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

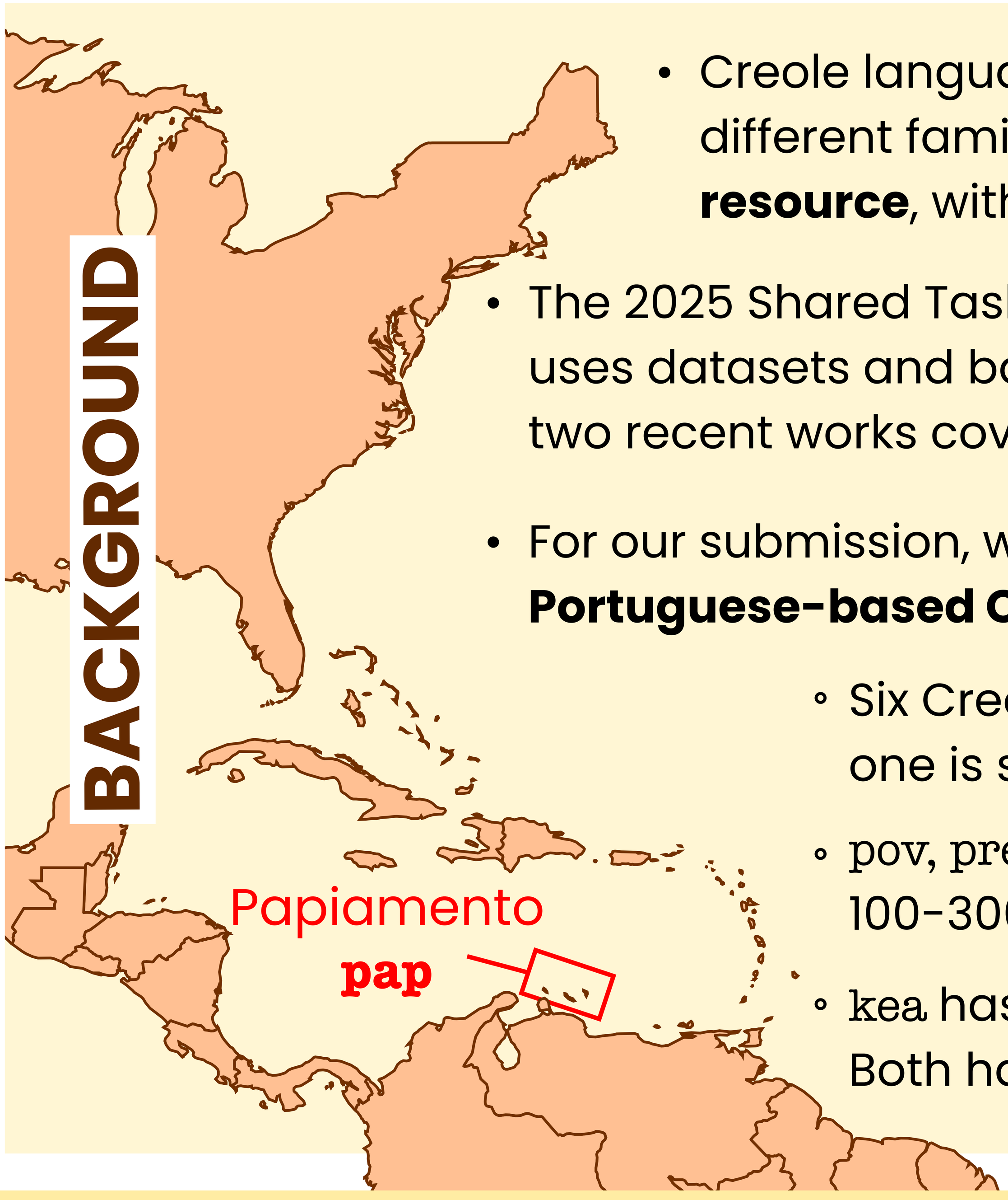
Jacqueline Rowe, Ona de Gibert, Mateusz Klimaszewski, Coleman Haley, Alexandra Birch, Yves Scherrer

