Multilingual Encoder Knows more than You Realize: Shared Weights Pretraining for Extremely Low-Resource Languages

Anonymous ACL submission

Abstract

While multilingual language models like XLM-R have advanced multilingualism in NLP, they still perform poorly in extremely low-resource languages. This situation is exacerbated by the fact that modern LLMs such as LLaMA and Qwen support far fewer languages than XLM-R, making text generation models non-existent for many languages in the world. To tackle this challenge, we propose a novel framework for adapting multilingual encoders to text generation in extremely low-resource languages. By reusing the weights between the encoder and the decoder, our framework allows the model to leverage the learned semantic space of the encoder, enabling efficient learning and effective generalization in low-resource languages. Applying this framework to four Chinese minority languages, we present XLM-SWCM, and demonstrate its superior performance on various downstream tasks even when compared with much larger models.

1 Introduction

In recent years, with the development of multilingual pretrained models such as XLM-R (Conneau et al., 2020), mBART (Liu et al., 2020), and mT5 (Xue et al., 2021), language models have achieved significant progress in multilingual tasks, especially for high-resource languages. However, low-resource languages like Tibetan, Uyghur, Kazakh, and Mongolian—spoken by millions of people in China—remain critically underserved. Among these languages, Tibetan has over 10 million speakers, Uyghur over 11 million, Kazakh approximately 3 million, and Mongolian around 7 million, yet their representation in existing multilingual corpora is vastly inadequate. As illustrated in Figure 1, there is a significant disparity between the population sizes of these languages and the amount of available data in popular multilingual corpora such as OSCAR (Jansen et al., 2022). The situation

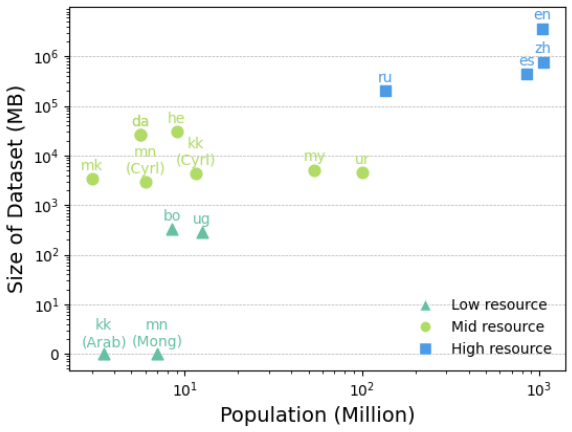

Figure 1: The relationship between population size and dataset size in OSCAR (y-axis, in MB) for various high-, middle-, and low-resource languages.

is especially dire for Kazakh and Mongolian, with virtually zero usable data, hindering their inclusion in mainstream multilingual models.

Despite claims of multilingual support for hundreds of languages, models like mBART and mT5 are not trained on Chinese minority languages. In comparison, more advanced multiglingual large language models such as LLaMA (Touvron et al., 2023) and Qwen (Yang et al., 2024) support even fewer languages.

This gap underscores the need for targeted solutions to address the challenges of text generation in extremely low-resource languages. To tackle this challenge, we propose a novel framework for efficiently extending a multilingual encoder into an encoder-decoder architecture. To address the scarce training data in low-resource languages, we introduce a weight-sharing mechanism between the encoder and the decoder by interleaving weights transferred from the encoder with randomly initialized ones, allowing for efficient adaptation to text generation in low-resource settings.

Extensive experiments on the aforementioned

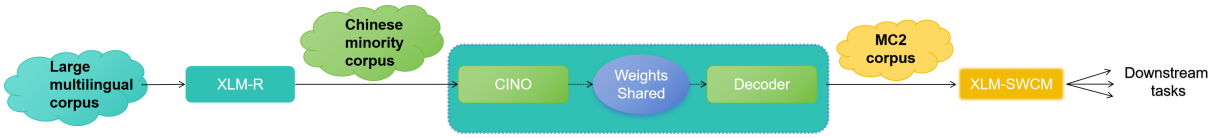

Figure 2: An overview of the shared weight framework for efficiently adapting multilingual encoders to text generation in low-resource languages.

four Chinese minority languages demonstrate the convincing advantages of our proposed method, with both faster convergence, better generalization, and strong cross-lingual transfer capabilities. Our model, **XLM-SWCM** (XLM-Shared Weight for Chinese Minorities), outperforms an mBART baseline by up to 199% on text summarization, 108% on reading comprehension, and also bests the much larger MC2-LLaMA 13B (Zhang et al., 2024b) in cross-lingual transfer settings.

