AfriNLLB: Efficient Translation Models for African Languages

Yasmin Moslem[*]
ADAPT Centre
Trinity College Dublin
Dublin, Ireland
yasmin.moslem@adaptcentre.ie

Aman Kassahun Wassie[*]
African Institute for
Mathematical Sciences (AIMS)
Addis Ababa, Ethiopia
awassie@aimsammi.org

Amanuel Gizachew Abebe[*]
Shaggar Institute of
Technology (SIT)
Addis Ababa, Ethiopia
amanuel.g.abebe1@gmail.com

Abstract

In this work, we present AfriNLLB, a series of lightweight models for efficient translation from and into African languages. AfriNLLB supports 15 language pairs (30 translation directions), including Swahili, Hausa, Yoruba, Amharic, Somali, Zulu, Lingala, Afrikaans, Wolof, and Egyptian Arabic, as well as other African Union official languages such as Arabic (MSA), French, Portuguese, and Spanish. Our training data covers bidirectional translation between English and 13 languages, and between French and two languages (Lingala and Wolof).

AfriNLLB models are based on NLLB-200-600M, which we compress using iterative layer pruning. We fine-tune the pruned models on curated parallel corpora we collected for African languages and knowledge distillation from a larger teacher model. We further apply model quantization for efficient inference. Our work aims at enabling efficient deployment of translation models for African languages in resource-constrained settings.

Our evaluation results demonstrate that AfriNLLB models achieve performance comparable to the baseline while being significantly faster. We release two versions of the AfriNLLB models, a Transformers version that allows further fine-tuning and a CTranslate2 version for efficient inference. Moreover, we release all the training data that we curated for fine-tuning the baseline and pruned models to facilitate further research.

1 AfriNLLB: Background & Motivation

Africa is a linguistically rich continent, with over 2,000 native languages (Grimes, 1996; Heine and Nurse, 2000). Although African languages have millions of native speakers, most of them are low-resource languages (Azime et al., 2024; Wassie, 2024; Adelani et al., 2025b; Farouq et al., 2025; Ojo et al., 2025). This results in a scarcity of African datasets and models for diverse natural language processing tasks, including machine translation (MT). Since MT resources for African languages are scattered across multiple sources, gathering these resources for fine-tuning open-source models is costly and time-consuming. Moreover, providing translation support for speakers of these low-resource languages in governmental and health sectors remains a significant challenge (Anastasopoulos et al., 2020; Wassie et al., 2025).

AfriNLLB seeks to bridge this gap by delivering efficient translation models and curated training data.[1,2] Language selection for AfriNLLB considered several factors, including the number of native speakers in Africa and dataset availability. The AfriNLLB models are based on NLLB-200 (Costa-jussà et al., 2022), and support 15 language pairs (30 translation directions), including 10 native African languages: Swahili, Hausa, Yoruba, Amharic, Somali, Zulu, Lingala, Afrikaans, Wolof, and Egyptian Arabic (cf. Table 1). Additionally, we include 5 of the official languages of the African Union, namely Arabic (MSA), English, French, Portuguese, and Spanish. Since several African languages share some lexicon with these languages due to historical contact, multilingual models can leverage this linguistic overlap through transfer learning from high-resource languages to enhance the performance of low-resource languages (Liu et al., 2020; Fan et al., 2021).

AfriNLLB is a series of efficient multilingual open-source models for African languages, motivated by multiple goals:

- Gathering and curating bilingual training datasets for African languages
- Building lightweight MT models specialized in translating African languages, utilizing compression approaches such as pruning and quantization

[*]Equal contribution

[1]https://github.com/AfriNLP

[2]https://huggingface.co/AfriNLP

Family	Subfamily	Name	Code	Regions
Afro-Asiatic	Chadic	Hausa	hau_Latn	West Africa (Nigeria, Niger)
	Cushitic	Somali	som_Latn	Horn of Africa (Somalia, Ethiopia, Djibouti, Kenya)
	Semitic	Amharic	amh_Ethi	Horn of Africa (Ethiopia)
	Semitic	Egyptian Arabic	arz_Arab	North Africa (Egypt)
Indo-European	Germanic	Afrikaans	afr_Latn	Southern Africa (South Africa, Namibia)
Niger-Congo	Atlantic	Wolof	wol_Latn	West Africa (Senegal, Gambia, Mauritania)
	Bantu	Lingala	lin_Latn	Central Africa (Congo)
	Bantu	Swahili	swh_Latn	East Africa (Tanzania, Kenya)
	Bantu	Zulu	zul_Latn	Southern Africa (South Africa)
	Volta-Niger	Yoruba	yor_Latn	West Africa (Nigeria, Benin)

Table 1: African Languages in AfriNLLB

Family	Subfamily	Name	Code	Regions
Afro-Asiatic	Semitic	Arabic, Modern Standard	arb_Arab	North Africa (formal use)
Indo-European	Germanic	English	eng_Latn	Southern Africa (South Africa)
	Romance	French	fra_Latn	Africa-wide (mostly L2)
	Romance	Portuguese	por_Latn	Southern Africa (Angola, Mozambique)
	Romance	Spanish	spa_Latn	Central Africa (Equatorial Guinea)

Table 2: Non-Native Languages in AfriNLLB

- Open-sourcing the code, training data, and models we have created
- Sharing our approaches and lessons learned to facilitate future work in this area

2 Data

We employ multi-stage fine-tuning before and after model pruning. First, we fine-tune the baseline NLLB-200 600M to improve the performance for African languages. Afterwards, we fine-tune the pruned models again to restore the translation performance. For this purpose, we collect datasets primarily in African languages (Swahili, Hausa, Yoruba, Amharic, Somali, Zulu, Lingala, Afrikaans, Wolof, and Egyptian Arabic) and a few relevant high-resource languages (Arabic (MSA), French, Spanish, Portuguese).

2.1 Data Sources

We mainly collect the datasets from OPUS (Tiedemann, 2012) and Huggingface, with additional data from GitHub and other publicly available online sources. This results in a total of 1,2M samples for 11 African language pairs (9 from/into English, and 2 from/into French). For high-resource languages (Arabic, French, Spanish, Portuguese), we focus on collecting only 1.5M for processing, filter the data, and then sample 200k from each language pair for training. Table 3 summarizes data before and after filtering, while Table 6 elaborates on data sources.

2.2 Data Processing

To ensure the quality of data, we process the datasets in a four-stage pipeline: (i) rule-based filtering, (ii) semantic filtering, (iii) language detection, and (iv) quality estimation. While rule-based filtering uses predefined rules, the other pipeline stages employ a model to generate scores and filter the data based on a threshold. We experimented with different threshold values and found 0.6 to be a reasonable choice.

Rule-based filtering involves deduplication, dropping empty segments, and removing HTML tags. We also filter out sentence pairs with lengths less than 3 or greater than 200 characters. Moreover, to avoid misaligned segments, we remove translation pairs exceeding the 2x source-target length ratio.

Language detection discards segments that are unlikely to be in the expected language. We use two language detector models, AfroLID (Adebara et al., 2022) for the African languages and fastText (Joulin et al., 2017) for the rest of the languages.

Semantic filtering evaluates the translation pairs with cosine similarity scores derived from sentence embedding models, using the Sentence-Transformers library (Reimers and Gurevych, 2019). To handle all the languages, we employ different embedding models based on language support. We use *DistilUSE* for all high-resource language pairs and *LaBSE* for African languages.

We apply semantic filtering for all languages except Lingala as we could not find an embedding model that supports it.

