EdinHelsOW WMT 2025 CreoleMT System Description: Improving Lusophone Creole Translation through Data Augmentation, Model Merging and LLM Post-editing

Jacqueline Rowe[1], Ona de Gibert[2], Mateusz Klimaszewski[3], Coleman Haley[1], Alexandra Birch[1], Yves Scherrer[4]
[1]University of Edinburgh [2]University of Helsinki
[3]Warsaw University of Technology [4]University of Oslo
Correspondence: jacqueline.rowe@ed.ac.uk

Abstract

In this work, we present our submissions to the unconstrained track of the System subtask of the WMT 2025 Creole Language Translation Shared Task. Of the 52 Creole languages included in the task, we focus on translation between English and seven Lusophone Creoles. Our approach leverages known strategies for low-resource machine translation, including back-translation and distillation of data, fine-tuning pre-trained multilingual models, and post-editing with large language models and lexicons. We also demonstrate that adding high-quality parallel Portuguese data in training, initialising Creole embeddings with Portuguese embedding weights, and strategically merging best checkpoints of different fine-tuned models all produce considerable gains in performance in certain translation directions. Our best models outperform the baselines on the Task test set for eight out of fourteen translation directions. When evaluated on decontaminated test sets, they surpass the baselines in all directions.

1 Introduction

The introduction of the first Shared Task for Creole language machine translation (MT) (Robinson et al., 2025) is emblematic of the increased attention that Creole languages have received in the field of Natural Language Processing in recent years, both as individual languages (Robinson et al., 2022; Dabre et al., 2014; Lent et al., 2021; Dabre and Sukhoo, 2022; Rowe et al., 2025) and in multilingual modeling efforts (Robinson et al., 2024; Lent et al., 2024). Building on the latter, this Shared Task covers over 50 Creole languages from a range of geographical and linguistic contexts. Some are relatively high-resourced; for example, Haitian Creole and Papiamento are supported in Google Translate and many others are institutionalised as official or educational languages (Robinson et al., 2024). Others are extremely low-resource languages or even critically endangered or extinct.

The Shared Task invites submissions of data and systems serving MT between any of the Creole languages and either English or French, with the existing Kreyòl-MT (Robinson et al., 2024) and CreoleVal (Lent et al., 2024) translation models serving as baselines. In this submission, we develop systems to translate between English (`eng`) and seven Lusophone[1] Creoles: Angolar (`aoa`), Annobonese (`fab`), Guinea-Bissau Creole (`pov`), Kabuverdianu (`kea`), Papiamento (`pap`), Principense (`pre`) and Sãotomense (`cri`).[2] This set includes relatively high-resource Creoles (like `pap` and `kea`) and extremely low-resource ones (like `aoa`, `fab` and `pre`).

In our submission, we utilise known strategies for low-resource MT as well as techniques designed to leverage the linguistic relationship between our seven Creoles of focus and Portuguese (`por`). In particular, we contribute the following:

- We collate additional parallel and monolingual data for `pap`, `pov`, `kea` and `cri` (Sections 3.1.2 and 3.1.3).
- We augment the training data with high-quality parallel `eng-por` data, synthetic parallel data created via back-translation, and distilled data created via forward-translation (Section 3.2).
- We fine-tune three pretrained multilingual base models with different combinations of data and initialisation strategies (Section 4.2).
- We apply model merging to further improve translation performance (Section 4.3).
- We post-edit system outputs using LLMs and bilingual lexicons, improving performance for five translation directions (Section 4.4).

We release our code in our Github repository.[3]

[1]Creoles which are related to Portuguese.

[2]We focus on translation between these seven Creoles and English due to availability of test data, but future work could expand to Creole–Portuguese translation.

[3]`https://github.com/JacquelineRowe/EdinHelsOW_CreolesMT`. Due to the copyright terms of most of our data sources, we do not publicly share our dataset. It is available

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 924–940
November 8–9, 2025. ©2025 Association for Computational Linguistics

2 Related Work

Robinson et al. (2024) release four versions of Kreyòl-MT (**KMT**), a translation model which supports all seven of our Creole languages of focus. The four versions are created by training on both public and private datasets, and training both from-scratch and fine-tuning an existing model. For fine-tuning, they use many-to-many (m2m) **mBART-50** (Tang et al., 2021), a multilingual version of mBART (Liu et al., 2020) fine-tuned for translation between 50 languages, as the base model. mBART is a sequence-to-sequence denoising auto-encoder pre-trained on large-scale monolingual corpora in many languages using the BART objective (Lewis et al., 2020). Lent et al. (2024) also fine-tune m2m **mBART-50** on a different set of Creole languages including pap.

While the models released in Robinson et al. (2024) and Lent et al. (2024) are the strongest baselines for MT for Creoles in general, some of our seven languages of focus are also included in other prior work on MT. The No Language Left Behind (**NLLB**) translation model excels at translation of low-resource languages, and supports pap and kea (as well as three other Creoles not included in our study) (NLLB Team et al., 2022). The training data curated as part of the NLLB effort include less than 10 bitexts for each Creole, but 28M monolingual sentences in pap and 300k in kea. The FLORES-200 evaluation dataset was also translated into both Creoles in this context.

Both kea and pap are featured in **PanLex**,[4] a massive, open-access online lexicon covering over 5,000 languages (Kamholz et al., 2014); but only pap is supported in **GATITOS**, a smaller, higher-quality parallel lexicon for low-resource languages developed by Jones et al. (2023). These lexical resources have been used to improve low-resource MT performance for Creoles. Following prior work using LLMs to post-edit machine translation system outputs to correct errors (Xu et al., 2024; Chen et al., 2024; Hus et al., 2025), Nielsen et al. (2025) showed that including the entire GATITOS lexicon in such post-editing prompts can improve ChrF scores and reduce lexical confusion, including for pap-eng MT. Similarly, Hus and Anastasopoulos (2024) showed improvements of over 15 ChrF++ in eng-kea MT by post-editing using an LLM with prompts including parallel words and sentences extracted from the kea PanLex dataset.

The question of how training data from related languages can improve MT for Creoles remains open (Lent et al., 2022). Ma et al. (2025) showed that the speech foundation model Whisper (Radford et al., 2023) performs surprisingly well on kea-eng speech translation (despite having not been trained on kea speech) when the por language code is used for decoding, which they hypothesise is due to pronunciation similarities between the two languages. Conversely, Fekete et al. (2025) demonstrated that parameter efficient fine-tuning via language adapters improves MT for three Creoles (including pap) regardless of whether the adapters were trained on related languages, unrelated languages, or even random noise, indicating that language adapters improve performance due to regularization rather than cross-lingual transfer.

3 Data

In this section, we briefly describe the data provided by the task organisers and the additional data we collect and create for model training. Our novel data sources are documented in full in Table 6 in Section A.

3.1 Data Collection

3.1.1 Organiser-Provided Data

To train their models, Robinson et al. (2024) gathered data for 43 Creoles from multilingual datasets, extracting parallel and monolingual texts from websites, Wikipedia collections, educational materials, religious texts and other sources where available. Some of their data remains private due to copyright reasons, but their public training and development splits ($\text{Train}_{\text{KMT}}$ and Val_{KMT}) form the official training data for the Shared Task. Robinson et al. (2024) also provide a public test split (Test_{KMT}), which we do not use as training data.[5]

For our seven Creoles of focus, the publicly available resources parallel with eng from Robinson et al. (2024) vary in size and domain. The datasets for pov, pre, aoa, cri and fab have between 170 and 450 parallel aligned sentences from

to academic researchers for non-commercial purposes upon request; please contact the lead author for license agreement and access.

[4]At the time we conducted our study, PanLex was not accessible online and so we did not use this resource for kea in our work.

[5]While we did not use Test_{KMT} data to train our models, we did evaluate our models' performance on this public test split in order to make modelling decisions, prior to the announcement that the official Shared Task test set would be identical to Test_{KMT}.

educational materials, collected from the APiCS corpus (Michaelis et al., 2013). In contrast, the parallel datasets for kea and pap are larger and more diverse, both drawing data from FLORES-200 dev and NLLB train (NLLB Team et al., 2022) as well as APiCS (Michaelis et al., 2013). The public pap dataset[6] also includes bitexts from the Online library of The Church of Jesus Christ of Latter-day Saints[7], LegoMT (Yuan et al., 2023), Tatoeba[8], and Wikipedia, as well as a bilingual lexicon.[9] Parallel sentences with languages other than eng are available for pap and kea, but we include only parallel data with eng.

