hyperparameters for both languages are consistent, as detailed in Appendix Table 3.

We preserve the pretrained XLM-R encoder parameters $\{L_1, L_2, \ldots, L_n\}$ and adapt only the embedding layer $E$, initializing it with a language-specific vocabulary $\mathcal{V}_{\text{morph}}$ tailored to each target language. Pretraining is performed on monolingual Tigrinya and Amharic corpora, applying a dynamic masking probability of 15% to sequences that are either truncated or padded to a maximum length of 256 tokens.

## 7 Evaluation

We evaluate our proposed models on a suite of downstream tasks in two low-resource, morphologically rich languages that use the Ge'ez script: **Amharic** and **Tigrinya**. These languages were selected due to the availability of supervised, morphologically annotated data as well as curated evaluation datasets. We conduct experiments on three key tasks: text classification, question answering, and named entity recognition. The hyperparameters used for all evaluation tasks are provided in Appendix B Table 3.

**Text Classification:** We address the task of assigning predefined labels to input texts, with a specific focus on evaluating the quality of educational content.

**Educational Quality Classification Dataset:** To support this task, we introduce a new benchmark dataset comprising 2,500 human-annotated samples in **Amharic** and **Tigrinya**. The dataset was developed in close collaboration with local linguistic communities to ensure cultural and linguistic relevance. Data collection proceeded in two stages: initially, a diverse set of texts was sourced from publicly available educational materials, including manuals and blog posts; subsequently, each text was annotated on a 1–6 scale reflecting perceived educational quality. Comprehensive dataset statistics and illustrative examples are provided in Appendix A. For model training and evaluation, the dataset was carefully curated and split into 80% for training, 10% for development, and 10% for testing.

**Named Entity Recognition (NER):** We perform NER experiments using the balanced train-dev-test splits of the **MasakhaNER** dataset (Adelani et al., 2021) for Amharic and For the **Tigrinya NER dataset** (Yohannes and Amagasa, 2022), where no official data split is provided, we create a

Table 1: Performance of XLM-R across three NLP tasks in Tigrinya and Amharic. F1 score is used for QA and NER; Accuracy is used for TC. All results are reported as mean ± standard deviation over five runs. The best performance per task is highlighted in bold.

| Model | Task Category | Task | Metric | Tigrinya | Amharic | Avg |
|---|---|---|---|---|---|---|
| XLM-R (off-the-shelf) | Question Answering | QA | F1 | 61.3 ± 0.4 | 71.4 ± 0.9 | 66.35 |
| | Text Classification | TC | AC | 63.2 ± 0.7 | 70.1 ± 0.6 | 66.65 |
| | Named Entity Recognition | NER | F1 | 66.4 ± 0.6 | 70.2 ± 0.8 | 68.3 |
| XLM-R + LAPT | Question Answering | QA | F1 | 70.5 ± 0.8 | 74.9 ± 0.5 | 72.7 |
| | Text Classification | TC | AC | 69.4 ± 0.5 | 71.0 ± 0.4 | 67.8 |
| | Named Entity Recognition | NER | F1 | 69.8 ± 0.5 | 75.0 ± 0.6 | 70.4 |
| XLM-R + Random + LAPT | Question Answering | QA | F1 | 68.7 ± 0.6 | 71.3 ± 0.8 | 70 |
| | Text Classification | TC | AC | 69.9 ± 0.6 | 70.8 ± 0.8 | 70.35 |
| | Named Entity Recognition | NER | F1 | 70.3 ± 0.7 | 74.0 ± 0.7 | 72.15 |
| XLM-R + FOCUS + LAPT | Question Answering | QA | F1 | 75.5 ± 0.3 | 77.8 ± 1.0 | 76.65 |
| | Text Classification | TC | AC | 72.4 ± 0.4 | 76.5 ± 0.9 | 74.45 |
| | Named Entity Recognition | NER | F1 | 77.5 ± 0.4 | 78.1 ± 0.9 | 77.8 |
| XLM-R + LGSE + LAPT | Question Answering | QA | F1 | **78.0 ± 0.4** | **78.5 ± 0.4** | **78.25** |
| | Text Classification | TC | AC | **75.2 ± 0.5** | **77.8 ± 0.3** | **76.5** |
| | Named Entity Recognition | NER | F1 | **79.0 ± 0.3** | **79.4 ± 0.4** | **79.2** |

consistent partition by randomly splitting the data into **80%** for training, **10%** for development, and **10%** for testing. Model selection is based on performance on the development set, and final results are reported on the test set.

