TopXGen: Topic-Diverse Parallel Data Generation for Low-Resource Machine Translation

Armel Randy Zebaze, Benoît Sagot, Rachel Bawden

TopXGen: Topic-Diverse Parallel Data Generation for Low-Resource Machine Translation

Armel Zebaze **Benoît Sagot** **Rachel Bawden**
Inria, Paris, France
firstname.lastname@inria.fr

Abstract

LLMs have been shown to perform well in machine translation (MT) with the use of in-context learning (ICL), rivaling supervised models when translating into high-resource languages (HRLs). However, they lag behind when translating into low-resource language (LRLs). Example selection via similarity search and supervised fine-tuning help. However the improvements they give are limited by the size, quality and diversity of existing parallel datasets. A common technique in low-resource MT is synthetic parallel data creation, the most frequent of which is back-translation, whereby existing target-side texts are automatically translated into the source language. However, this assumes the existence of good quality and relevant target-side texts, which are not readily available for many LRLs. In this paper, we present TopXGen, an LLM-based approach for the generation of high quality and topic-diverse data in multiple LRLs, which can then be backtranslated to produce useful and diverse parallel texts for ICL and fine-tuning. Our intuition is that while LLMs struggle to translate into LRLs, their ability to translate well into HRLs and their multilinguality enable them to generate good quality, natural-sounding target-side texts, which can be translated well into a high-resource source language. We show that TopXGen boosts LLM translation performance during fine-tuning and in-context learning. Code and outputs are available at https://github.com/ArmelRandy/topxgen.

1 Introduction

The performance of Machine Translation (MT) models has considerably evolved through the years, with new models increasingly supporting more languages (Bapna et al., 2022; Costa-jussà et al., 2022; Yang et al., 2023). However, performance remains unequal across languages, which themselves vary greatly in terms of available resources and representation in NLP research (Joshi et al., 2020). MT models perform much better in high-resource languages (HRLs), such as English, French and German, compared to low-resource languages (LRLs) such as Hausa, which lack large quantities of high-quality parallel data. Decoder-based LLMs can perform MT without relying on parallel data (in a zero-shot fashion). However they lag behind supervised MT baselines when translating into LRLs (Hendy et al., 2023). In-Context Learning (ICL;

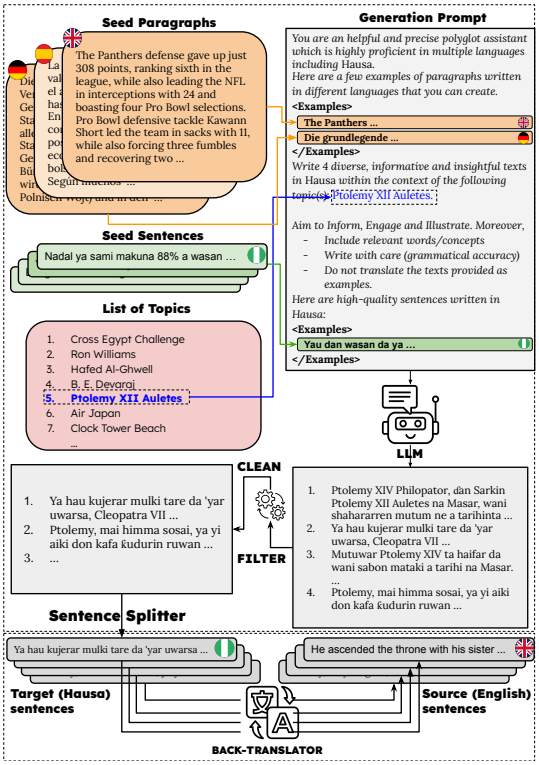

Figure 1: Overview of TopXGen. An LLM is used to write a diverse set of paragraphs in an LRL guided by topics, example sentences in the LRL and example paragraphs in HRLs. The generated paragraphs are later cleaned and divided into sentences that are back-translated into the source language to create a sentence-level parallel dataset.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 22358–22381
November 4-9, 2025

Brown et al., 2020), which involves using a few high-quality demonstrations, has been shown to improve performance, especially when they are similar to the sentence to be translated (Moslem et al., 2023; Zebaze et al., 2025b), highlighting the importance of parallel datasets even in this setting. A common approach is to synthesize parallel data using forward translation or back-translation (Schwenk, 2008; Bojar and Tamchyna, 2011; Sennrich et al., 2016). Forward translation comes with issues such as low translation fidelity and the neglect of cultural nuances in the target language. Back-Translation (BT; Sennrich et al., 2016) typically relies on a good quality monolingual corpus in the LRL, which can be difficult to obtain. In this work, we explore the construction of synthetic datasets for MT into LRLs by automatically generating text in the LRL target language and then backtranslating into a HRL source language. Marie and Fujita (2021) first proposed an approach with a similar aim, using an LLM to generate in-domain monolingual data, which they then used to perform BT alongside parallel data. They generate their monolingual data using an LLM fine-tuned on each of their domain-specific datasets. However, their approach focuses on English and requires domain-specific fine-tuning (and datasets) in addition to relying on parallel data to perform BT, making it impractical for translation into LRLs. To address this, we introduce TopXGen (Figure 1), a pipeline that exploits LLMs' multilinguality and instruction-following capabilities. Unlike prior work, we generate monolingual data (sentences) beyond English to cover numerous LRLs. Rather than domain-specific fine-tuning, we directly prompt an LLM and guide its generations towards a predefined list of topics in order to encourage diversity in the outputs. The quality of sentences stems from the ability of state-of-the-art multilingual LLMs to produce coherent text in LRLs (Enis and Hopkins, 2024)—even if they struggle to translate accurately into them. We then backtranslate the generated sentences into an HRL (English in this work) using a reverse translation model (a supervised MT system or the same LLM). This target-aware generation helps mitigate the cultural loss often observed in LRLs with standard forward translation techniques. Moreover, translating into a HRL offers a practical advantage, as HRLs are generally easier to translate into with high fidelity.