In summary, the main contributions of this paper are:

1) a weight-sharing framework for efficiently adapting multilingual encoders to text generation in low-resource languages;

2) a model XLM-SWCM trained with this method for multiple Chinese minority languages;

3) extensive experiments showcasing the superior performance of XLM-SWCM compared with similar-sized baselines and much larger LLMs, confirming the feasibility of our framework.

Our code and models will be released upon publication.

2 Related Works

2.1 Multilingual Corpus

The evolution of multilingual large language models (LLMs) has been enabled by the release of extensive multilingual corpora such as CC100, mC4, OSCAR, CulturaX, and Madlad-400 (Wenzek et al., 2020; Raffel et al., 2019; Jansen et al., 2022; Nguyen et al., 2024; Kudugunta et al., 2023). While these resources cover a selection of low-resource languages to some extend, there remains a recognized gap in the representation for China's minority languages, primarily due to significant differences in writing systems.

China's minority languages often use different writing systems from the same language family used elsewhere in the world. For example, Uyghur is primarily written in the Arabic script (UEY—Uyghurche Ereb Yëziqi) in China, with the Latin script (ULY—Uyghurche Latin Yëziqi) used as a supplementary form. In contrast, Uyghur in Russia and Central Asia is written in the Cyrillic script (USY—Uyghurche Shilir Yëziqi). When collecting data for minority languages, the aforementioned multilingual corpora either do not distinguish between such different writing systems, or only contain data from one system, as shown in Figure 1.

Recently, the release of the Multilingual Corpus of Minority Languages in China (MC2, Zhang et al., 2024b) breaks the gap in the availability of Chinese minority language pretraining corpora, covering four underrepresented languages: Tibetan, Uyghur, Kazakh, and Mongolian. This dataset is used as the primary pretraining corpus in our work.

2.2 Development of Multilingual Language Models

In the past few years, multilingual variants of pretrained language models have been proposed in the NLP community, such as mBART (Liu et al., 2020) and mT5 (Xue et al., 2021), supporting up to 100 languages and demonstrating powerful cross-lingual transfer capabilities. More recently, the emergence of large language models (LLMs) has revolutionized multilingual natural language processing. Models like PaLM (Chowdhery et al., 2023) and BLOOM (Scao et al., 2022) have made significant strides in multilingual capabilities, while the LLaMA family (Touvron et al., 2023) and its multilingual variants have democratized access to multilingual LLMs. Some specialized models represented by XGLM and NLLB (Lin et al., 2022; Costa-jussà et al., 2022) have focused on expanding language coverage and improving cross-lingual transfer capabilities across hundreds of low-resource languages. However, few of these models support Chinese minority languages.

2.3 NLP for Minority Languages in China

To enhance the accessibility of minority languages in China, prior studies have primarily focused on curating annotated datasets for various NLP tasks. These efforts have mainly concentrated on three

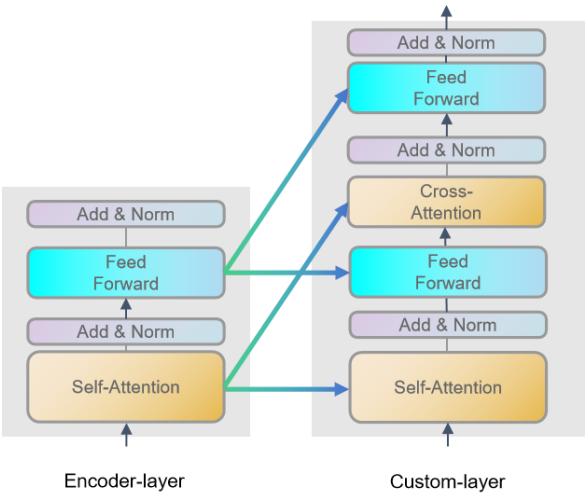

Figure 3: The weight initialization schemes for the CustomDecoderLayer. The colored arrows indicate the initialization of weights between the different components.

key task categories: text classification (Qun et al., 2017; Sun et al., 2021a; Shi et al., 2023), question answering (Sun et al., 2021b), and machine translation (Zhang et al., 2024a). Prominent models specifically trained for these languages include CINO (Yang et al., 2022), MiLMo (Deng et al., 2023), and TiBert (Liu et al., 2022). However, despite such progress, none of these models have released their pre-training corpora, and there is still a notable gap in the availability of models capable of text generation in these languages.

3 Method

3.1 Adapting Encoders to Text Generation

3.1.1 Framework Overview

In this section, we introduce the Shared Weights Framework, which leverages shared weights between the encoder and decoder for efficiently adapting multilingual encoders to text generation in low-resource languages.

