UNMUTE: Understanding Multilingual Transfer Learning through Encipherment

Anonymous ACL submission

Abstract

Cross-lingual transfer learning has shown promise for low-resource translation, but its effectiveness for extremely low-resource languages, such as indigenous and ancient languages, remains under-explored. This limitation stems from a circular challenge: insufficient data and limited understanding of linguistic features and grammar prevent a thorough analysis, which in turn hinders the development of effective methods. This paper identifies key challenges in this domain and introduces a novel analysis technique, **UNMUTE** (**Un**derstanding **MU**ltilingual **T**ransferability through **E**ncipherment), which enciphers well-studied and high-resource text to simulate the challenges posed by extremely low-resource languages. Our framework enables us to systematically and precisely study factors such as training data amount and the proportion of unseen characters or (sub)words. Using UNMUTE, we investigate the techniques that enable and constrain effective transfer learning for extremely low-resource machine translation.

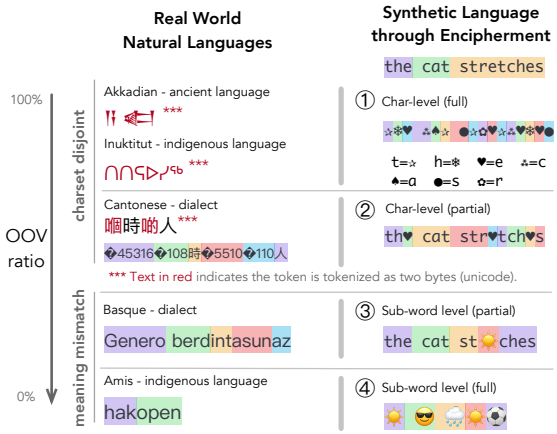

Figure 1: The left panel illustrates various OOV scenarios in low-resource languages, with colored regions showing subword tokenization from titoken[1]. The right panel demonstrates four encipherment methods. We use synthetic encipherment on high-resource languages to simulate challenges faced in low-resource languages, including **disjointed charsets** and **meaning mismatch**. The encipherment process breaks the tokenization of modern pre-trained models. For detail, check §2.1.

1 Introduction

Multilingual pre-trained models have demonstrated substantial benefits for NLP tasks in low-resource languages through cross-lingual transfer learning (NLLBTeam et al., 2024; Üstün et al., 2024). For instance, multilingual machine translation systems reportedly achieve BLEU scores of around 20 with only a few thousand parallel sentences. However, recent research challenges these results: Silva et al. (2024) reveals potential data contamination inflating performance metrics.

The true effectiveness of these models for extremely low-resource languages, which we define as those lacking both substantial monolingual data and meaningful subword unit overlap with high-resource languages, remains understudied. This gap in our understanding is significant: as Joshi et al. (2020) notes, fewer than 1% of the world's approximately 7,000 languages can be effectively processed using modern pre-training and fine-tuning pipelines.

Chen et al. (2025) demonstrate severe limitations in cross-lingual transfer learning for indigenous and ancient languages due to out-of-vocabulary (OOV) issues. As illustrated in Figure 1 (left), these OOV challenges manifest at different levels. In extreme cases, such as ancient and indigenous languages, the entire script is unknown to pre-trained models. In other cases, like the Basque and Amis languages, the challenges are more subtle—while the text can be tokenized, the resulting tokens may have nothing to do with their overlapping high-resource counterparts, making transfer ineffective and even potentially disadvantageous. We later refer to this phenomenon as **meaning mismatch**.

In this work, we analyze these OOV issues in

low-resource machine translation through synthetic encipherment. The characteristics of extremely low-resource languages differ substantially from modern languages, making it unclear whether performance gaps stem from corpus domain, sentence length, training data volume, or writing system differences. To address these challenges, we propose UNMUTE, a novel experimental framework using enciphered parallel sentences to systematically investigate translation challenges by disentangling these factors (Ebrahimi et al., 2024; Ojha et al., 2024; George et al., 2024).

From our analysis, we derive the following contributions:

1. We identified and categorized two major types of OOV issues and proposed a synthetic encipherment framework to systematically evaluate and understand the challenges for low-resource languages resulting from differences in writing systems. Our findings suggest that meaning mismatch, i.e. coincidentally overlapping tokens that have no real semantic correspondence, leads to similar performance drops as disjointed charset, which has not been directly demonstrated in the past. We also demonstrated that transliteration (romanization) of a non-Latin script performs similarly to enciphering with a substitution cipher.

2. We provided a lower bound on the data quantity requirement on extremely low-resource translation. Our results show that transfer learning becomes ineffective with datasets smaller than 100k tokens when there is no monolingual data or sibling languages available and dialect or indigenous languages requires at least 10k tokens. This finding, while intuitive, has important implications for the field: many languages falling into this category are currently overlooked in research and development efforts. We emphasize the need for increased attention to these truly low-resource languages, particularly in data collection and methodology development.