To evaluate TopXGen, we generate synthetic parallel data between English and ten low-resource languages—Basque, Hausa, Igbo, Kinyarwanda, Nepali, Somali, Sundanese, Swahili, Urdu, and Xhosa—using `Gemma-3-27B-It` as the generator and `NLLB-200-3.3B` as the back-translator. We then assess both in-context learning and fine-tuning setups across small translation models. Our experiments show that TopXGen consistently outperforms other data generation methods and achieves performance comparable to human-translated datasets.

2 Related Work

Low-resource Machine Translation with LLMs. LLMs encounter many languages during their training but in various proportions (Abadji et al., 2022; Penedo et al., 2024). Through ICL (Brown et al., 2020), they can perform a wide variety of tasks including MT. Decoder-based LMs are on par with supervised MT models —such as M2M100 (Fan et al., 2021) and NLLB (Goyal et al., 2022)—when translating between high-resource languages but still lag behind when translating into low-resource languages (Hendy et al., 2023; Zhu et al., 2024). Many works have emerged to bridge the gap at inference time by either using similarity-based in-context example selection (Moslem et al., 2023; Tanzer et al., 2024; Zebaze et al., 2025b) or more advanced prompting strategies (He et al., 2024; Briakou et al., 2024; Zebaze et al., 2025a). Another line of work involves fine-tuning LLMs. While most works focus on mid- to high-resource languages (Xu et al., 2024b,c, 2025), a few of them explore fine-tuning on LRLs. They generally use bitexts mined from the internet (Schwenk et al., 2021a; El-Kishky et al., 2020; Schwenk et al., 2021b), out-of-domain benchmarks written by native speakers in LRLs (Muennighoff et al., 2023; Üstün et al., 2024; Uemura et al., 2024) or continual pretraining on monolingual data covering multiple LRLs to improve few-shot MT performance (Buzaaba et al., 2025). The scarcity of high-quality parallel data is the major bottleneck for low-resource MT, leading to studies exploring methods to generate such data using LLMs.

Parallel Data Generation. Generating data using LLMs has emerged as a popular alternative to costly human annotation, primarily for instruction datasets. One of the first approaches to demonstrate the effectiveness of this paradigm is Self-Instruct (Wang et al., 2023; Taori et al., 2023). It consists in bootstrapping a small set of seed in-

structions into a bigger collection with the help of ICL. Building on this, Kou et al. (2024) proposed KNN-INSTRUCT which replaces the random selection of ICL demonstrations with nearest neighbor retrieval in an embedding space. Subsequent advances introduced multilingual instruction generation (Cui et al., 2024; Wei et al., 2023), generating increasingly complex and diverse instructions across domains (Xu et al., 2024a; Zeng et al., 2024; Chaudhary, 2023; Luo et al., 2025, 2024) and step-by-step explanations within responses (Mukherjee et al., 2023; Gunasekar et al., 2023; Li et al., 2023b). Other approaches use unlabeled human-written corpora as sources for generating instruction responses (Wei et al., 2024b,a; Li et al., 2024; Ben Allal et al., 2024). For MT and LRLs, many works generate datasets by simply machine translating existing ones into the languages of interest (The Aya Collection; Singh et al., 2024, the XLLMs-100 collection; Lai et al., 2024 and Bactrian-X; Li et al., 2023a). One of the most common and early strategies was to backtranslate monolingual target side data into the source language to create synthetic parallel data (Schwenk, 2008; Bertoldi and Federico, 2009; Bojar and Tamchyna, 2011; Sennrich et al., 2016; Caswell et al., 2019; Burlot and Yvon, 2018; Bogoychev and Sennrich, 2020; Marie et al., 2020). However, BT's reliance on high-quality monolingual data on the target side presents a challenge when such data is hard to obtain. To address this, Marie and Fujita (2021) proposed using LLMs to synthesize monolingual data by fine-tuning GPT-2 (Radford et al., 2019) on multiple domains and generating multi-domain English data. Combined with substantial amounts of parallel data, they apply BT to train a model to translate from a target language into English. We propose to also generate synthetic parallel datasets, but for low-resource MT instead of HRL translation, the challenge being the generation of high quality data in an LRL. Instead of fine-tuning an LLM for multi-domain adaptation, our approach uses more recent strategies involving LLM prompting to generate high quality and diverse data, which we then use to perform ICL and fine-tuning for translation into LRLs.