Quality estimation is the final stage of the filtering pipeline, in which we apply reference-free evaluation of the translation and exclude segments that are lower than the threshold. We use COMET (Rei et al., 2020) for high-resource language pairs, and Masakhane's model AfriCOMET-QE-STL (Wang et al., 2024) for African languages.

After thoroughly processing the dataset, we merge the datasets and deduplicate the combined dataset to avoid repetition from different sources. We ended up with a total of 6.4M. However, to mitigate data imbalance, we downsampled the high-resource languages to only 200k per language pair. This results in a total of 1.6M samples (3.2M bidirectional samples, after reversing the dataset), which we use for training. The dataset size for each language direction is presented in Table 3, and elaborated in Table 6.

2.3 Validation and Test Data

We use Flores200[3] (Costa-jussà et al., 2022) for validation and test, as it covers all the languages in our experiments. We use the *dev* split (997 segments) of Flores200 for validation during training, and for layer importance evaluation as part of iterative layer pruning (cf. Section 3), and use the *devtest* split (1,012 segments) for testing and evaluation of our models.

Language Pair		Initial	Processed	Sampled
eng_Latn	afr_Latn	192,541	161,644	161,644
	amh_Ethi	156,739	85,010	85,010
	arz_Arab	85,942	84,170	84,170
	hau_Latn	222,387	155,881	155,881
	som_Latn	87,521	43,657	43,657
	swh_Latn	286,687	181,045	181,045
	wol_Latn	34,956	31,170	31,170
	yor_Latn	34,720	22,626	22,626
	zul_Latn	38,532	33,189	33,189
	arb_Arab	1,526,102	1,424,237	200,000
	fra_Latn	1,500,000	1,483,951	200,000
	por_Latn	1,500,000	1,401,671	200,000
	spa_Latn	1,500,000	1,324,681	200,000
fra_Latn	wol_Latn	10,745	9,071	9,071
	lin_Latn	8126	1,948	1,948
	Total	7,184,998	6,443,951	1,609,411

Table 3: Parallel corpus sizes before and after processing from and into English and French. Since all data is reversed to create the opposite translation direction, the final dataset size is effectively doubled.

[3] https://hf.co/datasets/facebook/flores

3 Methodology

In our experiments, we apply iterative layer pruning to the *NLLB-200 600M* model after fine-tuning it on the training dataset. This approach incrementally identifies and removes layers with minimal contribution to translation quality, one layer at a time. The pruned models resulting from this process are then fine-tuned again to restore most of the translation quality of the baseline model. The resulting models are smaller and faster while retaining or outperforming the quality of the baseline. The following points elaborate on the process.

Layer importance evaluation: We conduct layer importance evaluation by measuring translation performance without each layer. In this greedy layer pruning approach (Peer et al., 2022; Rostami and Dousti, 2024; Moslem et al., 2025; Moslem, 2025), to prune $n + 1$ layers, only a single optimal layer to prune must be added to the already known solution for pruning n layers. After identifying and removing the least critical layer, we repeat the layer importance evaluation on the remaining layers until reaching our n pruning target. We observe that while removing certain layers of the model (e.g. the first or last layer) substantially degrades translation performance, others result in minimal performance drops. Following Moslem (2025), we use the chrF++ metric for layer importance evaluation for both better efficiency and quality. We use the dev split of the Flores200 dataset, mainly where African languages are the target, to improve their translation quality. In the future, we plan to experiment with using both directions.

Layer pruning: We iteratively prune one decoder layer at a time, selecting the layer whose removal has the least negative impact on translation quality, measured by chrF++ scores. At each iteration, we evaluate the translation performance of the pruned model on the dev split of the Flores200 dataset, after removing each candidate layer. The layer whose removal yields the best performance is eventually pruned. This process continues until a predefined number of layers (4 layers) have been removed. By iteratively removing the least important layers, this performance-guided method produces a more compact model that can be fine-tuned further to recover the translation quality of the original model. We also experimented with middle layer pruning and found that iterative layer pruning yields better results (cf. Section 4.1).

Direction	Model	BLEU ↑	chrF++ ↑	COMET ↑	Throughput (toks/s) ↑	Time (s) ↓
xx-en	NLLB 600M (Baseline)	33.81	56.22	71.11	1469.96	21.02
	NLLB 600M + FT	35.15	57.61	71.87	1530.94	20.39
	Pruned + FT	34.01	56.98	71.20	1807.61	17.38
	Pruned + FT (FP16)	34.05	56.99	71.19	3513.32	8.96
en-xx	NLLB 600M (Baseline)	22.70	47.89	69.36	1530.10	28.09
	NLLB 600M + FT	24.28	49.97	70.91	1610.23	26.98
	Pruned + FT	24.17	50.05	70.37	1946.61	22.51
	Pruned + FT (FP16)	24.15	50.06	70.41	3732.72	11.98
xx-fr	NLLB 600M (Baseline)	16.41	38.83	17.34	1475.48	26.46
	NLLB 600M + FT	17.91	40.45	18.47	1412.85	26.12
	Pruned + FT	17.43	40.21	14.52	1845.09	21.61
	Pruned + FT (FP16)	17.38	40.18	14.53	3569.23	11.17
fr-xx	NLLB 600M (Baseline)	9.44	33.42	19.25	1047.18	49.92
	NLLB 600M + FT	10.98	35.68	21.33	1081.84	51.56
	Pruned + FT	10.20	35.21	20.04	1261.66	49.91
	Pruned + FT (FP16)	10.11	35.13	20.03	2313.85	31.15

Table 4: Average Performance by Translation Direction. The category en ↔ xx includes 13 language pairs (26 translation directions), while the category fr ↔ xx includes 2 language pairs for Lingala and Wolof (4 translation directions). The pruned models are 23% faster than the baseline without quantization, and 57% faster with float16 quantization. While more efficient, the translation quality of the compressed models is comparable with the fine-tuned NLLB-200 model.

Fine-tuning: We employ multi-stage fine-tuning. First, we fine-tune the baseline NLLB-200 model on the training dataset to improve its quality for African languages. Since pruning the fine-tuned models results in performance degradation, the pruning step is followed by fine-tuning the pruned model for 1 epoch using the training dataset (cf. Section 2). During training, we use a learning rate of 5e-5, a batch size of 8, gradient accumulation steps of 4, and early stopping with a patience value of 10 evaluation runs. The evaluation takes place every 1000 training steps. The final saved model is the best model based on the evaluation loss score. The training is conducted on one A40 48GB GPU. We use the *Transformers* framework[4] (Wolf et al., 2020) for training. As illustrated by Table 4, this fine-tuning step successfully recovers the translation quality of the baseline model.

Knowledge distillation: To improve the quality of our models, we employ sequence-level knowledge distillation (Kim and Rush, 2016; Crego and Senellart, 2016; Gandhi et al., 2023), where the student model is fine-tuned on a combination of authentic data and synthetic data generated by the teacher model for the same training dataset. In this case, the teacher model is the NLLB-200 3.3B baseline while the students are the NLLB-200 600M baseline and then the pruned models based on our fine-tuned version. After generating the data, we filter it by removing duplicates (exact matches in the target side of the authentic data), and we follow the filtering pipeline we use for processing the original training data (cf. Section 2). The knowledge distillation data after filtering is 568k segments for African languages.

4 Evaluation and Results

For inference, we use CTranslate2[5] (Klein et al., 2020), with beam size of 3 and batch size of 1024 tokens, on a A40 48GB GPU.