3. The Relative Drop Performance (RDP) on enciphered text serves as a robust metric to evaluate transferability across different pre-trained models. A well-performing transfer learning model or method should exhibit a smaller RDP on the **UNMUTE** dataset, which consists of enciphered text. This metric provides a reliable assessment of a model's ability to transfer knowledge from high-resource languages and writing systems to low-resource ones. We found that byte-level model such as byT5 has the overall smallest RDP and indicates its strong generalization ability on unseen languages.

In §2, we discussed the OOV issues of low-resource languages in detail; In §3, we described our synthetic encipherment approach. §4 described the experimental setup and results; in §5 we analyzed popular pre-trained models for machine translation in low-resource settings on our UNMUTE synthetic data. Finally, we also present two real-world cases in §6 to demonstrate the effectiveness of our framework.

2 Background and Related Work

2.1 Out-of-Vocabulary Scenarios in Low-Resource Languages

Recent large-scale models are trained on nearly all available internet text data, but for languages not represented online, we identify two major out-of-vocabulary (OOV) scenarios: **disjointed charsets** and **meaning mismatch** (see Figure 1).

Disjointed charsets occur when a language's character set has no overlap with the model's training data and cannot be easily normalized. This is common in ancient extinct languages (De Cao et al., 2024; Gutherz et al., 2023; Chen et al., 2024), where the writing system is completely different from modern high-resource scripts. For example, ancient languages like Sumerian and Egyptian hieroglyphs use logographic scripts that have no direct correspondence to modern alphabets, making it challenging to apply standard tokenization methods. Additionally, living languages such as Cantonese (a dialect whose speakers are more than 85 million) (Liu, 2022) and Inuktitut (indigenous languages) (Roest et al., 2020), also suffers from similar problem.

Meaning mismatch occurs when tokenization creates misleading semantic associations, especially in low-resource languages where limited training data hinders overcoming incorrect initial associations. For instance, the Amis word hawopen (meaning "enclose") is segmented by standard BPE tokenizers into haw and open, creating false connections to English semantics that can impede learning the true meaning of the word (Zheng et al., 2024).

This issue is particularly problematic for languages with morphological structures that differ significantly from high-resource languages like English.

2.2 Transfer Learning in Machine Translation

Zoph et al. (2016) first explored transfer learning for low-resource translation, considering languages with fewer than 1M tokens. Recent studies show that large language models (LLMs) do not outperform traditional MT systems on high-resource languages (Robinson et al., 2023). Various solutions have been explored to address the challenges posed by disjoint character sets in ancient language processing (De Cao et al., 2024), including byte-level encoding (Xue et al., 2022), vocabulary expansion on multilingual models (Liu et al., 2020; Xue et al., 2021), and pixel-based text representations (Salesky et al., 2023) to tackle fertility issues in cross-lingual transfer.

3 UNMUTE: Enciphering the Text

As discussed previously, we encipher of high-resource languages to systematically investigate cross-lingual transfer learning for extremely low-resource languages in machine translation. For simplicity and consistency, we only study translating into English direction and encipher the source language side only, the target side (English) is not enciphered.

3.1 Encipherment and Tokenization

In our UNMUTE framework, we first encipher the text and then apply tokenization. We use **character**-level encipherment to simulate disjoint charsets, and **subword**-level encipherment to simulate the meaning mismatch case. A simple flowchart to show how the process works is shown in Figure 2.

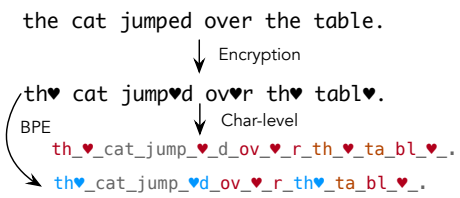

Figure 2: Visualization of how encipherment and tokenization work and the choice of tokenization after encipherment affects the number of new tokens. In this case we encipher one character e as ♡. If we directly use a character-level tokenizer, a common word such as the will be encoded as two tokens, but if we use BPE to re-train a new tokenizer, th♡ and ♡d (in blue) will be encoded as subword tokens.

I. Encipherment first. As shown in Figure 1, we consider two type of encipherment units: **char-level** and **subword-level**. For the encipherment ratios, i.e. how many percentages of the text are being enciphered, we select {0, .1, .5, .8, .9, 1} for controlled experiments. For example, if we choose encipher at the character level and the encipherment ratio is 0.5, we randomly select 50% of the characters and always encipher them for both training and testing. This is essentially equivalent to applying a 1:1 substitution cipher on 50% of the characters.