The overall pipeline is visually summarized in Figure 2. Starting from CINO (Yang et al., 2022), a continual-pretrained version of XLM-R for Chinese minority languages, we copy its weight to initialize the decoder layers for knowledge transfer, and tie some of the weights between encoder and dedocer to enable efficient training. This model, which we name XLM-SWCM, is pretrained on the MC2 corpus and then applied to downstream tasks, including both single-language finetuning and cross-lingual transfer.

3.1.2 Model Architecture

Like the vanilla Transformer, the proposed model has two main components:

Encoder: a pre-trained encoder-only model, specifically CINO, a variant of XLM-R enhanced for Chinese minority languages.

Decoder: a transformer decoder stack with a specialized weight transfer mechanism. To balance the knowledge acquired during the encoder's large-scale pretraining and new knowledge required for downstream generation tasks, we introduce two types of decoder layers: NormalDeocderLayer and CustomDecoderLayer, both maintaining the same hidden dimension, intermediate size, and number of attention heads as the encoder.

NormalDecoderLayer: A standard transformer decoder layer with randomly initialized weights. It follows a conventional architecture with sequential self-attention, cross-attention, and feed-forward network. These layers enable the model to learn generation-specific features from scratch, complementing the knowledge transfered from the encoder.

CustomDecoderLayer: A modified transformer decoder layer that inherits pre-trained weights from the encoder. It features an enhanced structure with two strategically positioned feed-forward networks: FFN1 between self-attention and cross-attention, and FFN2 following cross-attention, each with its own layer normalization and residual connection, as shown in Figure 3. CustomDecoderLayer inherits all its weights from the pre-trained encoder to reuse learned representations.

3.1.3 Weight Sharing Mechanism

In our framework, the pre-trained encoder consists of only self-attention and feed-forward blocks, while the decoder layers require both self-attention and cross-attention mechanisms for effective generation. Thus, special schemes are designed to initialize and reuse the weights, as shown in Figure 3.

For weight initialization of CustomDecoderLayers, weights of both self-attention and cross-attention in the decoder are initialized from the encoder's self-attention blocks. Similarly, weights of both two FFN blocks in a decoder layer are initialized from the FFN block in the corresponding encoder layer. This mechanism reduces the effective number of parameters to be learned, accelerating convergence and enabling effective transfer of

linguistic knowledge from the pre-trained encoder while maintaining model stability.

A key architectural decision in our framework is the insertion pattern of these layers. After every X CustomDecoderLayers, we insert one NormalDecoderLayer, so that an encoder with n layers would correspond to a decoder with $n + \lfloor n/X \rfloor$ layers. The value of X significantly impacts the model's generalization capabilities, and its optimal value varies across different model scales. Through extensive experimentation, we find that $X = 3$ yields the best performance, and a detailed analysis of how this choice affects the model's performance is discussed in Section 5.2.3.

3.2 Pretraining

3.2.1 Pretraining Tasks

We adopte a multi-task training approach for pretraining. The primary task involves self-supervised learning using mBART's **denoising auto-encoding (DAE)** strategy. This strategy helps with the model's transition from the encoder's word-level cloze tasks to sequence generation tasks by predicting the masked portions of the input sequence with a decoder.

Additionally, we incorporate **machine translation** as an auxiliary objective, particularly focusing on translation between Mandarin Chinese and various Chinese minority languages. Specifically, the training data includes bidirectional translation pairs between Mandarin Chinese and the minority languages. This auxiliary objective improves the model's cross-lingual transfer capability, thereby enhancing the model's performance in various low-resource language processing tasks.

3.2.2 Training Data

THUCNews (THU-NLP Group, 2016) is a Chinese news dataset, derived from historical data from the Sina News RSS feed between 2005 and 2011 and containing approximately 740,000 news articles. From this dataset, we extracted a subset of Simplified Chinese news articles.

MC2 (Zhang et al., 2024b) provides multilingual data for several Chinese minority languages, including Tibetan, Uyghur, Kazakh, and Mongolian. The specific data volumes are described in detail in Appendix A. Together with THUCNews, these monolingual datasets serve as training data for the DAE task.

For machine translation, we leveraged Google Translate to create bidirectional translation pairs between Chinese and the minority languages (Tibetan, Uyghur, Kazakh, and Mongolian). These translations were verified by native speakers to ensure accuracy. A total of 2,000 sentence pairs from each language pair were selected to form the supplementary training data.

Combining these three corpora, the integrated dataset allows the model to effectively handle both high-resource and low-resource languages, improving its cross-lingual transfer and multilingual capabilities.