To evaluate our systems, we calculated BLEU (Papineni et al., 2002), chrF++ (Popović, 2017), as implemented in the sacreBLEU library[6] (Post, 2018). For semantic evaluation, we use AfriCOMET (Wang et al., 2024) for African languages, and COMET (Rei et al., 2020) for Arabic and European languages.[7]

The process of iterative layer pruning of 4 decoder layers created a 548M model that is 23% faster in average than the baseline. Moreover, the quality degradation caused by pruning has been mitigated through fine-tuning and knowledge distillation. As demonstrated by Table 4 and elaborated by Table 7, by the end of the process, the

[4] https://github.com/huggingface/transformers

[5] https://github.com/OpenNMT/CTranslate2

[6] https://github.com/mjpost/sacrebleu

[7] In particular, we used the *"africomet-mtl"* model for AfriCOMET and the *"wmt22-comet-da"* model for COMET.

pruned model could recover most of the translation quality of the baseline model. Moreover, quantization (float16) of the pruned model further enhanced the inference performance, making the model 57% faster than the baseline.

4.1 Ablation Study

In this ablation study, we compare three scenarios: (i) removing middle layers[8] instead of iteratively determining the layers to remove based on layer importance evaluation (cf. Section 3), (ii) pruning both encoder and decoder layers instead of pruning decoder layers only, and (iii) pruning various values of the decoder layers, namely 4, 6, and 8 layers.

We observe that iterative layer pruning clearly outperforms middle layer pruning in both cases of removing decoder layers only or both encoder and decoder layers. Fine-tuning after pruning is crucial in all cases, as it mitigates the effect of pruning on performance. Figure 1 illustrates four pruned models, both before and after fine-tuning:

- Middle pruning, 4 decoder layers (Mid 548M)
- Middle pruning, 4 encoder layers and 4 decoder layers (Mid 498M)
- Iterative pruning, 4 decoder layers (Iter 548M)
- Iterative pruning, 4 encoder layers and 4 decoder layers (Iter 498M)

When it comes to removing encoder layers in addition to decoder layers, it is not clear to what extent this affects the quality. Obviously, removing encoder layers reduces the size of the model further, which can cause performance degradation. Keeping encoder layers intact was recommended by previous work on speech (Gandhi et al., 2023; Moslem, 2025), which poses the question whether the same concept applies to text-based encoder-decoder models such as NLLB-200. We intend to investigate this further in future work.

Furthermore, we thoroughly studied the effect of keeping all 12 encoder layers intact while iteratively removing different numbers of decoder layers. We experimented with three pruning configurations, removing 4, 6, or 8 decoder layers, resulting in models with 12 encoder layers and 8, 6, or 4 decoder layers, respectively. As illustrated in Figure 2 and Figure 3, the effect of the number of decoder layers removed varies across language pairs, although removing up to 6 layers (50%) yields similar or better performance compared to the NLLB-200 600M baseline, thanks to fine-tuning before and after pruning. Table 5 elaborates further on the performance results in terms of both translation quality and inference speed.

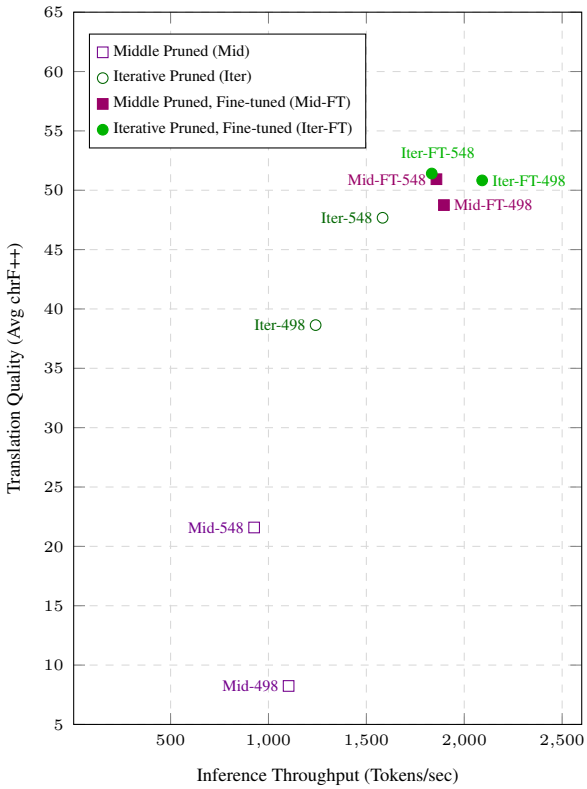

Figure 1: Quality-Efficiency Comparison. The iterative-pruned models demonstrate a superior balance of speed and quality compared to the middle-pruned variants. The 548M models include 12 encoder layers and 8 decoder layers (i.e. 4 decoder layers are pruned), while the 498M models include 8 encoder layers and 8 decoder layers (i.e. 8 layers are pruned, 4 from the encoder and 4 from the decoder). The chart reports the average chrF++ scores across all language pairs before and after fine-tuning the pruned models.

5 Conclusions and Future Work

In this work, we presented AfriNLLB, lightweight models for African languages, that achieve over 20–50% inference performance gains compared to their baseline NLLB-200 600M. We release models with various sizes to match different needs.

We have demonstrated that iterative layer pruning is an effective approach for model compression while retaining translation quality. The method relies on layer importance evaluation, followed by fine-tuning on a medium-sized dataset. This iterative layer pruning process reduces the model size and accelerates inference. We are open-sourcing AfriNLLB models and data. In addition, to en-

[8] For middle layer pruning, we remove layers 4 to 7 inclusively.

sure reproducibility, we are going to make all the processing and training code publicly available.

In future versions of AfriNLLB, we plan to add more languages. Research directions include investigating data augmentation approaches besides knowledge distillation, such as back-translation. Moreover, we plan to expand our approach to other architectures, such as autoregressive large language models and encoder-only models.

We hope that by releasing AfriNLLB models, training data, and code, we facilitate further research on African languages, and support the African community worldwide.

References

Idris Abdulmumin, Satya Ranjan Dash, Musa Abdullahi Dawud, Shantipriya Parida, Shamsuddeen Muhammad, Ibrahim Sa'id Ahmad, Subhadarshi Panda, Ondřej Bojar, Bashir Shehu Galadanci, and Bello Shehu Bello. 2022. Hausa visual genome: A dataset for multi-modal English to Hausa machine translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6471–6479, Marseille, France. European Language Resources Association.

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Inciarte. 2022. AfroLID: A neural language identification tool for African languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1958–1981, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Sackey Sackey, Bonaventure Dossou, Chris Emezue, Colin Leong, Fatoumata Michael, George Mierupa, Hameed Bolaji, Hector Mimouni, and 4 others. 2022. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070. Association for Computational Linguistics.

David Adelani, Dana Ruiter, Jesujoba Alabi, Damilola Adebonojo, Adesina Ayeni, Mofetoluwa Adeyemi, Ayodele Awokoya, and Cristina Espa na-Bonet. 2021. MENYO-20k: A multi-domain English-Yor'ub'a corpus for machine translation and domain adaptation. In *Proceedings of the Second Workshop on African Natural Language Processing*, pages 27–34. Association for Computational Linguistics.

David Ifeoluwa Adelani, Alison Chi, Simbiat Aderibigbe, Butoyi Beatrice, Tumaini Balikwisha, Barkwende Hugues Diallo, Tunde Oluwaseyi Ajayi, Joseph K. O. Oaminu, Ruqayya Nasir Iro, Abdoulaye Guindo, Bamigbade Daniel, Moussa Ibrahissa, Oumarou Sanda Soumana, Ayorinde Olugbenga, Mercy Wamriew, Opeyemi Adediran, Kula Kekana, Mpho Raborife, Zeenat Vallie, and 2 others. 2025a. AFRIDOC-MT: Document-level MT Corpus for African Languages. *arXiv preprint arXiv:2501.06374*.

