CUTE: A Multilingual Dataset for Enhancing Cross-Lingual Knowledge Transfer in Low-Resource Languages

Wenhao Zhuang[1,2], **Yuan Sun**[1,2,*]
[1]Minzu University of China, Beijing, China
[2]National Language Resource Monitoring & Research Center Minority Languages Branch
Emails: sdrz_zwh@163.com, sunyuan@muc.edu.cn
* Corresponding author: Yuan Sun

Abstract

Large Language Models (LLMs) demonstrate exceptional zero-shot capabilities in various NLP tasks, significantly enhancing user experience and efficiency. However, this advantage is primarily limited to resource-rich languages. For the diverse array of low-resource languages, support remains inadequate, with the scarcity of training corpora considered the primary cause. We construct and open-source CUTE (**C**hinese, **U**yghur, **T**ibetan, **E**nglish) dataset, consisting of two 25GB sets of four-language corpora (one parallel and one non-parallel), obtained through machine translation. CUTE encompasses two resource-rich languages (Chinese and English) and two low-resource languages (Uyghur and Tibetan). Prior to constructing CUTE, human assessment validates that the machine translation quality between Chinese-Uyghur and Chinese-Tibetan approaches that of Chinese-English translation. CUTE represents the largest open-source corpus for Uyghur and Tibetan languages to date, and we demonstrate its effectiveness in enhancing LLMs' ability to process low-resource languages while investigating the role of corpus parallelism in cross-lingual transfer learning. The CUTE corpus and related models are made publicly available to the research community[1].

1 Introduction

The current LLMs demonstrate remarkable capabilities in resource-rich languages. However, their performance is limited for numerous resource-poor languages (Ebrahimi et al., 2021; Chowdhery et al., 2023). Even powerful multilingual models such as XLM-R (Conneau et al., 2019), mT5 (Xue et al., 2020), and NLLB (Costa-jussà et al., 2022) support only approximately 100-200 languages, leaving nearly 7,000 low-resource languages untapped (van Esch et al., 2022). Among these are several low-resource languages with significant numbers of speakers. Uyghur and Tibetan, two low-resource minority languages in China, have over 13 million and 8 million speakers respectively. However, known LLMs have yet to achieve adequate support for these two languages.

Existing multilingual datasets, such as OSCAR (Abadji et al., 2022) and CulturaX (Nguyen et al., 2024), include Uyghur and Tibetan languages. The CC100 dataset (Conneau et al., 2019), used for training XLM-R, also contains a small amount of Uyghur text. However, these datasets still exhibit several limitations, including insufficient data volume, misidentification of languages, and imbalanced data distribution (Zhang et al., 2024b). MC^2 represents the largest open-source multilingual corpus of Chinese ethnic minority languages to date. It comprises crawled, integrated, and cleaned data from existing minority language sources, including Uyghur and Tibetan (Zhang et al., 2024b). Nevertheless, the scale of this dataset remains relatively small, not exceeding 3GB in size.

To enhance LLMs' ability to process low-resource languages, continued pre-training or adding adapters are common approaches (Yong et al., 2022; Zhang et al., 2024b; Cahyawijaya et al., 2023; Jin et al., 2022). However, continued pre-training typically requires substantial unlabeled text for learning language representations, while question-answer pairs in low-resource languages for fine-tuning are even more challenging to obtain. Existing low-resource corpora are often insufficient to effectively update LLM parameters (Cahyawijaya et al., 2024). To rapidly address the scale issue of low-resource corpora, one solution involves using machine translation models to translate training corpora from resource-rich languages like English and Chinese into low-resource languages. However, this approach raises two primary concerns: (1) the accuracy of the translation process may not be guaranteed (Ebing and Glavas, 2023). (2) cultural

[1]https://github.com/CMLI-NLP/CUTE

Proceedings of the 31st International Conference on Computational Linguistics, pages 10037–10046
January 19–24, 2025.

nuances inherent in the languages may be lost or erroneously propagated during translation (Zhang et al., 2024b; Liu et al., 2023).

LLMs demonstrate strong comprehension and instruction-following capabilities across multiple high-resource languages. This multilingual proficiency largely relies on cross-lingual sentence embeddings. Specifically, cross-lingual sentence embeddings encode multilingual text into a unified semantic representation space, where sentences with similar meanings in different languages are mapped to proximate vector locations (Conneau et al., 2019; Devlin et al., 2019; Lample and Conneau, 2019). However, significant representational disparities exist between cross-lingual word representations of low-resource languages and those of high-resource languages in current LLMs (Miao et al., 2024). Given the word-level representational differences, sentence-level cross-lingual representation alignment faces even more severe challenges. Achieving semantic representation alignment between low-resource and high-resource languages would provide more opportunities for transferring knowledge from high-resource languages to low-resource languages.Parallel corpora play a crucial role as bridges in transfer learning (Pham et al., 2024). However, accurately assessing their impact on LLMs and determining whether parallel corpora can facilitate more effective cross-lingual knowledge transfer requires a parallel corpus of sufficient scale and quality.

