Language verY Rare for All

Ibrahim Merad[1], Amos Wolf[2], Ziad Mazzawi[2], Yannick Léo[1,2]

[1]Kaukana Ventures, [2]Emerton Data,

Abstract

In the quest to overcome language barriers, encoder-decoder models like NLLB have expanded machine translation to rare languages, with some models (e.g., NLLB 1.3B) even trainable on a single GPU. While general-purpose LLMs perform well in translation, open LLMs prove highly competitive when fine-tuned for specific tasks involving unknown corpora. We introduce LYRA (Language verY Rare for All), a novel approach that combines open LLM fine-tuning, retrieval-augmented generation (RAG), and transfer learning from related high-resource languages. This study is exclusively focused on single-GPU training to facilitate ease of adoption. Our study focuses on two-way translation between French and Monégasque — a rare language unsupported by existing translation tools due to limited corpus availability. Our results demonstrate LYRA's effectiveness, frequently surpassing and consistently matching state-of-the-art encoder-decoder models in rare language translation.

1 Introduction

Machine translation has come a long way since its inception in the 1940s. The methodology evolved from the initial rule-based approach (Hutchins, 1986, 1997) to statistical machine translation (Brown et al., 1993; Koehn, 2009) and most recently adopted neural systems as the defacto approach yielding superior results (Bahdanau, 2014; Cho, 2014). An important breakthrough occurred with the advent of Transformers (Vaswani, 2017) whose attention-based architecture did not only allow for better translation but paved the way for an NLP revolution through LLMs (Brown, 2020; Radford, 2018; Minaee et al., 2024). The considerable progress observed on a wide range of NLP tasks is the combined result of the ingenuous Transformer neural architecture, the availability of large GPU compute resources and macroscopic amounts of training data. However, the uneven data amounts between different languages translate to varying performances on NLP tasks (Joshi et al., 2020; Blasi et al., 2022), including machine translation. Thus, contrary to widespread languages for which large text corpora are available including parallel data, lesser known languages suffer from data scarcity which makes it difficult to train deep learning models (Zhang and Zong, 2020). Moreover, compensating this inequality by obtaining data for low resource languages is expensive and logistically challenging (Nekoto et al., 2020; Kuwanto et al., 2023; Orife et al., 2020).

This work is concerned with training a neural machine translator between the French and Monégasque language. A very low resource language only spoken by around 5,000 people to date in the Principality of Monaco and which, to our knowledge, remains uncovered by any neural machine translator. We take on the task of creating a parallel French-Monégasque dataset enabling the training of translators and language models on this language. We finetune multiple models on this task and present our methodology called LYRA allowing to optimize results with limited data (about 10K parallel sentences and a dictionary).

2 Related works

Given the challenge it poses, the low-resource setting has received much attention in the literature (Haddow et al., 2022; Hedderich et al., 2020; Magueresse et al., 2020). The proposed strategies include targeted data gathering (Hasan et al., 2020), exploiting monolingual data (Gibadullin et al., 2019), backtranslation (Sennrich, 2015), transfer learning (Dabre et al., 2020; Zoph et al., 2016) and multilingual models (Johnson et al., 2017).

The most notable effort towards a model with high language coverage is NLLB (Costa-jussà et al., 2022) (No Language Left Behind). The latter translator was trained for pairs among over 200 dif-

Proceedings of the First Workshop on Language Models for Low-Resource Languages (LoResLM 2025), pages 166–174
January 20, 2025.

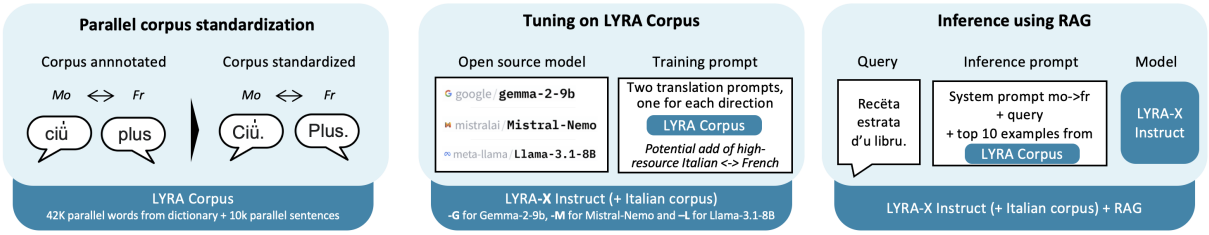

Figure 1: Illustration of our method for building LYRA.

ferent languages using a Sparsely Gated Mixture of Experts architecture. For this purpose, the authors created the Flores-200 dataset consisting of 3000 parallel sentences establishing a benchmark for multilingual machine translation. However, this effort did not include the Monégasque language.