3 Methodology

We propose TOPXGEN (Figure 1), a topic-guided and target-language centric method for automatically generating parallel sentence-level datasets between English and LRLs, with the end goal of improving MT into LRLs via ICL or supervised fine-tuning using the synthetic examples. It consists of two steps: data generation and back-translation.

Data generation. We generate data in the LRL of interest by prompting a multilingual LLM. We aim to produce data that is structurally and lexically diverse by generating it at the paragraph level data (which we then split into sentences before back-translation) and by guiding the generation with predefined topics. To generate a new paragraph in a given target LRL, we prompt the LLM with:

- **Topics:** To foster diversity in the output texts. We rely on the generator's instruction-following abilities and prompt it to generate content on a randomly selected topic drawn from a predefined list of Wikipedia topics (Ziadé, 2023) following (Li et al., 2023b).[1]
- **Seed paragraphs:** To have the generator understand what we expect in terms of length and format. We use the 240 paragraphs from XQuAD (Artetxe et al., 2020), which are written in 11 HRLs and perform cross-lingual ICL (X-ICL; Cahyawijaya et al., 2024).
- **Seed sentences.** These serve to illustrate how a sentence is structured in the LRL and to help ensure the outputs are generated in the correct script. We use the FLORES-200 dev set (Goyal et al., 2022; Costa-jussà et al., 2022).

Note that during generation, we automatically remove generations that overlap too much with previous ones with respect to the ROUGE[2] score (Lin, 2004) following (Wang et al., 2023). Given the collection of paragraphs produced by the generator, we build a collection of sentences by applying a sentence splitter.[3] We then perform language identification with fastText (Bojanowski et al., 2017; Costa-jussà et al., 2022) on each sentence and remove incorrectly labeled ones.

Back-Translation. We then use a multilingual back-translator (e.g., `NLLB-200-3.3B`; Costa-jussà et al., 2022) to translate the generated sentences into the HRL we want to learn to translate from. Translating into the HRL is likely to be of good quality given that MT models perform better in this direction than the reverse one.

[1] In contrast to Marie and Fujita (2021), who focus on broad domains like IT or Health, we use fine-grained topics (e.g., specific personalities or events).

[2] Version: 0.1.2|Pure python implementation of ROUGE-1.5.5

[3] https://github.com/mediacloud/sentence-splitter

	Basque	Hausa	Igbo	Kinyarwanda	Nepali	Somali	Sundanese	Swahili	Urdu	Xhosa
Paragraphs	16,829	14,981	18,518	8,900	14,490	14,623	10,483	11,489	13,923	15,781
Sentences	120,031	101,488	133,071	57,884	143,014	96,315	78,264	86,981	131,133	104,992
Sentences (After decon.)	120,031	101,466	133,063	57,884	142,681	96,315	78,257	86,981	131,118	104,979

Table 1: Statistics of the TOPXGEN dataset in terms of paragraphs and sentences.

3.1 The TOPXGEN dataset

We generate multiple paragraphs in 10 languages: Basque, Hausa, Igbo, Kinyarwanda, Nepali, Somali, Sundanese, Swahili, Urdu and Xhosa. Following Gunasekar et al. (2023) and Ben Allal et al. (2024), to avoid data contamination, we filtered the generated paragraphs to remove those containing a significant overlap (at least one 10-gram overlap) with FLORES (also NTREX 128 and TICO-19, whose results are given in Appendix B.4). We apply the same strategy to the English translations of the sentences to make sure that they do not resemble the seed paragraphs that we used (i.e. XQuAD; Artetxe et al., 2020). In Table 1, we report the number of paragraphs generated per language and the number of sentences before and after decontamination. Each language has between 50k and 150k sentences for a total of 1.05M sentences. We conduct further analysis in Appendices B.6, B.7 and B.8.

4 Experiments

4.1 Experimental Setup

We work on translation from English to 10 LRLs: Basque, Hausa, Igbo, Kinyarwanda, Nepali, Somali, Sundanese, Swahili, Urdu and Xhosa.

Datasets. FLORES-200 (Goyal et al., 2022; Costa-jussà et al., 2022) is a dataset consisting of translations from web articles into 204 languages. These sentences are divided into two splits: devtest and dev. We use the devtest set (1012 examples) as the evaluation set and the dev set (997 examples) as the selection pool for few-shot MT.