David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba Oluwadara Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, Chiamaka Ijeoma Chukwuneke, Happy Buzaaba, Blessing Kudzaishe Sibanda, Godson Koffi Kalipe, Jonathan Mukiibi, Salomon Kabongo Kabenamualu, Foutse Yuehgoh, Mmasibidi Setaka, Lolwethu Ndolela, and 8 others. 2025b. IrokoBench: A new benchmark for African languages in the age of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2732–2757, Albuquerque, New Mexico. Association for Computational Linguistics.

Rania Al-Sabbagh. 2024. Arzen-multigenre: An aligned parallel dataset of egyptian arabic song lyrics, novels, and subtitles, with english translations. *Data in Brief*, 54:110271.

Duarte Miguel Alves, Jose Pombal, Nuno M. Guerreiro, Pedro Henrique Martins, Joao Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, Jose G. C. de Souza, and Andre Martins. 2025. WMT24++: Expanding the Language Coverage of WMT24 to 55 Languages & Dialects. *arXiv preprint arXiv:2502.12404*.

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Franscisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020. TICO-19: the Translation Initiative for COvid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Israel Abebe Azime, Atnafu Lambebo Tonja, Tadesse Destaw Belay, Mitiku Yohannes Fuge, Aman Kassahun Wassie, Eyasu Shiferaw Jada, Yonas Chanie, Walelign Tewabe Sewunetie, and Seid Muhie Yimam. 2024. Walia-LLM: Enhancing Amharic-LLaMA by integrating task-specific and generative datasets. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 432–444, Miami, Florida, USA. Association for Computational Linguistics.

Isaac Caswell, Elizabeth Nielsen, Jiaming Luo, Colin Cherry, Geza Kovacs, Hadar Shemtov, Partha Talukdar, Dinesh Tewari, Baba Mamadi Diane, Koulako Moussa Doumbouya, Djibrila Diane,

Solo Farabado Cissé, Edoardo Ferrante, Alessandro Guasoni, Mamadou K. Keita, Sudhamoy DebBarma, Ali Kuzhuget, David Anugraha, Muhammad Ravi Shulthan Habibi, and 3 others. 2025. SMOL: Professionally translated parallel data for 115 under-represented languages. *arXiv preprint arXiv:2502.12301*.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, pages 261–268. European Association for Machine Translation.

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, John Hoffman, and 19 others. 2022. No Language Left Behind: Scaling human-centered machine translation. *arXiv [cs.CL]*.

Josep Crego and Jean Senellart. 2016. Neural Machine Translation from Simplified Translations. *arXiv [cs.CL]*.

Andreas Eisele and Yu Chen. 2010. MultiUN: A multilingual corpus from united nation documents. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA).

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Gabriel Sequeira, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, and 67 others. 2025. Mmteb: Massive multilingual text embedding benchmark. *arXiv preprint arXiv:2502.13595*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond English-Centric Multilingual Machine Translation. *Journal of Machine Learning Research*, 22(107):1–48.

Muhammad Hazim Al Farouq, Aman Kassahun Wassie, and Yasmin Moslem. 2025. Bemba Speech Translation: Exploring a Low-Resource African Language. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 354–359, Vienna, Austria (in-person and online). Association for Computational Linguistics.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – a benchmark for evaluating machine translation performance. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24. Association for Computational Linguistics.

Sanchit Gandhi, Patrick von Platen, and Alexander M Rush. 2023. Distil-Whisper: Robust knowledge distillation via large-scale pseudo labelling. *arXiv [cs.CL]*.

Barbara F. Grimes. 1996. *Ethnologue: Languages of the World*, 13th edition. SIL International, Dallas, TX. Summer Institute of Linguistics.

Bernd Heine and Derek Nurse, editors. 2000. *African Languages: An Introduction*. Cambridge University Press, Cambridge.

Andreea Iana, Goran Glavoči'c, and Heiko Paulheim. 2023. News without borders: Domain adaptation of multilingual sentence embeddings for cross-lingual news recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*. ACM.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

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.

Guillaume Klein, Dakun Zhang, Clément Chouteau, Josep Crego, and Jean Senellart. 2020. Efficient and high-quality neural machine translation with OpenNMT. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 211–217, Stroudsburg, PA, USA. Association for Computational Linguistics.

Laban Kumbuga, Joyce Nakatumba-Nabende, Jonathan Mukiibi, and Andrew Katumba. 2024. SALT: Sunbird African language translation corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5462–5472. ELRA and ICCL.

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.

Yasmin Moslem. 2025. Efficient speech translation through model compression and knowledge distillation. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 379–388, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yasmin Moslem, Muhammad Hazim Al Farouq, and John Kelleher. 2025. Iterative layer pruning for efficient translation inference. In *Proceedings of the Tenth Conference on Machine Translation*, pages 1022–1027, Suzhou, China. Association for Computational Linguistics.

Jessica Ojo, Odunayo Ogundepo, Akintunde Oladipo, Kelechi Ogueji, Jimmy Lin, Pontus Stenetorp, and David Ifeoluwa Adelani. 2025. AfroBench: How good are large language models on African languages? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19048–19095, Vienna, Austria. Association for Computational Linguistics.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

David Peer, Sebastian Stabinger, Stefan Engl, and Antonio Rodríguez-Sánchez. 2022. Greedy-layer pruning: Speeding up transformer models for natural language processing. *Pattern Recognit. Lett.*, 157:76–82.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. 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.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Pedram Rostami and Mohammad Javad Dousti. 2024. CULL-MT: Compression using language and layer pruning for machine translation. *arXiv [cs.CL]*.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

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.

Jiayi Wang, David Adelani, Sweta Agrawal, Marek Masiak, Ricardo Rei, Eleftheria Briakou, Marine Carpuat, Xuanli He, Sofia Bourhim, Andiswa Bukula, Muhidin Mohamed, Temitayo Olatoye, Tosin Adewumi, Hamam Mokayed, Christine Mwase, Wangui Kimotho, Foutse Yuehgoh, Anuoluwapo Aremu, Jessica Ojo, and 39 others. 2024. AfriMTE and AfriCOMET: Enhancing COMET to embrace under-resourced African 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 5997–6023, Stroudsburg, PA, USA. Association for Computational Linguistics.

Aman Kassahun Wassie. 2024. Machine translation for ge'ez language. *arXiv preprint arXiv:2311.14530*, arXiv:2311.14530.

Aman Kassahun Wassie, Mahdi Molaei, and Yasmin Moslem. 2025. Domain-specific translation with open-source large language models: Resource-oriented analysis. *arXiv preprint arXiv:2412.05862*, arXiv:2412.05862.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Stroudsburg, PA, USA. Association for Computational Linguistics.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639. Association for Computational Linguistics.