To advance research and development of Uyghur and Tibetan in LLMs, validate the reliability of machine translation in generating low-resource data, and explore the impact of parallel corpora in knowledge transfer from high-resource to low-resource languages, this paper introduces the CUTE (**C**hinese, **U**yghur, **T**ibetan, **E**nglish) dataset. It comprises two equal-sized four-language corpora: one parallel in content and the other non-parallel, totaling approximately 50GB. The Uyghur and Tibetan components are ten times larger than the current largest open-source MC^2 dataset.

The CUTE dataset addresses the common issues present in the aforementioned datasets that include Uyghur and Tibetan languages. Notably, the CUTE dataset offers several improvements. First, it boasts a significantly larger scale. Additionally, the use of machine translation in CUTE eliminates the problem of language misidentification. Furthermore, the dataset features a more balanced distribution of data across languages and content domains. The mutual parallelism among the four languages in CUTE also provides expanded opportunities for research in machine translation and cross-lingual knowledge transfer.

In summary, we make the following contributions:

- We construct and open-source the CUTE dataset, a large-scale multilingual corpus with parallel and non-parallel data in two high-resource languages (Chinese and English) and two low-resource languages (Uyghur and Tibetan), facilitating LLM training and evaluation for minority languages.
- We propose a novel approach for LLM training using parallel and non-parallel corpora through vocabulary expansion and embedding initialization. Using this method, we develop and release two versions of CUTE-Llama trained on different corpus types.
- We validate the effectiveness of parallel corpora in cross-lingual knowledge transfer through zero-shot experiments on various downstream tasks, showing that parallel data enables more effective knowledge transfer from high-resource to low-resource languages.

2 Related Works

Low-resource Language Corpora In recent years, several large-scale multilingual corpora have emerged to support NLP tasks for low-resource languages. Datasets such as OSCAR (Abadji et al., 2022), CulturaX (Nguyen et al., 2024), and MADLAD-400 (Kudugunta et al., 2023) provide rich cross-lingual text for multilingual model training through web crawling and multi-source integration. The ROOTS dataset emphasizes openness and traceability, implementing strict management in data collection and cleaning for low-resource languages (Laurençon et al., 2022). OPUS, as an open-source parallel corpus, offers extensive bilingual data for machine translation task (Tiedemann, 2012)s. Meta's FLORES-200 and NLLB projects focus on evaluating and improving translation performance for low-resource languages, covering 200 languages and significantly advancing cross-lingual knowledge transfer research (Costa-jussà et al., 2022). Although these corpora demonstrate excellent performance in processing high-resource languages, they still face challenges such as insufficient data volume, language misidentification, and

imbalanced content distribution for low-resource languages (e.g., Uyghur, Tibetan). Future research directions should prioritize increasing data collection for low-resource languages and improving the quality of existing corpora.

NLP Development for Low-Resource Languages NLP research for low-resource languages has made some progress in recent years but still faces numerous challenges. In terms of datasets, research primarily focuses on tasks such as text classification (Qun et al., 2017; Yang et al., 2022; Deng et al., 2023), machine reading comprehension (Sun et al., 2021), instruction following (Zhuang et al., 2024a), and machine translation (Zhang et al., 2024a). However, these datasets are typically limited in scale and cover a narrow range of languages. Regarding models, pre-trained models specifically developed for Chinese minority languages, such as CINO (Yang et al., 2022), MiLMo (Deng et al., 2023), CMPT (Li et al., 2022), and TiLamb (Zhuang et al., 2024b), have achieved certain breakthroughs in processing languages like Uyghur and Tibetan through techniques including multilingual pre-training and vocabulary expansion. Nevertheless, the pre-training corpora for these models are generally not publicly available, which limits the reproducibility and further development of research. Existing LLMs contribute minimally to improving the processing capabilities for low-resource languages, primarily due to the lack of high-quality instruction data. Knowledge distillation methods from teacher models prove ineffective for these languages, as current LLMs already perform inadequately in low-resource languages such as Uyghur and Tibetan. Moreover, finding suitable annotators capable of writing high-quality instruction samples is challenging due to the high requirements for creative thinking and professional expertise (Li et al., 2024). These factors collectively constrain the development of NLP for low-resource languages in China, highlighting urgent research needs in data collection, model optimization, and cross-lingual knowledge transfer.