Designing Responsible Natural Language Processing
UNIVERSITY OF HELSINKI

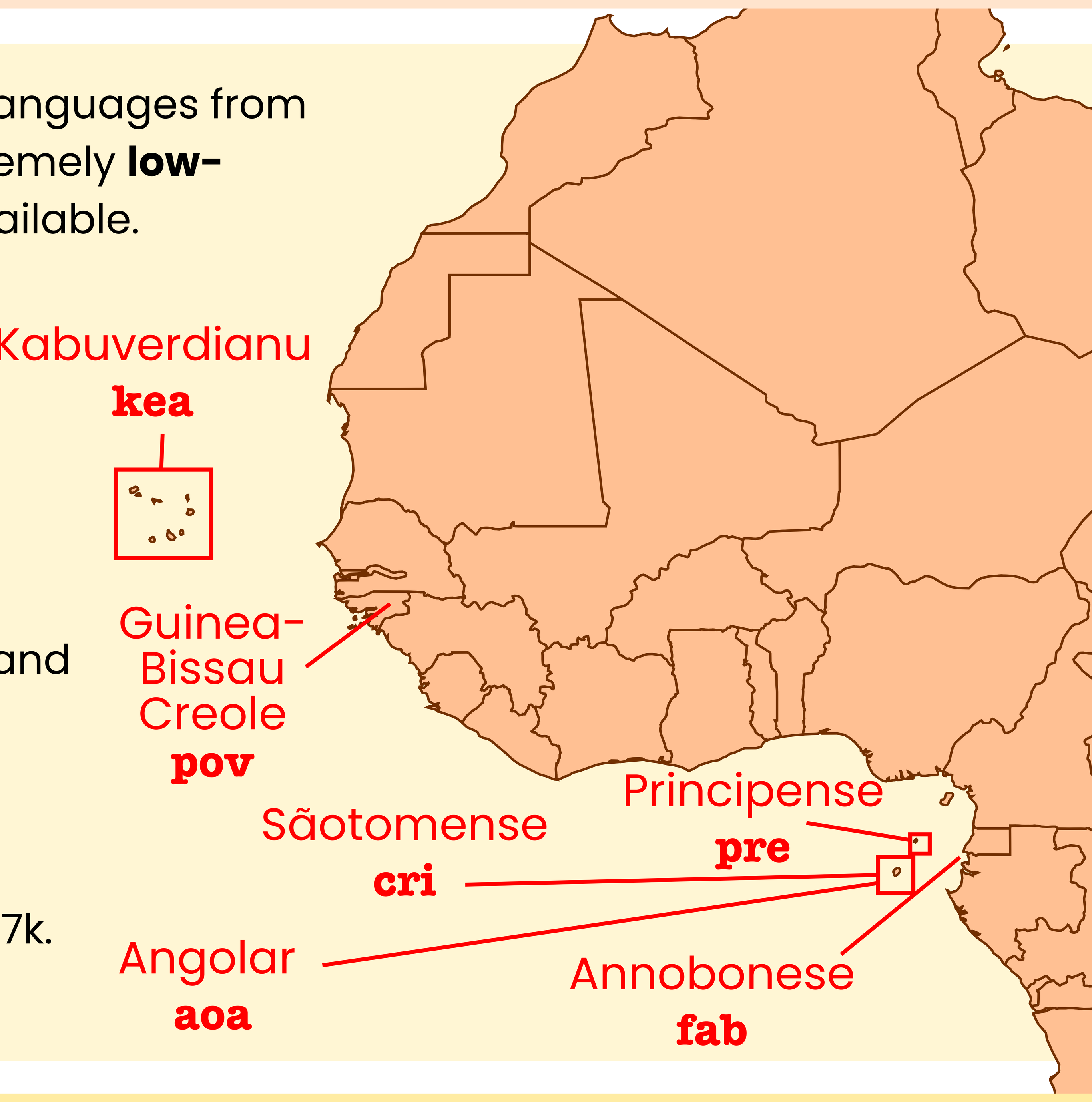
This research was funded in part by the UKRI AI Centre for Doctoral Training in Responsible and Trustworthy in-the-world Natural Language Processing (grant ref: EP/Y030656/1).

We present our submissions to the WMT 2025 Creole Translation Shared Task. For seven Portuguese-based Creole languages, we augment the Task datasets with extra parallel and monolingual data in four languages. We leverage a range of low-resource MT techniques including back-translation, data distillation and model merging, as well as cross-lingual transfer from Portuguese data and embedding weights.