II. Re-tokenization after encipherment. As shown in Figure 2, while our primary experiments employ substitution cipher, which is a lossless transformation from a human perspective, the subsequent tokenization process introduces information loss when processed by neural models. Consequently, the choice of tokenization method **after** encipherment significantly affects downstream performance. We can process the text using either **character-level** or **BPE** tokenization. By default, we use the **character-level** tokenization, a further study can be found in §4.3.

III. Handling new tokens. In the previous step, we introduce new tokens to a pre-trained model, and we can handle these tokens in either of two approaches:

1. **(Default) vocabulary expansion:** We map subwords to newly registered tokens, which are initialized from a multivariate normal distribution that has existing embeddings' mean and covariance (Hewitt, 2021).

2. **In-place substitution:** We randomly map a new token to another existing token. For example, the subword *haw* of the Amis language map to the English subword *haw*. This approach parallels real-world scenarios such as language romanization.

IV. Special encipherment case. We also implement a special **transliteration encipherment**, which is analogous to the relationship between Linear B and Ancient Greek (Chadwick, 1990). We apply such transliteration encipherment to non-alphabetic writing systems such as Chinese. We report this experiment on Chinese in Section 6.2.

3.2 Choice of Languages

To construct synthetic data via encipherment that effectively simulates the challenges encountered in extremely low-resource language settings as described previously, we implement encipherment while accounting for multiple linguistic dimensions: language families, writing systems (Sproat and Gutkin, 2021), and typological features, following the approach of Chen et al. (2025).

Since a major challenge for low-resource languages in machine learning is their disjoint character sets, we selected representative languages spanning diverse phonographic categories and writing systems. We chose five high-resource languages: Chinese, Finnish, Japanese, Hindi, and Arabic. These languages represent distinct writing systems and typological features while providing sufficient data for our encipherment approach. Detailed statistics for each language are presented in Table 1.

Lang	Writing	Dataset	Phonography	sent len
Chinese	Han	wmt18-zh-en	Syllabic	4.6
Finnish	Latin	wmt18-fi-en	Alphabetic	17.6
Arabic	Arabic	iwslt2017-ar-en	Abjad	14.6
Japanese	Han/Kana	iwslt2017-ja-en	Moraic	4.6
Hindi	Devanagari	IITB-hi-en	Abugida	17.0

Table 1: Statistics of different languages we choose for experiment. The sentence length is counted according to the number of English words.

3.3 Models

For cross-lingual transfer learning in machine translation, we study five different models that fall into two broad categories:

1. **Unsupervised denoising training:** Models pre-trained using masked language modeling (MLM) or next word prediction objectives, including **mT5** and **mBART-25**.

2. **Aligned:** Models pre-trained on parallel data, and **mBART-50** falls into this category.

mT5 and mBART-25 These multilingual seq-to-seq models are pre-trained using only unsupervised multilingual data. The primary difference between them lies in their training data: mT5 (Xue et al., 2021) uses the **mc4** dataset, while mBART-25 uses **common-crawl-25 (cc25)**.

mBART-50 Building upon mBART-25, mBART-50 first continues pre-training on monolingual corpora of 50 languages using the same unsupervised approach. Crucially, mBART-50 (Tang et al., 2020) then undergoes additional training on **supervised** parallel data across these 50 languages.

3.4 Overfitting Remedies

We observed overfitting is a big bottleneck for a large portion of our experiments, especially when the training data size is less than 10k or the encipherment ratio is higher than 0.5. Therefore, we revisit different approaches including dropout, label smoothing, and LoRA to avoid overfitting.

Dropout. Dropout is the de facto technique to prevent overfitting (Srivastava et al., 2014), by randomly disabling parameters in contributing to the model outputs. We have conducted extensive experiments between dropout rates and learning rate scheduling on their effects on preventing overfitting to the smaller-size training corpus.

Label Smoothing. A regularization technique that prevents neural networks from becoming overconfident in their predictions by replacing hard target probabilities (like 0 or 1) with smoothed values (Szegedy et al., 2016).

LoRA. Low-rank adaptation techniques have become the go-to method for saving training memory in the LLM era (Hu et al., 2021). However, its benefits in preventing catastrophic forgetting and preserving general knowledge in the models have often been ignored (Biderman et al., 2024).

4 Experiments and Result

4.1 Experiment Setup

We conducted experiments on A100, A6000, L40 and H100 GPUs with an effective batch size of 64. Models were trained for 31,250 steps and evaluated every 500 steps. Early stopping with patience of 10 evaluation steps was used to prevent overfitting and select the best-performing model. We set the beam size to 5.

We use mBART's default settings, learning rate = 3e-5 and dropout = 0.3, label smoothing = 0.2 for 1k, 10k, and 100k training set size, we additionally run a dropout=0.1 and no label smoothing setting for experiments using 1M data. Our code is based on Huggingfaces's Transformers with DDP. The wall time for training one experiment settings under mBart-50 is about 30 mins with 4 x NVIDIA A100.