4 Experiments

4.1 Pretraining

Training Configuration The models are trained for 8 epochs with a peak learning rate of 1e-4, AdamW (Loshchilov and Hutter, 2019) optimizer, global batch size 600, and a linear learning rate scheduler with a warmup proportion of 0.1. The maximum sequence length is set to 256 tokens, and mixed-precision is enabled to optimize memory usage and training efficiency. To ensure training stability, the norms of gradients are clipped to 1.0. The models are trained on two NVIDIA A800 GPUs, each with 80GB of memory, and the training process takes 92 hours.

Balanced Sampling Strategy To address the inherent data imbalance across different languages, we implemente a balanced sampling strategy similar to XLM-R. The sampling probability for each language is calculated as

$$p_i = \frac{q_i^\alpha}{\sum_j q_j^\alpha}, \tag{1}$$

where q_i represents the original proportion of language i in the dataset, and α (set to 0.3) is a smoothing parameter that balances between uniform sampling and size-proportional sampling. This approach ensures that low-resource languages receive adequate representation in the training process while maintaining the influence of larger datasets.

Model Adaptations We extende the model's vocabulary with special language tokens (<bo>, <kk>, <mn>, <ug>, <zh>) to handle our target languages (Tibetan, Kazakh, Mongolian, Uyghur, and Chinese). These language identifiers are directly added after the bos token <s> in the model inputs. This modification ensures that the model can effectively process and distinguish between different languages during both pre-training and downstream

task finetuning. The same approach is consistently applied in all subsequent experiments.

Based on the aforementioned settings, we trained a new seq2seq model - XLM-SWCM, utilizing CINO-base-v2 as the encoder, with 457 million parameters. The detailed architectural configuration is provided in Appendix B.

4.2 Downstream Tasks

4.2.1 Experiment Setting

To evaluate the capabilities of XLM-SWCM, we conduct fine-tuning experiments on three downstream tasks in both low-resource and high-resource languages: Text Summarization, Machine Reading Comprehension (MRC), and Machine Translation. These tasks are chosen to cover diverse areas of text generation in NLP.

Single-Language Fine-tuning Due to the scarcity of labeled data for low-resource languages, we focus primarily on Tibetan for single-language fine-tuning, which has several publicly available datasets:

- **Text Summarization**: For this task, we utilize the Ti-Sum dataset (Xiaodong, 2022) with 20,000 pairs of titles and articles.

- **MRC**: We mainly use the TibetanQA dataset (Sun et al., 2022) for this task, which claims to contain 20K examples. However, only 2K examples are publicly available. Thus we enrich it by integrating 5K examples from the TibetanSFT Corpus[1] and 3K examples translated from a Chinese MRC dataset (Cui et al., 2019a) using Google Translate. This approach enables us to create a comprehensive dataset consisting of 10K examples.

- **Machine Translation**: For Machine Translation, we also use the TibetanSFT Corpus, which is cleaned to generate 50,000 parallel Chinese-Tibetan sentence pairs.

Cross-lingual Transfer In addition to single-language fine-tuning, we also conduct cross-lingual transfer experiments to test XLM-SWCM's ability to generalize across multiple low-resource languages. This experiment aims to assess the model's performance in Tibetan, Uyghur, Mongolian, and Kazakh after being fine-tuned on a high-resource language (Simplified Chinese) and a very small number of samples in the target languages.

- **Text Summarization**: For Mandarin Chinese, we use the publicly available LCSTS dataset (Hu et al., 2015), which contains 100K samples scraped from various Chinese portals. For the four minority languages, approximately 3K cleaned samples per language are scraped from language-specific news portals, using the news titles as their summarization.

- **MRC**: For Chinese, we employ the CMRC 2018 dataset (Cui et al., 2019b), which consists of 10K samples. For Tibetan, we use 500 samples extracted from the publicly available TibetanQA dataset. For the other three minority languages (Uyghur, Mongolian, Kazakh), we utilize machine translation tools to translate and clean MRC data, ultimately selecting 500 samples per language.

Baseline Models We employ two baseline models to ensure broad coverage and robust performance in handling Chinese minority languages. The first model builds upon LLaMA2-Chinese and is fine-tuned on the MC2 dataset, resulting in the *MC2-LLaMA-13B* model. The second baseline, referred to as *mBART-CM*, is an adaptation of mBART-cc25. Its vocabulary is expanded to include tokens specific to our minority languages, followed by further pretraining on MC2.

Training settings Both XLM-SWCM and mBART-CM are sequence-to-sequence models that are fine-tuned using standard training configurations. Each of these models is trained for 50 epochs with a batch size of 200 samples to ensure comprehensive learning and optimal performance. MC2-LLaMA-13B model is trained using LoRA (Hu et al., 2022) with a rank of 8 for 3 epochs.

