KozKreolMRU WMT 2025 CreoleMT System Description: Koz Kreol: Multi-Stage Training for English–Mauritian Creole MT

Hemkeshsing Y. Rajcoomar
Independent Researcher
yush2398@live.com

Abstract

Mauritian Creole (Kreol Morisyen), spoken by approximately 1.5 million people worldwide, faces significant challenges in digital language technology due to limited computational resources. This paper presents "Koz Kreol," a comprehensive approach to English-Mauritian Creole machine translation using a three-stage training methodology: monolingual pretraining, parallel data training, and LoRA fine-tuning. We achieve state-of-the-art results with 28.82 BLEU score for EN→MFE translation, representing a 74% improvement over ChatGPT-4o. Our work addresses critical data scarcity through use of existing datasets, synthetic data generation, and community-sourced translations. The methodology provides a replicable framework for other low-resource Creole languages while supporting digital inclusion and cultural preservation for the Mauritian community. This paper consists of both a systems and data subtask submission as part of a Creole MT Shared Task.

1 Introduction

Mauritian Creole [1] is spoken by individuals from Mauritius, Rodrigues, Agalega and the Chagos Archipelago. Over the course of its history, Mauritius was visited by the Arabs, colonized by the Dutch, French and the British. Originally, it is a language made of French and Afro-Malagasy languages which was used as a means of communication between slaves and their French masters (Piat, 1999). Over time, as the English rule began and indentured labourers arrived from India, more words infiltrated the existing Mauritian Creole lexicon. With this deep diversity of linguistic families, Mauritian Creole has words that can etymologically be traced back to France, England, Madagascar and both north and south India (Eriksen, 2007). A defining characteristic of creole languages is their dynamic lexicon, which often exhibits clear phonetic and semantic shifts from their source languages (Kouwenberg and Singler, 2009). For example, in Mauritian Creole, kalamindas = candy floss. However, the exact etymology of the word "kalamindas" is unknown. This multilingual substrate influence is characteristic of Creole formation processes, where multiple source languages contribute to the emerging Creole's lexicon and structure (DeGraff, 2001).

Mauritian Creole is considered to be a low-resource language since it lacks digital computational resources for language technology applications (Lent et al., 2022). Despite having a vibrant community of speakers, Mauritian Creole, like many Creole languages, faces social stigmatization and is often perceived as linguistically inferior or underdeveloped compared to its lexifier languages.(Kouwenberg and Singler, 2009).Throughout the years, several efforts have been made by the government of Mauritius to enforce Mauritian Creole as a language rather than a dialect, part of a broader movement for creole language recognition and standardization (DeGraff, 2005). Mauritian Creole has become part of the school curriculum when teaching languages at an early age. However since most Mauritian Creole media are through traditional sources like newspapers or magazines, few digital resources exist. Creoles are generally under represented in language research since they are generally low resource languages whose datasets are seldom publicly available. This digital divide creates significant barriers for the Mauritian community's participation in the modern digital economy and limits access to language technologies that could support cultural preservation and digital inclusion. (Team et al., 2022)

This work addresses these challenges by developing "Koz Kreol," a comprehensive machine translation system for English-Mauritian Creole translation. We present a three-stage training

[1] Also referred to as "Kreol Morisyen"

Proceedings of the Tenth Workshop on Machine Translation (WMT), pages 1183–1190

methodology combining monolingual pretraining, parallel data training, and Low Rank Adaptation (LoRA) fine-tuning that achieves state-of-the-art performance on this language pair. Our approach strategically combines existing datasets from previous research efforts with high-quality community-sourced translations and synthetic data generation. The resulting system not only advances the state of machine translation for Mauritian Creole but also provides a replicable framework for developing MT systems for other low-resource creole languages, contributing to broader efforts in digital language inclusion and cultural preservation. This paper consists of both a systems and data subtask submission as part of a Creole MT Shared Task (Robinson et al., 2025).