Cross-lingual Knowledge Transfer Optimizing cross-lingual knowledge transfer is a key strategy for addressing NLP challenges in low-resource languages. Multilingual pre-trained models such as XLM-R (Conneau et al., 2019) and mBERT (Devlin et al., 2019) currently serve as powerful tools for effective cross-lingual transfer, yet they still face challenges when applied to low-resource languages. In recent years, ICL (In-Context Learning) (Brown et al., 2020) and few shot learning have shown potential to adapt large language models to new tasks. However, their effectiveness in helping models understand under-trained low-resource languages remains limited. To further enhance the efficacy of cross-lingual knowledge transfer, this study focuses on two critical directions: evaluating the effectiveness of machine-translated low-resource language data in model training, and investigating the impact of parallel data on cross-lingual knowledge transfer.

3 CUTE Dataset and CUTE-Llama

This section provides a detailed description of the scale of the CUTE dataset, which encompasses Chinese, Uyghur, Tibetan, and English. It represents the largest open-source dataset for Uyghur and Tibetan languages in China to date. It is important to note that Chinese and English, as resource-rich languages, are well-supported in the majority of LLMs. In contrast, Uyghur and Tibetan, as low-resource languages, often exhibit suboptimal performance in nearly all LLMs.

Additionally, this section elucidates the construction and training process of CUTE-Llama, as well as the model's evolving adaptability to Uyghur and Tibetan languages during the training phase.

3.1 Machine Translation for Chinese Minority Languages

Low-resource languages exhibit a significant disparity in data acquisition methods and scale compared to resource-rich languages. One viable approach to build large datasets for low-resource languages is using machine translation. This method converts existing training corpora from resource-rich languages into low-resource language data. However, ensuring translation quality becomes a critical consideration in this process.

The combined number of Uyghur and Tibetan language users in China exceeds 20 million. Neural machine translation technologies between various minority languages, with Chinese as the pivot, have developed over an extended period and have now reached maturity. Given that bidirectional translation between Chinese and English, both resource-rich languages, has attained a considerable level of reliability, we have established scoring criteria to evaluate the quality of machine translation from Chinese to Uyghur and Tibetan. Native speakers

of these respective minority languages have been invited to conduct the assessment.

Evaluation Criteria We have established unified translation standards with Chinese as the source language and Uyghur, Tibetan, and English as the target languages, as shown in Table 1.

Human Evaluation We randomly select 500 Chinese texts from the SkyPile-150B dataset (Wei et al., 2023) and generate corresponding parallel sentence pairs in Uyghur, Tibetan, and English through machine translation, with 500 pairs for each language. For Chinese-English translation, we employ the Google Translate system, while specialized machine translation models are used for Chinese-Uyghur and Chinese-Tibetan translations.

To evaluate the translation quality, we invite three Uyghur and three Tibetan graduate students as evaluators. These evaluators are native speakers of their respective ethnicities and are also proficient in Chinese. For the assessment of Chinese-English translations, we recruit three graduate students with excellent English communication skills. After standardizing the translation scoring criteria, they independently complete their respective translation scoring tasks. The final human evaluation results for Chinese translations into Uyghur, Tibetan, and English are shown in Figure 1.

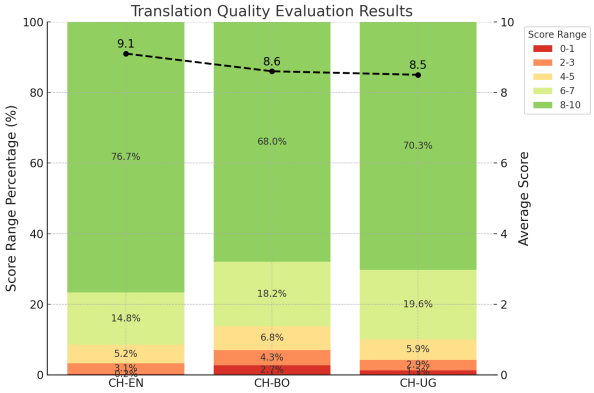

Figure 1: Human evaluation results of Chinese to English, Tibetan, and Uyghur translations. The stacked bar chart displays the distribution of translation quality scores across five score ranges (0-1, 2-3, 4-5, 6-7, 8-10) for each language pair, while the black dashed line represents the average scores.