3.1.2 Additional Parallel Data

To augment the official task data, we collect additional data parallel with eng for pap, pov and kea.[10] As is common with low-resource languages, much of the publicly-available parallel data sources we could find for each language are religious in nature (Siddhant et al., 2022). We collect aligned Bible verses (pap and pov) and aligned sentences from available editions of Jehovah's Witnesses Watchtower (JWW) series[11] (pap, pov and kea). We also collect non-religious parallel sentences from a random sentence generator (pap), an article about internet access (pov), and the glosses from a por-pov bilingual dictionary (we translate the por glosses into eng using Google Translate).

Portuguese Since our focus is on Lusophone Creoles, we hypothesise that adding high-quality eng-por data can improve transfer learning. We download the eng-por Tatoeba Translation Challenge Dataset (Tiedemann, 2020), which is a collection of all data in OPUS, shuffled and deduplicated. We use the corresponding Bicleaner-AI (Zaragoza-Bernabeu et al., 2022) scores[12] to aggressively filter the dataset. Bicleaner-AI is a neural metric that estimates how likely it is that a sentence pair is a translation. We keep only sentence pairs with a Bicleaner-AI score of 1.0 to ensure high quality, leaving us with a seed dataset of 112k sentences (representing 0.03% of the total Tatoeba dataset).

3.1.3 Additional Monolingual Data

We also collect monolingual Creole data, including a high school textbook (kea), a blog series (kea), glosses from an unpublished monolingual dictionary (pov) and transcriptions of a documentary (pov). The JWW Series (see Section 3.1.2) in cri is hosted on a different website from the eng, pap, pov and kea versions; as this makes it impossible to align the cri data with the eng data, we instead collect JWW as a monolingual resource for cri.

3.1.4 Lexicons

In order to experiment with post-editing with LLMs and lexicons, as demonstrated in Nielsen et al. (2025), we collect bilingual lexicons for each of our seven Creoles of focus. For aoa, we could not find a publicly-available lexicon, and instead manually curate a small set of parallel lexical items using word-aligned entries from IMT Vault.[13] For pap, we use both the GATITOS lexicon (Jones et al., 2023) and a newly collected traditional lexicon.

3.2 Synthetic Data

We backtranslate all sources of monolingual data into eng using the KMT model that scores the highest ChrF on the KMT test set for that language pair.[14] We also use kreyol-mt (the single best KMT model) as a 'teacher' model, using it to forward translate the eng sentences from the pap, kea, pov and cri parallel datasets into each Creole via Sequence-Level Distillation (Seq-KD) (Kim and Rush, 2016).[15] These distilled datasets allow us to train models which better imitate the distribution output of the KMT model at sentence-level.

[6]The private pap-eng dataset (used for model training but not publicly released) includes additional parallel data from CreoleVal (Lent et al., 2024), a textbook, the JHU bible corpus (McCarthy et al., 2020), the QED corpus (Lamm et al., 2021) and Ubuntu texts from the OPUS corpus.

[7]https://www.churchofjesuschrist.org/study?lang=pap. This dataset was shared directly by the organisers as it is not on HuggingFace yet.

[8]https://tatoeba.org/en/downloads

[9]https://www.scribd.com/document/119363393/Parleremo-English-Papiamento-Papiamento-English-Dictionary-1ed

[10]We later found small parallel resources for aoa, fab and pre; while it was too late to include these sources in our model training, we list these sources in Table 6 for future reference.

[11]JWW is a monthly Bible study resource which is mostly about religious matters but also includes some discussion of more general topics.

[12]https://github.com/Helsinki-NLP/Tatoeba-Challenge/blob/d34a89ac102fd236503a1911dd1050564bf4e682/BicleanerScores.md

[13]https://imtvault.org/?languageiso6393%5B0%5D=aoa

[14]kreyol-mt for cri and kea; kreyol-mt-pubtrain for pap and pre; and kreyol-mt-scratch for pov. We used the publicly available Test_{KMT} set to select which models to use for back-translation before realising that the Shared Task test set would be identical to the publicly available test set.

[15]We do not use distillation for aoa, fab and pre because the KMT model demonstrates ChrF scores which are too low to generate reasonable forward translations.

3.3 Data Pre-processing

We use all novel collected data for training and evaluation, except the bilingual lexicons which we reserve for post-editing experiments. We first remove any pairs of parallel sentences from our novel datasets where either the source (src) or target (tgt) sentence is in that language pair's Test_{KMT} dataset, to ensure we do not train on any test data. We then split out 10% of our novel cri, kea, pap and pov data (up to a limit of 1,000 sentences) for both validation and test data. We combine our training and validation splits with $\text{Train}_{\text{KMT}}$ and Val_{KMT} respectively, but keep Test_{KMT} separate from our own test data ($\text{Test}_{\text{Ours}}$) for evaluation purposes.

To clean each data split, we remove duplicates, empty or identical src/tgt pairs, and pairs where src or tgt have more than 150 or fewer than three words. We also discard pairs where the ratio of the length of the src to the tgt sentence is unusually high or low, following Robinson et al. (2024). Finally, we normalise special characters like quotes, dashes, and Unicode Hex codes.

We noted that several sentences[16] from $\text{Train}_{\text{KMT}}$ and $\text{Validation}_{\text{KMT}}$ included multiple eng glosses for a single Creole sentence. For example, the cri sentence "*Ê tava ka vivê ni Libôkê.*" has the eng gloss "*He was living in Libôkê. OR: He used to live in Libôkê.*" To reduce ambiguity at train time, we split each of these double glosses into two separate eng sentences. For $\text{Train}_{\text{KMT}}$, we duplicate each Creole sentence and use each eng gloss to create two pairs of parallel sentences; for $\text{Validation}_{\text{KMT}}$, we retain only the first gloss as the eng translation of each Creole sentence.

Table 1 shows the combined dataset sizes after pre-processing. For complete details on the train, validation, and test splits for each language, including both our data and the organizer-provided data before and after cleaning, see Tables 7, 8 and 9.

	Train	Val.	Test_{KMT}	$\text{Test}_{\text{Ours}}$	All
pap	105,805	1,085	1,967	1,000	109,857
pov	43,699	1,027	33	1,000	45,759
kea	9,438	1,084	163	1,000	11,685
cri	1,376	189	33	155	1,753
pre	105	36	36	0	177
aoa	71	35	39	0	145
fab	61	31	38	0	130

Table 1: Numbers of Parallel Sentences (with eng) for each language pair, ordered by size of dataset.

[16]Specifically, those collected from the APiCS data source.

4 Models

To create our MT systems, we fine-tune the three multilingual pre-trained translation models described in Section 2: **KMT** (Robinson et al., 2024), **mBART-50** (Tang et al., 2021), and **NLLB** (NLLB Team et al., 2022). We explain our approach for fine-tuning each model below, listing additional training configuration details in Appendix E.

4.1 Baselines

The baseline models specified by the organisers for the unconstrained track of the Systems Subtask were CreoleM2M (Lent et al., 2024) and kreyol-mt (Robinson et al., 2024).[17] Both were created by fine-tuning m2m mBART-50 (Tang et al., 2021) on private datasets. While CreoleM2M performs slightly better than kreyol-mt on pap-eng and eng-pap translation, it does not support our other six Creoles of focus, and so for simplicity we use kreyol-mt as our experimental baseline.

4.2 Our Models

Fine-tuned KMT We first explore whether we can improve the performance of the baseline kreyol-mt model[18] by fine-tuning it further on our datasets using PyTorch Lightning (Falcon and team, 2019). We use kreyol-mt's existing language tags and embeddings for each Creole.[19] Like mBART-50, kreyol-mt has 611M parameters and a SentencePiece (Kudo and Richardson, 2018) vocabulary of 250k subwords.