4.2.2 Experimental Results

As illustrated in Table 1, XLM-SWCM consistently outperforms the baseline models across all three tasks. Despite having fewer parameters, XLM-SWCM demonstrates a substantial margin of superiority over mBART-CM and even surpasses the much larger MC2-LLaMA-13B.

Notably, XLM-SWCM achieves an impressive **198.8% improvement in F1-score for Text Summarization** over mBART-CM, along with a significant **107.6% F1 improvement in MRC**. These remarkable gains are a direct result of XLM-SWCM's efficient weight sharing framework to maximize the utilization of pre-trained encoder features in resource-constrained scenarios. Even under equivalent seq2seq structures and identical training corpora, XLM-SWCM demonstrates greater

[1] https://huggingface.co/datasets/shajiu/ParallelCorpusSFT

Model	Size	Sum			MRC			MT		
		F	P	R	F	P	R	F	P	R
MC2-LLaMA-13B	13B	16.1	12.3	15.5	13.2	11.7	13.1	15.1	12.2	16.8
mBART-CM	611M	8.6	11.2	15.2	7.9	6.1	5.6	11.5	7.3	9.3
XLM-SWCM (ours)	457M	**25.7**	**29.1**	**24.2**	**16.4**	**29.5**	**16.2**	**24.5**	**26.3**	**24.3**

Table 1: Performance metrics of the baseline models, evaluated using three ROUGE-L sub metrics: **F** (F1-score), **P** (precision), and **R** (recall). Size refers to the number of parameters in each model.

Model	Zh		Bo		Ug		Mn		Kk	
	Sum	MRC	Sum	MRC	Sum	MRC	Sum	MRC	Sum	MRC
MC2-LLaMA-13B	47.1	43.5	9.5	6.1	3.5	2.4	3.7	2.2	2.6	3.9
MC2-LLaMA-13B*	**47.3**	**44.7**	13.1	**11.5**	11.7	10.1	9.7	**10.2**	2.9	4.6
mBART-CM	32.7	25.6	6.8	2.1	2.7	2.2	3.1	1.7	0.2	0.1
XLM-SWCM (ours)	33.1	23.5	**17.1**	11.1	**12.5**	**11.1**	**13.5**	7.2	**5.6**	**6.9**

Table 2: Cross-lingual Transfer performance of different models on Text Summarization (Sum) and Machine Reading Comprehension (MRC) tasks, evaluated using ROUGE-L. The best results for each task are highlighted. * indicates explicitly prompting MC2-LLaMA-13B with the language to be used in the response during evaluation.

efficiency and learning capacity.

In comparison to MC2-LLaMA-13B, which benefits from richer pretraining corpora and larger-scale parameters, XLM-SWCM achieves a **59% higher F1-score in Text Summarization**, a **24.1% F1 improvement in MRC**, and a **62.3% higher F1-score in MT**. These results underscore the effectiveness of XLM-SWCM's shared weight framework in resource-constrained environments, making it a superior choice for tasks involving Chinese minority languages.

Table 2 highlights the performance of XLM-SWCM and baseline models in cross-lingual transfer settings. For the primary source language (Zh), the baseline models demonstrate better performance, which stems from their larger parameter sizes and more extensive pretraining corpora in Simplified Chinese. However, when it comes to **generalization to minority languages**, XLM-SWCM showcases exceptional adaptability, significantly outperforming the baseline models. mBART-CM, for instance, struggles to distinguish between languages and often defaults to outputs in the primary language (Zh), even when language-specific labels are present. Similarly, MC2-LLaMA-13B exhibits language-related errors, though its performance improves when explicitly informed of the current language type, as seen with MC2-LLaMA-13B*.

In Text Summarization, XLM-SWCM outperforms all baselines. Specifically, XLM-SWCM achieves significant improvements of **30.5%, 6.8%, and 39.1%** for Tibetan (Bo), Uyghur (Ug), and Mongolian (Mn) respectively over MC2-LLaMA-13B*, the best-performing baseline. For MRC, XLM-SWCM also demonstrates competitive performance across most languages, being only slightly weaker than MC2-LLaMA-13B* for Tibetan and Mongolian.

Overall, these experiments indicate that XLM-SWCM can effectively leverage the shared weight mechanism to maximally reuse the semantic space of the pre-trained encoder, demonstrating excellent performance in Chinese minority language applications with limited data and parameter size.