Results Analysis The Chinese-English (CH-EN) translation performs best, achieving an average score of 9.1, with 76.7% of translations falling within the 8-10 score range, demonstrating high accuracy and fluency. Notably, the quality of Chinese-Uyghur (CH-UG) and Chinese-Tibetan (CH-BO) translations closely approaches that of Chinese-English, with average scores of 8.5 and 8.6 respectively, also attaining the level of "generally accurate and naturally fluent." All three language pairs exhibit a pronounced right-skewed distribution, with over 90% of translations scoring above 6, indicating that the vast majority of translations accurately convey the overall meaning of the original text. These results highlight the balanced high-level performance of current machine translation systems in handling translations between Chinese and English, Uyghur, and Tibetan. In particular, the quality of Chinese-Uyghur and Chinese-Tibetan translations now approaches that of Chinese-English translation.

3.2 CUTE Dataset Language Distribution

The CUTE dataset utilizes machine translation to translate a small portion of the SkyPile-150B dataset into Uyghur, Tibetan, and English. SkyPile-150B is a dataset specifically designed for pre-training large-scale Chinese language models, containing approximately 150 billion tokens primarily sourced from a wide range of Chinese internet web content. This dataset undergoes rigorous deduplication and sensitive information filtering to ensure data quality and safety.

CUTE comprises two sets of corpora, each containing four languages. The first set consists of parallel corpora in four languages, achieving a 99.98% similarity in content parallelism. The second set includes non-parallel corpora in four languages, with the English portion identical to the first set, while the remaining three languages differ. The specific scale of CUTE is presented in Table 2.

	CUTE-P		CUTE-NP	
Lang.	**Lines**	**Size**	**Lines**	**Size**
ZH	933,946	2.62	1,000,609	2.64
EN	933,989	3.49	933,989	3.49
UG	934,002	7.37	1,010,381	7.77
BO	934,140	11.22	989,723	11.90
Total	3,736,077	24.70	3,934,702	25.80

Table 2: Distribution of CUTE dataset. CUTE-P: parallel corpus, CUTE-NP: non-parallel corpus. Language codes: ZH (Chinese), EN (English), UG (Uyghur), BO (Tibetan). Size in GB.

Score	Description
10	Perfect translation, accurate and fluent, fully consistent with the original style and meaning.
8–9	Generally accurate, natural and fluent, with only minor errors or improprieties.
6–7	Overall meaning is understandable, but with noticeable errors or awkward phrasing that affect partial comprehension.
4–5	Partially understandable, but with serious errors that impact overall comprehension.
2–3	Mostly incomprehensible, with only a few words correctly translated.
0–1	Completely incomprehensible or unrelated to the original text.

Table 1: Machine Translation Quality Evaluation Criteria: Universal Scoring Standards for Chinese-Uyghur, Chinese-Tibetan, and Chinese-English Translation Directions in the CUTE Dataset

3.3 Analysis of Document Length Distribution

Figure 2 illustrates the document length distribution for four languages (Chinese, Uyghur, Tibetan, and English) in the CUTE dataset. By analyzing the document lengths for each language, we observe significant variations in the average document length across languages. Uyghur exhibits the highest average document length at 1,094.23 tokens, substantially exceeding the other three languages. In contrast, English, Tibetan, and Chinese demonstrate relatively similar average document lengths of 955.36, 883.01, and 879.28 tokens, respectively.

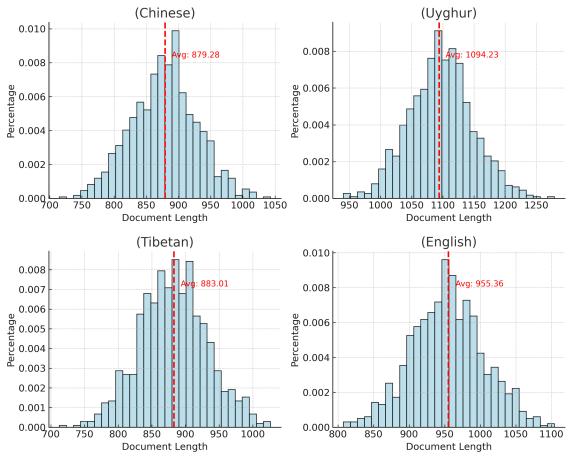

Figure 2: Document length distribution across Chinese, Uyghur, Tibetan, and English in the CUTE dataset, based on token counts calculated using the CUTE-Llama tokenizer. The red dashed lines represent the average document length for each language.

3.4 CUTE-Llama Vocabulary Training

CUTE-Llama is based on the Llama2 model architecture (Touvron et al., 2023). The original tokenization model of Llama2 is trained using the SentencePiece (Kudo and Richardson, 2018) library, employing the Byte Pair Encoding (BPE) algorithm. This algorithm constructs the vocabulary by merging common byte pairs, enabling effective processing of various languages. Considering that Llama2 is well-adapted for English but lacks sufficient support for Chinese, Uyghur, and Tibetan, we train separate vocabularies of 6,000 tokens each for these three languages. The training content, while distinct from the CUTE dataset, is of comparable scale. These newly trained vocabularies are subsequently merged with the Llama2 vocabulary. The parameters for vocabulary training are presented in Table 3.