BACKGROUND



- Creole languages arise through contact between two or more languages from different families over a short space of time. They are often extremely **low-resource**, with few datasets, NLP tools or translation systems available.
- The 2025 Shared Task on Machine Translation for Creole Languages uses datasets and baseline models (**KreyòlMT** and **CreoleVal**) from two recent works covering over 50 Creole languages between them [1].
- For our submission, we focus on translation between **seven Portuguese-based Creoles** and English (eng):
  - Six Creoles are spoken in West and Central Africa (right) and one is spoken in the Dutch Caribbean (left).
  - pov, pre, aoa, cri and fab are extremely low-resource, with 100-300 parallel sentences provided by the organisers
  - kea has 1.5k organiser-provided sentences and pap has 67k. Both have featured in other MT and NLP modelling efforts.



SYSTEM DESIGN

1. **Collect Data** (Parallel, Monolingual and bilingual lexicons)
2. **Backtranslate** monolingual data / **Forward translate** some parallel data
3. **Combine, clean and re-split datasets**
4. **Finetune (FT) 3 pre-trained multilingual translation models**: KreyòlMT (KMT), mBART, NLLB
5. **Find best FT setup for each base model**
   - which size of base model?
   - FT for specific translation directions? (m2m, m2o, o2m)
   - use distilled data instead of normal data?
   - with high-quality Portuguese data in training?
   - initialise Creole weights with Portuguese embeddings?
   - *evaluate using best checkpoints per language, as per validation sets*
6. **Merge** best checkpoints together
7. Prompt LLMs to **post-edit** system outputs, including bilingual lexicons in prompt context (MISTRAL AI_, OpenAI, Gemini)
8. **Select best models for submissions:**
   1. Best XX-eng/eng-XX models
   2. Best models per language pair (LP)
   3. Best post-edited systems

DATA

Numbers of parallel sents with eng in each Creole*

Creole	Sentences
pap	110k
pov	46k
kea	12k
cri	2k
pre	0.2k
fab	0.1k
aoa	0.1k

*including all train/val/test splits and both KreyòlMT data and our data (including back-translations).