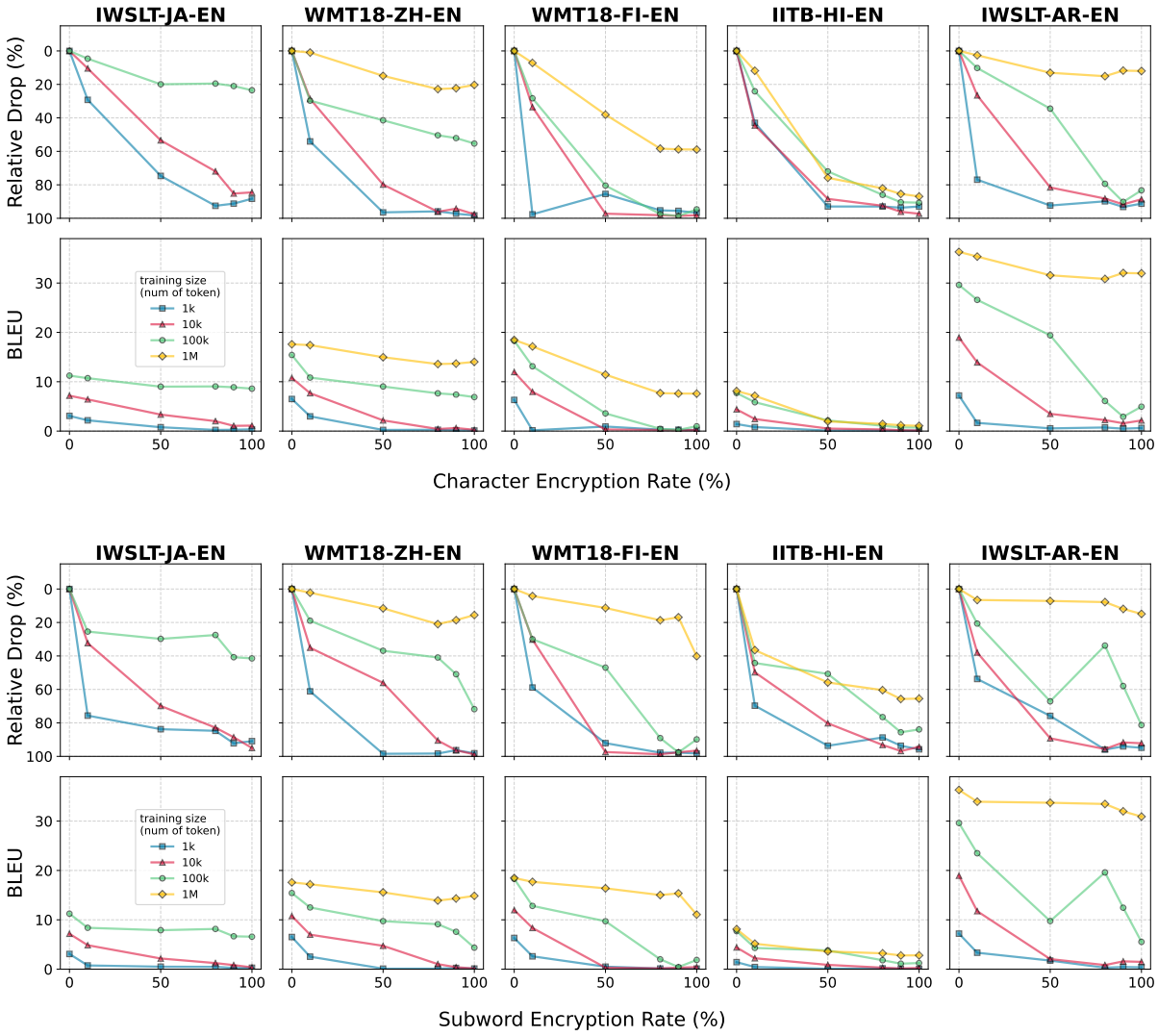

Figure 3: Relative performance drop of different languages when changing the encipherment ratios from 0% to 100%. (top) is character-level encipherment and (bottom) is subword-level encipherment. Different lines show trending of encipherment rate given different number of tokens {1k, 10k, 100k, 1M} in training.

4.2 Main Results and Discussion

The main results of our experiments are shown in Figure 3. We observed several key findings:

Non-linear performance degradation The relationship between the encipherment ratio and relative performance drop is not linear. Most language pairs show relatively modest degradation up to around 50% encipherment, followed by a sharp decline between 50-80%. This pattern suggests there may be a critical threshold of familiar tokens needed to maintain reasonable translation quality, beyond which performance rapidly deteriorates.

Similar performance degradation across OOV types At 100% encipherment ratio, both character- and subword-level encipherment show significant performance degradation compared to the 0% encipherment baseline. This supports our earlier claim that even when a new unknown language uses Latin script, transfer learning effectiveness is limited by insufficient data and lack of semantic overlap. The absence of shared semantic meaning between source and target languages fundamentally constrains word-level transferability.