Parameter	Value
vocab_size	6,000
model_type	bpe
split_digits	True
max_sentence_length	10,000
byte_fallback	True

Table 3: SentencePiece training parameters for Chinese, Tibetan, and Uyghur vocabularies. Vocabulary sizes after merging with the original Llama model: 36,820 (Chinese), 42,353 (Tibetan), and 47,905 (Uyghur) tokens, from an initial 32,000.

3.5 CUTE-Llama Training

We trained the CUTE-Llama model using key hyperparameters as shown in Table 4. The training process was conducted on 8 NVIDIA H800 GPUs for approximately 18 hours to obtain one CUTE-Llama model.

3.6 Perplexity Analysis Across Training Stages

To evaluate the model's performance across different languages throughout the training process, we

Hyperparameter	Value
Max Sequence Length	4,096
Batch Size	256
Learning Rate	1e-4
Warmup Steps	100
Epoch	1
Data Type	BF16

Table 4: Key hyperparameters for CUTE-Llama training.

Stage	Tibetan	Uyghur	Chinese	English
Original	3.24	5.24	5.23	7.24
Post-Exp	50,633	16,317	822	7.25
CUTE-P	12.00	5.50	9.93	4.75
CUTE-NP	11.84	5.40	10.41	4.67

Table 5: Perplexity scores across training stages. Stages: Original (initial Llama2), Post-Exp (after vocabulary expansion), CUTE-P (after parallel corpus training), CUTE-NP (after non-parallel corpus training).

conducted a perplexity (PPL) analysis using 1,000 samples for each language that were not included in the training set. Table 5 presents the perplexity scores at various stages of model development.

The results reveal an intriguing phenomenon in the original Llama2 model, where Tibetan, Uyghur, and Chinese show lower perplexity than English. This counterintuitive outcome likely stems from the model's treatment of unfamiliar scripts as sequences of unknown tokens, leading to simplified character-level predictions. The vocabulary expansion initially causes a sharp increase in perplexity for these languages, reflecting the model's adjustment to the new token distribution. Subsequent training on both parallel and non-parallel corpora significantly improves performance across all languages, resulting in a more balanced multilingual model. This process demonstrates the effectiveness of our approach in adapting Llama2 to handle Tibetan, Uyghur, and Chinese while maintaining its English capabilities.

4 Evaluation Tasks and Results

To evaluate the practical value of the CUTE dataset and investigate the impact of multilingual parallel corpora on cross-lingual knowledge transfer, we construct two foundation models based on the Llama2-7B architecture: CUTE-Llama-P (trained with parallel corpora) and CUTE-Llama-NP (trained with non-parallel corpora). The vocabularies of both models are expanded to include Chinese, Uyghur, and Tibetan, with the embeddings of newly added tokens initialized using mean values. Subsequently, we conduct continuous pre-training on these models using parallel and non-parallel corpora from CUTE, respectively.

Our experimental design is as follows: We first fine-tune these two foundation models using downstream task data from resource-rich languages, then assess the cross-lingual zero-shot transfer learning capabilities of the fine-tuned models on low-resource languages. This process aims to validate the efficacy of the CUTE dataset and compare the differences between parallel and non-parallel corpora in facilitating cross-lingual knowledge transfer.

4.1 Evaluation Datasets

Public test sets for Uyghur and Tibetan are limited in quantity and narrow in domain coverage. To address this issue, we identify corresponding datasets with similar training tasks in resource-rich languages. **WCM-v2** (Yang et al., 2022) is a multilingual dataset containing 10 categories of text classification tasks, including classification tasks for Chinese, Uyghur, and Tibetan. As the training set of WCM-v2 only contains Chinese data, we fine-tune the model using the Chinese training set and evaluate it on test sets in Chinese, Uyghur, and Tibetan. **TibetanQA** (Sun et al., 2021) is a Tibetan machine reading comprehension dataset, primarily used to assess model performance on extractive reading comprehension tasks. **CMRC** (Cui et al., 2019), as a Chinese machine reading comprehension dataset, shares the same task type as TibetanQA. Therefore, we fine-tune CUTE-Llama using CMRC and utilize TibetanQA to test the model's transfer ability from Chinese to Tibetan. **SQuAD** (Rajpurkar, 2016) is a widely used English machine reading comprehension dataset containing question-answer pairs from Wikipedia articles. We fine-tune CUTE-Llama using SQuAD and evaluate its transfer ability from English to Tibetan through TibetanQA. The Chinese relation extraction dataset released by Baidu, which we refer to as **Baidu-KG** for convenience, encompasses type definitions for 50 relation extraction tasks. Based on this, we construct a Tibetan relation extraction dataset, which we name **Tibetan-KG**, containing 11 relation types to test the model's performance. The **Flores-200** (Costa-jussà et al., 2022) dataset includes machine translation tasks for Tibetan and

Uyghur. We employ a few-shot prompting approach to complete this task.