2 Related Work

Low Resource Machine Translation (LRMT) has evolved from early transfer learning approaches (Zoph et al., 2016) and backtranslation techniques (Sennrich et al., 2016) to sophisticated multilingual pre-trained models like mBART (Liu et al., 2020), which achieved up to 12 BLEU[2] (Papineni et al., 2002) points improvement for low-resource pairs[3]. The paradigm for low-resource languages was further established by mBART-50 (Tang et al., 2020), which scaled multilingual pre-training to 50 languages and became the standard approach for many low-resource translation tasks. Early work by Tanzer et al. (2024) established benchmarks for learning translation from minimal linguistic resources, demonstrating how grammar books alone can provide sufficient structural information for basic translation capabilities in truly low-resource scenarios.

Creole languages present distinctive challenges beyond typical low-resource scenarios due to their genealogical complexity, orthographic variability, and historical stigmatization, requiring multi-source transfer learning rather than conventional single-source approaches. Recent creole MT research has made substantial progress across multiple fronts. Dabre and Sukhoo (2022) established foundational baselines with KreolMorisienMT, creating the first comprehensive parallel corpus for Mauritian Creole with 21,810 sentence pairs and demonstrating effective transfer learning from pre-trained multilingual models. Robinson et al. (2024) dramatically scaled Creole coverage with Kreyòl-MT, presenting 14.5 million unique creole sentences across 41 languages supporting 172 translation directions. Lent et al. (2024) introduced CreoleVal, the first comprehensive benchmark spanning 8 NLP tasks across 28 creole languages. Fekete et al. (2025) explored parameter-efficient approaches through adapter architectures for cross-lingual transfer, while Adelani et al. (2022) demonstrated that strategic fine-tuning of large pre-trained models with small amounts of high-quality data can achieve significant improvements.

3 Dataset Construction and Methodology

3.1 Data Sources

3.1.1 Existing Parallel Corpora

High-quality Mauritian Creole data is scarce, particularly parallel translations. Our training data includes monolingual and bilingual resources from KreolMorisienMT (Dabre and Sukhoo, 2022) and parallel bitext from Kreyol-MT (Robinson et al., 2024), mostly drawn from translated Bibles and local dictionaries, totaling around 40K bilingual sentences of generally acceptable quality[4].

The monolingual data was downsampled to 18,145 sentences (~500K tokens), as empirical testing showed this size outperformed larger sets (250K, 1M, 2M tokens), likely mitigating catastrophic forgetting (McCloskey and Cohen, 1989) and avoiding repetitive degeneration (Holtzman et al., 2020) during pretraining.

3.1.2 Synthetic Data Generation

Although our primary goal was to fine-tune an LLM for machine translation, we enriched the training data with greater diversity and nuance by including 2,023 Massively Multilingual Language Understanding (MMLU) questions, 961 Question Answering (QA) items, 692 Topic Classification sentences, and 1,225 grammar prompts. We also enhanced Claude's (Anthropic, 2024) context with 150 high-quality parallel bitexts from Flores Dev and created grammar exercises using Gramer Kreol Morisien (Carpooran, 2005). The resulting synthetic dataset comprises 4,901 sentences.

3.1.3 Community Sourced Bitext

To address the shortage of high-quality parallel data for Mauritian Creole, we launched a community-

[2]BLEU: Bilingual Evaluation Understanding

[3]Translation pairs where atleast one language is low resource

[4]Not grammatically consistent throughout.

driven data collection initiative using a web-based annotation platform. Native speakers contributed English–Mauritian Creole translations in both directions, based on Claude-generated English sentences containing at least 15 words to ensure sufficient context and complexity. A two-stage validation process ensured quality, with each translation reviewed by another native speaker. This effort yielded approximately 300 high-quality parallel sentence pairs to supplement our training corpus.

3.1.4 FLORES-200

FLORES-200 extends the original FLORES-101 benchmark by incorporating 200 languages with comprehensive evaluation datasets, providing standardized dev and devtest splits of approximately 1,000 sentences each for multilingual machine translation evaluation. The FLORES data for Mauritian Creole was sourced through a rigorous translation process involving two qualified native speakers who translated the English sentences into Mauritian Creole. Each translated sentence underwent review by the other translator, ensuring high linguistic accuracy and cultural authenticity through this collaborative validation approach.