Fine-tuned mBART-50 We then explore whether we can recreate our own version of kreyol-mt by fine-tuning the m2m version of mBART-50 on our novel datasets using Fairseq (Ott et al., 2019). As the English-centric many-to-one (m2o) and one-to-many (o2m) versions of mBART-50 have been shown to outperform their m2m counterpart (Liu et al., 2020), we also use these models for fine-tuning. All three mBART-50

[17]While these baselines were listed on the Shared Task website, organisers clarified afterwards that kreyol-mt has been trained on portions of text from Test_{KMT}, and that the intended baseline was, in fact, kreyol-mt-pubtrain.

[18]We chose to fine-tune kreyol-mt without realising that its training data included text from the public Test_{KMT} set. The results of these models on Test_{KMT} are therefore inflated.

[19]We note that kreyol-mt was trained with src language tags appended to the end of each training src sentence (in contrast to traditional mBART-50 language tagging in which the src tag is prepended to the beginning of the src sentence). We replicate the kreyol-mt tagging system for tokenising the training, validation and test data.

models share the same SentencePiece (Kudo and Richardson, 2018) vocabulary of 250k subwords. We repurpose existing language tags for our unseen language pairs following Robinson et al. (2024), initialising their embeddings randomly. To compensate for the imbalance in dataset sizes across languages, we use temperature-based sampling with $\tau = 2$, which increases the relative sampling probability of low-resource languages and promotes more balanced training.

Fine-tuned NLLB As a state-of-the-art translation model designed specifically to perform well on low-resource languages, NLLB (NLLB Team et al., 2022) is also commonly fine-tuned for unseen language pairs in specific translation contexts (Ebrahimi et al., 2023; De Gibert et al., 2025). The largest NLLB model is a 54.5B parameter sparsely-gated mixture of experts model; we use two smaller distilled versions of this model (`distilled-1.3B` and `distilled-600M`) for our experiments. While `pap` and `kea` are already supported in NLLB, we add additional language tags for the other five languages and initialise their embeddings randomly. We use PyTorch Lightning for training as described for fine-tuning `kreyol-mt`, except for fine-tuning NLLB where we implement a maximum of 30 training epochs to keep total training time feasible.

Fine-tuning Experiments We first fine-tune `kreyol-mt`, the three different versions of mBART-50 and the two different versions of NLLB on our dataset for three translation directions; all Creoles into `eng` (`XX-eng`), `eng` into all Creoles (`eng-XX`), and both of these directions simultaneously (`XX-XX`). We select the best overall setup for each of the three base models for translation both into and out of `eng`, and then repeat each of those best setups for the following experiments:

1. Initialising embeddings for Creole language tags with existing embeddings in each model for `por`, instead of using existing Creole embeddings (for `kreyol-mt` models) or random initialisation (for NLLB and mBART-50).[20]
2. Including `eng-por` or `por-eng` as an additional training direction, leveraging the high-quality parallel data collected from Tatoeba (see Section 3.1.2).
3. Using `kreyol-mt` distilled data (see Section 3.2) as target side translations for fine-tuning on `pap`, `kea`, `pov` and `cri`.

For each of these fine-tuned models, we find the checkpoint with the highest scores across all languages on the validation set, and then use this best checkpoint to evaluate that model's performance on Test_{KMT}. Where any two experimental settings show improvements on the basic setup for a given base model, we also combine them together.

4.3 Model merging

To obtain most of our final models we applied model merging using Arcee's MergeKit framework (Goddard et al., 2024), specifically the `linear` method (Wortsman et al., 2022). We define three different merging strategies: (i) averaging different checkpoints of the same training run, (ii) merging different (our) models or (iii) merging our models with the `kreyol-mt` baseline model (i.e. federated learning, as the training set of `kreyol-mt` is not public). While the two first options were applied to fine-tuned mBART-50 and NLLB models (described in Section 4.2), the last option was applied to the fine-tuned KMT models (Section 4.2). In our experiments we merge between 3 and 5 checkpoints, mostly from our internal finetuned models (selecting based on best-performance on the validation dataset for specific language pairs), but also – in the case of (iii) – external models. We note that most of the time, this procedure meant averaging three last checkpoints of our finetuned models.

4.4 Post-editing

With the lexicons we collected for each Creole and the system outputs of the best models for each language pair on the Test_{KMT} dataset, we implement post-editing with three LLMs; Gemini 1.5 Pro, Mistral Large 2.1 and Open AI's GPT 3.5 Turbo.[21] Following Nielsen et al. (2025), our first prompting strategy (P1) includes only the source sentence and the system translation, while our second prompting strategy (P2) includes the translations as well as the entire lexicon for the relevant language pair. For each of these two strategies, we experiment with using the exact prompt proposed in Nielsen et al. (2025) as well as our own prompt construction. All four prompts are listed in full in Table 11 in Section B. For `pap`, we repeat the experiment with both the traditional bilingual lexicon and the GATITOS lexicon (Jones et al., 2023).

[20]For NLLB, as `pap` and `kea` are already supported languages in the pre-trained model, we do not reset the embedding weights for these language tags in the same fashion.

[21]Due to resource limitations, we did not use the paid OpenAI model to post-edit the `pap` Test_{KMT} dataset, which is over ten times as long as the test sets for the other six languages.

ID	XX→eng								eng→XX							
	pap	pov	kea	cri	pre	fab	aoa	all	pap	pov	kea	cri	pre	fab	aoa	all
kreyol-mt	75.1	89.0	94.0	83.1	10.6	11.3	11.0	53.4	66.4	91.8	91.8	80.0	8.38	6.65	8.56	50.5
KMT1	75.4	69.2	**91.1**	73.5	31.8	14.7	**19.9**	**53.7**	68.0	56.4	71.2	64.2	19.4	12.9	17.5	44.2
A. + por embeddings	**75.9**	68.0	89.6	66.3	32.7	15.2	19.0	52.4	66.9	61.4	74.5	45.7	18.7	**14.0**	**17.6**	42.7
B. + por data	75.6	68.7	89.8	67.3	29.7	14.7	19.3	52.2	**67.6**	56.2	70.4	55.2	**21.5**	12.1	**17.6**	42.9
C. + distilled data	73.1	**75.5**	89.7	**80.1**	25.3	13.7	18.0	53.6	65.6	66.6	**84.5**	**75.0**	0.0	0.0	0.0	41.7
D. + A + C	71.8	72.5	86.5	64.5	**36.9**	**15.4**	18.3	52.3	63.0	**71.9**	81.6	52.5	16.6	12.3	15.5	**44.8**
MB1/MB2	76.1	49.6	63.3	33.9	**50.7**	20.8	26.7	**46.2**	**73.1**	32.4	**44.1**	**26.5**	26.4	17.0	**28.3**	**35.4**
A. + por embeddings	**76.4**	50.0	**63.5**	32.9	50.2	20.1	27.3	45.8	**73.1**	33.4	43.1	25.6	27.2	**17.5**	26.0	35.1
B. + por data	75.6	50.4	63.3	34.9	47.7	**22.1**	**27.9**	46.0	71.3	29.6	40.1	21.7	**28.6**	**17.5**	25.2	33.4
C. + distilled data	74.4	50.8	62.2	**36.6**	43.9	20.4	25.2	44.8	71.3	**36.7**	39.1	22.0	23.9	15.3	20.1	32.6
D. + A + B	75.6	**54.0**	63.2	35.9	48.4	19.4	27.1	**46.2**	71.5	29.3	39.5	22.2	24.1	17.1	24.8	32.6
NLLB1/NLLB2	**83.3**	**55.5**	70.5	24.8	35.6	20.4	21.0	**44.4**	77.1	52.5	56.3	28.1	23.4	**18.4**	**24.6**	**40.1**
A. + por embeddings	82.6	51.3	68.2	**27.0**	**39.9**	20.3	20.2	44.2	74.2	49.5	52.5	24.8	24.2	14.9	20.9	37.3
B. + por data	83.1	49.9	68.6	24.0	37.9	**20.7**	**25.1**	44.0	75.5	53.0	56.6	**31.9**	**28.3**	16.2	18.4	40.0
D. + A + B	83.0	49.7	**72.0**	24.6	31.4	20.6	19.5	43.0	**77.3**	**53.8**	**56.7**	26.1	26.0	**18.4**	22.0	40.0

Table 2: Results of fine-tuning experiments (A) initialising language embeddings with por embeddings; (B) adding high-quality por data to training data; (C) using distilled data as training data for pap, kea, pov and cri, and (D) any relevant combinations of the three conditions. Results calculated on Test_{KMT} dataset, using single best checkpoint for each model (as evaluated on validation set). Results in **bold** indicate best results for that language pair out of all experimental settings for that base model; highlighted results are best out of all fine-tuned models (green = beats kreyol-mt baseline).