RESULTS

1. Different base models and FT settings benefit different LPs

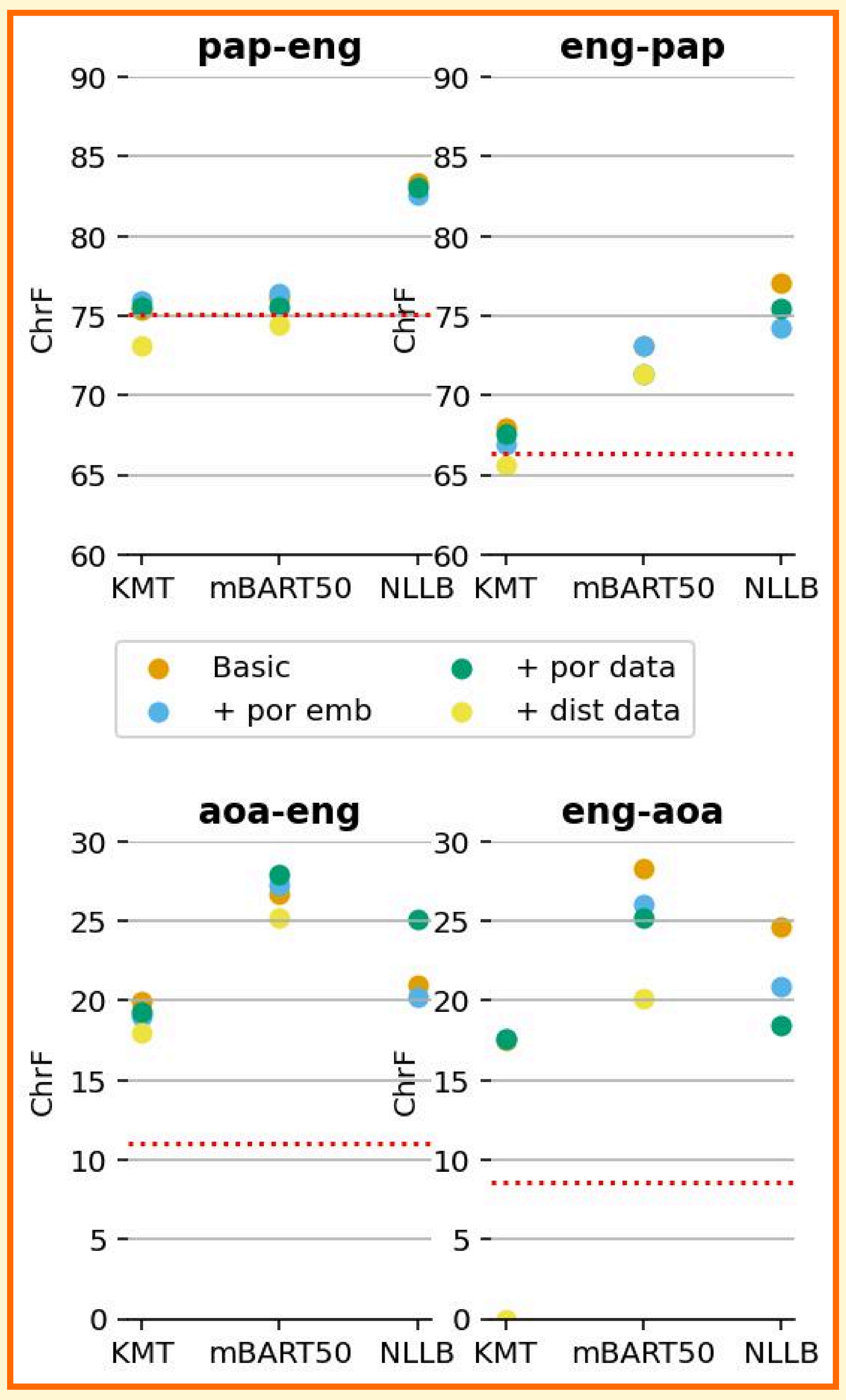


Fig 1: Results of different fine-tuned based models (basic), trained with (1) initialising Creoles with Portuguese embeddings, (2) adding por data and (3) using distilled data). Results shown for pap (top) and aoa (bottom). KreyòlMT baselines in red.