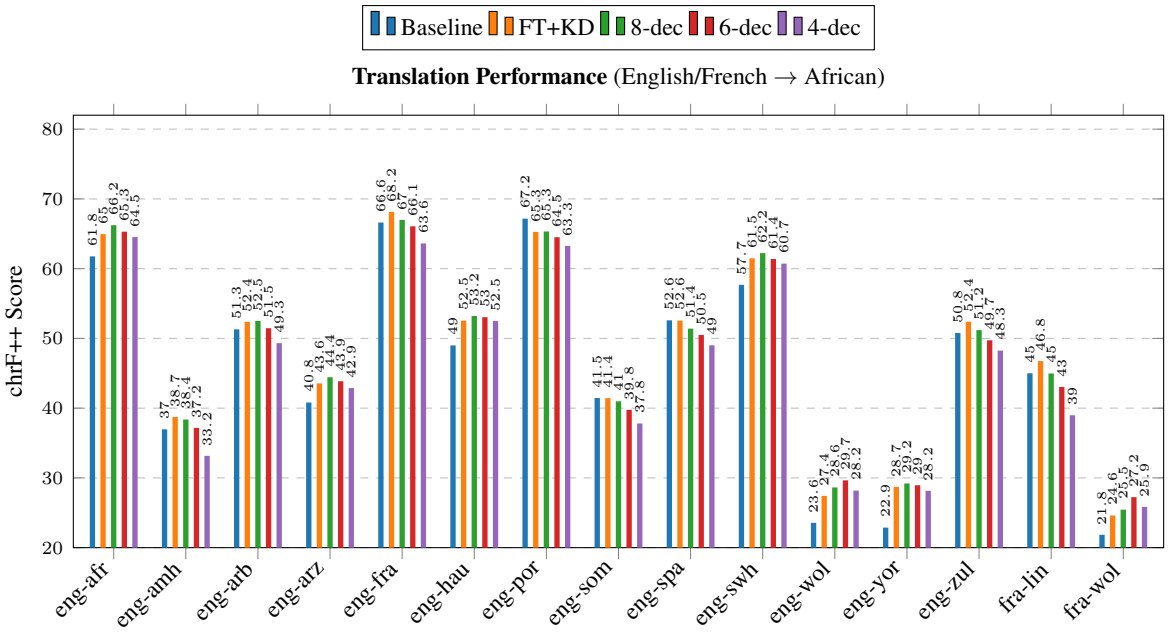

Figure 2: Translation performance (chrF++) from English/French to target African languages.

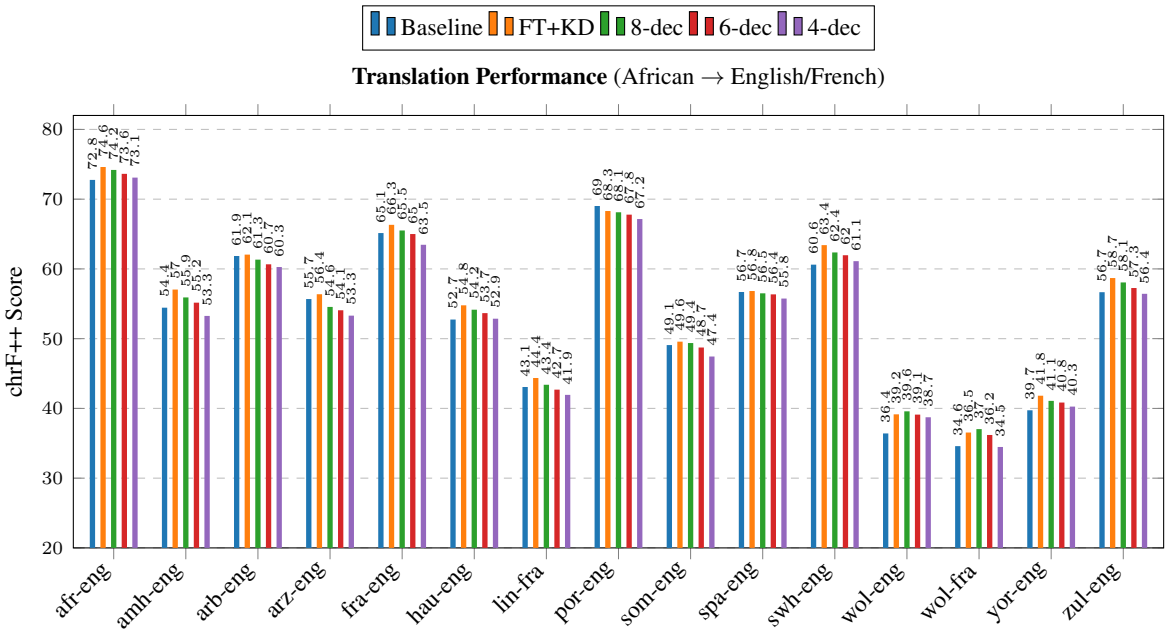

Figure 3: Translation performance (chrF++) from African languages to English/French.

Performance Comparison: Layer Pruning Configurations

Translation quality (BLEU, chrF++, COMET) and efficiency (throughput, inference time) across baseline, fine-tuned, and pruned configurations with optional FP16 quantization

Lang	Model	Enc	Dec	Quant	BLEU ↑	chrF++ ↑	COMET ↑	Throughput ↑	Time ↓
xx-en	NLLB	12	12	–	33.81	56.22	71.11	1469.96	21.02
				FP16	33.80	56.22	71.13	2834.69	10.92
	NLLB + FT	12	12	–	35.15	57.61	71.87	1530.94	20.39
				FP16	35.10	57.61	71.87	2808.90	11.15
	AfriNLLB	12	8	–	34.01	56.98	71.20	1807.61	17.38
				FP16	34.05	56.99	71.19	3513.32	8.96
		12	6	–	33.35	56.48	70.79	2028.18	15.41
				FP16	33.32	56.45	70.79	4000.25	7.82
		12	4	–	32.03	55.62	69.71	2257.03	13.77
				FP16	32.01	55.60	69.71	4589.42	6.79
		8	8	–	30.89	54.32	68.08	1852.13	17.05
				FP16	30.86	54.30	68.08	3550.50	8.91
en-xx	NLLB	12	12	–	22.70	47.89	69.36	1530.10	28.09
				FP16	22.68	47.88	69.38	2898.38	15.33
	NLLB + FT	12	12	–	24.28	49.97	70.91	1610.23	26.98
				FP16	24.14	49.84	70.90	2811.34	18.82
	AfriNLLB	12	8	–	24.17	50.05	70.37	1946.61	22.51
				FP16	24.15	50.06	70.41	3732.72	11.98
		12	6	–	23.48	49.34	68.98	2265.87	18.50
				FP16	23.49	49.35	69.00	4428.68	9.65
		12	4	–	21.77	47.80	65.68	2489.35	17.31
				FP16	21.77	47.81	65.68	4954.62	9.09
		8	8	–	23.59	49.64	69.90	2015.53	21.34
				FP16	23.58	49.63	69.88	3851.13	11.34
xx-fr	NLLB	12	12	–	16.41	38.83	17.34	1475.48	26.46
				FP16	16.33	38.83	17.23	2850.66	13.71
	NLLB + FT	12	12	–	17.91	40.45	18.47	1524.32	26.12
				FP16	17.83	40.42	18.37	2749.45	14.68
	AfriNLLB	12	8	–	17.43	40.21	14.52	1845.09	21.61
				FP16	17.38	40.18	14.53	3569.23	11.17
		12	6	–	16.52	39.44	11.78	2044.27	19.21
				FP16	16.54	39.42	11.68	3953.51	9.92
		12	4	–	14.96	38.21	5.67	2340.99	16.77
				FP16	14.90	38.17	5.71	4766.12	8.24
		8	8	–	14.42	37.05	3.14	1866.26	21.84
				FP16	14.34	36.97	3.14	3448.51	11.83
fr-xx	NLLB	12	12	–	9.44	33.42	19.25	1047.18	49.92
				FP16	9.52	33.40	19.38	1920.41	29.05
	NLLB + FT	12	12	–	10.98	35.68	21.33	1081.84	51.56
				FP16	10.48	35.05	21.49	1700.25	51.31
	AfriNLLB	12	8	–	10.20	35.21	20.04	1261.66	49.91
				FP16	10.11	35.13	20.03	2313.85	31.15
		12	6	–	10.07	35.14	19.83	1416.33	30.89
				FP16	9.99	35.08	19.78	2465.60	18.68
		12	4	–	7.57	32.42	14.16	1207.06	38.75
				FP16	7.57	32.38	14.29	2069.52	23.25
		8	8	–	9.75	35.23	20.05	1222.83	45.33
				FP16	9.84	35.31	20.11	2186.73	25.97

Table 5: Comprehensive performance evaluation across translation directions. AfriNLLB models use various encoder-decoder layer configurations (12-8, 12-6, 12-4, 8-8) with and without FP16 quantization.