Models. We use `Gemma-3-27b-It` (Team et al., 2025) as the generator, and `NLLB-200-3.3B` (Costa-jussà et al., 2022) as the back-translator in order to reduce the computational cost, as translating in a few-shot setting with `Gemma-3-27b-It` would be significantly more expensive. We use `LLaMA-2-7B` (Touvron et al., 2023) and `LLaMA-3-8B` (Dubey et al., 2024) during fine-tuning and compare the resulting models against strong multilingual LLMs including `LLaMA-3.1-8B-It` & `LLaMA-3.1-70B It` (Dubey et al., 2024), `Gemma-2-9B-It` & `Gemma-2-27B-It` (Gemma Team et al., 2024), `Aya-expanse-8B` & `Aya-expanse-32B` (Dang et al., 2024), `Qwen-2.5-7B-It` & `Qwen-2.5-32B-It` (Yang et al., 2024; Team, 2024) and `Command-R7B` (Cohere et al., 2025).

Evaluation Metrics. We mainly evaluate using MetricX-24 (Juraska et al., 2024). We use the reference-based version `MetricX-24-Hybrid-XXL` (which supports the same 101 languages as mT5 (Xue et al., 2021)). MetricX assigns a score ranging from 0 to 25, with higher scores indicating more errors in the translation. We also use n-gram matching metrics via sacreBLEU (Post, 2018), namely BLEU[4] (Papineni et al., 2002) and chrF++[5] (Popović, 2015; Popović, 2017) and report these results for transparency reasons in Appendix B.3.

Implementation Details. The generator's temperature of generation is set to 1.0. We back-translate with beam search (beam size = 5). For the topics, we use a list of 67,573 Wikipedia topics curated by Ziadé (2023). We fine-tune unidirectional models for 5k steps (about 3 hours on 1 H100 80G) with a learning rate of 1e-5, a batch size of 4 with 4 gradient accumulation steps and a maximum sequence length equal to 512. The multidirectional model requires 100k steps (about 30 hours on 1 H100 80G) and in both cases we choose the last checkpoint as the final model. To assess statistical significance, we follow (Koehn, 2004) and use paired bootstrap resampling with 300 samples of 500 sentences and a p-value threshold of 0.05. All models are evaluated in a zero-shot fashion with greedy decoding unless stated otherwise. See Appendices A.1 and A.2 for additional details.

5 Results

5.1 Fine-tuning

We fine-tune two small models (`LLaMA-2-7B` and `LLaMA-3-8B`) on our constructed dataset and com-

[4] nrefs:1|case:mixed|eff:no|tok:flores200|smooth:exp|version:2.4.2

[5] nrefs:1|case:mixed|eff:yes|nc:6|nw:2|space:no|version:2.4.2

Models	Basque		Hausa		Igbo		Kinyarwanda		Nepali	
	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX
Toplines										
NLLB-200-3.3B	24.96	6.11	28.82	2.46	19.92	4.89	23.35	4.58	26.64	6.60
Gemma-3-27B-It	27.17	4.81	19.03	4.84	15.37	7.56	12.61	7.92	25.74	3.47
Baselines										
Gemma-2-27B-It	23.33	6.21	17.54	5.54	11.71	10.89	7.00	14.83	22.63	4.00
LLaMA-3.1-70B It	26.06	5.15	19.01	5.90	15.59	8.37	8.25	13.45	25.33	4.31
Gemma-2-9B-It	16.69	9.33	13.76	7.13	9.06	14.65	4.33	19.99	18.91	4.80
Command-R7B	3.16	13.34	1.88	20.37	2.12	21.46	2.19	22.60	5.37	9.12
Aya-expanse-32B	8.35	17.06	4.73	17.41	4.51	21.35	3.51	21.18	9.96	7.91
Qwen-2.5-32B-It	7.27	18.99	4.26	16.68	5.76	20.03	2.99	22.84	9.86	9.26
LLaMAX3-8B Alpaca	12.14	10.57	17.50	6.33	13.57	9.34	4.06	19.29	21.20	5.31
LLaMA-2-7B 5-SHOT BM25	3.42	23.02	1.71	22.16	2.03	23.08	2.58	13.92	3.22	15.40
LLaMA-3-8B 5-SHOT BM25	18.15	8.48	12.28	10.24	8.32	16.15	4.47	21.11	17.71	6.71
Our Models										
LLaMA-2-7B uni.	13.00	14.76	13.11	8.57	12.30	10.97	7.30	15.87	15.11	7.98
LLaMA-2-7B uni. *beam size=5*	14.93	12.55	13.77	8.42	12.90	10.04	7.61	14.73	16.08	6.31
LLaMA-3-8B uni.	23.70	6.25	19.65	5.20	16.28	7.31	11.76	9.91	21.88	4.21
LLaMA-3-8B uni. *beam size=5*	**25.64**	**5.27**	**20.52**	**5.07**	**17.02**	**6.54**	**13.60**	**8.51**	**23.24**	**3.77**
LLaMA-3-8B multi.	21.77	6.95	18.22	6.05	15.54	7.96	9.76	12.96	20.84	4.52
LLaMA-3-8B multi. *beam size=5*	24.07	5.68	19.36	5.76	16.09	7.13	11.62	11.40	22.37	3.94