5 Results and Discussion

In this section, we report and discuss the results of our fine-tuning experiments, model merging and post-editing with LLMs. All results are calculated using the ChrF metric[22] (Popović, 2015) implemented in the SacreBLEU library (Post, 2018).[23]

Fine-tuning We find through initial fine-tuning on our dataset that the best overall models for translation into eng are kreyol-mt fine-tuned for XX-XX translation (KMT1), mBART-50 m2m fine-tuned for XX-eng translation (MB1) and NLLB distilled-1.3B fine-tuned for XX-eng translation (NLLB1). We find the best overall models for translation out of eng are kreyol-mt fine-tuned for XX-XX translation (KMT1), mBART-50 o2m fine-tuned for eng-XX translation (MB2) and NLLB distilled-1.3B fine-tuned for eng-XX translation (NLLB2). For each of these best setups, we then implement our initial set of experiments by retraining each model using por embeddings, por data or distilled data, and then the combinations of the two best settings for each base model.

The results in Table 2 show that different strategies work best for different base models, directions and language pairs – there is no single experimental setting that shows across-the-board advantages. NLLB-based models (NLLB1/NLLB2) show the strongest performance on translation to and from pap, which is not surprising given that this is one of NLLB's supported languages and that the model has seen large amounts of pap data during pretraining. However, using distilled data does not improve the NLLB1/NLLB2 results for pap nor any other language pairs, therefore we exclude the results for this setting. The mBART-50-based models (MB1/MB2) outperform the other fine-tuned models on aoa, fab and pre, except for eng-fab translation. Their high performance on these languages (the lowest-resourced in the set) is likely due to the temperature sampling strategy utilised in our fine-tuning setup for mBART. Conversely, the fine-tuned kreyol-mt model (KMT1) performs better than the other fine-tuned models on kea, pov and cri in both translation directions, particularly when training on distilled data.

Our best model for kea, pov and cri (fine-tuned kreyol-mt) does not beat the kreyol-mt baseline in these languages, so we experiment further with fine-tuning kreyol-mt. We therefore repeat the three experiments while fine-tuning kreyol-mt only for one translation direction at a time (XX-eng or eng-XX), as well as fine-tuning on only the highest-resource languages (cri, pov, kea and pap). To further improve scores, we find each model's best checkpoint for each language pair on the validation set and then use this checkpoint to translate Test_{KMT} for that language pair. Any of these new models which improve on our previous best results for a given language pair are included in Table 13 in Section D, along with the per-language checkpointed scores for the other best models per language pair from Table 2.

[22] Note that we use ChrF but the official Shared Task proceedings uses ChrF++.

[23] nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.5.1

	XX→eng									eng→XX							
ID	pap	pov	kea	cri	pre	fab	aoa	all	ID	pap	pov	kea	cri	pre	fab	aoa	all
kreyol-mt	75.1	89.0	94.0	83.1	10.6	11.3	11.0	53.4		66.4	91.8	91.8	80.0	8.38	6.65	8.56	50.5
H1	73.9	57.4	67.4	36.4	**55.3**	**28.2**	**35.1**	50.5	H4	66.2	67.5	**90.4**	**78.3**	0.0	0.0	0.0	**43.2**
H2	**84.4**	68.3	72.3	37.6	39.5	22.3	21.9	49.6	H5	**77.6**	52.5	57.1	27.3	**27.8**	**18.7**	**25.9**	41.0
H3	76.8	80.9	**93.6**	**82.3**	24.0	12.7	19.1	**55.6**	H6	59.3	**73.7**	76.0	47.0	16.9	10.2	14.3	42.5
H4	73.9	**81.4**	92.9	76.3	22.9	13.7	17.8	54.1									

Table 3: Results of model merging, calculated on Test$_{KMT}$ dataset. Results in **bold** indicate best results for that language pair across all merged models; highlighted results are better than all other fine-tuned models (green = beats kreyol-mt baseline).

For translation into eng, fine-tuning kreyol-mt for XX-eng translation only gave best results for kea-eng and cri-eng (KMT2). Fine-tuning kreyol-mt for XX-XX translation but with distilled data and only with the higher-resource Creoles (pap, pov, kea and kea) improved results for pov-eng translation (KMT3). For translation out of eng, fine-tuning kreyol-mt for eng-XX translation only gave best results for eng-kea and eng-pov, using distilled data for the former (KMT4) and distilled data plus initialisation with por embeddings for the latter (KMT5). Despite these improvements, no models beat kreyol-mt scores for pov, kea and cri in either translation direction; and KMT1C remains our best-performing model for eng-cri.

Model Merging We create a total of six new models by merging different combinations of our fine-tuned and base models. Results across all language pairs and translation directions are displayed in Table 3. To improve performance on the lowest-resource languages (aoa, fab and pre) we first combine the best checkpoints of MB1B (2 checkpoints) and MB1C (3 checkpoints), obtaining the H1 model. For pap-eng we try averaging the last three checkpoints of NLLB1 (H2) and for eng-pap we take the same approach for NLLB2D (H5). For pov, kea and cri, for XX-eng we try averaging the last three checkpoints of KMT2, but find no improvements on our best scores and so exclude this model from our results. For eng-XX we average the last three checkpoints of KMT5 (H6), obtaining a new best-score for eng-pov translation. Finally, we explore whether incorporating the base kreyol-mt model directly in the merging can improve scores, combining the last three checkpoints of KMT2 with kreyol-mt (H3) and the last three checkpoints of KMT1C with kreyol-mt (H4). Our six model merges beat our existing best scores on all language directions except eng-aoa, eng-fab and eng-pre; yet our new best scores for translation from and into kea, pov and cri still do not beat the kreyol-mt baseline.

Post-editing Finally, we take our best models for each language direction and post-edit their Test$_{KMT}$ outputs with different LLMs. We include a full list of results in Table 14 in Section D. In most cases, the LLM-edited outputs are worse than the original system outputs, but we obtain modest improvements for fab-eng, eng-fab, pre-eng, eng-pre and eng-aoa translation. For every translation direction, post-editing with the lexicon gives better results than post-editing without the lexicon, even for aoa which has only a small, hand-crafted lexicon. For pap, we obtain better results using the traditional bilingual lexicon than the GATITOS lexicon, despite the fact that the GATITOS lexicon is over three times larger than the former, potentially indicating that the lexical items included in the former are more useful for this test set domain.

Final models Out of all our finetuning, merging and post-editing experiments, we select the best systems to submit to the Shared Task, reporting the performance of each system on the test set in Table 4. The first submissions are generated by the single best model for XX-eng translation (merged model H3) and eng-XX translation (best overall checkpoint of MB2, a finetuned mBART-50 model).[24] The second submissions are generated by the best models or checkpoints for each individual language pair, except for eng-kea and eng-cri where there is no better model or checkpoint than Submission 1. We also include a third submission for translation directions where the LLM post-editing resulted in improvements on the second submission outputs.

6 Data Contamination

At the end of the Shared Task, Organisers communicated with us that the kreyol-mt model, one of

[24] We selected MB2 because, when evaluated on each language with the best checkpoint per language, it showed the highest average performance across all language directions. However, we realised in hindsight that the best *single* checkpoint across all language pairs was actually from KMT1D.