5 Ablation Studies

In this section, we present a series of ablation experiments aimed at evaluating the impact of key components in our framework that play essential roles in enhancing the model's multilingual capabilities and improving its generalization to low-resource languages. We perform ablation experiments on the Tibetan finetuning tasks, maintaining a consistent finetuning setting with Section 4.2.1.

Removing Module	Sum	MRC	MT
None (XLM-SWCM)	**25.7**	**16.4**	**24.5**
MT	25.6	15.1	20.3
DAE	22.4	12.2	18.7
WS	17.1	11.7	18.2
MT + DAE	22.5	12.3	17.7
MT + WS	17.5	11.3	18.4
DAE + WS	15.2	11.9	17.1
MT + DAE + WS	15.9	10.8	16.5

Table 3: Objective ablation results, evaluated using ROUGE-L. The experiments involve removing different combinations of training components, such as Machine Translation (MT), DAE (Denoising Auto-Encoding), and Weight Sharing (WS).

5.1 Objective Ablation

We first focus on three critical aspects of the model: DAE pretraining, machine translation, and weight initialization by removing each and combinations of them. The results are shown in Table 3. Removing any of the three components is detreimental to performance, specifically:

- Machine Translation (MT): Removing machine translation has a relatively small impact on performance across tasks, as shown by both individual removal (maintaining 25.6 in Sum) and combined removals (MT+DAE vs DAE showing similar scores);

- Denoising Auto-Encoding (DAE): The removal of DAE pretraining causes considerable performance drops across all three downstream tasks, and its impact becomes more pronounced in combined removals (DAE+WS), indicating its fundamental importance in establishing the model's basic text generation capabilities.

- Weight Sharing (WS): The removal of weight sharing demonstrates the most significant impact among all modules, showing the largest performance drops in individual removal and maintaining this substantial negative effect across all combined removal scenarios, establishing it as the most crucial component for the model's effectiveness in low-resource settings.

In short, while all three components contribute positively to the model's performance, weight sharing emerges as the most critical component. This finding highlights the importance of weight sharing as a key architectural choice for multilingual models, especially in resource-constrained scenarios.

5.2 Structure Ablation

We also perform experiments to evaluate the impact of different structural components in our proposed framework. These experiments aim to understand how the initialization of decoder weights and the insertion of normal layers affect model performance.

5.2.1 Impact of Weight Initialization

Firstly, we train a baseline model called **Cino-Transformer**. Unlike XLM-SWCM, the decoder of this model is randomly initialized, and also matches the number of encoder layers. The model is pretrained using the same DAE and MT tasks as XLM-SWCM but without weight sharing, and then finetuned on downstream tasks in the same setting as XLM-SWCM.

Model	Sum	MRC	MT
Cino-Transformer	18.9	13.5	18.7
XLM-SWCM (ours)	**25.7**	**16.4**	**24.5**

Table 4: Performance metrics of the Ablation of Weight Initialization, evaluated using the ROUGE-L metric.

Model	Sum	MRC	MT
BASE-A	13.7	10.3	15.7
BASE-B	16.3	14.1	21.1
XLM-SWCM (ours)	**25.7**	**16.4**	**24.5**

Table 5: Performance metrics of the Ablation of Normal Layers, evaluated using the ROUGE-L metric. **BASE-A** has fewer layers and does not include any normal layers, while **BASE-B** maintains the same number of layers as XLM-SWCM but uses weight duplication instead of normal layers.

The results in Table 4 demonstrate the effectiveness of our weight initialization scheme. By transferring weights from the encoder to the decoder, XLM-SWCM can be efficiently adapted to text generation with limited training data, outperforming Cino-Transformer on all tasks.

5.2.2 Impact of Randomly Initialized Layers

Secondly, we explore the impact of inserting normal layers among the custom layers in the decoder. To assess the effectiveness of this modification, we use two baseline models for comparison:

- **Baseline A (XLM-SWCM without normal layers)**: This model is identical to XLM-SWCM but without any normal layers inserted into the custom layer architecture. The absence of normal layers leads to a reduced total number of layers in the decoder.

- **Baseline B (Weight duplication model)**: Instead of inserting normal layers, this model simply copies the weights of the preceding layer to maintain consistency in the number of model parameters. This results in identical weights across consecutive layers, allowing us to isolate the impact of inserting randomly initialized normal layers.

The results in Table 5 demonstrate the significant impact of inserting normal layers into the decoder. BASE-A, which has fewer layers, performs the worst across all tasks. BASE-B, which maintain the same number of layers as XLM-SWCM but lacks randomly initialized weights, shows some improvement but still underperforms.

Overall, these findings indicate that randomly initialized normal layers is also crucial for adapting encoders to text generation.