**Question Answering (QA):** QA performance is evaluated on the **TIGQA** train-dev-test splits balanced dataset (Teklehaymanot et al., 2024), which contains expert-annotated question-answer pairs in Tigrinya. For Amharic, we use the **AmQA**, train-dev-test splits dataset (Taffa et al., 2024), developed for low-resource QA benchmarking. The final results are reported on the test set for both Amharic and Tigriyna QA datasets.

We report **F1 scores** for NER, QA, and Text classification. Each experiment is repeated **five times** with different random seeds. We report the **mean and standard deviation** of results. Full training configurations and hyperparameter settings are presented in Appendix B Table 3.

We compare our approach against three strong baselines:

- **XLM-R (off-the-shelf) (Conneau et al., 2020)** with no modifications,
- **Random embedding initialization**, and
- **FOCUS** (Dobler and de Melo, 2023), a recent subword-level initialization strategy using overlapping token combinations.

Unlike these baselines, our method (LGSE) explicitly incorporates **morpheme-level structure**, which we argue is essential for capturing the deep semantics of morphologically complex languages such as Amharic and Tigrinya.

## 8 Results and Discussion

The results in Table 1 demonstrate a clear and consistent performance improvement across all three tasks **Question Answering (QA)**, **Text Classification (TC)**, and **Named Entity Recognition (NER)** as additional training strategies and embedding initialization methods are applied to the XLM-R model.

### 8.1 Baseline Performance

The off-the-shelf XLM-R model yields the lowest performance across all tasks. This is expected, as the model has not been adapted to the specific languages or domains involved. For instance, it achieves an average QA F1 score of **66.35** and NER F1 score of **68.30**, indicating limited ability to generalize to Tigrinya and Amharic without further adaptation.

### 8.2 Impact of Language-Adaptive Pretraining (LAPT)

Applying **Language-Adaptive Pretraining (LAPT)** substantially improves performance across all tasks. QA and NER scores increase by approximately 6-7 points on average, confirming the benefit of continued pretraining on language-specific data for low-resource scenarios.

### 8.3 Effect of Embedding Initialization Methods

Beyond LAPT, we examine the impact of different subword embedding initialization methods: **Random** , **FOCUS**, and our proposed **Lexically Grounded Subword Embedding Initialization (LGSE)**.

The **FOCUS + LAPT** configuration outperforms the **Random + LAPT** baseline, achieving a QA F1 score of **76.65** and NER F1 of **77.80**. This indicates that more informed subword representations can lead to better convergence and improved performance.

### 8.4 Effectiveness of LGSE and Cross-Language Impact

The proposed method, **LGSE + LAPT**, which integrates Language-Adaptive Pretraining with **Lexically Grounded Subword Embedding Initialization (LGSE)**, achieves the best overall performance, obtaining QA F1 of **78.25**, TC accuracy of **76.50**, and NER F1 of **79.20**. LGSE employs a **morpheme-aware tokenizer** that captures linguistically meaningful units, offering improved representations for morphologically rich and low-resource languages such as Amharic and Tigrinya. Unlike conventional subword-based approaches, this method aligns with the underlying morphological structure of these languages, thereby enhancing semantic fidelity and reducing segmentation errors.

Our analysis further reveals that **vocabulary overlap** plays a non-trivial role in cross-lingual embedding transfer. Despite Tigrinya's absence in pretraining corpora, we observe approximately 1,280 shared morphemes with Amharic, largely driven by code-mixing rather than strict linguistic similarity. While this overlap facilitates partial transfer, it also introduces potential semantic drift. To address rare and out-of-vocabulary morphemes, LGSE leverages **FastText-based character n-gram embeddings**, enabling compositional representations and robust initialization, which are crucial for improving generalization in low-resource settings.