The resulting datasets comprise 997 sentences in the dev split and 1,012 sentences in the devtest split, providing a total of 2,009 high-quality parallel sentence pairs for English-Mauritian Creole translation. Given the extremely scarce number of high quality parallel bitext available for Mauritian Creole, our final model underwent finetuning on both the dev and devtest portions to maximize the utilization of these linguistic resources.

3.1.5 Evaluation Dataset

Since we fine-tune on the Flores-200 Devtest, we created a 100-sentence evaluation set to monitor BLEU (Papineni et al., 2002) and ChrF[5] (Popović, 2015) during training and fine-tuning. The hold-out test data was sourced from LALIT[6] newspaper, focusing on global geopolitics to assess performance on news content. Source sentences in English were translated into Mauritian Creole, with only sentences over 15 words included. Each translation was validated by another fluent native speaker. Aware of domain-specific evaluation limitations, we report Flores-200 Devtest results in the appendix (Table 3) where the fine-tuned model uses only Flores-200 Dev.

[5]Character F-score.

[6]lalitmauritius.org

English–Kreol Morisien					
Split	**L**	**AL-en**	**AL-mfe**	**U-en**	**U-mfe**
train	46,160	7.3	6.7	31,195	32,106
dev	997	21.0	21.4	6,695	6,195
devtest	1,012	21.6	21.9	7,054	6,413
lalit (test)	102	48.5	45.2	1,874	1,762
Kreol Morisien Monolingual					
Split	**L**	**AL**	**–**	**U**	**–**
mono	18,145	87.13	–	27,967	–

Table 1: Dataset statistics. L: total sentences/pairs; AL-en/AL-mfe: average word counts for English/Mauritian Creole; U-en/U-mfe: unique word counts for English/Mauritian Creole.

3.2 Dataset Statistics

Table 1 presents comprehensive statistics for our datasets across different splits[7]. Our training dataset exhibits diverse characteristics across different data sources. The training split contains a substantial number of single-word entries representing 1-1 translations sourced from previous lexical datasets sourced by Dabre and Sukhoo (2022), contributing to the lower average word counts (7.3 for English, 6.7 for Mauritian Creole) compared to the evaluation sets.

The FLORES evaluation sets show significantly higher average word counts (21.0-21.6 for source, 21.4-21.9 for target), likely reflecting the more complex sentence structures typical of the FLORES benchmark. With higher average word counts of 48.5 for English and 45.2 for Mauritian Creole, we assume the hold-out test set to have even higher linguistic complexity, consistent with the discourse of news content covering global geopolitics.

4 Experiments

4.1 Experimental Design

In our comprehensive arsenal, we now have an extensive collection of resources including monolingual Creole data, valuable parallel bitext published by previous researchers, the robust Flores-200 training data, and carefully generated synthetic data. Additionally, we have meticulously curated high quality parallel sentences sourced directly from native local speakers through an ambitious community outsourcing project we launched around a year ago. This community-driven approach ensures authentic linguistic representation and cultural ac-

[7]More details in Section B of the Appendix.

curacy in our training data. We will break our carefully designed training recipe down to three distinct important steps: Pretraining, Training and Finetuning. For this comprehensive study, we will use Llama 3.1-8B model and tokenizer as our robust backbone LLM here.

4.2 Training Setup

According to the findings of Xu et al. (2024), there's significant and demonstrable benefit in pretraining a large language model with a language it is previously unfamiliar with. This crucial step helps the model build a rich internal vocabulary, as well as, develop a deep understanding of the intricate semantics of a language. However, this process has to be done extremely carefully and with precise control since it can lead to the detrimental phenomenon of catastrophic forgetting (McCloskey and Cohen, 1989) when fed too much overwhelming data. To maintain this delicate balance, we use a carefully measured 500K monolingual tokens of authentic Mauritian Creole, complemented by 100K tokens of English and French each. We employ full-weight finetuning for this critical foundational portion.