Models	Somali		Sundanese		Swahili		Urdu		Xhosa	
	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX
Toplines										
NLLB-200-3.3B	17.40	4.82	21.44	4.97	36.20	4.59	28.41	4.44	23.24	3.80
Gemma-3-27B-It	13.58	5.50	17.01	4.77	35.16	3.80	26.24	3.13	12.82	7.62
Baselines										
Gemma-2-27B-It	8.93	10.77	14.60	7.18	35.99	3.78	23.28	4.01	9.27	12.25
LLaMA-3.1-70B It	9.74	10.52	17.34	5.13	34.26	4.46	27.37	3.64	6.67	16.01
Gemma-2-9B-It	6.32	14.67	12.54	8.48	29.63	5.23	19.26	5.25	6.99	18.32
Command-R7B	1.85	20.20	8.41	5.53	6.51	18.85	3.75	13.14	2.33	22.77
Aya-expanse-32B	5.88	15.13	9.81	10.84	9.37	17.25	11.14	7.72	4.84	22.00
Qwen-2.5-32B-It	4.11	19.35	8.12	15.01	9.46	17.44	11.72	9.31	4.13	22.32
LLaMAX3-8B Alpaca	11.12	7.41	11.63	8.84	26.76	6.63	19.94	5.72	11.01	10.31
LLaMA-2-7B 5-SHOT BM25	2.05	21.97	6.43	16.65	2.85	22.86	2.65	19.10	2.34	23.42
LLaMA-3-8B 5-SHOT BM25	4.74	18.45	14.17	8.90	22.61	8.63	17.17	6.47	3.22	22.53
Our Models										
LLaMA-2-7B uni.	8.89	11.73	14.70	7.19	19.19	11.42	14.89	8.43	9.25	15.24
LLaMA-2-7B uni. *beam size=5*	8.20	10.63	15.42	5.75	20.96	9.52	16.33	6.63	7.67	13.09
LLaMA-3-8B uni.	13.20	7.00	16.84	4.88	30.99	5.23	22.43	4.16	12.39	9.33
LLaMA-3-8B uni. *beam size=5*	**13.70**	**6.17**	**18.16**	**4.26**	**33.49**	**4.51**	**23.51**	**3.78**	**13.93**	**7.77**
LLaMA-3-8B multi.	12.03	7.94	16.24	5.66	28.53	6.03	21.65	4.54	11.74	10.56
LLaMA-3-8B multi. *beam size=5*	12.71	7.05	17.32	4.83	31.11	5.07	22.92	3.96	13.15	8.61

Table 2: BLEU and MetricX scores for 10 English → X directions from FLORES 200. Best results after fine-tuning are highlighted in bold.

pare them to state-of-the-art models of various sizes in a zero-shot setting.[6] We evaluate two setups: a unidirectional setting with one model per translation direction (English↔X), and a multidirectional setting with a single model trained on all 10 directions. Results are shown in Table 2[7]. Fine-tuning LLaMA-2-7B, whose performance is close to random in all directions, turns it into a model that outperforms Aya-expanse-32B, Qwen-2.5-32B and Command-R7B. Unidirectional fine-tuning of LLaMA-3-8B outperforms Gemma-2-27B-It and LLaMA-3.1-70B-It. With beam search (Freitag and Al-Onaizan, 2017), we get even better results with unidirectional models, closing the gap with the generator. Fine-tuning a model to support the ten languages together leads to a small drop of performance of about 1 BLEU in all languages, i.e. there is no positive cross-lingual transfer. We provide more baseline comparisons in Appendix B.1. Additional results from fine-tuning NLLB-200-3.3B and Gemma-3-27B-PT on the TOPXGEN dataset are provided in Appendix B.5.

5.2 In-Context Learning

We compare the performance obtained when doing 5-shot MT with example selection via similarity search in the FLORES dev set and in the

[6]We use 5-shot learning with BM25 selection in the FLORES dev set for base models as they cannot follow instructions.

[7]See Appendix B.2 for MT into English.

TopXGen dataset with `LLaMA-3.1-8B-It`. As shown in Table 3, retrieval in the TopXGen dataset yields superior results compared to zero-shot performance, showing that its content is of good enough quality to help the model during the translation. Moreover, it also works better than retrieval in FLORES, particularly in terms of the MetricX score. While TopXGen has its size and diversity as an advantage, the FLORES dev set is in-domain with respect to the evaluation set as they come from the same research effort on top of being written by professional translators.

5.3 Comparison to existing approaches

We investigate the impact of changing the data generation pipeline from our TopXGen approach to Self-Instruct (Wang et al., 2023) and KNN-Instruct (Kou et al., 2024). We keep the same parameters (Generator, Seed sentences, Back-translator etc.) and compare the result of fine-tuning `LLaMA-2-7B` on 20K sentence pairs generated by each strategy in Sundanese and Somali. For KNN-Instruct, we use the multilingual SONAR embeddings (Duquenne et al., 2023) to compute similarity between sentences and we perform 3 rounds with $K = 8$. In Figure 2, we report the zero-shot (beam size = 5) BLEU and MetricX scores every 200 steps and compare the values across methods. TopXGen consistently outperforms Self-Instruct and KNN-Instruct in terms of BLEU and MetricX at each checkpoint.