5.2.3 Impact of Insertion Frequency of Normal Layers

Thirdly, we thoroughly investigate the impact of insertion frequency of normal layers in the decoder, and how this interacts with varying dataset sizes. This experiment is designed along two dimensions:

- **Insertion Frequency of Normal Layers**: we explore values of X where a normal layer is inserted after every X custom layers, with X ranging from 1 to 6. All these models are pretrained in the same setting as XLM-SWCM.

- **Effect of Finetuning Dataset Size**: we evaluate the model's performance on datasets of varying sizes, including 10K, 20K, and 50K samples. As the existing Ti-SUM dataset only has 20K samples, we supplement it by crawling and cleaning 30K additional news articles from various major Chinese websites. This dimension allows us to examine the interaction between the amount of available data and the frequency of normal layers.

The results are plotted in Figure 4:

- For the small dataset (10k), larger X results in better performance, as smaller decoders generalize more effectively when data is limited. In contrast, smaller X (i.e. larger decoders) leads to overfitting.

- For the medium dataset (20k), performance peaks at $X = 3$. This indicates that a moderate decoder size strikes a balance between capacity and data availability.

- For the large dataset (50k), smaller X achieve the highest F1-scores, as the larger decoder capacity enables the model to fully exploit the larger dataset.

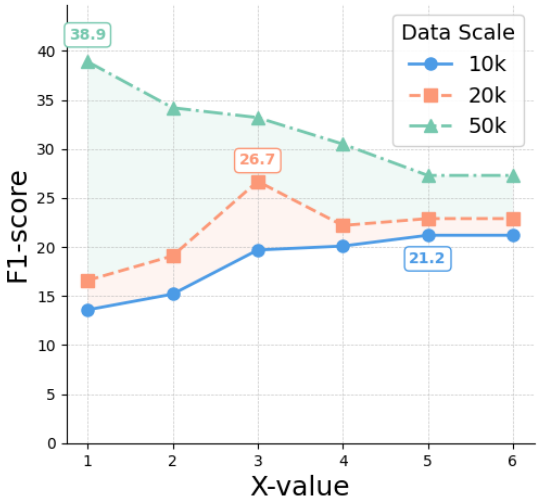

Figure 4: ROUGE-L scores on Tibetan summarization for different X-values (insertion frequency of normal layers). The three lines correspond to different dataset sizes.

Overall, these results demonstrate the flexibility of our framework, where the insertion frequency of normal layers can be adjusted based on the task-specific dataset size. Larger X (fewer layers) is better suited for small datasets, while smaller X (more layers) performs best on larger datasets.

6 Conclusion

In this work, we proposed a novel pretraining framework tailored for low-resource languages, with a particular focus on Chinese minority languages. Our framework leverages a shared weight mechanism between the encoder and decoder, which allows for the efficient adaptation of multilingual encoders to generation tasks without the need to start from scratch. Experimental results demonstrate that our model XLM-SWCM significantly outperforms traditional baselines on various text generation tasks for Tibetan, Uyghur, Kazakh, and Mongolian, which have long been underserved in NLP research. Our approach opens up new possibilities for developing robust models for these extremely low-resource languages, and also provides a promising method for the integration of resources across similar languages.

7 Limitations

Due to the availability of pretrained language models for Chinese minority languages and high-quality corpora, our study focused on only four minority languages. Our single-language finetuning experiments are further constrained to Tibetan given the lack of relevant datasets, limiting the scope of our exploration.

Thus, we hope that future work will put more focus on the development of high-quality datasets in these minority languages and beyond, enabling a more thorough exploration of underrepresented languages in the LLM era.

References

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1171–1179.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113.

Alexis Conneau, Kartikay 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, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Y. Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2019a. A span-extraction dataset for Chinese machine reading comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5886–5891, Hong Kong, China. Association for Computational Linguistics.

Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2019b. A span-extraction dataset for Chinese machine reading comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5886–5891, Hong Kong, China. Association for Computational Linguistics.

Junjie Deng, Hanru Shi, Xinhe Yu, Wugedele Bao, Yuan Sun, and Xiaobing Zhao. 2023. Milmo: Minority multilingual pre-trained language model. In *IEEE International Conference on Systems, Man, and Cybernetics, SMC 2023, Honolulu, Oahu, HI, USA, October 1-4, 2023*, pages 329–334. IEEE.