4.2 Compared Models

We compare the CUTE-Llama model with CINO (Yang et al., 2022), Llama2-7B (Touvron et al., 2023), BLOOM7.1B (BigScience et al., 2022), and Llama3.1-8 (Dubey et al., 2024). CINO is a pre-trained model for ethnic minority languages in China, incorporating training data from Chinese, Uyghur, and Tibetan languages. BLOOM-7.1B is pre-trained on more than 45 languages, while both Llama2-7B and Llama3.1-8B are multilingual models developed by Meta.

4.3 Experimental Setup and Results

We compare and analyze the experimental results for text classification, relation extraction, machine reading comprehension, and translation in this section.

Text Classification We evaluate the models' performance on the WCM-v2 dataset, which includes Chinese (zh), Tibetan (bo), and Uyghur (ug) languages. The dataset comprises 32,000 Chinese samples for training. For testing, we use 4,000 Chinese, 1,110 Tibetan, and 300 Uyghur samples. Models are fine-tuned on the Chinese training data and tested on all three languages to assess zero-shot transfer capabilities. Table 6 presents the classification results.

Relation Extraction For the relation extraction task, we use Baidu-KG (194,747 samples) as the training set and our self-constructed Tibetan-KG (3,510 samples) as the test set. This setup evaluates the models' ability to transfer knowledge from Chinese to Tibetan in relation extraction. Table 7 shows the results.

Model	Precision	Recall	F1
Llama2-7B	0.2379	0.1338	0.1614
BLOOM-7.1b	0.2707	0.1435	0.1760
Llama3.1-8B	0.2781	0.1712	0.1982
CUTE-Llama-NP	0.7038	0.4006	0.4718
CUTE-Llama-P	**0.7312**	**0.4118**	**0.4843**

Table 7: Performance comparison on the Tibetan-KG relation extraction task.

Machine Reading Comprehension We evaluate the models' performance on the TibetanQA dataset to assess their machine reading comprehension capabilities in Tibetan. The models are fine-tuned on Chinese (CMRC) or English (SQuAD) datasets and tested on TibetanQA to measure cross-lingual transfer. Table 8 presents the results using Exact Match (EM) and F1 scores.

Model	CMRC-trained		SQuAD-trained	
	EM	F1	EM	F1
Llama2-7B	0.0767	0.6646	0.0612	0.6103
BLOOM-7.1b	0.0568	0.6513	0.0437	0.5924
Llama3.1-8B	0.0065	0.4859	0.0041	0.4213
CUTE-Llama-NP	0.1455	0.7927	0.1187	0.7352
CUTE-Llama-P	**0.1674**	**0.8071**	**0.1346**	**0.7489**

Table 8: Performance comparison on the TibetanQA machine reading comprehension task, with models trained on CMRC (Chinese) and SQuAD (English) datasets.

Translation For the translation task, we evaluate the models' performance on Chinese-to-Tibetan (zh-bo) and Chinese-to-Uyghur (zh-ug) translation using the Flores-200 dataset. We employ few-shot prompting with 3 examples for each language pair. Table 9 presents the results using BLEU, chrF, and TER (Translation Edit Rate) metrics.

4.4 Analysis of Results

The experimental results demonstrate the significant potential of the CUTE dataset. In the text classification task, the CUTE-Llama-P model exhibits exceptional cross-lingual zero-shot transfer capabilities, with particularly noteworthy performance in Tibetan and Uyghur languages. Compared to Llama3.1-8B, our model shows accuracy improvements of 13.87 and 18.67 percentage points for these two languages, respectively. Even more promising is the model's performance in machine reading comprehension tasks, where CUTE-Llama-P excels even in cross-family language transfer from English to Tibetan. In translation tasks, the performance of CUTE-Llama models further corroborates the importance of parallel corpora. Even in few-shot scenarios, models trained on parallel corpora consistently outperform their counterparts trained on non-parallel data in Chinese-to-Tibetan and Chinese-to-Uyghur translations, with significant improvements in BLEU scores. These results highlight the high quality and practical value of the CUTE dataset while emphasizing the crucial role of parallel corpora in enhancing cross-lingual transfer learning effectiveness.