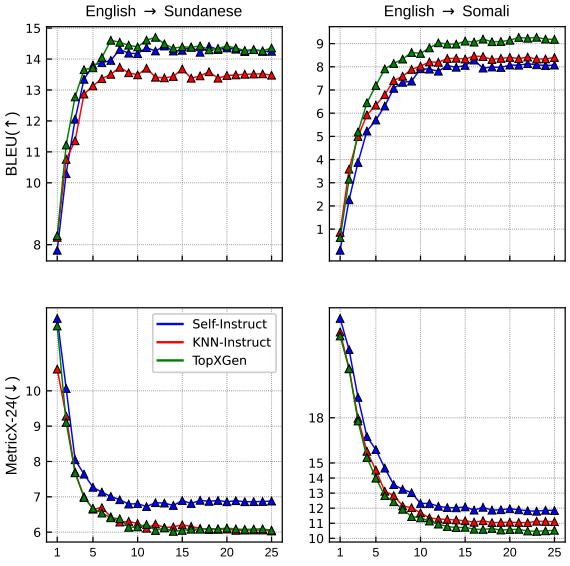

Figure 2: TopXGen vs Self-Instruct & KNN-Instruct.

We also compare fine-tuning on TopXGen to fine-tuning on high-quality professionally translated data such as SmolSent (Jones et al., 2023; Caswell et al., 2025) (863 parallel sentences) and the FLORES dev set (997 samples). We select the first 863 samples from each dataset and choose `LLaMA-3-8B` as the base model due to the limited data. We fine-tune one model for English ↔ Hausa and another for English ↔ Igbo and report the scores every 100 steps on Figure 3. As expected, TopXGen does not consistently work better than professional translations. However it outperforms SmolSent when translating in Hausa and competes with FLORES in Igbo. Moreover, TopXGen has the advantage of its scale; as shown in Table 2, fine-tuning on the full TopXGen dataset (which is 100 times bigger) outperforms the best SmolSent and FLORES checkpoint by at least 3 BLEU points.

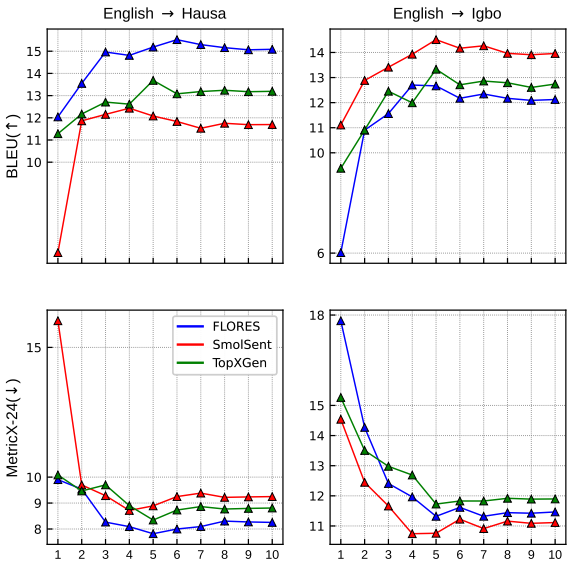

Figure 3: TopXGen vs FLORES & SmolSent.

6 Ablation Studies

Impact of the generator and the number of topics. In this section, we focus on Sundanese and study 2 setups. First, we use `gpt-4o-mini-2024-07-18` (OpenAI, 2024) as the generator LLM and analyze the performance of small models (`LLaMA-2-7B` and `LLaMA-3-8B`) fine-tuned on 55k sentences. Second, we reduce the number of topics from 67,573 to a curated list of 509 elements and analyze how the fine-tuned models perform for both generator(s). As shown in Table 4, a stronger generator (here `gpt-4o-mini`)

Models	Basque		Hausa		Igbo		Kinyarwanda		Nepali	
	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX
LLaMA-3.1-8B-It ZERO-SHOT	15.47	11.15	9.61	11.75	7.75	17.39	3.73	20.71	12.21	7.66
LLaMA-3.1-8B-It 5-SHOT FLORES	17.53	9.08	11.85	9.98	9.88	14.51	**5.97**	19.57	**17.18**	6.37
LLaMA-3.1-8B-It 5-SHOT TOPXGEN	**18.05**	**8.61**	**12.29**	**9.02**	**9.93**	**13.66**	5.69	**19.16**	17.16	**6.02**
Methods	**Somali**		**Sundanese**		**Swahili**		**Urdu**		**Xhosa**	
	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX
LLaMA-3.1-8B-It ZERO-SHOT	4.20	19.63	10.30	10.71	20.44	9.21	19.24	5.93	3.41	23.41
LLaMA-3.1-8B-It 5-SHOT FLORES	6.04	17.12	**14.06**	8.56	23.47	8.69	20.75	5.54	**4.67**	22.00
LLaMA-3.1-8B-It 5-SHOT TOPXGEN	**6.50**	**15.39**	13.28	**8.14**	**24.14**	**8.22**	**21.06**	**5.26**	4.63	**21.08**

Table 3: BLEU and MetricX scores for 10 English → X directions from FLORES 200 with ICL. Best results are highlighted in bold. Scores that are statistically equivalent to the best are underlined.

results in stronger students. Moreover, we observe that fine-tuning with more topics achieves stronger results, suggesting that diversity matters for building strong student models. The number of topics is also an important aspect when you generate data on a larger scale, beyond the numbers that we explore in this paper.