While NLLB uses an encoder-decoder architecture specifically intended for translation, decoder-only models also reached competitive performance on multiple tasks including translation (Hendy et al., 2023; Wei et al., 2022; Ouyang et al., 2022). This motivated works to improve results with such models (Xu et al., 2024; Yang et al., 2023; Alves et al., 2024) since they offer a far more interesting option due to their higher flexibility and impressive multitasking abilities (Reynolds and McDonell, 2021; Kojima et al., 2022; Perez et al., 2021). Moreover, decoder-only models can leverage strategies like RAG to improve performance and enjoy greater attention in the literature leading to faster progress. Finally, these models hold the same potential for multilingual translation and transfer learning. Nonetheless, these references did not consider low-resource languages.

Most recently, both model types were combined by GenTranslate (Hu et al., 2024) which uses a Seq2Seq model to sample translations that are fed into an LLM to combine them into an improved answer. Note however that this work assumes high compute resources with multiple GPUs.

In this work, NLLB as well as a few open LLMs are finetuned using LYRA on a newly created French-Monégasque dataset using only a single GPU machine. We compare their performances on this translation task and showcase the benefits of LYRA in the low-resource setting.

3 Data

Since we are unaware of any preexisting parallel corpus involving Monégasque, we created a French-Monégasque dataset using OCR from a few sources including: A French-Monégasque dictionary, a Monégasque grammar book, as well as a few literary works available in both languages. These include works such as the poem collection "Lettres de mon moulin", the play "Antigone" and some Tintin comics. The acquired inputs were later combined into parallel entries via manual annotation.

The dataset contains a total number of 10,794 parallel French-Monégasque sentences in addition to 42,698 entries from the dictionary and the grammar book which includes verb conjugations and proverbs. The test set was constituted by selecting sentences with high quality translation in order to ensure a reliable basis for evaluation.

The fact that this unique existing dataset has under 100K pairs makes the Monégasque language a very low resource language based on the conventions adopted by Costa-jussà et al. (2022).

4 Methods

The LYRA methodology, illustrated on Figure 1, aims to maximize translation quality in the low data context using three main strategies.

Leveraging related high-resource languages Previous works demonstrated the benefits of knowledge transfer in multilingual neural machine translation (Dabre et al., 2020; Zoph et al., 2016; Maimaiti et al., 2019). In order to take advantage of this phenomenon, we perform a preliminary finetuning phase on translation between French and Italian, which is a high resource language pair, before finetuning on French-Monégasque translation. The idea is to exploit the grammatical similarity between Monégasque and Italian. Thus, in the preliminary phase, the model learns to transition between French and Italian-like grammatical structures on plentiful data which facilitates the subsequent finetuning on French-Monégasque translation.

Data standardization As often emphasized, training models for NLP applications considerably depends on data quality to achieve high performance (Tokpanov et al., 2024; Hoffmann et al.,

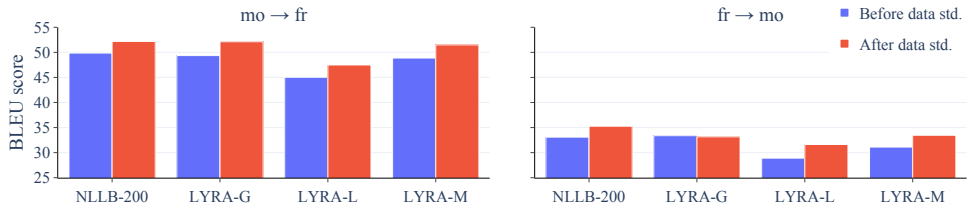

Figure 2: Comparison of models' translation performance in both directions in terms of BLEU scores before and after data standardization. The latter uniformly improves translation performance across all models.

2022; Rae et al., 2021). This aspect is all the more important when data is scarce. We measure the impact of careful data curation in the current setting by training the candidate models on two versions of the French-Monégasque dataset. The initial raw version featured some issues of inconsistent capitalization and punctuation and used various quotation marks. The impact of these details on downstream performance should not be underestimated since they can confuse the model by causing irregular tokenization.