For the next step in our pipeline, we use our extensive parallel bitext sourced by other dedicated researchers as well as our synthetic data, comprising around 46K sentences. These valuable sources are mostly drawn from the carefully translated Bible and comprehensive local dictionaries. Remarkably, one single pass of the data onto a powerful 3.1-8B Llama backbone is already sufficient to see vast and encouraging improvements in translation performance. Training the model with a learning rate of 1e-5 and the AdamW (Loshchilov and Hutter, 2019) optimizer, we stop after 2 complete epochs to prevent overfitting.

For the final and most refined step in our training methodology, we use our mix of Flores 200 dev and devtest sets, for efficient Low-Rank Adaptation (LoRA) (Hu et al., 2021) Finetuning for a maximum of 3 epochs. We conduct a hyperparameter sweep over the critical rank parameter, the scaling factor named "alpha" and the target modules. The LoRA hyperparamters for our best performing results based on BLEU and ChrF metrics are $\alpha = 8$, $r = 16$, target modules = query and value projections.

5 Results

For the sake of our experiment, since Mauritian Creole data is scarce, we are training on "devtest" and we use the LALIT test set for evaluation.

5.1 Baseline Comparisons

Model / Setup	BLEU	CHRF
EN → MFE		
Zero-Shot (Llama 3.1-8B)	4.22	35.37
ChatGPT 4o	16.55	53.58
Mono Only	22.54	51.67
Mono + Train	26.76	59.55
Mono + Train + LoRA	**28.82**	**60.86**
MFE → EN		
Zero-Shot (Llama 3.1-8B)	28.4	57
ChatGPT 4o	**46.63**	**71.22**
Mono Only	43.32	69.23
Mono + Train	41.78	68.60
Mono + Train + LoRA	43.14	70.21

Table 2: BLEU and ChrF scores for different model configurations in EN ↔ MFE translation.

From Table 2, we observe a significant improvement in translation performance when incorporating 500K monolingual tokens into the model. However, the model still lacks the ability to translate the language effectively from English to Mauritian Creole. The training portion of our approach provides the largest performance boost, with a 18.7% increase in BLEU score and an 15.2% increase in ChrF score. Following this stage, we perform fine-tuning using Low-Rank Adaptation (LoRA) using the peft package. (Mangrulkar et al., 2022) Fine-tuning on 2,000 sentences in both translation directions contributes to approximately 10% improvement in BLEU score.

When examining the reverse direction (MFE → EN), the model performs significantly better than in the forward direction, a common finding when translating Low Resource Langauges to English (Neubig and Hu, 2018). The model performs better on BLEU score after passing in 500K tokens of Mauritian Creole only sentences, and the performance slightly declines when training. The improvement from LoRA fine-tuning is more modest compared to training, likely because the model has already achieved strong performance in back-translation and is approaching a performance plateau.

When comparing our model to a frontier model such as ChatGPT-4o (OpenAI et al., 2024), we observe that our model performs considerably better on forward translations (English→Mauritian Creole). However, for reverse translations (Mauritian Creole→English), ChatGPT-4o achieves slightly superior performance on both BLEU and ChrF scores. This asymmetry can likely be attributed to ChatGPT-4o's extensive multilingual training across numerous language pairs, enabling it to leverage cross-lingual priors for improved Mauritian Creole decoding.

Haitian Creole and Mauritian Creole share significant linguistic similarities as French-based creoles developed under similar colonial conditions. They exhibit overlapping vocabulary, grammar, and simplification patterns compared to French (Déprez, 2019), enabling cross-linguistic transfer. Models trained on Haitian Creole can leverage this overlap when processing Mauritian Creole. Given Haitian Creole's much larger training corpus, this likely influences ChatGPT-4o's performance—improving reverse translation but degrading forward translation, as the model tends to apply Haitian grammar to Mauritian output. We observed the same behavior using the Flores-200 Devtest set (see Appendix).

Large language models outperform traditional encoder-decoder architectures in data-scarce scenarios due to their ability to extract linguistic patterns from limited examples. Pre-trained on multilingual corpora, LLMs provide rich contextual representations adaptable to new language pairs with minimal fine-tuning, unlike encoder-decoder models that need substantial parallel data. This advantage is especially important for creole languages, where complex lexifier and substrate influences are better captured by the nuanced knowledge in large-scale pre-trained models.