Baotian Hu, Qingcai Chen, and Fangze Zhu. 2015. LCSTS: A large scale chinese short text summarization dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1967–1972. The Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Tim Jansen, Yangling Tong, Victoria Zevallos, and Pedro Ortiz Suarez. 2022. Perplexed by Quality: A Perplexity-based Method for Adult and Harmful Content Detection in Multilingual Heterogeneous Web Data. *arXiv e-prints*, arXiv:2212.10440.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine

Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: A multilingual and document-level large audited dataset. *Preprint*, arXiv:2309.04662.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona T. Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 9019–9052. Association for Computational Linguistics.

Sisi Liu, Junjie Deng, Yuan Sun, and Xiaobing Zhao. 2022. Tibert: Tibetan pre-trained language model. In *IEEE International Conference on Systems, Man, and Cybernetics, SMC 2022, Prague, Czech Republic, October 9-12, 2022*, pages 2956–2961. IEEE.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Trans. Assoc. Comput. Linguistics*, 8:726–742.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2024. CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4226–4237, Torino, Italia. ELRA and ICCL.

Nuo Qun, Xing Li, Xipeng Qiu, and Xuanjing Huang. 2017. End-to-end neural text classification for tibetan. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data - 16th China National Conference, CCL 2017, - and - 5th International Symposium, NLP-NABD 2017, Nanjing, China, October 13-15, 2017, Proceedings*, volume 10565 of *Lecture Notes in Computer Science*, pages 472–480. Springer.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Yuan Sun, Zhengcuo Dan, Sisi Liu, and et al. 2022. Tibetanqa: Tibetan dataset for machine reading comprehension[ds/ol]. v3. CSTR:31253.11.sciencedb.j00001.00351. Science Data Bank, accessed 2025-01-02.

Yuan Sun, Sisi Liu, Chaofan Chen, Zhengcuo Dan, and Xiaobing Zhao. 2021a. Teaching machines to read and comprehend tibetan text. *Journal of Computer and Communications*, 9(09):143–152.

Yuan Sun, Sisi Liu, Chaofan Chen, Zhengcuo Dan, and Xiaobing Zhao. 2021b. Teaching machines to read and comprehend tibetan text. *Journal of Computer and Communications*, 9(09):143–152.

THU-NLP Group. 2016. THUCNews: A Chinese News Dataset from Sina News RSS (2005-2011). Technical report, Tsinghua University.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu,

Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Yan Xiaodong. 2022. Ti-sum[ds/ol]. v3. CSTR:31253.11.sciencedb.j00001.00352. Science Data Bank, accessed 2025-01-02.

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, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 483–498. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Ziqing Yang, Zihang Xu, Yiming Cui, Baoxin Wang, Min Lin, Dayong Wu, and Zhigang Chen. 2022. CINO: A chinese minority pre-trained language model. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 3937–3949. International Committee on Computational Linguistics.

Chen Zhang, Xiao Liu, Jiuheng Lin, and Yansong Feng. 2024a. Teaching large language models an unseen language on the fly. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 8783–8800. Association for Computational Linguistics.

Chen Zhang, Mingxu Tao, Quzhe Huang, Jiuheng Lin, Zhibin Chen, and Yansong Feng. 2024b. Mc^2: Towards transparent and culturally-aware NLP for minority languages in china. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 8832–8850. Association for Computational Linguistics.

A Dataset Details

For pretraining of XLM-SWCM and other baseline models, we used a combination of Simplified Chinese data from THUCNews and minority languages from MC2. The breakdown of their distribution is given in Table 6.

Language	Data Size	Number of Samples
Tibetan	2.2 GB	184,045
Uyghur	736 MB	90,441
Kazakh	937 MB	57,827
Mongolian	970 MB	171,847
Simplified Chinese	2.1 GB	836,075

Table 6: Statistics of our pretraining dataset.

B Training Details

In addition to the settings presented in the main paper, here we detail other parameters used during pre-training XLM-SWCM for complete reproduction:

Hardware and Software Configuration

- **Hardware:** NVIDIA Tesla A800 GPU, 80 GB RAM * 2, Intel i7 CPU.
- **Software:** Ubuntu 20.04, CUDA 11.7, PyTorch 2.3

Training Configurations

- **Total Training Samples:** 1,340,235
- **Local Batch Size:** 75
- **Gradient Accumulation Steps:** 4
- **Global Batch Size:** 600
- **Epochs:** 8
- **Total Training Steps:** 17,864
- **Optimizer:** AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$
- **Learning Rate:** 1e-4
- **Warm-up:** Linear warm-up for the first epoch, gradually increasing the learning rate from 1e-5 to 1e-4.
- **Scheduled Sampling:** In the first epoch, teacher forcing is applied to guide the model. Subsequently, the teacher forcing ratio is gradually decreased in a linear fashion, transitioning to scheduled sampling (Bengio et al., 2015).