The importance of training data size Our **UNMUTE** analysis reveals that model performance deteriorates rapidly when parallel data decreases from 1M to 100k tokens, with an even steeper decline at 10k tokens. This finding has two important implications: First, it empirically demonstrates that data collection remains the most effective solution

for improving machine translation quality in low-resource settings. Second, it suggests that claims of exceptional performance on very small datasets (around 2,000 sentences) without advanced techniques like back-translation or data augmentation should be scrutinized carefully (Silva et al., 2024; Chen et al., 2025). The fundamental challenges inherent to under-represented languages make achieving high-quality machine translation with such limited data highly improbable.

Language-specific sensitivity to encipherment Different language pairs exhibit varying levels of resilience to encipherment. For instance, ar-en maintains relatively better performance at high encipherment rates compared to fi-en, which shows a more dramatic drop. This suggests that the impact of OOV issues varies significantly across language pairs, possibly due to underlying linguistic similarities or differences in the base tokenization.

Impact of writing system characteristics Finnish, known to have the highest morphological complexity among non-logographic languages (Sproat and Gutkin, 2021), shows the largest gap between character and subword-level encipherment performance. In contrast, Chinese shows minimal impact from different encipherment methods, suggesting that highly logographic writing systems may not benefit from subword tokenization.

Counter-intuitive effects of partial overlap Surprisingly, more than half of the experiments group (a line plot in Figure 3) performance at 80-90% encipherment is better than at 100% encipherment. This suggests that having very few overlapping (sub)words may actually be more detrimental to adaptation or fine-tuning than having no overlap at all, possibly because minimal lexical overlap creates misleading linguistic signals.

4.3 Re-tokenization: Character or BPE?

Tokenization	FI	AR	HI	ZH
char	7.59	31.96	0.36	14.02
BPE (5000)	12.62	31.33	2.11	14.07

Table 2: Different re-tokenization approaches will highly affect the performance of enciphered data.

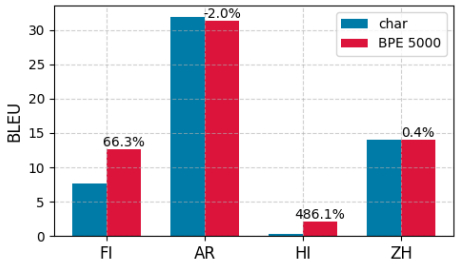

Figure 5: Test BLEU scores on four languages with character level encipherment and comparison between character- and BPE-level re-tokenization. Training data is 1M token and the encryption ratio is 100%.

Figure 5 demonstrates that the choice between BPE and character-level tokenization significantly impacts performance, with effects varying across languages. First of all, Chinese (ZH) maintains a very high performance of BLEU 14.02 even using character-level encipherment. For languages with limited character sets (fewer than 100 characters, such as Finnish, Hindi, and Arabic), retraining a subword tokenizer to expand the existing vocabulary is crucial for performance. In these cases, BPE tokenization after encipherment improves performance by 3.92 BLEU points, representing a 51.6% relative improvement. Logographic languages such as Chinese (ZH) with larger character sets and inherent semantic subword units do not show comparable benefits from this approach. Interestingly, Arabic (AR) shows a slightly degenerated performance when changing from character to BPE tokenization.

4.4 Fighting against Overfitting

Encipherment Ratio	0.5	0.9	1.0
Default	0.26	0.17	0.79
Dropout + Label Smoothing	1.96	1.18	1.06
Dropout only	1.71	1.10	**1.62**
Label Smoothing only	1.26	**1.27**	1.15
$\text{Lora}_{\text{Rank16\&Alpha32}}$	0.94	0.12	0.21
$\text{Lora}_{\text{Rank64\&Alpha128}}$	**2.14**	0.33	0.30

Table 3: Effectiveness of methods in mitigating overfitting. The experiments are all HI-EN translations with 100% character-level encipherment.

Unsurprisingly, models trained with less than 1M tokens suffer from overfitting. We tried different combinations of remedies described in §3.4 and found the suggested learning rate and hyperparam settings from mBART's is the best settings. The two settings are shown in Figure 6. Polynomial