6 Conclusions

In this paper, we present a new state-of-the-art model for English–Mauritian Creole translation, along with several novel datasets used for training and evaluation. These include: (1) the Flores-200 Dev and Devtest sets, (2) synthetically generated data, (3) a test set from the LALIT newspaper, and (4) community-sourced parallel bitext. This work provides a strong baseline for future model development, which can be improved by collecting more high-quality parallel data. Our results with limited data suggest that incremental augmentation will boost performance, supporting sustainable Mauritian Creole MT development. Additionally, our model can generate high-quality synthetic translations, enabling continual learning through iterative data generation and refinement.

Several promising directions emerge for future research. The inclusion of French parallel bitext represents a particularly valuable avenue, given Mauritian Creole's French lexifier heritage and the abundance of high-quality French-English parallel corpora that could enhance transfer learning effectiveness (Robinson et al., 2023). Incorporating pivot languages; intermediate languages that share linguistic features with both English and Mauritian Creole, could provide additional pathways for cross-lingual knowledge transfer and improved translation quality.

Finally, the systematic generation of synthetic training data through back-translation, paraphrasing, and multilingual data augmentation techniques offers scalable approaches to address the persistent data scarcity challenges that characterize creole language processing. These future developments, building upon the foundation established in this work, promise to advance Mauritian Creole machine translation toward broader practical applicability and community benefit.

Limitations

Our work has several important limitations that should be acknowledged. First, we train our final model on both the FLORES-200 dev and devtest splits, which raises potential evaluation concerns regarding data contamination. While we mitigate this through the use of an independent hold-out evaluation set sourced from LALIT newspaper, the limited size of available high-quality parallel data necessitated this approach to maximize training effectiveness.

Second, our evaluation is constrained by the relatively small size of our test sets (100-1,000 sentences), which may limit the statistical significance and generalizability of our results. The scarcity of Mauritian Creole digital resources inherently constrains the scale of evaluation possible for this language pair.

Finally, our synthetic data generation approach, while innovative, relies on a single large language model (Claude Sonnet 4) and may introduce systematic biases or artifacts that could affect model

performance. The quality and cultural authenticity of synthetically generated Mauritian Creole content, while supplemented with expert knowledge, may not fully capture the nuanced variations present in natural language use.

Acknowledgments

Special thanks to Aishani Rajarai who helped me create and review the high quality parallel bitext datasets. We'd like to also thank Professor David Ifeoluwa Adelani from the Masakhane Community for his guidance.

References

David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, and 1 others. 2022. A few thousand translations go a long way! leveraging pre-trained models for african news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070. Association for Computational Linguistics.

Anthropic. 2024. Claude sonnet 4. Artificial Intelligence Model. Large language model, version as of June 2024.

Arnaud Carpooran. 2005. *Gramer Kreol Morisien*. Editions Bartholdi, Mauritius. First comprehensive grammar of Mauritian Creole in the language itself.

Raj Dabre and Aneerav Sukhoo. 2022. KreolMorisienMT: A dataset for mauritian creole machine translation. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 22–29, Online only. Association for Computational Linguistics.

Michel DeGraff. 2001. On the origin of creoles: A cartesian critique of neo-darwinian linguistics. *Linguistic Typology*, 5(2/3):213–310.

Michel DeGraff. 2005. Linguists' most dangerous myth: The fallacy of creole exceptionalism. *Language in Society*, 34(4):533–591.

Viviane Déprez. 2019. Plurality and definiteness in mauritian and haitian creoles. *Journal of Pidgin and Creole Languages*, 34(2).

Thomas Hylland Eriksen. 2007. Creolization in anthropological theory and in mauritius. In Charles Stewart, editor, *Creolization: History, Ethnography, Theory*, pages 153–177. Left Coast Press, Walnut Creek, CA.

Marcell Fekete, Nathaniel Romney Robinson, Ernests Lavrinovics, Djeride Jean-Baptiste, Raj Dabre, Johannes Bjerva, and Heather Lent. 2025. Limited-resource adapters are regularizers, not linguists. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 222–237, Vienna, Austria. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. *Preprint*, arXiv:1904.09751.

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. *Preprint*, arXiv:2106.09685.