Datasets Sources and Sizes

Names, sources, and sizes of our training datasets before and after filtering for each language pair

Dataset	fra-eng	spa-eng	por-eng	arb-eng	swh-eng	amh-eng	som-eng	hau-eng	yor-eng	zul-eng	afr-eng	arz-eng	wol-fra	wol-eng	lin-fra
OPUS Datasets															
Tatoeba (Tiedemann, 2020)	–	–	–	–	–	213/188	9/5	259/183	423/421	72/170	2.4K/2.1K	6.5K/1.3K	–	–	555/120
translatewiki	–	–	–	–	–	–	–	–	–	6.2K/244	–	111K/23K	1.7K/243	–	–
wikimedia	1.4M/1.1M	–	822K/610K	621K/374K	16.3K/11.3K	942/425	1.1K/718	190K/121K	12.5K/4.8K	9.3K/5.5K	78.5K/66.5K	–	690/169	21/5	–
GNOME	–	–	21.2K/15.3K	150/41	40/43	57.1K/26.9K	753/1.1K	5.5K/110	1K/590	4.5K/7.7K	12.7K/27.8K	–	–	–	–
Ubuntu	–	–	–	6K/2.5K	–	–	–	–	141/0	–	–	–	220/38	222/26	–
GlobalVoices	–	–	–	–	32.3K/26.9K	1.8K/1.2K	–	–	136/61	–	–	–	–	–	–
bible-uedin	–	–	–	62.2K/16.3K	–	6.1K/46.6K	6.2K/49.5K	–	–	–	62.1K/50.6K	–	7.9K/648	15.8K/2.6K	–
NeuLab-TedTalks	212K/185K	215K/190K	81.2K/55K	–	–	–	–	–	–	–	–	–	–	–	–
EMEA	–	1.1M/235K	1.1M/223K	–	–	–	–	–	–	–	–	–	–	–	–
EUbookshop	–	–	4.2M/610K	–	1.7K/110	–	–	–	–	–	–	–	–	–	–
ELRC-wiki_health	4.4K/3.7K	–	–	15.1K/14.4K	608/501	–	–	–	–	–	404/312	–	–	–	–
News-Commentary	156K/125K	–	–	–	–	–	–	–	–	–	–	–	–	–	–
JRC-Acquis	814K/65.3K	806K/398K	–	–	–	–	–	–	–	–	–	–	–	–	–
TED2020	–	–	–	408K/341K	9.7K/80.8K	1K/1.7K	2K/1.3K	27/21	–	–	2.3K/1.8K	–	–	–	–
KDE4	–	–	–	116K/25.6K	–	–	–	149/66	–	–	64.3K/29.8K	–	–	–	–
ELRC-EMEA	–	777K/614K	–	–	–	–	–	–	–	–	–	–	–	–	–
Books	–	93.5K/63.4K	–	–	–	–	–	–	–	–	–	–	–	–	–
Tanzil	–	–	–	–	138K/96.7K	93.5K/50.5K	93.8K/10.5K	128K/63.4K	–	–	–	–	–	–	–
OpenSubtitles	–	–	–	–	94.6K/95.8K	3K/1.6K	531/446	–	–	–	969K/11.8K	–	–	–	–
tico-19 (Anastasopoulos et al., 2020)	–	–	–	–	3.1K/2.8K	3.1K/3.1K	3.1K/1.2K	3.1K/2.1K	–	3.1K/2.3K	–	–	–	–	2.9K/544
ELRC_2922	–	–	–	–	607/498	–	–	–	–	–	403/310	–	–	–	–
ELRC-3073-wiki_health	–	–	–	–	608/501	–	–	–	–	–	–	–	–	–	–
infopankki	–	–	–	–	–	–	47.2K/89.8K	–	–	–	–	–	–	–	–
qed	–	–	–	–	–	–	–	–	–	–	28.8K/17.5K	–	–	–	–
spc	–	–	–	–	–	–	–	–	–	–	57.4K/47.3K	–	–	–	–
ELRC-monumentos	–	–	–	–	–	–	–	–	–	–	54/41	–	–	–	–
ELRC-Museus	–	–	–	–	–	–	–	–	–	–	32/0	–	–	–	–
HuggingFace Datasets															
smol (Caswell et al., 2025)	–	–	–	–	863/719	863/712	862/551	863/548	863/153	863/552	863/610	–	–	7.4K/570	–
mafand (Adelani et al., 2022)	–	–	–	–	34.4K/29.9K	1.9K/1.4K	–	5.9K/4.4K	6.6K/4K	3.5K/2K	–	–	–	–	–
mafand-dev	–	–	–	–	–	–	–	1.3K/971	6.6K/4K	1.2K/636	–	–	–	–	–
mafand-test	–	–	–	–	–	–	–	1.5K/1.2K	6.6K/4K	998/596	–	–	–	–	–
Pontoon-Translations	–	–	–	–	6.1K/2.8K	–	–	3.2K/1.2K	4.4K/553	3.3K/735	13.1K/2.7K	–	–	6.8K/802	–
Weblate-Translations	–	–	–	–	17.2K/7.2K	8K/2.1K	1.6K/310	143/54	164/533	66/53	23.2K/1.8K	–	–	–	–
ntrex (Federmann et al., 2022)	–	–	–	–	2K/1.7K	2K/1.5K	2K/1.9K	2K/1.2K	2K/602	2K/1K	2K/1.8K	–	–	–	–
AfriDocMT-health (Adelani et al., 2025a)	–	–	–	–	7K/6.6K	7K/6.6K	–	7K/5.9K	7K/3.7K	7K/5.4K	–	–	–	–	–
AfriDocMT-doc_health	–	–	–	–	240/4	240/6	–	–	–	–	–	–	–	–	–
AfriDocMT-doc_health_2	–	–	–	–	540/96	540/104	–	–	–	–	–	–	–	–	–
AfriDocMT-doc_health_5	–	–	–	–	1.5K/1.5K	1.5K/1.5K	–	–	–	–	–	–	–	–	–