**Cross-Language Impact.** Although Amharic benefits from relatively larger resources, LGSE substantially reduces the performance gap with Tigrinya. This improvement underscores the effectiveness of **linguistically informed tokenization and embedding strategies** in supporting cross-lingual generalization under severe resource constraints, particularly for morphologically complex languages.

## 9 Conclusion

We propose a Lexically Grounded Subword Embedding Initialization (LGSE) framework for morphologically rich, low-resource languages, focusing on Amharic and Tigrinya. By combining morpheme-aware tokenization with FastText-based compositional embeddings and Language-Adaptive Pretraining (LAPT), LGSE consistently improves performance across multiple downstream tasks. These results underscore the benefits of incorporating lexical and morphological structure into multilingual NLP models.

## Ethical Considerations and Limitations

**Limitations and Future Work.** While the proposed framework demonstrates promising improvements, it faces several limitations. First, it depends on morphologically annotated resources, which remain scarce for many low-resource languages, constraining its applicability in truly multilingual settings. Second, the current design targets encoder-based architectures such as XLM-R, limiting direct integration with decoder-based or sequence-to-sequence models widely used in machine translation and other generative tasks. Third, the incorporation of Lexically Grounded Subword Embedding Initialization introduces additional computational overhead compared to frequency-driven subword segmentation methods, which may impact scalability for very large vocabularies or low-resource deployment environments.

As future work, we plan to extend the framework to decoder-based and encoder–decoder architectures, enabling its use in machine translation and generative modeling. Additionally, we aim to investigate vocabulary *replacement* versus *expansion* strategies under these settings to better understand their trade-offs in terms of efficiency and performance across diverse language families.

**Ethical Considerations.** This work uses only publicly available datasets, with all sources properly cited to ensure transparency. ChatGPT was used only for paraphrasing and language clarity no scientific content was generated. The Amharic and Tigrinya annotated datasets, models, and code will be released under an open-access license to support research equity and inclusivity. No personally

identifiable information (PII) or sensitive content is involved. All research activities adhere to established ethical guidelines for NLP, with attention to linguistic and cultural sensitivity in underrepresented language communities. Our goal is to promote responsible and inclusive cross-lingual NLP development.

## References


David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, and 1 others. 2021. Masakhaner: Named entity recognition for african languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David Mortensen, Noah Smith, and Yulia Tsvetkov. 2023. 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.

Duygu Ataman and Marcello Federico. 2018. Compositional representation of morphologically-rich input for neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 305–311, Melbourne, Australia. Association for Computational Linguistics.

Abhinand Balachandran. 2023. Tamil-llama: A new tamil language model based on llama 2. *arXiv preprint arXiv:2311.05845*.

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, Online. Association for Computational Linguistics.

Ethan C. Chau, Lucy H. Lin, and Noah A. Smith. 2020. Parsing with multilingual BERT, a small corpus, and a small treebank. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1324–1334, Online. Association for Computational Linguistics.

ChangSu Choi, Yongbin Jeong, Seoyoon Park, Inho Won, HyeonSeok Lim, SangMin Kim, Yejee Kang, Chanhyuk Yoon, Jaewan Park, Yiseul Lee, HyeJin Lee, Younggyun Hahm, Hansaem Kim, and Kyung-Tae Lim. 2024a. Optimizing language augmentation for multilingual large language models: A case study on Korean. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12514–12526, Torino, Italia. ELRA and ICCL.

ChangSu Choi, Yongbin Jeong, Seoyoon Park, Inho Won, HyeonSeok Lim, SangMin Kim, Yejee Kang, Chanhyuk Yoon, Jaewan Park, Yiseul Lee, and 1 others. 2024b. Optimizing language augmentation for multilingual large language models: A case study on korean. *arXiv preprint arXiv:2403.10882*.

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.

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

Wietse de Vries and Malvina Nissim. 2021. As good as new. how to successfully recycle English GPT-2 to make models for other languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 836–846, Online. Association for Computational Linguistics.