Considering the potential performance gain, we invest the effort of standardizing the sentences in the first version of the dataset to fix these issues and obtain a curated second version.

Retrieval Augmented Generation For decoder-only models, the training data can be used to improve test-time performance by including the most similar sentence or word pairs into the prompt. Note that this is akin to few-shot prompting but using embeddings to retrieve the most similar examples. Since the Monégasque language is unknown to the available embedding models, the French parts are used to generate an embedding for each instance. For this purpose, words and sentences are encoded using a high performing model on French retrieval tasks. The latter is available on the HuggingFace Hub under the reference `BAAI/bge-multilingual-gemma2`. Retrieval of the nearest neighbors is then carried out based on cosine similarity. The number of retrievals is fixed to 10 instances.

5 Experiments

The impact of each strategy on translation quality is evaluated by testing them sequentially. The effect of data standardization is measured prior to testing the other strategies. Performance is measured using the BLEU score (Papineni et al., 2002) as well as METEOR which is more correlated with human assessment (Banerjee and Lavie, 2005). We also provide evaluations using the chrF++ metric (Popović, 2015) in Appendix A.

Models The focus is set on single-GPU training to make the experiments more relevant for the low resource context. We fine-tune some high-performing models on French-Monégasque translation and assess the performance gains from each strategy. The distilled model `nllb-200-distilled-1.3B` was chosen as a representative of the NLLB encoder-decoder model family since it outperforms the 3B model and reaches close performance to the original 54B model at much lower computational costs (Costa-jussà et al., 2022). As for decoder-only models, the candidates are the public LLMs : `Llama-3.1-8B` (Dubey et al., 2024) (LYRA-L), `gemma-2-9b` (LYRA-G) and `Mistral-Nemo-Instruct-2407` 12B (LYRA-M). This choice targets high performing models which have benefited from multilingual pretraining, including French and Italian (to which Monégasque is related), while keeping our compute budget in mind. The LLMs are finetuned using LoRA (Hu et al., 2021) with the efficient implementation of the `unsloth` library.

Given that Monégasque was not among the languages covered by NLLB, the `nllb-200-distilled-1.3B` model is fine-tuned using the French-Monégasque data. In order to maximize downstream performance, we use NLLB's Ligurian tokenizer on Monégasque sentences. The rationale behind this choice is that Ligurian (another low resource language related to genoese) is an even closer language to Monégasque than Italian. Therefore, using the Ligurian tokenizer is likely to yield a more useful representation of Monégasque text. All the presented experiments use greedy decoding.

Effect of Data standardization The candidate models are trained on both versions of the French-

Monégasque dataset and evaluated on translation in both directions. Figure 2 compares the performances reached by each model by training on the dataset before and after undergoing standardization. We observe that all models improve their scores by a significant amount thanks to the standardized data.

We also note that translation quality is clearly superior towards the French language. This is explained by the fact that most models were pretrained on plentiful amounts of French text allowing them to master this high-resource language beforehand. On the other hand, they only discover the Monégasque language through our small dataset which limits the proficiency they are able to reach.

Effect of RAG As previously mentioned, the `BAAI/bge-multilingual-gemma2` model is used in order to generate embeddings of the French sentences. This is done for the train and test sets and the embeddings are used to improve test-time performance by retrieving, for each test sample, the 10 nearest train samples and including them in the prompt. Obviously, this can only be done for LLMs and not for NLLB. The models are trained on the standardized data and their BLEU and METEOR scores with and without RAG are reported on Table 1.

Significant improvements in BLEU scores are seen for translation towards French across all models after the addition of RAG. However, LYRA-G is the only one to benefit from RAG for the fr→mo direction while LYRA-M suffers a significant degradation of its score. These observations may be explained by the fact that the embeddings are based on the French part of the data only and that the embedding model is originally based on Gemma 2.

Effect of French-Italian finetuning We finally evaluate the effect of a preliminary finetuning phase on French-Italian translation before training on the French-Monégasque data. This recipe is tested using the `opus-books` dataset (Tiedemann, 2009) which contains high quality French-Italian parallel sentences. NLLB is excluded from this experiment since it is considered to have already benefited from transfer learning. Indeed, NLLB was pretrained on over 200 languages including French, Italian and Ligurian which is even closer to Monégasque.