AfriDocMT-doc_health_10	–	–	–	–	812/366	812/440	–	–	–	–	–	–	–	–	–
quran_multilingual	–	–	–	–	–	6.2K/1K	6.2K/740	6.2K/3.8K	6.2K/1K	–	–	–	–	–	–
nazimali-quran	–	–	–	–	6.2K/5K	6.2K/3.7K	6.2K/5K	–	–	–	–	–	–	–	–
opus-100 (Zhang et al., 2020)	–	–	–	–	–	–	–	–	10.4K/2.3K	–	–	–	–	–	–
opus-100-dev	–	–	–	–	–	–	–	–	10.4K/2.3K	–	–	–	–	–	–
opus-100-test	–	–	–	–	–	–	–	–	10.4K/2.3K	–	–	–	–	–	–
menyo20k_mt-train (Adelani et al., 2021)	–	–	–	–	–	–	–	–	10.1K/4.6K	–	–	–	–	–	–
menyo20k_mt-dev	–	–	–	–	–	–	–	–	3.4K/1.4K	–	–	–	–	–	–
menyo20k_mt-test	–	–	–	–	–	–	–	–	6.6K/3.7K	–	–	–	–	–	–
yoruba_audio_trans	–	–	–	–	–	–	–	–	9.2K/1.9K	–	–	–	–	–	–
arz-en-parallel	–	–	–	–	–	–	–	–	–	–	–	25K/22.6K	–	–	–
news-comm-eng-arz (Moslem et al., 2025)	–	–	–	–	–	–	–	–	–	–	–	832K/83.3K	–	–	–
mteb/tatoeba-bitext (Enevoldsen et al., 2025)	–	–	–	–	–	–	–	–	–	–	–	8.9K/2.9K	–	–	–
fr-wolof-trans-gs	–	–	–	–	–	–	–	–	–	–	–	–	10.4K/1.6K	–	–
wolof_en_fr	–	–	–	–	–	–	–	–	–	–	–	–	26.6K/6.5K	26.6K/7.6K	–
english_wolof_trans	–	–	–	–	–	–	–	–	–	–	–	–	–	84.7K/17.2K	–
comet_score_en_wo	–	–	–	–	–	–	–	–	–	–	–	–	–	7.5K/4K	–
wolof_en_bible	–	–	–	–	–	–	–	–	–	–	–	–	–	13.4K/2.2K	–
MultiUN (Eisele and Chen, 2010)	–	–	–	9.8M/9.8M	–	–	–	–	–	–	–	–	–	–	–
ted_talks_iwslt-14 (Cettolo et al., 2012)	–	–	–	–	52/42	–	–	–	–	–	–	–	–	–	–
ted_talks_iwslt-15	–	–	–	–	68/53	–	–	–	–	–	–	–	–	–	–
ted_talks_iwslt	–	–	–	–	–	–	188/730	–	–	–	–	–	–	–	–
wmt24pp (Alves et al., 2025)	–	–	–	–	998/691	–	–	–	–	–	–	–	–	–	–
sunbird-salt (Kumbuga et al., 2024)	–	–	–	–	24.9K/23.1K	–	–	–	–	–	–	–	–	–	–
HausaVG (Abdulmumin et al., 2022)	–	–	–	–	–	–	–	28.9K/7.9K	–	–	–	–	–	–	–
polynews-parallel (Iana et al., 2023)	–	–	–	–	–	–	–	5.7K/4.4K	–	3.4K/2K	–	–	–	–	–
Quran	–	–	–	–	–	–	–	6.2K/3.7K	–	–	–	–	–	–	–
hf-spc	–	–	–	–	–	–	–	–	–	–	57.4K/47.4K	–	–	–	–
lingvanex_test	–	–	–	–	–	–	–	–	–	–	1.1K/649	–	–	–	–
subscene	–	–	–	–	–	–	900/0	–	–	–	–	–	–	–	–
opus_infopankki	–	–	–	–	–	–	47.2K/89.8K	–	–	–	–	–	–	–	–
other sources															
ArzEn-MultiGenre (Al-Sabbagh, 2024)	–	–	–	–	–	–	–	–	–	–	–	25K/6.6K	–	–	–
ethiopian-legal	–	–	–	–	–	5.4K/3.7K	–	–	–	–	–	–	–	–	–
ethiopian-history	–	–	–	–	–	1.3K/737	–	–	–	–	–	–	–	–	–
ethiopian-news	–	–	–	–	–	5.4K/1.1K	–	–	–	–	–	–	–	–	–
ethiopian-ebible	–	–	–	–	–	6.5K/3.8K	–	–	–	–	–	–	–	–	–
ethiopian-ethio_bible	–	–	–	–	–	11.7K/5.7K	–	–	–	–	–	–	–	–	–
ethiopian-jw_bible	–	–	–	–	–	31.1K/25.2K	–	–	–	–	–	–	–	–	–
ethiopian-jw_daily	–	–	–	–	–	4.7K/4.3K	–	–	–	–	–	–	–	–	–
horn-mt	–	–	–	–	–	2K/2K	–	–	–	–	–	–	–	–	–
mt-eval-am-amen	–	–	–	–	–	997/712	–	–	–	–	–	–	–	–	–
mt-eval-am-enam	–	–	–	–	–	1.9K/1.4K	–	–	–	–	–	–	–	–	–
ukuxhumana	–	–	–	–	–	–	–	–	–	26.7K/13.8K	–	–	–	–	–
zenodo-training	–	–	–	–	–	–	–	–	–	4.7K/2.6K	–	–	–	–	–
zenodo-eval	–	–	–	–	–	–	–	–	–	998/596	–	–	–	–	–
Gamayun-fr-ln	–	–	–	–	–	–	–	–	–	–	–	–	–	–	5K/1.4K
Total (Origin)	2.6M	3M	6.2M	11M	399K	319K	275K	398K	124K	113K	619K	234K	47.5K	162K	8.1K
After Filter	1.5M	1.5M	1.5M	10.5M	287K	157K	87.5K	222K	34.7K	38.5K	174K	85.9K	9.2K	35K	2K
After Dedup	1.48M	1.32M	1.4M	1.42M	181K	85K	87.5K	156K	22.6K	33.2K	174K	84.2K	9.1K	31.2K	1.9K