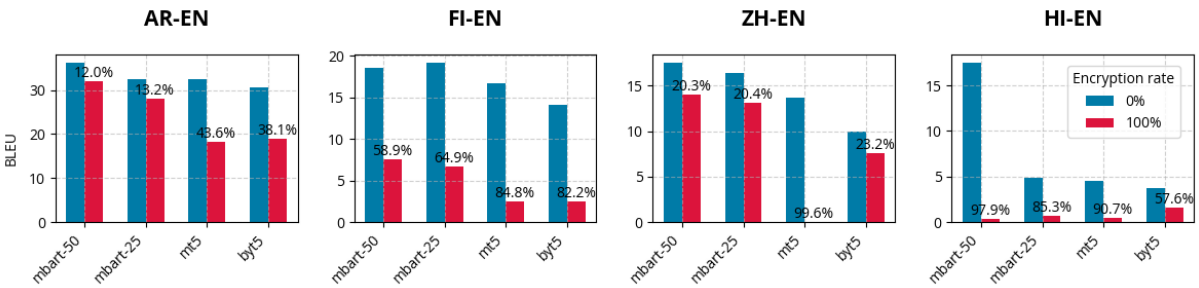

Figure 4: BLEU scores of 4 languages using 0% and 100% encipherment rate on four popular models: mBart-50, mbart-25, mT5 and byT5. Training used 1M tokens.

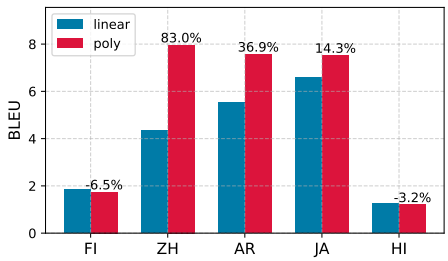

Figure 6: BLEU scores of two learning rate scheduler on different languages. **linear**: Linear decay without warmup steps; **poly:** polynominal decay with 5% warmup step. Trained on 100k tokens and use subword-level encipherment with a encipherment ratio of 100%.

decay with warmup steps is in general better for fine-tuning.

We observe that the enciphered Hindi (HI) text is prone to overfitting. In contrast, `hi` performs the worst. Although `HI` and `ZH` exhibit similar BLEU scores when token_size = 1M and ratio = 0, the BLEU score for `hi` drops abruptly to nearly zero as the ratio increases, whereas `ZH` remains stable. We hypothesize that this discrepancy may result from `ZH` being a more compact language that inherently includes "subwords," whereas `hi` lacks this property. Additionally, `HI` requires a high dropout rate and smoothing even when token_size = 1M, indicating that it is inherently noisy and prone to overfitting.

Given the ineffectiveness of the above methods in mitigating over-fitting, we further experiment with LoRA, which preserves backbone model parameters. We follow the common practice of a double Lora alpha to LoRA rank and study 2 hyperparameter sets across 4 enciphered ratios on `HI`. A higher LoRA rank allows higher adaptation capability to the new language constructed through encipherment but higher risks of over-fitting, and vice versa. As shown in Table 3, LoRA yields a BLEU score of 2.14 on 50% enciphered `HI`.

5 Different Pre-trained Models on UNMUTE data

While the advantages of multilingual models have been extensively studied in previous work, their adoption has not yet become universal in the field. Our experiments provide compelling evidence for their effectiveness, particularly in handling unseen languages. We conducted comparative experiments using encipherment rates of 0% and 100% across four widely-used models. As illustrated in Figure 4, the performance degradation from 0% to 100% encipherment reveals important insights about model capacity for handling under-represented languages:

Model Performance Comparison mBART-50 demonstrates superior performance across all language pairs, both in terms of absolute BLEU scores and relative performance degradation (55.92% drop). The performance gap is particularly pronounced for low-resource language pairs (AR-EN and HI-EN). Notably, mBART-25, despite being trained solely on unsupervised data from CC-25, shows remarkable resilience under high encipherment settings.

Architecture-Specific Analysis mT5 exhibits the most severe performance degradation, with BLEU scores plummeting from 13 to 2 (91.93% drop) under 100% encipherment. The stark contrast between mBART and mT5 performance suggests that architectural choices and pre-training objectives significantly impact cross-lingual transfer capability. The relative stability of mBART-25 under encipherment indicates that the choice of the pre-training objective may be more crucial than the size of training data for robust cross-lingual transfer.

Language-Specific Patterns The impact of encipherment varies notably across language pairs.

AR-EN and ZH-EN show more graceful degradation compared to FI-EN and HI-EN. This pattern suggests that language family relationships and script similarities may influence model robustness. These findings are further corroborated by our case study on Akkadian-English translation (§6.1). Importantly, our results challenge the current practice in WMT challenges where mT5 and mBART are often treated as interchangeable options. This equivalence assumption may mislead researchers and practitioners in developing more effective multilingual translation systems, particularly for low-resource scenarios.

6 Validating Synthetic Experiments with Real-World Cases

To demonstrate the practical applicability of our UNMUTE framework, we examine two representative case studies: Akkadian, an extinct language from the ancient Middle East, and Chinese, a widely used modern language with distinct linguistic properties. These cases serve to validate our framework's findings in both historical and contemporary contexts.