Models	Less Topics		Usual Topics	
	BLEU	MetricX	BLEU	MetricX
Gemma 3 27B It	17.01	4.77	17.01	4.77
LLaMA-2-7B uni. Gemma	14.03	6.04	15.31	5.72
LLaMA-3-8B uni. Gemma	16.97	4.46	**18.15**	**4.35**
gpt-4o-mini-2024-07-18	24.17	3.46	24.17	3.46
LLaMA-2-7B uni. GPT	17.65	5.79	18.25	5.65
LLaMA-3-8B uni. GPT	19.79	4.21	**20.51**	**4.09**

Table 4: Results for English → Sundanese direction on FLORES 200 (BLEU and MetricX scores) for different generators and topics.

Impact of the Back-translator. Throughout the paper, we used NLLB-200-3.3B as the back-translator. In this section, we study two setups. First, we analyze how the results change when the generator is also used as the back-translator. Back-translation is performed with greedy 5-shot with BM25 retrieval (Robertson et al., 1995; Lù, 2024) from the FLORES-200 dev set following (Zebaze et al., 2025b). We fine-tune separate models on Hausa and Sundanese, and compare their results with using NLLB-200-3.3B as the back-translator in Table 5. Using the generator as the back-translator works almost as well as using NLLB-200-3.3B. This is a direct by-product of the fluency and translation literacy of state-of-the-art LLMs in English.

Second, we use the [fine-tuned] student as the back-translator (also in 5-shot) with the idea of performing iterative self-improvement (Zelikman et al., 2022; Chen et al., 2024) via iterative back-translation (He et al., 2016; Lample et al., 2018; Artetxe et al., 2018). For a language l, we start by using M_0 = LLaMA-2-7B to back-translate Y^l (generated via TOPXGEN) and create X_0. We create M_1 by fine-tuning M_0 on $(Y^l \rightarrow X_0) \cup (X_0 \rightarrow Y^l)$. Similarly, we create X_1 by back-translating Y^l with M_1, build M_2 by fine-tuning M_0 on $(Y^l \rightarrow X_1) \cup (X_1 \rightarrow Y^l)$ and so on. The same model performs the translation into both directions. We apply that pipeline with LLaMA-2-7B and LLaMA-3-8B on data generated by gpt-4o-mini in Sundanese (Table 6). At round 0—when the base student model also serves as the back-translator—we already observe improvements in both translation directions, though they are smaller than when using NLLB as the back-translator. After one iteration, LLaMA-2-7B shows further gains, with a +3 BLEU increase and a -1 MetricX point drop for English-to-Sundanese, and a modest improvement in the reverse direction. However, the self-improvement process soon plateaus, largely because performance on the English side stagnates. In contrast, LLaMA-3-8B fails to significantly improve Sundanese-to-English performance at round 0, preventing the iterative process from taking off. We hypothesize that self-improvement stalls when the model's English-side performance nears that of the generator on the Sun-

Methods	Hausa		Sundanese	
	BLEU	MetricX	BLEU	MetricX
LLaMA-2-7B uni. NLLB	**13.77**	**8.42**	**15.42**	**5.75**
LLaMA-2-7B uni. GEMMA	13.14	9.31	15.19	5.83
LLaMA-3-8B uni. NLLB	**20.52**	**5.07**	**18.16**	**4.26**
LLaMA-3-8B uni. GEMMA	19.58	5.64	17.98	4.31

Table 5: Full quantitative results for English ↔ Hausa and English ↔ Sundanese on FLORES 200 (BLEU and MetricX scores). Best results are highlighted in bold. Scores that are statistically equivalent to the best are underlined.

danese side. Initializing the pipeline with higher-quality Sundanese data than provided by TopXGen could therefore potentially support more improvement and additional self-iteration rounds.

Models	English to Sundanese		Sundanese to English	
	BLEU	MetricX	BLEU	MetricX
`gpt-4o-mini`	24.17	3.46	39.66	2.67
`LLaMA-2-7B`	6.43	16.65	16.30	8.76
Round 0	10.78	6.86	21.84	6.41
Round 1	14.08	5.82	**22.36**	**6.37**
Round 2	**14.30**	5.79	22.25	6.53
Round 3	14.10	**5.69**	21.85	6.52
`LLaMA-3-8B`	14.17	8.90	33.77	3.83
Round 0	19.71	**3.92**	**35.55**	3.48
Round 1	19.70	4.00	35.37	3.48
Round 2	19.75	3.94	35.40	**3.46**
Round 3	**19.84**	3.95	35.44	3.47

Table 6: Results for English ↔ Sundanese direction on FLORES-200 (BLEU and MetricX scores) in 5-shot during Self-Improvement. Best results are highlighted in bold. Scores that are statistically equivalent to the best are underlined.