Model	Classification (Accuracy / F1)			Average	
	zh	bo	ug	Minorities	All
CINO-base	- / 78.0	- / 36.2	- / 33.4	- / 34.8	- / 47.6
CINO-large	- / 79.2	- / 40.6	- / 28.8	- / 34.7	- / 48.4
Llama2-7B	90.0 / 90.02	26.13 / 25.43	78.0 / 82.42	52.07 / 53.93	76.23 / 76.34
BLOOM-7.1b	90.025 / 89.99	25.23 / 27.10	35.67 / 49.86	30.45 / 38.48	73.72 / 74.86
Llama3.1-8B	90.225 / 90.13	37.21 / 39.85	68.33 / 77.92	52.77 / 58.89	78.13 / 79.14
CUTE-Llama-NP	90.1 / 89.98	49.91 / 48.44	86.33 / 87.97	68.12 / 68.21	81.65 / 81.34
CUTE-Llama-P	**90.25 / 90.17**	**51.08 / 48.46**	**87.0 / 89.08**	**69.04 / 68.77**	**82.03 / 81.56**

Table 6: Performance comparison of different models on the WCM-v2 dataset for text classification tasks. The best scores are in **bold**, with the second best underlined. For CINO models, only F1 scores are available. The Minorities average for CINO models is calculated as the mean of bo and ug F1 scores, while for other models it's the average of both Accuracy and F1 scores for bo and ug.

Model	Chinese-to-Tibetan (zh-bo)			Chinese-to-Uyghur (zh-ug)		
	BLEU↑	chrF↑	TER↓	BLEU↑	chrF↑	TER↓
BLOOM-7.1b	4.2	0.297	0.881	4.9	0.319	0.862
Llama2-7B	4.7	0.311	0.865	5.4	0.334	0.847
Llama3.1-8B	6.8	0.364	0.811	7.5	0.376	0.798
CUTE-Llama-NP	8.3	0.401	0.773	9.0	0.419	0.762
CUTE-Llama-P	**9.5**	**0.427**	**0.745**	**10.2**	**0.443**	**0.738**

Table 9: Translation performance comparison on Flores-200 dataset using few-shot prompting (3 examples). ↑: higher is better, ↓: lower is better. The best scores are in **bold**, with the second best underlined.

5 Conclusion

This study constructs and open-sources the CUTE dataset, providing the largest open resource to date for Uyghur and Tibetan NLP research. The CUTE-Llama model developed based on this dataset demonstrates excellent multilingual processing capabilities, particularly excelling in Uyghur and Tibetan tasks. The experimental results not only validate the effectiveness of machine translation in generating training data for low-resource languages but also highlight the crucial role of parallel corpora in facilitating cross-lingual knowledge transfer. The public release of the CUTE dataset and CUTE-Llama model opens up new possibilities for NLP research and applications in China's minority languages.

Limitations

The CUTE dataset provides rich resources for low-resource language research; however, its large scale inevitably leads to some errors in the translation process, especially in sentences with complex grammar or significant cultural differences. The heavy reliance on machine translation may also result in the loss of cultural-specific expressions and linguistic features unique to Uyghur and Tibetan languages. While the selection of data emphasizes diversity and balance, coverage of certain domains may still be inadequate, limiting the model's performance in specific fields. Furthermore, although CUTE-Llama demonstrates outstanding performance in handling low-resource language tasks, its performance in more complex language understanding tasks (such as deep reasoning or generation tasks) still requires further evaluation and optimization.

Acknowledgements

This work is supported by the National Social Science Foundation (22&ZD035), the National Nature Science Foundation (61972436), and the Minzu University of China Foundation (GRSCP202316, 2023QNYL22, 2024GJYY43).

References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a cleaner document-

oriented multilingual crawled corpus. *ArXiv*, abs/2201.06642.

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

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 1877–1901.

Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024. Llms are few-shot in-context low-resource language learners. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 405–433.

Samuel Cahyawijaya, Holy Lovenia, Tiezheng Yu, Willy Chung, and Pascale Fung. 2023. Instructalign: High-and-low resource language alignment via continual crosslingual instruction tuning. In *Proceedings of the First Workshop in South East Asian Language Processing*, pages 55–78.

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

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. In *Annual Meeting of the 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, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2019. 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 5883–5889.