	XX→eng							eng→XX						
	pap	pov	kea	cri	pre	fab	aoa	pap	pov	kea	cri	pre	fab	aoa
kreyol-mt	75.1	**89.0**	**94.0**	**83.1**	10.6	11.3	11.0	66.4	**91.8**	**91.8**	**80.0**	8.38	6.65	8.56
Sub. 1 (H3/MB2)	76.8	80.9	93.6	82.3	24.0	12.7	19.1	73.1	32.4	44.1	26.5	26.4	17.0	28.3
Sub. 2 (best per LP)	**84.4**	81.4			55.3	28.2	**35.1**	**77.6**	73.7	90.4	78.0	41.7	25.7	33.6
Sub. 3 (Sub. 2 + LLM)					**57.1**	**28.7**						**44.2**	**26.6**	**33.6**

Table 4: ChrF scores for system submissions from best single models per translation direction (Sub. 1), best models per language pair (Sub. 2) and best models per language pair + LLM post-editing (Sub. 3) on the Test$_{\text{KMT}}$ dataset (Bold = best score, green highlight = beats kreyol-mt baseline). Unfortunately, XX-eng model outputs for Submission 2 (grey) were not submitted to the Shared Task due to administrative error.

		XX→eng							eng→XX						
		pap	pov	kea	cri	pre	fab	aoa	pap	pov	kea	cri	pre	fab	aoa
Test$_{\text{KMT-D}}$	kreyol-mt	68.4	42.8	57.9	37.3	6.00	11.0	10.4	60.3	29.7	51.6	27.4	8.93	5.47	9.55
	Submission 1	**67.3**	**50.7**	**61.9**	**39.4**	26.4	21.9	26.7	48.4	27.3	**45.8**	36.0	26.0	**41.2**	**46.2**
	Submission 2	64.4	39.7	-	-	**60.0**	**48.4**	**50.1**	**59.6**	**51.3**	27.4	**40.5**	**26.5**	39.0	31.3
Test$_{\text{Ours}}$	kreyol-mt	39.5	29.8	-	-	-	-	-	38.8	20.1	-	-	-	-	-
	Submission 1	45.8	28.6	-	-	-	-	-	26.9	**44.2**	-	-	-	-	-
	Submission 2	**67.6**	**46.2**	-	-	-	-	-	**49.5**	18.4	-	-	-	-	-

Table 5: Results for kreyol-mt baseline model compared to our Submission 1 and Submission 2 models on Test$_{\text{KMT-D}}$ and Test$_{\text{Ours}}$. Bold = best score, green highlight = beats kreyol-mt baseline.

the specified baseline models for the unconstrained systems track, had been trained on some of the Shared Task public test data; and the intended baseline was kreyol-mt-pubtrain. This explains why kreyol-mt scored so highly on the official test set for certain language pairs (kea, pov and cri), and why our models cannot beat it in these directions despite additional data and modelling efforts.

For our submission, this clarification impacted our experimental baseline and our finetuned or merged models which use kreyol-mt as a base model. This means a substantial proportion of our submissions were affected.[25] To address this, we re-evaluated both the kreyol-mt baseline and our Submission 1 and Submission 2 models[26] on two further test sets:

- A decontaminated version of the KMT test datasets (Test$_{\text{KMT-D}}$) provided by the organisers, with data not seen during training of either kreyol-mt or kreyol-mt-pubtrain (see dataset sizes in Table 10).
- Test$_{\text{Ours}}$, which is made of pap and pov data we collected but did not use for training, including data from domains not seen during training of kreyol-mt (see dataset sizes in Table 7).[27]

The results (Table 5) show that our Submission 1 and Submission 2 models outperform kreyol-mt in 12 out of 14 translation directions (all except pap-eng and eng-pap) on Test$_{\text{KMT-D}}$. On Test$_{\text{Ours}}$, our Submissions beat kreyol-mt in all four translation directions, including pap-eng and eng-pap. These results provide a more realistic picture of the performance of the baseline and our own models on the different language pairs, without inflation on a contaminated test set. Furthermore, kreyol-mt performs considerably worse on the FLORES benchmark (Goyal et al., 2022) for pap and kea (see Appendix C) than on either Test$_{\text{KMT}}$ or Test$_{\text{KMT-D}}$. These results indicate that, aside from the issue of data contamination, the kreyol-mt model seems to be heavily overfitted to KMT-style data and less good at generalising to novel domains. We note that this may have also degraded the quality of our backtranslated training data, since we use three kreyol-mt models to backtranslate monolingual Creole data from different domains into English (see Section 3.2).

[25] Specifically, our finetuned and merged models which used kreyol-mt as a base model included H3, H4, H5 and H6, used for Submission 1 and Submission 2 for several language pairs.

[26] Due to resource and time constraints, we were not able to repeat our LLM-post editing techniques (creating Submission 3) on the new test sets.

[27] We split out this test data *after* synthetically creating parallel data by using kreyol-mt-pubtrain and kreyol-mt-scratch models to backtranslate monolingual data (see Section 3.2). As a result, 13% and 15% of our pap and pov test sets are made up of synthetic data. We also have our own test data for kea and cri (see Table 7) but because a much higher proportion of these splits are synthetic (63% and 100% respectively), we do not evaluate on this data here.

7 Conclusion

Our submissions to the WMT 2025 Creoles MT Systems Subtask utilise a range of known MT techniques, including fine-tuning three pre-trained multilingual translation models on both task data and additional data, merging best models and checkpoints and post-editing system outputs using LLMs. While no single fine-tuning, merging or post-editing strategy emerged as best amongst all language pairs, we observed considerable gains over the baseline KMT model performance on the Test_{KMT} dataset for `pap`, `aoa`, `fab` and `pre` by combining different approaches, including oversampling the lowest-resource languages in the training data via temperature sampling. While some of our results are unreliable due to the fact that Test_{KMT} is contaminated with `kreyol-mt` training data, we demonstrate the robustness of our model's performance using alternative test sets, and show that `kreyol-mt` appears to be overfitted to KMT-style data in general. Future work could explore whether the techniques and strategies we have utilised here to improve performance are also useful for other Creole language pairs and across data from a broader variety of different domains.

Limitations

The official Shared Task test sets for these languages are identical to the test sets which are publicly available on Hugging Face, meaning that the gold labels were available at the point of submission. We ensured that no samples from these test sets were in our own training data. However, before we realised that the official test set would be identical to the public one, we made modelling and design decisions based on performance on the publicly-available test set. For example, we selected the best of the four `kreyol-mt`, `kreyol-mt-pubtrain`, `kreyol-mt-scratch` and `kreyol-mt-pubtrain-scratch` models for forward translation and backward translation of our training data based on their performance on the publicly available test set, both per language and overall. We also selected our models for submission based on their performance on this test set, given that the gold labels were freely available. This biases our model development process towards this particular test set, potentially reducing generalisability or robustness of the overall MT systems and potentially giving us an advantage in the context of the Shared Task.

A key limitation of our work is that our modelling decisions and comparisons were initially guided by the `kreyol-mt` model, which was mistakenly announced as the Shared Task baseline. The organisers later clarified that this model had been trained on portions of the Test_{KMT} set, meaning not only that the baseline we were comparing to was trained on the data we were testing on, but also that our models which use it as a base model are also likely inflated. We address this in Section 6 but reiterate here that the results for our KMT-based models in Table 2, and the results for H3, H4, H5 and H6 in Table 3 and Table 4 are likely inflated.

In addition, `kreyol-mt` was trained using a non-standard tagging scheme, appending `src` language tags to the end of source sentences rather than prepending them as in standard mBART-50. Our models inherit this convention, which may limit comparability with other mBART-based systems.

Acknowledgments

This research was supported by the UK Research and Innovation (UKRI) AI Centre for Doctoral Training in Designing Responsible Natural Language Processing (grant number EP/Y030656/1); the National Science Centre, Poland (2023/49/N/ST6/02691); the EU Horizon Europe Research and Innovation programme (GA No 101070350) and UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (GA No 10052546); the OpenEuroLLM project, co-funded by the Digital Europe Programme (GA No 101195233); and EU Horizon Europe Research and Innovation programme (GA No 101070631) and UK Research and Innovation under the UK government's Horizon Europe funding guarantee (GA No 10039436).

For the purpose of Open Access, the authors have applied a Creative Commons Attribution (CC-BY) public copyright licence to any Author Accepted Manuscript version arising from this submission.

We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Center: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2025/018209. We also thank Bhavitvya Malik of the StatMT group at the University of Edinburgh for sharing his code for fine-tuning NLLB with PyTorch Lightning, which we adapted for our experiments.

References

Vanessa Pinheiro de Araújo and Gabriel Antunes de Araujo. 2013. Um dicionário principense-português. Master's thesis, Faculdade de Filosofia, Letras e Ciências Humanas, University of São Paulo.

Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2024. Iterative translation refinement with large language models. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 181–190, Sheffield, UK. European Association for Machine Translation (EAMT).

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.

Raj Dabre, Aneerav Sukhoo, and Pushpak Bhattacharyya. 2014. Anou tradir: Experiences in building statistical machine translation systems for mauritian languages – creole, English, French. In *Proceedings of the 11th International Conference on Natural Language Processing*, pages 82–88, Goa, India. NLP Association of India.

Ona De Gibert, Robert Pugh, Ali Marashian, Raul Vazquez, Abteen Ebrahimi, Pavel Denisov, Enora Rice, Edward Gow-Smith, Juan Prieto, Melissa Robles, Rubén Manrique, Oscar Moreno, Angel Lino, Rolando Coto-Solano, Aldo Alvarez, Marvin Agüero-Torales, John E. Ortega, Luis Chiruzzo, Arturo Oncevay, and 3 others. 2025. Findings of the AmericasNLP 2025 shared tasks on machine translation, creation of educational material, and translation metrics for indigenous languages of the Americas. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 134–152, Albuquerque, New Mexico. Association for Computational Linguistics.

Abteen Ebrahimi, Manuel Mager, Shruti Rijhwani, Enora Rice, Arturo Oncevay, Claudia Baltazar, María Cortés, Cynthia Montaño, John E. Ortega, Rolando Coto-solano, Hilaria Cruz, Alexis Palmer, and Katharina Kann. 2023. Findings of the AmericasNLP 2023 shared task on machine translation into indigenous languages. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 206–219, Toronto, Canada. Association for Computational Linguistics.

William Falcon and The PyTorch Lightning team. 2019. PyTorch Lightning. `https://zenodo.org/records/3828935`. Accessed: 2025-08-06.

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.

Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vladimir Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. 2024. Arcee's MergeKit: A toolkit for merging large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 477–485, Miami, Florida, US. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Tjerk Hagemeijer, Philippe Maurer-Cecchini, and Armando Zamora Segorbe. 2020. *A Grammar of Fa d'Ambô*. De Gruyter Mouton, Berlin, Boston.

Jonathan Hus and Antonios Anastasopoulos. 2024. Back to school: Translation using grammar books. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20207–20219, Miami, Florida, USA. Association for Computational Linguistics.

Jonathan Hus, Antonios Anastasopoulos, and Nathaniel Krasner. 2025. Machine translation using grammar materials for LLM post-correction. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 92–99, Albuquerque, New Mexico. Association for Computational Linguistics.

Alexander Jones, Isaac Caswell, Orhan Firat, and Ishank Saxena. 2023. GATITOS: Using a new multilingual lexicon for low-resource machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 371–405, Singapore. Association for Computational Linguistics.

David Kamholz, Jonathan Pool, and Susan Colowick. 2014. PanLex: Building a resource for panlingual lexical translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3145–3150, Reykjavik, Iceland. European Language Resources Association (ELRA).

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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, Brussels, Belgium. Association for Computational Linguistics.

Matthew Lamm, Jennimaria Palomaki, Chris Alberti, Daniel Andor, Eunsol Choi, Livio Baldini Soares, and Michael Collins. 2021. QED: A framework and dataset for explanations in question answering. *Transactions of the Association for Computational Linguistics*, 9:790–806.

Heather Lent, Emanuele Bugliarello, Miryam de Lhoneux, Chen Qiu, and Anders Søgaard. 2021. On language models for creoles. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 58–71, Online. Association for Computational Linguistics.

Heather Lent, Emanuele Bugliarello, and Anders Søgaard. 2022. Ancestor-to-creole transfer is not a walk in the park. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pages 68–74, Dublin, Ireland. Association for Computational Linguistics.

Heather Lent, Kushal Tatariya, Raj Dabre, Yiyi Chen, Marcell Fekete, Esther Ploeger, Li Zhou, Ruth-Ann Armstrong, Abee Eijansantos, Catriona Malau, Hans Erik Heje, Ernests Lavrinovics, Diptesh Kanojia, Paul Belony, Marcel Bollmann, Loïc Grobol, Miryam de Lhoneux, Daniel Hershcovich, Michel DeGraff, and 2 others. 2024. CreoleVal: Multilingual multitask benchmarks for creoles. *Transactions of the Association for Computational Linguistics*, 12:950–978.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

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. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Rao Ma, Mengjie Qian, Yassir Fathullah, Siyuan Tang, Mark Gales, and Kate Knill. 2025. Cross-lingual transfer learning for speech translation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 33–43, Albuquerque, New Mexico. Association for Computational Linguistics.

Arya D. McCarthy, Rachel Wicks, Dylan Lewis, Aaron Mueller, Winston Wu, Oliver Adams, Garrett Nicolai, Matt Post, and David Yarowsky. 2020. The Johns Hopkins University Bible corpus: 1600+ tongues for typological exploration. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2884–2892, Marseille, France. European Language Resources Association.

Susanne Maria Michaelis, Philippe Maurer, Martin Haspelmath, and Magnus Huber, editors. 2013. *APiCS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Elizabeth Nielsen, Isaac Rayburn Caswell, Jiaming Luo, and Colin Cherry. 2025. Alligators all around: Mitigating lexical confusion in low-resource machine translation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 206–221, Albuquerque, New Mexico. Association for Computational Linguistics.

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.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53.

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.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International*

Conference on Machine Learning, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Nathaniel Robinson, Raj Dabre, Ammon Shurtz, Rasul Dent, Onenamiyi Onesi, Claire Monroc, Loïc Grobol, Hasan Muhammad, Ashi Garg, Naome Etori, Vijay Murari Tiyyala, Olanrewaju Samuel, Matthew Stutzman, Bismarck Odoom, Sanjeev Khudanpur, Stephen Richardson, and Kenton Murray. 2024. Kreyòl-MT: Building MT for Latin American, Caribbean and colonial African creole languages. 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 3083–3110, Mexico City, Mexico. Association for Computational Linguistics.

Nathaniel Robinson, Cameron Hogan, Nancy Fulda, and David R. Mortensen. 2022. Data-adaptive transfer learning for translation: A case study in Haitian and jamaican. In *Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022)*, pages 35–42, Gyeongju, Republic of Korea. Association for Computational Linguistics.

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*.

Jacqueline Rowe, Edward Gow-Smith, and Mark Hepple. 2025. Limitations of religious data and the importance of the target domain: Towards machine translation for Guinea-Bissau creole. In *Proceedings of the Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2025)*, pages 183–200, Albuquerque, New Mexico, U.S.A. Association for Computational Linguistics.

Aditya Siddhant, Ankur Bapna, Orhan Firat, Yuan Cao, Mia Xu Chen, Isaac Caswell, and Xavier Garcia. 2022. Towards the next 1000 languages in multilingual machine translation: Exploring the synergy between supervised and self-supervised learning. *Preprint*, arXiv:2201.03110.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.

Jörg Tiedemann. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23965–23998. PMLR.

Wenda Xu, Daniel Deutsch, Mara Finkelstein, Juraj Juraska, Biao Zhang, Zhongtao Liu, William Yang Wang, Lei Li, and Markus Freitag. 2024. LLMRefine: Pinpointing and refining large language models via fine-grained actionable feedback. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1429–1445, Mexico City, Mexico. Association for Computational Linguistics.

Fei Yuan, Yinquan Lu, Wenhao Zhu, Lingpeng Kong, Lei Li, Yu Qiao, and Jingjing Xu. 2023. Lego-MT: Learning detachable models for massively multilingual machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11518–11533, Toronto, Canada. Association for Computational Linguistics.

Jaume Zaragoza-Bernabeu, Gema Ramírez-Sánchez, Marta Bañón, and Sergio Ortiz Rojas. 2022. Bicleaner AI: Bicleaner goes neural. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 824–831, Marseille, France. European Language Resources Association.