Silvia Kouwenberg and John Victor Singler. 2009. *The Handbook of Pidgin and Creole Studies*. John Wiley & Sons.

Heather Lent, Kelechi Ogueji, Miryam de Lhoneux, Orevaoghene Ahia, and Anders Søgaard. 2022. What a creole wants, what a creole needs. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6439–6449, Marseille, France. European Language Resources Association.

Heather Lent, Kushal Tatariya, Raj Dabre, Yiyi Chen, and 1 others. 2024. CreoleVal: Multilingual multitask benchmarks for creoles. *Transactions of the Association for Computational Linguistics*, 12:950–978.

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. *Preprint*, arXiv:2001.08210.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. PEFT: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Michael McCloskey and Neal J. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, 24:109–165.

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. *arXiv preprint arXiv:1808.04189*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Denis Piat. 1999. *Sur la Route des Épices: L'Île Maurice*. Les Editions du Pacifique.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Nathaniel R. Robinson, Raj Dabre, Ammon Shurtz, Rasul Dent, and 1 others. 2024. Kreyòl-mt: Building mt for latin american, caribbean and colonial african creole languages. *Preprint*, arXiv:2405.05376.

Nathaniel R. Robinson, Claire Bizon Monroc, Rasul Dent, Stefan Watson, Raj Dabre, Kenton Murray, Andre Coy, and Heather Lent. 2025. Findings of the first shared task for creole language machine translation at wmt25. In *Proceedings of the Tenth Conference on Machine Translation*.

Nathaniel Romney Robinson, Matthew Dean Stutzman, Stephen D. Richardson, and David R Mortensen. 2023. African substrates rather than european lexifiers to augment african-diaspora creole translation. In *4th Workshop on African Natural Language Processing*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. *Preprint*, arXiv:1511.06709. [link].

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. *arXiv preprint arXiv:2008.00401*.

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2024. A benchmark for learning to translate a new language from one grammar book. *Preprint*, arXiv:2309.16575.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. A paradigm shift in machine translation: Boosting translation performance of large language models. *Preprint*, arXiv:2309.11674.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. *Preprint*, arXiv:1604.02201.

A Training and Evaluation with Flores-200

In this section, we will evaluate the model's performance on FLORES-200 devtest across three modalities: (i) monolingual only, (ii) monolingual + training data, and (iii) monolingual + training + FLORES-200 dev LoRA finetune. We evaluate both translation directions to quantify the additional performance lift from the FLORES dev finetune and assess the validity of using the FLORES-200 devtest set as an evaluation dataset. The LoRA finetune was performed in both translation directions.

Model / Setup	BLEU	CHRF
EN → MFE		
Kreyòl-MT	17.28	49.07
ChatGPT-4o	17.48	48.40
Mono Only	11.94	39.78
Mono + Train	25.76	55.71
Mono + Train + LoRA	**26.83**	**57.68**
MFE → EN		
Kreyòl-MT	28.31	57.29
ChatGPT-4o	**43.08**	**68.76**
Mono Only	33.80	60.88
Mono + Train	40.75	66.65
Mono + Train + LoRA	41.79	67.73

Table 3: BLEU and ChrF scores for different model configurations in EN ↔ MFE translation.

B Dataset Specifics

The training data consisted of KreolMorisienMT (21,810 sentences), KreyolMT (19,149 sentences), Flores Dev/Devtest (2,009 sentences), 300 community-sourced sentences, and synthetic data generated by Claude Sonnet 4. We created parallel bitext datasets (MMLU, QA, and Topic Classification totaling 3,676 sentences) and conversational prompts (1,225 grammar-specific sentences from parsed Mauritian Creole grammar books). For parallel bitext, we used a dual prompt strategy: one prompt asking the model to translate between English and Mauritian Creole, and another presenting questions directly in Mauritian Creole with options and answers in Mauritian Creole.

For monolingual pretraining we only used the monolingual dataset provided by KreolMorisienMT. For the training stage we use both parallel datasets from Kreyol-MT, KreolMorisienMT

and the community sourced bitext. For the fine-tuning stage, we use the flores dev and devtest datasets.