Junjie Deng, Hanru Shi, Xinhe Yu, Wugedele Bao, Yuan Sun, and Xiaobing Zhao. 2023. Milmo: minority multilingual pre-trained language model. In *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 329–334. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Benedikt Ebing and Goran Glavas. 2023. To translate or not to translate: A systematic investigation of translation-based cross-lingual transfer to low-resource languages. *ArXiv*, abs/2311.09404.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John E. Ortega, Ricardo Ramos, Annette Rios Gonzales, Ivan Vladimir, Gustavo A. Gim'enez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando A. Coto Solano, Ngoc Thang Vu, and Katharina Kann. 2021. Americasnli: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. *ArXiv*, abs/2104.08726.

Xisen Jin, Xiang Ren, Daniel Preotiuc-Pietro, and Pengxiang Cheng. 2022. Dataless knowledge fusion by merging weights of language models. *ArXiv*, abs/2212.09849.

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

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier García, 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. *ArXiv*, abs/2309.04662.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *ArXiv*, abs/1901.07291.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, et al. 2022. The bigscience roots corpus: A 1.6 tb composite multilingual dataset. *Advances in Neural Information Processing Systems*, 35:31809–31826.

Bin Li, Yixuan Weng, Bin Sun, and Shutao Li. 2022. A multi-tasking and multi-stage chinese minority pre-trained language model. In *CCMT*.

Chong Li, Wen Yang, Jiajun Zhang, Jinliang Lu, Shaonan Wang, and Chengqing Zong. 2024. X-instruction: Aligning language model in low-resource languages with self-curated cross-lingual instructions. *ArXiv*, abs/2405.19744.

Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2023. Are multilingual llms culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. *ArXiv*, abs/2309.08591.

Zhongtao Miao, Qiyu Wu, Kaiyan Zhao, Zilong Wu, and Yoshimasa Tsuruoka. 2024. Enhancing cross-lingual sentence embedding for low-resource languages with word alignment. *ArXiv*, abs/2404.02490.

Thut Nguyen, Chien Van Nguyen, Viet Dac Lai, Hiu Mn, 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.

Trinh Pham, Khoi M. Le, and Anh Tuan Luu. 2024. Unibridge: A unified approach to cross-lingual transfer learning for low-resource languages. In *Annual Meeting of the Association for Computational Linguistics*.

Nuo Qun, Xing Li, Xipeng Qiu, and Xuanjing Huang. 2017. End-to-end neural text classification for tibetan. In *China National Conference on Chinese Computational Linguistics*.

P Rajpurkar. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Y Sun, S Liu, C Chen, Z Dan, and X Zhao. 2021. Construction of high-quality tibetan dataset for machine reading comprehension. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 208–218.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

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

Daan van Esch, Tamar Lucassen, Sebastian Ruder, Isaac Caswell, and Clara Rivera. 2022. Writing system and speaker metadata for 2,800+ language varieties. In *International Conference on Language Resources and Evaluation*.

Tianwen Wei, Liang Zhao, Lichang Zhang, Bo Zhu, Lijie Wang, Haihua Yang, Biye Li, Cheng Cheng, Weiwei Lü, Rui Hu, Chenxia Li, Liu Yang, Xilin Luo, Xuejie Wu, Lunan Liu, Wenjun Cheng, Peng Cheng, Jianhao Zhang, Xiaoyu Zhang, Lei Lin, Xiaokun Wang, Yutuan Ma, Chuanhai Dong, Yanqi Sun, Yifu Chen, Yongyi Peng, Xiaojuan Liang, Shuicheng Yan, Han Fang, and Yahui Zhou. 2023. Skywork: A more open bilingual foundation model. *Preprint*, arXiv:2310.19341.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. In *North American Chapter of the Association for Computational Linguistics*.

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 *International Conference on Computational Linguistics*.

Zheng-Xin Yong, Hailey Schoelkopf, Niklas Muennighoff, Alham Fikri Aji, David Ifeoluwa Adelani, Khalid Almubarak, M Saiful Bari, Lintang Sutawika, Jungo Kasai, Ahmed Baruwa, Genta Indra Winata, Stella Biderman, Dragomir R. Radev, and Vassilina Nikoulina. 2022. Bloom+1: Adding language support to bloom for zero-shot prompting. *ArXiv*, abs/2212.09535.

Chen Zhang, Xiao Liu, Jiuheng Lin, and Yansong Feng. 2024a. Teaching large language models an unseen language on the fly. *ArXiv*, abs/2402.19167.

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. *arXiv preprint arXiv:2311.08348*.

Wenhao Zhuang, Dawa Cairen, and Yuan Sun. 2024a. Tifd: Tibetan instruction-following dataset for large language models supervised fine-tuning. *Data Intelligence*.

Wenhao Zhuang, Yuan Sun, and Xiaobing Zhao. 2024b. Tilamb (tilamb: A tibetan large language model based on incremental pre-training). In *Proceedings of the 23th Chinese National Conference on Computational Linguistics*.