A Data Collection

Data Type	L1	L2	Description	Source	No. items
Parallel	por	eng	Tatoeba Translation Challenge	Tiedemann (2020)	112,376
	pap	eng	Bible data	Bible.com	29,367
	pap	eng	Watchtower Series†	The Jehovah's Witnesses	4,275
	pap	eng	Online Random Sentence Generator	Sapaté, na bo sapatu!	5,936
	pov	eng	Bible data	Bible.com	29,876
	pov	eng	Watchtower Series	The Jehovah's Witnesses	8,685
	pov	por	Bilingual dictionary gloss sentences	Dicionário Bilíngue	1,603
	pov	eng	Article on internet access	Open Global Rights	18
	kea	eng	Watchtower Series	The Jehovah's Witnesses	4,273
	fab	eng	Translated stories	Hagemeijer et al. (2020)	430
	pre	por	Bilingual dictionary gloss sentences	Araújo and Araujo (2013)	81
	aoa	eng	IMT Vault sentences	IMT Vault	46
Monolingual	pov	-	Monolingual dictionary gloss sentences	Amarilio Da Mata*	6,930
	pov	-	Documentary Subtitles	Language and Society in Guinea-Bissau	254
	pov	-	Song Lyrics	Tino Trimó via Letras	177
	pap	-	Song Lyrics	Lyrics Translate‡	5,803
	kea	-	School Textbook	Língua e Cultura Cabo-verdiana 10º ano	2,688
	kea	-	Blogposts	Odju d'Agu	2,357
	kea	-	Song Lyrics	Cesária Évora via Letras	2,317
	cri	-	Watchtower Magazine	The Jehovah's Witnesses	1,554
Lexical	cri	por	Bilingual Lexicon	Dicionário livre santome/português	4,929
	pap	eng	Bilingual Lexicon	GATITOS	4,001
	pap	eng	Bilingual Lexicon	Parleremo	1,307
	pov	por	Bilingual Lexicon	Dicionário Bilíngue	1,983
	kea	eng	Bilingual Lexicon	Disonariu Kabuverdianu	1,763
	pre	por	Bilingual Lexicon	Araújo and Araujo (2013)	1,684
	fab	eng	Bilingual Lexicon	Hagemeijer et al. (2020)	473
	aoa	eng	Bilingual Lexicon	IMT Vault§	68

†Corsou dialect.
‡ We collected only lyrics which were tagged exclusively with the pap language tag and no other language tags.
* This is an unpublished manuscript shared privately with the lead author. Lexical items and their definitions were made into full sentences for the purposes of model training by appending each lexical item + 'i' *(is)* + definition.
§ For aoa, we could not find an official lexicon and therefore manually curated a small set of parallel lexical items using the word-aligned entries in the IMT Vault resource.

Table 6: Raw data sources and sizes. Rows shaded in gray were collected too late in the Shared Task period for us to use for model training, but are included here in case useful for future research.

	Train	$\text{Train}_{\text{Clean}}$	Validation	Test	All†	avg. length
pov	44,275	43,419	1,000	1,000	45,419	26.2
pap	43,381	40,850	1,000	1,000	42,850	23.3
kea	8,501	8,099	1,000	1,000	10,099	27.5
cri	1,244	1,218	155	155	1,528	14.4
aoa	0	0	0	0	0	0
fab	0	0	0	0	0	0
pre	0	0	0	0	0	0

†Calculated using cleaned training data.

Table 7: Numbers of parallel sentences for each language pair from **our** data, ordered from highest to lowest resourced. For training data, we show the numbers of raw and cleaned sentences (e.g. after pre-processing). Average length is calculated as average number of words per sentence across all data splits.

	Train	$\text{Train}_{\text{Clean}}$	Validation	Test	All†	avg. length
pap	65,094	64,983	85	1,967	67,035	22.1
kea	1,470	1,340	84	163	1,587	16.6
pov	389	284	27	33	344	5.8
cri	209	155	34	33	222	6.0
pre	147	105	36	36	177	5.8
fab	109	61	31	38	130	5.4
aoa	99	71	35	39	145	6.5

†Calculated using cleaned training data.

Table 8: Numbers of parallel sentences for each language pair from **Organiser-Provided** data, ordered from highest to lowest resourced. For training data, we show the numbers of raw and cleaned sentences (e.g. after pre-processing). Average length is calculated as average number of words per sentence across all data splits.

	Train	$\text{Train}_{\text{Clean}}$	Validation	Test	All†	avg. length
pap	108,475	105,698	1,085	2,967	109,750	22.6
pov	44,664	43,701	1,027	1,033	45,761	26.1
kea	9,971	9,439	1,084	1,163	11,686	26.0
cri	1,453	1,375	189	188	1,752	13.4
pre	147	105	36	36	177	5.8
fab	109	61	31	38	130	5.4
aoa	99	71	35	39	145	6.5

†Calculated using cleaned training data.

Table 9: Numbers of parallel sentences for each language pair from **our and Organiser-Provided** data, ordered from highest to lowest resourced. For training data, we show the numbers of raw and cleaned sentences (e.g. after pre-processing). Average length is calculated as average number of words per sentence across all data splits.

	Test	avg. length
pap	1,896	17.9
pov	23	2.9
kea	34	15.2
cri	33	4.8
pre	36	3.7
fab	34	6.8
aoa	35	6.2

Table 10: Numbers of parallel sentences for each language pair from the **Decontaminated Organiser-Provided Test set**, ordered from highest to lowest resourced. Average length is calculated as average number of words per sentence across all data splits.

B Prompts used for LLM post-editing

Condition	Nielsen et al. 2025	Ours
1. Post-editing without lexicon	**P1A:** You are asked to edit the following translation from {src_code} into {tgt_code}. The proposed translation is high-quality, but may have some incorrect words. Please output only the translation of the text without any other explanation. {src_code}: {source} {tgt_code}: {model_translation}	**P1B:** You are given a source sentence and a translation. Improve the translation from {src_code} into {tgt_code}. You must return ONLY the corrected translation sentence, without explanation or extra text. Source: {source} Translation: {model_translation}
2. Post-editing with lexicon	**P2A:** You are asked to edit the following translation from {src_code} into {tgt_code}. The proposed translation is high-quality, but may have some incorrect words. Note the following translations: Lexicon: {lexicon_str} Please output only the translation of the text without any other explanation. {src_code}: {source} {tgt_code}: {model_translation}	**P2B:** You are given a source sentence, a translation and a lexicon. Improve the translation from {src_code} into {tgt_code}. You must return ONLY the corrected translation sentence, without explanation or extra text. Source: {source} Translation: {model_translation} Lexicon: {lexicon_str}

Table 11: Prompts used in LLM post-editing experiments.

C FLORES Evaluation

Model	**pap→eng**		**eng→pap**		**kea→eng**		**eng→kea**	
	Test_{KMT}	FLORES	Test_{KMT}	FLORES	Test_{KMT}	FLORES	Test_{KMT}	FLORES
kreyol-mt-pubtrain	79.84	54.39	69.94	60.14	80.66	45.65	52.54	52.16
kreyol-mt	75.10	63.12	66.39	57.27	93.94	55.46	91.76	52.33
kreyol-mt-scratch-pubtrain	74.68	47.17	69.36	55.54	70.23	37.22	49.46	46.98
kreyol-mt -scratch	71.82	60.73	67.19	55.06	89.85	50.83	81.67	49.04
nllb-200-distilled-600M	46.50	59.18	53.18	50.09	59.36	63.04	38.27	41.67
nllb-200-1.3B	58.40	68.88	56.58	55.08	62.68	65.86	41.09	43.02
nllb-200-distilled-1.3B	55.30	69.20	58.02	55.40	59.28	64.89	39.75	42.09
nllb-200-3.3B	60.90	69.16	58.78	55.66	63.69	67.46	43.92	45.76

Table 12: ChrF scores for each `kreyol-mt` model across language directions, evaluated on both Test_{KMT} and FLORES test sets.