Konstantin Dobler and Gerard de Melo. 2023. FOCUS: Effective embedding initialization for monolingual specialization of multilingual models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13440–13454, Singapore. Association for Computational Linguistics.

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.

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

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities. *arXiv preprint arXiv:2404.17790*.

Takuro Fujii, Koki Shibata, Atsuki Yamaguchi, Terufumi Morishita, and Yasuhiro Sogawa. 2023. How

do different tokenizers perform on downstream tasks in scriptio continua languages?: A case study in Japanese. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 39–49, Toronto, Canada. Association for Computational Linguistics.

Fitsum Gaim, Wonsuk Yang, and Jong C. Park. 2021. Tlmd: Tigrinya language modeling dataset. [Dataset].

Viktor Hangya, Silvia Severini, Radoslav Ralev, Alexander Fraser, and Hinrich Schütze. 2023. Multilingual word embeddings for low-resource languages using anchors and a chain of related languages. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 95–105, Singapore. Association for Computational Linguistics.

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.

Valentin Hofmann, Hinrich Schuetze, and Janet Pierrehumbert. 2022. An embarrassingly simple method to mitigate undesirable properties of pretrained language model tokenizers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–393, Dublin, Ireland. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium. Association for Computational Linguistics.

Celio Larcher, Marcos Piau, Paulo Finardi, Pedro Gengo, Piero Esposito, and Vinicius Caridá. 2023. Cabrita: closing the gap for foreign languages. *arXiv preprint arXiv:2308.11878*.

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.

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

Xin Liu, Baosong Yang, Dayiheng Liu, Haibo Zhang, Weihua Luo, Min Zhang, Haiying Zhang, and Jinsong Su. 2021. Bridging subword gaps in pretrain-finetune paradigm for natural language generation. 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 6001–6011, Online. 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.

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.

Md Mofijul Islam, Gustavo Aguilar, Pragaash Ponnusamy, Clint Solomon Mathialagan, Chengyuan Ma, and Chenlei Guo. 2022. A vocabulary-free multilingual neural tokenizer for end-to-end task learning. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 91–99, Dublin, Ireland. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Nandini Mundra, Aditya Nanda Kishore Khandavally, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, and Mitesh M Khapra. 2024. 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.

Arijit Nag, Soumen Chakrabarti, Animesh Mukherjee, and Niloy Ganguly. 2025. Efficient continual pretraining of LLMs for low-resource languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 304–317, Albuquerque, New Mexico. Association for Computational Linguistics.

Arijit Nag, Bidisha Samanta, Animesh Mukherjee, Niloy Ganguly, and Soumen Chakrabarti. 2023. Entropy-guided vocabulary augmentation of multilingual language models for low-resource tasks. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8619–8629, Toronto, Canada. Association for Computational Linguistics.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Zhiqiang Hu, Chenhui Shen, Yew Ken Chia, Xingxuan Li, Jianyu Wang, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. 2024. SeaLLMs - large language models for Southeast Asia. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 294–304, Bangkok, Thailand. Association for Computational Linguistics.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, Thomas Wolf, and 1 others. 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. 2023. Language model tokenizers introduce unfairness between languages. *Advances in neural information processing systems*, 36:36963–36990.

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.

Trinh Pham, Khoi Le, and Anh Tuan Luu. 2024. UniBridge: A unified approach to cross-lingual transfer learning for low-resource languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3168–3184, Bangkok, Thailand. 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.

Jimin Sun, Patrick Fernandes, Xinyi Wang, and Graham Neubig. 2023. 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.

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.

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.

Hai Wang, Dian Yu, Kai Sun, Jianshu Chen, and Dong Yu. 2019. 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.

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

# A Appendix A

## Educational Quality Classification Dataset

The scarcity of labeled data is a particularly notable issue in low-resource languages. With this in mind, we collected human-annotated educational content benchmark data for training and evaluating models in two low-resource languages, Amharic and Tigriyna. The data is collected from October 2024 to April 2025. 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 split 80,10,10 to train-dev-test. This dataset also serves to evaluate models on their ability to identify and select high-quality educational content from web-based sources using a variety of methods.