The scores of models trained in this fashion and tested with RAG are reported on Table 1 (omitting RAG led to inferior results). A clear benefit is observed on fr→mo scores for LYRA-L and LYRA-G which lets the latter virtually match NLLB's BLEU score. However, LYRA-M still attains its best fr→mo score in the base setting. On the other hand, some performance is lost in the mo→fr direction. We posit that the LLMs' pretrained proficiency in French slightly degrades after undergoing a finetuning procedure involving two other languages.

Model	BLEU		METEOR	
	fr→mo	mo→fr	fr→mo	mo→fr
NLLB-200 1.3B	**35.27**	52.18	48.17	63.55
LYRA-L Instruct	31.62	47.49	49.35	65.20
+ RAG	31.32	52.67	49.45	70.04
++ Italian corpus	32.83	51.95	50.79	69.07
LYRA-G Instruct	33.16	52.12	51.47	69.40
+ RAG	34.42	**58.10**	52.91	**74.31**
++ Italian corpus	35.25	57.23	**53.19**	73.36
LYRA-M Instruct	33.46	51.49	51.77	69.02
+ RAG	30.69	56.75	48.38	72.38
++ Italian corpus	32.31	54.88	49.31	70.97

Table 1: Translation performance in both directions as measured by BLEU and METEOR scores using the standardized data and other methods. Bold numbers represent best scores among all models.

6 Conclusion

In this work, we presented LYRA, a methodology to boost machine translation performance despite scarce data. We saw that enhancing data quality effectively improved results in general. RAG also showed significant potential although some model specific adaptation may sometimes be necessary. Finally, we have also seen that models can reach higher proficiency in a low resource language thanks to transfer learning. Further gains will likely be possible by finetuning future higher performing LLMs. Finally, data augmentation is another interesting research avenue to deal with the low-resource setting.

7 Limitations

Although the results confirm the benefits of the presented methodology, the latter still has its limitations. For example, data curation cannot improve performance beyond a certain point and should be combined with data augmentation to alleviate data scarcity. Moreover, RAG can only help performance if train data are diverse enough and include

relevant examples. Finally, not all low-resource languages are related to high resource ones so that transfer learning will not always be useful.

Acknowledgments

We would like to extend our special thanks to the annotation teams from Afuté and Isahit for their hard work, the Monégasque experts from the Comité des Traditions, the Government of Monaco—particularly the Délégation Interministérielle chargée de la Transition Numérique—and the FAIR team at Meta, including Alexandre Mourachko, for their invaluable advice on the NLLB project.

References

Duarte Miguel Alves, José Pombal, Nuno M Guerreiro, Pedro Henrique Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and Andre Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. In *First Conference on Language Modeling*.

Dzmitry Bahdanau. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Peter F Brown, Stephen A Della Pietra, Vincent J Della Pietra, and Robert L Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Kyunghyun Cho. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A survey of multilingual neural machine translation. *ACM Computing Surveys (CSUR)*, 53(5):1–38.

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

Ilshat Gibadullin, Aidar Valeev, Albina Khusainova, and Adil Khan. 2019. A survey of methods to leverage monolingual data in low-resource neural machine translation. *arXiv preprint arXiv:1910.00373*.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

Tahmid Hasan, Abhik Bhattacharjee, Kazi Samin, Masum Hasan, Madhusudan Basak, M Sohel Rahman, and Rifat Shahriyar. 2020. Not low-resource anymore: Aligner ensembling, batch filtering, and new datasets for bengali-english machine translation. *arXiv preprint arXiv:2009.09359*.

Michael A Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2020. A survey on recent approaches for natural language processing in low-resource scenarios. *arXiv preprint arXiv:2010.12309*.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are GPT models at machine translation ? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

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

Yuchen Hu, Chen Chen, Chao-Han Huck Yang, Ruizhe Li, Dong Zhang, Zhehuai Chen, and Eng Siong Chng. 2024. GenTranslate: Large Language Models are Generative Multilingual Speech and Machine Translators. *arXiv preprint arXiv:2402.06894*.

John Hutchins. 1997. From first conception to first demonstration: the nascent years of machine translation, 1947–1954. a chronology. *Machine Translation*, 12:195–252.

William John Hutchins. 1986. *Machine translation: past, present, future*. Ellis Horwood Chichester.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Philipp Koehn. 2009. *Statistical machine translation*. Cambridge University Press.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Garry Kuwanto, Afra Feyza Akyürek, Isidora Chara Tourni, Siyang Li, Alex Jones, and Derry Wijaya. 2023. Low-resource machine translation training curriculum fit for low-resource languages. In *Pacific Rim International Conference on Artificial Intelligence*, pages 453–458. Springer.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264*.