Table 6: Dataset statistics for all language pairs. Values shown as Original/Final (K=thousand, M=million), and "–" indicates dataset not used.

Performance Comparison: NLLB-200 600M vs. AfriNLLB 548M Models

Comparison of NLLB-200 600M baseline, Pruned (Iterative) + Fine-tuned, and Pruned (Iterative)+ Fine-tuned (float16 quantization) across BLEU, chrF++, COMET, and Output Throughput (out toks/sec)

Lang Pair	NLLB 600M (baseline)				AfriNLLB 548M (Iterative FT)				AfriNLLB 548M (Iterative FT FP16)			
	BLEU	chrF++	COMET	Throughput	BLEU	chrF++	COMET	Throughput	BLEU	chrF++	COMET	Throughput
en-af	35.82	61.76	75.10	1,672	41.17 ↑14.9%	66.23 ↑7.2%	75.86 ↑1.0%	2,147 ↑28.4%	41.08 ↑14.7%	66.20 ↑7.2%	75.85 ↑1.0%	4,222 ↑152.5%
af-en	54.18	72.76	74.02	1,484	56.09 ↑3.5%	74.19 ↑2.0%	74.28 ↑0.4%	1,912 ↑28.8%	56.08 ↑3.5%	74.17 ↑1.9%	74.30 ↑0.4%	3,758 ↑153.3%
en-am	12.12	36.96	69.24	1,407	12.82 ↑5.8%	38.37 ↑3.8%	70.72 ↑2.1%	1,816 ↑29.1%	12.86 ↑6.1%	38.42 ↑4.0%	70.85 ↑2.3%	3,137 ↑123.0%
am-en	30.02	54.44	67.88	1,542	31.56 ↑5.1%	55.91 ↑2.7%	67.22 ↓1.0%	1,797 ↑16.5%	31.57 ↑5.2%	55.89 ↑2.7%	67.20 ↓1.0%	3,454 ↑124.0%
en-ar	22.86	51.30	84.78	1,648	24.16 ↑5.7%	52.51 ↑2.4%	84.94 ↑0.2%	2,104 ↑27.7%	24.18 ↑5.8%	52.50 ↑2.3%	84.92 ↑0.2%	4,111 ↑149.5%
ar-en	39.00	61.85	85.99	1,415	37.00 ↓5.1%	61.33 ↓0.8%	85.74 ↓0.3%	1,732 ↑22.4%	37.11 ↓4.8%	61.33 ↓0.8%	85.73 ↓0.3%	3,345 ↑136.4%
en-arz	11.87	40.81	80.16	1,525	14.94 ↑25.9%	44.43 ↑8.9%	81.98 ↑2.3%	2,063 ↑35.3%	14.95 ↑25.9%	44.46 ↑8.9%	81.99 ↑2.3%	4,042 ↑165.0%
arz-en	30.64	55.68	82.65	1,382	28.69 ↓6.4%	54.55 ↓2.0%	82.11 ↓0.7%	1,753 ↑26.8%	28.77 ↓6.1%	54.58 ↓2.0%	82.11 ↓0.7%	3,422 ↑147.6%
en-es	26.71	52.59	85.30	1,696	24.78 ↓7.2%	51.40 ↓2.3%	84.38 ↓1.1%	2,152 ↑26.9%	24.71 ↓7.5%	51.37 ↓2.3%	84.40 ↓1.1%	4,210 ↑148.2%
es-en	29.91	56.69	86.11	1,571	27.99 ↓6.4%	56.50 ↓0.3%	86.05 ↓0.1%	1,903 ↑21.1%	28.00 ↓6.4%	56.46 ↓0.4%	86.03 ↓0.1%	3,738 ↑138.0%
en-fr	46.70	66.61	86.78	1,700	46.16 ↓1.2%	66.99 ↑0.6%	86.25 ↓0.6%	2,118 ↑24.6%	46.25 ↓1.0%	67.05 ↑0.7%	86.26 ↓0.6%	4,161 ↑144.8%
fr-en	43.15	65.14	88.19	1,454	41.73 ↓3.3%	65.51 ↑0.6%	88.17 ↓0.0%	1,819 ↑25.1%	41.88 ↓2.9%	65.56 ↑0.6%	88.18 ↓0.0%	3,537 ↑143.3%
en-ha	23.69	48.99	63.93	1,596	27.64 ↑16.7%	53.21 ↑8.6%	65.01 ↑1.7%	1,894 ↑18.7%	27.61 ↑16.5%	53.22 ↑8.6%	65.12 ↑1.9%	3,583 ↑124.6%
ha-en	31.06	52.74	65.97	1,514	32.36 ↑4.2%	54.15 ↑2.7%	66.37 ↑0.6%	1,907 ↑26.0%	32.48 ↑4.6%	54.22 ↑2.8%	66.31 ↑0.5%	3,754 ↑147.9%
fr-ln	15.15	45.00	35.84	1,419	15.88 ↑4.8%	44.97 ↓0.1%	34.07 ↓4.9%	1,789 ↑26.1%	15.72 ↑3.8%	44.87 ↓0.3%	33.85 ↓5.6%	3,498 ↑146.5%
ln-fr	19.85	43.07	36.61	1,559	20.06 ↑1.1%	43.39 ↑0.7%	30.12 ↓17.7%	1,906 ↑22.3%	20.00 ↑0.8%	43.35 ↑0.7%	30.09 ↓17.8%	3,663 ↑135.0%
en-pt	46.45	67.17	88.56	1,726	42.72 ↓8.0%	65.33 ↓2.7%	87.74 ↓0.9%	2,150 ↑24.6%	42.56 ↓8.4%	65.23 ↓2.9%	87.72 ↓0.9%	4,224 ↑144.7%
pt-en	48.08	69.02	88.95	1,580	46.48 ↓3.3%	68.12 ↓1.3%	88.57 ↓0.4%	1,949 ↑23.4%	46.62 ↓3.0%	68.15 ↓1.3%	88.59 ↓0.4%	3,832 ↑142.5%
en-so	11.38	41.45	61.63	1,600	11.05 ↓2.9%	40.98 ↓1.1%	57.92 ↓6.0%	1,799 ↑12.4%	11.03 ↓3.1%	40.98 ↓1.1%	58.00 ↓5.9%	3,400 ↑112.5%
so-en	26.20	49.08	61.31	1,371	26.36 ↑0.6%	49.37 ↑0.6%	60.14 ↓1.9%	1,718 ↑25.3%	26.42 ↑0.8%	49.41 ↑0.7%	60.10 ↓2.0%	3,328 ↑142.7%
en-sw	31.50	57.68	70.14	1,780	36.77 ↑16.7%	62.23 ↑7.9%	71.72 ↑2.3%	2,138 ↑20.1%	36.78 ↑16.8%	62.25 ↑7.9%	71.73 ↑2.3%	4,168 ↑134.2%
sw-en	39.47	60.61	70.12	1,672	41.40 ↑4.9%	62.36 ↑2.9%	70.38 ↑0.4%	1,974 ↑18.1%	41.41 ↑4.9%	62.40 ↑3.0%	70.42 ↑0.4%	3,875 ↑131.8%
en-wo	5.06	23.56	17.27	923	6.97 ↑37.7%	28.63 ↑21.5%	22.65 ↑31.2%	992 ↑7.5%	6.99 ↑38.1%	28.75 ↑22.0%	22.75 ↑31.7%	1,678 ↑81.8%
wo-en	14.90	36.41	36.16	1,403	17.11 ↑14.8%	39.58 ↑8.7%	38.37 ↑6.1%	1,596 ↑13.8%	16.98 ↑14.0%	39.53 ↑8.6%	38.38 ↑6.1%	3,011 ↑114.6%
wo-fr	12.96	34.59	-1.93	1,392	14.79 ↑14.1%	37.03 ↑7.1%	-1.09 ↑43.5%	1,784 ↑28.2%	14.76 ↑13.9%	37.00 ↑7.0%	-1.03 ↑46.6%	3,475 ↑149.6%
fr-wo	3.73	21.84	2.66	676	4.52 ↑21.2%	25.45 ↑16.5%	6.01 ↑125.9%	735 ↑8.7%	4.49 ↑20.4%	25.38 ↑16.2%	6.21 ↑133.5%	1,130 ↑67.2%
en-yo	4.32	22.87	51.89	1,002	8.03 ↑85.9%	29.20 ↑27.7%	59.51 ↑14.7%	1,820 ↑81.6%	8.05 ↑86.3%	29.19 ↑27.6%	59.63 ↑14.9%	3,437 ↑243.1%
yo-en	17.61	39.73	49.68	1,295	18.62 ↑5.7%	41.08 ↑3.4%	50.45 ↑1.5%	1,590 ↑22.8%	18.72 ↑6.3%	41.18 ↑3.6%	50.35 ↑1.3%	3,014 ↑132.7%
en-zu	16.68	50.78	66.95	1,616	16.98 ↑1.8%	51.19 ↑0.8%	66.12 ↓1.2%	2,116 ↑30.9%	16.92 ↑1.4%	51.14 ↑0.7%	66.07 ↓1.3%	4,152 ↑157.0%
zu-en	35.32	56.66	67.45	1,427	36.77 ↑4.1%	58.06 ↑2.5%	67.80 ↑0.5%	1,848 ↑29.5%	36.56 ↑3.5%	57.96 ↑2.3%	67.76 ↑0.5%	3,606 ↑152.7%
Average	**26.21**	**49.93**	**63.31**	**1468.2**	**27.05** ↑3.2%	**51.41** ↑3.0%	**63.65** ↑0.53%	**1833.95** ↑24.9%	**27.05** ↑3.2%	**51.41** ↑3.0%	**63.66** ↑0.55%	**3532.15** ↑140.57%

Table 7: Detailed evaluation of AfriNLLB models for each language direction. Overall, the compressed models achieve comparable or improved translation quality while achieving significant inference throughput gains over the baseline NLLB-200 600M.