2. Model merging and LLM post-editing with lexicons boosts performance on some LPs

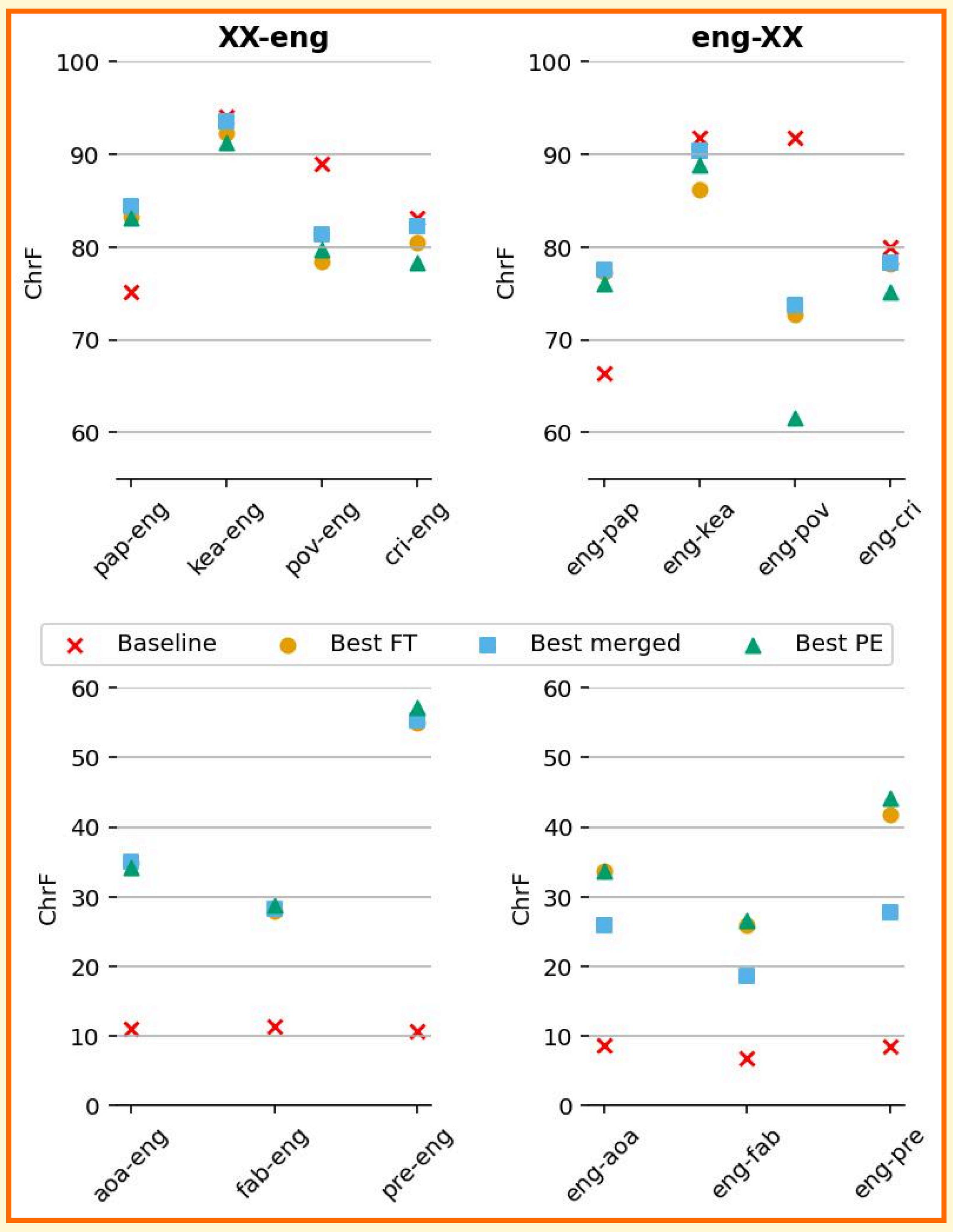


Fig 2: Results of our best finetuned model checkpoints (FT), our best merged models, and our best post-edited (PE) system outputs on high-resource (top) and low-resource (bottom) LPs on KreyòlMT test set.

3. Baseline model is overfitted to 3 LPs

None of our models could beat the KreyòlMT baseline model scores on the KreyòlMT test set for kea, pov or cri, which were suspiciously high. When we re-evaluate on our own test set (larger and more diverse), our models beat the baseline model in all LPs but one.

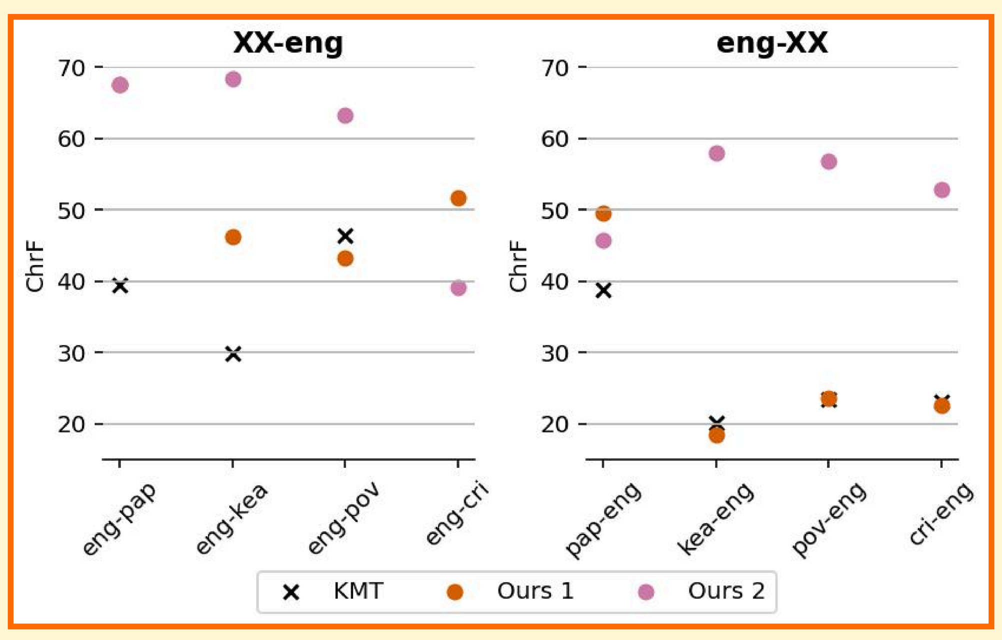


Fig. 3: Our models and baseline KreyòlMT model evaluated on our test set. Ours 1: best finetuned or merged models per LP on KMT test set. Ours 2: best eng-XX and XX-eng models on our test set.

[1] Lent, H., Tatariya, K., Dabre, R., Chen, Y., Fekete, M., Ploeger, E., Zhou, L., Armstrong, R.A., Eijansantos, A., Malau, C. and Heje, H.E., 2024. CreoleVal: Multilingual multitask benchmarks for creoles. Transactions of the Association for Computational Linguistics, 12, pp.950-978; Robinson, N.R., Dabre, R., Shurtz, A., Dent, R., Onesi, O., Monroc, C.B., Grobol, L., Muhammad, H., Garg, A., Etori, N.A. and Tiyyala, V.M., 2024. Kreyòl-MT: Building MT for Latin American, Caribbean and Colonial African Creole Languages. arXiv preprint arXiv:2405.05376.