Mieradilijiang Maimaiti, Yang Liu, Huanbo Luan, and Maosong Sun. 2019. Multi-round transfer learning for low-resource nmt using multiple high-resource languages. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 18(4):1–26.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade Abbott, Vukosi Marivate, Salomon Kabongo, et al. 2020. Masakhane–machine translation for africa. *arXiv preprint arXiv:2003.11529*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

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.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. In *Advances in Neural Information Processing Systems*, volume 34, pages 11054–11070. Curran Associates, Inc.

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.

Alec Radford. 2018. Improving language understanding by generative pre-training.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, New York, NY, USA. Association for Computing Machinery.

Rico Sennrich. 2015. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Jörg Tiedemann. 2009. *News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces*, volume V, pages 237–248.

Yury Tokpanov, Beren Millidge, Paolo Glorioso, Jonathan Pilault, Adam Ibrahim, James Whittington, and Quentin Anthony. 2024. Zyda: A 1.3 T Dataset for Open Language Modeling. *arXiv preprint arXiv:2406.01981*.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

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. In *The Twelfth International Conference on Learning Representations*.

Wen Yang, Chong Li, Jiajun Zhang, and Chengqing Zong. 2023. Bigtranslate: Augmenting large language models with multilingual translation capability over 100 languages. *arXiv preprint arXiv:2305.18098*.

Jiajun Zhang and Chengqing Zong. 2020. Neural machine translation: Challenges, progress and future. *Science China Technological Sciences*, 63(10):2028–2050.

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

A Additional results

The performances of the trained models as measured by the chrF++ metric (Popović, 2015) are reported on Table 2. These figures mostly agree with BLEU scores when comparing the models.

Figure 3 displays the evolution of BLEU scores on translation in both directions through training epochs. One can observe that, apart from NLLB, most models quickly overfit the data due to their limited quantity.

B Additional data details

We provide below a list of the sources used to constitute the French-Monégasque parallel dataset on which the models were trained:

Model	chrF++	
	fr→mo	mo→fr
NLLB-200 1.3B before std.	55.61	65.59
NLLB-200 1.3B	**57.90**	67.05
LYRA-L before std.	50.87	61.47
LYRA-L	53.26	63.90
+ RAG	53.78	68.03
++ Italian corpus	54.81	66.99
LYRA-G before std.	55.32	66.51
LYRA-G	55.48	67.87
+ RAG	57.32	**71.89**
++ Italian corpus	57.16	71.55
LYRA-M before std.	53.63	65.19
LYRA-M	55.44	67.55
+ RAG	52.11	69.75
++ Italian corpus	54.02	69.42

Table 2: Translation performance in both directions as measured by chrF++ scores using the standardized data and other methods. Bold numbers represent best scores among all models. After preliminary finetuning on French-Italian data, all models achieved superior results using RAG rather than without.

- A French-Monégasque dictionary containing two-way translations of single words as well as proverbs.
- A Monégasque grammar book (Monégasque Bescherelle) containing verb conjugations and their translations into French.
- The "Üntra Nui" stories which is a Monégasque chronicle.
- Poems & Fables from Monégasque culture.
- The play "Antigone" written by Jean Anouilh.
- The collection of short stories titled "Lettres de mon moulin" by Alphonse Daudet.
- A collection of Monégasque songs.
- 3 chapters of Tintin comics available in both languages. Namely :
  - "Le secret de la Licorne".
  - "Le trésor de Rackham le Rouge".
  - "Les bijoux de la Castafiore".

Table 3 shows a few examples of sentence pairs before and after undergoing standardization. These illustrate the fixed issues including excessive use of ellipsis, non standard quotes, digital instead of literal numbers and arbitrary onomatopoeia.