D Model Results

		kreyol-mt	kreyol-mt-pubtrain	Ours Best	Model ID	Base model	Fine-tuning Direction	Additional setup
XX-eng	pap	75.1	79.8	83.3	NLLB1$_{pap}$	NLLB 1.3B	XX-eng	-
	kea	94.0	80.7	92.3	KMT2$_{kea}$	KMT	XX-eng	-
	pov	87.8	63.4	78.4	KMT3$_{pov}$	KMT	XX-XX	distilled data + HRLs Only
	aoa	10.9	17.0	34.8	MB1C$_{aoa}$	mBART-50 m2m	XX-eng	distilled data
	cri	83.1	31.7	80.5	KMT2$_{cri}$	KMT	XX-eng	-
	fab	11.3	13.7	27.9	MB1B$_{fab}$	mBART-50 m2m	XX-eng	por data
	pre	10.6	16.2	55.0	MB1B$_{pre}$	mBART-50 m2m	XX-eng	por data
	all	53.2	43.2	55.0	KMT2$_{all}$	KMT	XX-eng	-
eng-XX	pap	66.4	70.0	77.3	NLLB2D$_{pap}$	NLLB15 1.3B	eng-XX	por embeddings + por data
	kea	91.8	52.5	86.2	KMT4$_{kea}$	KMT	eng-XX	distilled data
	pov	91.8	51.6	72.8	KMT5$_{pov}$	KMT	eng-XX	por embeddings + distilled data
	aoa	8.6	13.6	33.6	MB2B$_{aoa}$	mBART02m	eng-XX	por data
	cri	80.0	32.1	78.2	KMT1C$_{cri}$	KMT	XX-XX	distilled data
	fab	6.7	9.3	25.9	MB2$_{fab}$	mBART-50 o2m	eng-XX	-
	pre	8.38	10.7	41.7	MB2$_{pre}$	mBART-50 o2m	eng-XX	-
	all	50.5	34.3	46.1	MB2$_{all}$	mBART-50 o2m	eng-XX	-

Table 13: Settings and results of best-performing model checkpoints for each language. Results are calculated on Test$_{KMT}$ dataset, using the best model checkpoint per language pair based on performance on the validation dataset, as indicated with subscript. For evaluation of all translation directions, we report the models with the best average scores using the best checkpoints for each language pair. New models not previously included in Table 2 are highlighted in gray. Green = beats kreyol-mt and kreyol-mt-pubtrain baselines.

		XX→eng								eng→XX							
	Prompt	pap	kea	pov	aoa	cri	fab	pre	all	pap	kea	pov	aoa	cri	fab	pre	all
Submission 2 models	-	84.4	93.6	80.9	35.1	82.3	28.2	55.3	65.7	77.6	90.4	73.7	33.6	78.0	25.9	41.7	60.1
GPT 3.5 Turbo	P1A	–	76.2	54.0	33.4	50.4	25.2	46.5	47.6	–	74.0	46.2	30.9	56.0	24.9	34.8	44.5
	P1B	–	74.2	55.1	31.5	51.1	25.4	47.4	47.5	–	68.7	41.6	29.6	50.3	25.2	32.4	41.3
	P2A	–	**87.1**	**69.2**	31.2	**78.3**	**28.7**	**51.7**	**57.7**	–	**83.2**	**59.8**	32.5	**75.2**	**25.7**	**41.8**	**53.0**
	P2B	–	81.1	63.5	**32.3**	62.9	24.6	46.6	51.8	–	75.2	55.8	**33.2**	67.0	25.1	40.4	49.5
Mistral Large 2.1	P1A	79.4	84.0	57.2	31.6	57.2	26.2	48.3	54.9	71.5	82.6	47.6	31.6	67.0	23.5	36.0	51.4
	P1B	78.1	81.1	56.2	33.0	55.6	25.2	44.3	53.4	70.6	79.0	46.8	31.5	61.3	25.0	36.7	50.1
	P2A	**83.1**	**91.3**	**79.7**	29.8	**66.1**	24.0	**57.1**	**61.6**	74.1	**88.9**	**61.6**	32.1	64.2	25.8	**42.4**	55.6
	P2B	81.7	85.7	68.6	**34.2**	57.8	**28.1**	51.1	58.2	**76.0**	88.3	58.9	**32.8**	**74.7**	**26.6**	**41.9**	**57.9**
	P2A$_{GAT}$	71.7	–	–	–	–	–	–	–	59.5	–	–	–	–	–	–	–
	P2B$_{GAT}$	70.8	–	–	–	–	–	–	–	66.2	–	–	–	–	–	–	–
Gemini 1.5 Pro	P1A	83.1	84.3	58.4	30.7	52.0	26.6	50.7	55.1	74.0	75.5	46.3	24.3	46.1	22.8	29.2	45.5
	P1B	78.8	75.1	47.8	27.5	44.2	23.7	40.0	48.2	70.0	63.5	38.4	24.5	37.7	25.1	26.8	40.8
	P2A	**83.2**	**86.0**	**66.3**	**32.7**	**57.5**	**27.3**	**53.8**	**58.1**	**75.2**	79.2	48.7	30.5	49.8	**26.0**	**44.2**	50.6
	P2B	81.7	84.4	57.2	31.4	48.1	26.4	45.0	53.4	74.0	**82.7**	**52.7**	**33.6**	**62.2**	25.8	41.3	**53.2**
	P2A$_{GAT}$	82.4	–	–	–	–	–	–	–	73.8	–	–	–	–	–	–	–
	P2B$_{GAT}$	80.1	–	–	–	–	–	–	–	72.0	–	–	–	–	–	–	–

Table 14: Results from post-editing best model outputs with three LLMs. P1 is post-editing without lexicon and P2 is post-editing with lexicon (see Table 11). Baseline scores are from models of Submission 2 for each language pair (Table 4). Results in **bold** are best results for each LLM for each language pair; highlighted results = best out of all LLMs (green = also beats Submission 2 baselines). We do not apply post-editing for GPT 3.5 Turbo for pap (which has an extremely large test set) due to resource constraints.

Submitted models Table 15 documents which models we use to generate our Shared Task submissions:

- For Submission 1, we select the single best model for XX-eng translation (H3) and eng-XX translation (best overall checkpoint of MB2). We selected MB2 because, when evaluated on each language with the best checkpoint per language, it showed the highest average performance across all language directions. However, we realised in hindsight that the best *single* checkpoint across all language pairs was actually from KMT1D.
- For Submission 2, where a model has multiple checkpoints we submit the best checkpoint for that language pair, as indicated with subscripts (except for eng-kea and eng-cri where there is no better model or checkpoint than Submission 1).
- For Submission 3 we submit the best system outputs after post-editing with LLMs when this showed improvements on Submission 2. We indicate which LLM and which prompting strategy (see Table 11) was applied in parentheses.

Due to administrative error, our Submission 2 models for the XX-eng direction were not submitted to the official Shared Task.

		Sub. 1	Sub. 2	Sub. 3
XX-eng	pap	H3	H2	
	pov	H3	H4	
	kea	H3	H3	
	cri	H3	H3	
	pre	H3	H1	+ Mistral (P2A)
	fab	H3	H1	+ GPT (P2A)
	aoa	H3	H1	
eng-XX	pap	MB2	H5	
	pov	MB2	H6	
	kea	MB2	H4	
	cri	MB2	H4	
	pre	MB2	MB2_{pre}	+ Gemini (P2A)
	fab	MB2	MB2_{fab}	+ Mistral (P2B)
	aoa	MB2	MB2B_{aoa}	+ Gemini (P2B)

Table 15: Model IDs for final system submissions.

E Fine-tuning Hyperparameters

KMT & NLLB We fine-tune KMT & NLLB models using PyTorch Lightning (Falcon and team, 2019) on a single GH200 GPU (bf16). We set the batch size to 32, use the Adam optimizer (Kingma and Ba, 2015) with a learning rate $5e-5$, a warm-up phase of 500 updates and maximum training length of 30 epochs. The model performance is validated using ChrF every 5,000 steps, early stopping after three consecutive validations with no improvement in ChrF score.

mBART We fine-tune mBART-50 using fairseq (Ott et al., 2019) with a multi-gpu (4 A100 GPUs, fP1A6). The data loader has used temperature-based sampling ($\tau = 2$). We set the batch size to maximum of 1024 tokens, use the Adam optimizer with a learning rate $3e{-}5$, a warm-up phase of 2500 updates and maximum training length of 40,000 updates. Moreover, we applied label smoothing with $\epsilon_{\text{ls}} = 0.2$, dropout of 0.3, and attention dropout of 0.1. The three best checkpoints were retained according to validation performance (based on the validation loss value), with early stopping after 10 validation intervals.