6.1 Case study I: Akkadian Machine Translation

Our first case study focuses on Akkadian, an extinct Semitic language that was predominantly used in ancient Mesopotamia until approximately 1,000 BCE (Gutherz et al., 2023). This language presents unique challenges for machine translation due to its historical nature and limited available corpus. The choice of Akkadian is particularly relevant as it represents an extreme case of an under-resourced language, allowing us to test the robustness of our findings in a real-world scenario.

Our experimental results on synthetic data (Table 4) predicted that mBART would demonstrate superior performance compared to other models when handling heavily enciphered or unfamiliar scripts. The actual performance on Akkadian translation tasks strongly aligns with these predictions, providing empirical validation of our framework's predictive capabilities.

test/BLEU	train size	mBART	from scratch
Akkadian	140k	**54.60**	37.47
Arabic (100% encipher)	100k	**5.53**	3.94

Table 4: mBART-50 works better than model trained from scratch. The Arabic text is 100% enciphered.

6.2 Case II: Romanization on Chinese

test/BLEU	10k	100k	1M
no encipher	10.77	15.42	17.59
char-encipher	0.26	6.89	14.02
subword-encipher	0.12	7.96	14.87
romanization	0.41	6.81	15.26

Table 5: Comparison on WMT18-ZH-EN dataset with a different number of training size (tokens). The encipherment ratio is 100% for both char- and subword-level encipherment.

For non-Latin languages, researchers sometimes use Latin transliteration (romanization) to handle the character set disjoint problem (Nguyen et al., 2023). For example, 自然语言处理 can be transliterated into `ziran yuyan chuli`. The experimental results shown in Table 5 demonstrate that the romanization of Chinese performs similarly to a random character-level 1:1 substitution cipher.

This finding suggests that without strong cross-lingual semantic sharing, transliteration may appear to resolve the out-of-vocabulary (OOV) problem, but in reality, it performs comparably to expanding the vocabulary. Transliteration alone does not address the lack of semantic information transfer between the source and target languages, which is crucial for effective machine translation.

These results highlight the importance of developing more sophisticated methods that go beyond simple character-level mappings to improve cross-lingual transfer learning in low-resource and ancient language settings. Approaches that incorporate semantic information, such as pixel-based representations or byte-level encoding, may be more promising for handling the challenges posed by disjoint character sets and limited training data.

7 Conclusion

In this paper, we introduce UNMUTE, a novel framework that enciphers high-resource languages to systematically analyze the challenges faced by under-represented low-resource languages in machine translation. By disentangling factors such as training data volume and writing system differences, UNMUTE enables a more comprehensive understanding of the barriers to effective cross-lingual transfer learning for low-resource languages in machine translation.

Limitations

While the UNMUTE framework provides a valuable tool for systematically analyzing the challenges faced by low-resource languages in machine translation, it is important to acknowledge its limitations.

Firstly, our framework relies on a simple 1:1 substitution cipher to simulate the out-of-vocabulary (OOV) challenges in low-resource languages. This approach does not fully capture the linguistic complexities of real-world low-resource languages, such as morphological richness, syntactic variations, and language-specific features. Incorporating more sophisticated linguistic features into the encipherment process could provide a more realistic simulation of low-resource language characteristics.

Secondly, the UNMUTE framework focuses primarily on the impact of disjoint character sets and meaning mismatch on cross-lingual transfer learning. However, other factors, such as the domain of the corpus, and the linguistic typology of the languages involved, also play crucial roles in the success of machine translation systems. Future work could extend the UNMUTE framework to investigate the interplay between these factors and the OOV challenges addressed in this study.

Ethics Statement

This work highlights the challenges faced by extremely low-resource languages in machine translation, which we define as those with fewer than 1M training tokens. By emphasizing this definition, we aim to underscore the need for more effective cross-lingual transfer learning approaches that can operate in data-scarce scenarios. We acknowledge that using enciphered modern languages as a proxy for low-resource languages is an imperfect approximation. However, we believe that this approach provides valuable insight into low-resource language processing while respecting the unique context of ancient languages. Throughout this research, we have strived to ensure that our methodology and findings do not perpetuate biases or stereotypes associated with any particular language or language family. We are committed to conducting ethical and responsible research with the ultimate goal of advancing NLP in a direction that benefits all languages and communities, regardless of their availability of resources.

References

Dan Biderman, Jose Gonzalez Ortiz, Jacob Portes, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, et al. 2024. Lora learns less and forgets less. *arXiv preprint arXiv:2405.09673*.

John Chadwick. 1990. *The decipherment of Linear B*. Cambridge University Press.

Danlu Chen, Ka Sing He, Chenghao Xiao, Zhaofeng Wu, Freda Shi, and Taylor Berg-Kirkpatrick. 2025. Translation or recitation? on the performance of transfer learning for underrepresented and extremely low-resource languages. *ArXiv*.