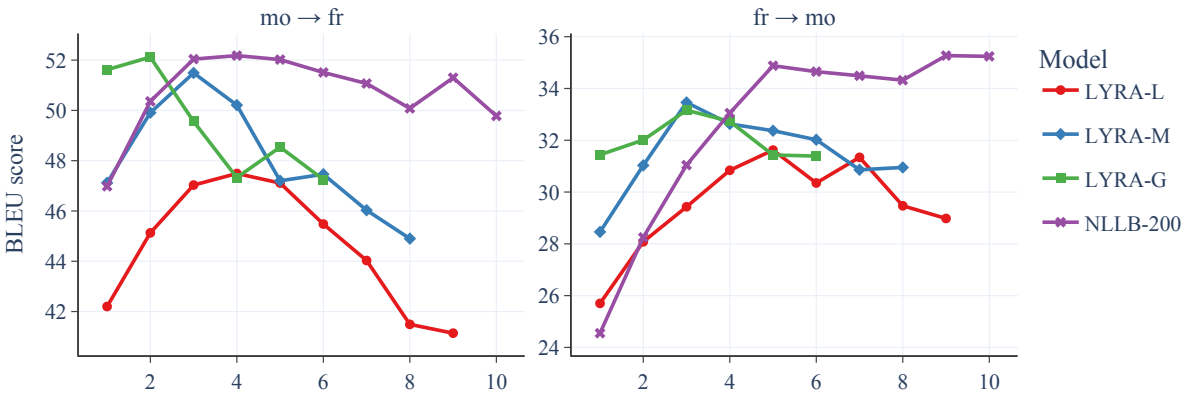

Figure 3: Evolution of translation performance in both directions for the considered models through training epochs as measured by the BLEU score. The training of certain models was stopped early due to overfitting.

Monégasque	French
Ah!... M' asperavi?... Savi dunca perche sun aiçi?..	Ah?... Vous m'attendiez? Vous connaissez donc le but de ma visite?..
Ah ! M' asperavi ? Savi dunca perche sun aiçi ?	Ah ? Vous m'attendiez ? Vous connaissez donc le but de ma visite ?
A grafia e tamben ë tradüçiue d'i testi d'achëstu calendari sun de l'autu sarvu a tradüçiun d'u puema «O belu Munegu»	La graphie ainsi que les traductions des textes de ce calendrier sont de l'auteur excepté la traduction du poème «Ô Monaco la belle»
A grafia e tamben ë tradüçiue d'i testi d'achëstu calendari sun de l'autu sarvu a tradüçiun d'u puema "O belu Munegu".	La graphie ainsi que les traductions des textes de ce calendrier sont de l'auteur excepté la traduction du poème "Ô Monaco la belle".
Ancœi, a Cumpagnia e cumpusa de trei ufiçiali, düjanœve suta-ufiçiali e nuranta sete surdati	Actuellement son effectif est de trois officiers, 19 sous-officiers et 97 hommes du rang
Ancœi, a Cumpagnia e cumpusa de trei ufiçiali, düjanœve suta-ufiçiali e nuranta sete surdati.	Actuellement son effectif est de trois officiers, dix-neuf sous-officiers et quatre vingt dix-sept hommes du rang.
E a fau tanta paciara, De « ci, ci », e de ci, cia » Ch'ün caciaire, che passava Gh'a futüu üna füsiya !	Et il fit tellement de potin, Des « ci, ci » et des « ci, cia », Qu'un chasseur qui passait, L'abattit d'un coup de fusil.
E a fau tanta paciara, ch'ün caciaire, che passava gh'a futüu üna füsiya.	Et il fit tellement de potin qu'un chasseur qui passait, l'abattit d'un coup de fusil.

Table 3: Example instances from the French-Monégasque dataset before (red cells) and after standardization (green cells).

The full dataset (before and after standardization) can be found in the following github repository https://github.com/EmertonData/lyra.

C Experimental details

All the models were trained using a single Nvidia A100 40 GB GPU. NLLB-200 1.3B was finetuned with learning rate: 10^{-5} and batch size 32.

Regarding the LLMs, the 4 bit quantized versions provided by `unsloth` were used as starting points and finetuned with this library using LoRA with the following configuration:

- `r = 16`
- `lora_alpha = 16`
- `lora_dropout = 0.0`
- `bias = "none"`
- `target_modules = ["q_proj", "k_proj", "v_proj", "o_proj", "gate_proj", "up_proj", "down_proj"]`
- `use_rslora = True`
- `loftq_config = None`

A learning rate equal to 3e-5 was used for LYRA-G and 1e-5 for LYRA-L and LYRA-M. Apart from that, the following training parameters are common:

- `batch_size = 48`
- `packing = False`
- `warmup_steps = 100`
- `optim = "adamw_8bit"`
- `weight_decay = 0.01`
- `lr_scheduler_type = "cosine"`
- `max_seq_length = 2048`

All LLMs were trained on completions only using the appropriate data collator. Training was launched for 10 epochs but early stopping was performed based on validation loss as seen on Figure 3.