## Language Selection

To compile the dataset, we employ a methodology akin to that used in the FineWeb-Edu datasets by (Penedo et al., 2024). 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 (Penedo et al., 2024), which relies solely on web content scraped through Common Crawl and often includes unstructured, noisy, and low-quality material. We enhance our dataset with structured data from online manuals and public educational blogs to improve quality and diversity by focusing on two specific languages, Tigrinya and Amharic.

| **Total Context** | |
|---|---|
| Amharic | 1,250 |
| Tigrinya | 1,250 |
| **Both Languages** | **2,500** |
| **Total No. Words** | |
| Amharic | 65,795 |
| Tigrinya | 73,683 |
| **Both Languages** | **139,478** |
| **Total Unique Tokens** | 42,991 |
| **Data Split (No. Samples)** | |
| Training Set (80%) | 2,000 |
| Development Set (10%) | 250 |
| Test Set (10%) | 250 |

Table 2: Statistics for the Educational Quality Classification dataset in Amharic and Tigrinya, comprising 2,500 human-annotated samples and split into training, development, and test sets.

As shown in Table 2, we constructed a dataset comprising 2,500 human-annotated samples for the task of educational quality classification in Amharic and Tigrinya. The dataset was developed using a diverse selection of source texts, including academic blogs and publicly available educational manuals, to ensure broad topical and stylistic coverage. Designed as a benchmark resource for low-resource language research, this dataset enables the training and evaluation of models capable of identifying high-quality educational content from general web-based material.

## Preprocessing

Following the collection of raw texts, we applied several preprocessing steps to ensure data quality. First, we removed non-target-language content using a FastText-based language identification tool (Bojanowski et al., 2017); for Tigrinya, which is not supported by FastText, manual verification was conducted. To further refine the dataset, we filtered out texts containing abusive language using a set of rule-based heuristics. Entries containing URLs or emojis were also excluded to maintain textual clarity and relevance. Finally, the cleaned text was tokenized, segmented into sentences, and further divided into individual words.

## Annotation

Each text entry in the dataset was annotated by five coders, all of them are from Ethiopia who are profeccent in the language, 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.).

## Annotation Guidline

Guidelines for Rating Educational Value of the Content. It comprises six categories: None, Minimal, Basic, Good, Excellent, and 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 A 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 A 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.

| **Hyperparameter** | **Value** |
|---|---|
| Maximum sequence length | 256 |
| Batch size | 32 |
| Number of training epochs | 10 |
| Learning rate | 5e-5 |
| Learning rate schedule | Constant |
| MLM probability | 0.15 |
| Weight decay | 0.01 |
| Optimizer | Adam |
| Adam $\epsilon$ | $1 \times 10^{-8}$ |
| Adam $\beta_1$ | 0.9 |
| Adam $\beta_2$ | 0.999 |
| Mixed precision (fp16) | True |

Table 3: Hyperparameter settings used for further pretraining with morpheme-aware tokenization.

# B Appendex B

## Model and training parameters

**Training Details** To better align the model's embeddings and internal representations with the linguistic characteristics of the target language, we employ Language Adaptive Pretraining (LAP), which continues training a multilingual model on monolingual data from the target language using the masked language modeling objective.

All expanded and initialized XLM-R models are trained following the hyperparameters in (Dobler and de Melo, 2023), employing the masked language modeling (MLM) objective with a mask probability of 15%. We use the Adam optimizer (Kingma and Ba, 2014) with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1 \times 10^{-8}$, and an initial learning rate of $5 \times 10^{-5}$ with a constant learning rate schedule. Unlike the original FOCUS approach, where batch size may vary, we fix the batch size at 32 to better utilize available hardware resources. Additionally, we set the maximum sequence length to 256 tokens and train for 10 epochs, allowing the model to capture better morpheme-level tokenized inputs, which tend to be shorter but more granular. We apply a weight decay of 0.01 to help regularize the training and enable mixed-precision training (fp16), which improves computational efficiency without sacrificing model stability.