Danlu Chen, Freda Shi, Aditi Agarwal, Jacobo Myerston, and Taylor Berg-Krikpatrick. 2024. Logogramnlp: Comparing visual and textual representations of ancient logographic writing systems for nlp. *ACL*.

Mattia De Cao, Nicola De Cao, Angelo Colonna, and Alessandro Lenci. 2024. Deep learning meets egyptology: a hieroglyphic transformer for translating Ancient Egyptian. In *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024)*, pages 71–86, Hybrid in Bangkok, Thailand and online. Association for Computational Linguistics.

Abteen Ebrahimi, Ona de Gibert, Raul Vazquez, Rolando Coto-Solano, Pavel Denisov, Robert Pugh, Manuel Mager, Arturo Oncevay, Luis Chiruzzo, Katharina von der Wense, and Shruti Rijhwani. 2024. Findings of the AmericasNLP 2024 shared task on machine translation into indigenous languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 236–246, Mexico City, Mexico. Association for Computational Linguistics.

Gideon George, Olubayo Adekanmbi, and Anthony Soronnadi. 2024. Tangalenlp: Building po tangle to english parallel corpora and machine translation of the tangle (tangale) language. In *5th Workshop on African Natural Language Processing*.

Gai Gutherz, Shai Gordin, Luis Sáenz, Omer Levy, and Jonathan Berant. 2023. Translating akkadian to english with neural machine translation. *PNAS Nexus*, 2(5):gad096.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*.

Chunxi Liu, Qiaochu Zhang, Xiaohui Zhang, Kritika Singh, Yatharth Saraf, and Geoffrey Zweig. 2020. Multilingual graphemic hybrid ASR with massive data augmentation. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 46–52, Marseille, France. European Language Resources association.

Evelyn Kai-Yan Liu. 2022. Low-resource neural machine translation: A case study of Cantonese. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 28–40, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Hoang Nguyen, Chenwei Zhang, Tao Zhang, Eugene Rohrbaugh, and Philip Yu. 2023. Enhancing cross-lingual transfer via phonemic transcription integration. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9163–9175, Toronto, Canada. Association for Computational Linguistics.

NLLBTeam et al. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841.

Atul Kr. Ojha, Chao-hong Liu, Ekaterina Vylomova, Flammie Pirinen, Jade Abbott, Jonathan Washington, Nathaniel Oco, Valentin Malykh, Varvara Logacheva, and Xiaobing Zhao, editors. 2024. *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*. Association for Computational Linguistics, Bangkok, Thailand.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.

Christian Roest, Lukas Edman, Gosse Minnema, Kevin Kelly, Jennifer Spenader, and Antonio Toral. 2020. Machine translation for English–Inuktitut with segmentation, data acquisition and pre-training. In *Proceedings of the Fifth Conference on Machine Translation*, pages 274–281, Online. Association for Computational Linguistics.

Elizabeth Salesky, Neha Verma, Philipp Koehn, and Matt Post. 2023. Multilingual pixel representations for translation and effective cross-lingual transfer. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13845–13861, Singapore. Association for Computational Linguistics.

Ana Silva, Nikit Srivastava, Tatiana Moteu Ngoli, Michael Röder, Diego Moussallem, and Axel-Cyrille Ngonga Ngomo. 2024. Benchmarking low-resource machine translation systems. In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 175–185, Bangkok, Thailand. Association for Computational Linguistics.

Richard Sproat and Alexander Gutkin. 2021. The taxonomy of writing systems: How to measure how logographic a system is. *Computational Linguistics*, 47(3):477–528.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Francis Zheng, Edison Marrese-Taylor, and Yutaka Matsuo. 2024. Improving low-resource machine translation for formosan languages using bilingual lexical resources. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11248–11259, Bangkok, Thailand. Association for Computational Linguistics.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016*

Conference on Empirical Methods in Natural Language Processing, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

A Example Appendix

This is a section in the appendix.

Lang	**1k**	**10k**	**100k**	**1M**
fi	63	537	5,666	56,924
ja	220	2,165	21,695	–
ar	62	523	6,546	68,641
hi	53	576	5,844	58,744
zh	217	2,241	21,126	217,541

Table 6: The number of sentences that correspond to the number of tokens across different languages.

tokens=1M, ratio=1				
Model	FI	AR	HI	ZH
char	7.59	31.96	0.36	14.02
BPE (new_token_size=1000)	11.51	29.07	1.89	14.43
BPE (new_token_size=5000)	12.62	31.33	2.11	14.07
BPE (new_token_size=10000)	12.22	31.93	1.63	14.60
tokens=100k, ratio=1				
char	0.98	4.95	0.12	6.89
BPE (new_token_size=1000)	0.35	5.48	0.74	3.68
BPE (new_token_size=5000)	0.68	7.09	1.04	6.79

Table 7: Performance comparison of different tokenization approaches across languages under different token settings