Impact of the temperature of the generation of the generator. A high temperature generally correlates with more diversity in the generations, but working with LRLs, there is a risk of going off track in terms of quality. Here we consider four temperatures and analyze the zero-shot (beam size = 5) performance on FLORES-200 every 200 steps after respectively fine-tuning `LLaMA-3-8B` on 50k and 40k sentence pairs generated by TopXGen with `Gemma-3-27B-It` respectively in Sundanese and Hausa. As illustrated in Figure 4, generating with a temperature of $T = 1.0$ leads to a stronger model after fine-tuning, yielding up to 3 BLEU points more than $T = 0.0$ and $T = 0.5$, while maintaining comparable MetricX scores. Using a temperature of $T = 1.2$ does not yield improvements over $T = 1.0$ in terms of BLEU, and results in lower MetricX scores compared to all other temperature values. Our attempts to generate content with a temperature greater than 1.2 (typically 1.5, 2.0 and 5.0) resulted in garbage: nonsensical or invented "sentences" that mix morphemes, word fragments, and orthographic features from multiple scripts and languages.

7 Conclusion

We presented TopXGen, a scalable pipeline for constructing synthetic parallel datasets for MT into LRLs. Our method generates diverse monolingual data directly in a wide range of LRLs using topic-guided prompting (He et al., 2024; Aycock and Bawden, 2024), which we then back-translate into a HRL, building synthetic parallel datasets. Our experiments demonstrate that models trained on data generated via TopXGen achieve strong performance across both unidirectional and multidirectional scenarios, approaching the generator's performance while requiring only small data volumes and limited GPU resources. Furthermore, existing generation strategies such as Self-Instruct and KNN-Instruct are compatible with our pipeline but fall short in terms of performance. TopXGen enables the creation of synthetic datasets that are comparable to professionally written text for ICL and fine-tuning, offering a practical and effective solution for low-resource multilingual MT.

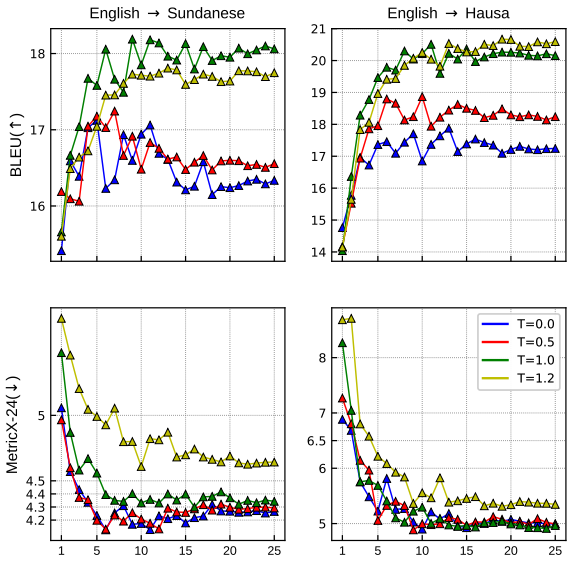

Figure 4: Impact of the temperature of the Generator.

Limitations

In this paper, we demonstrate that it is possible to use synthetic data to get smaller language models to improve their translation capabilities into low-resource languages. However, one limitation is the requirement of a language model (generator) that "understands pretty well" the languages of interest. As suggested in the paper, using monolingual data crawled from the internet can alleviate such a limitation. However it can be hard to find or manually build high-quality monolingual data covering a wide variety of topics in some languages.

Acknowledgments

This work was partly funded by Rachel Bawden and Benoît Sagot's chairs in the PRAIRIE insti-

tute, now PRAIRIE-PSAI, funded by the French national agency ANR, respectively as part of the "Investissements d'avenir" programme under the reference ANR-19-P3IA-0001 and as part of the "France 2030" strategy under the reference ANR-23-IACL-0008. It was also partly funded by the French *Agence Nationale de la Recherche* (ANR) under the project TraLaLaM ("ANR-23-IAS1-0006"). This work was granted access to the HPC resources of IDRIS under the allocation 2025-AD011015933 made by GENCI.

References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a Cleaner Document-Oriented Multilingual Crawled Corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4344–4355, Marseille, France. European Language Resources Association.

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.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. Unsupervised Statistical Machine Translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3632–3642, Brussels, Belgium. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the Cross-lingual Transferability of Monolingual Representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

AI at Meta. 2025. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. `https://ai.meta.com/blog/llama-4-multimodal-intelligence/`.

Seth Aycock and Rachel Bawden. 2024. Topic-guided Example Selection for Domain Adaptation in LLM-based Machine Translation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 175–195, St. Julian's, Malta. Association for Computational Linguistics.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, Theresa Breiner, Vera Axelrod, Jason Riesa, Yuan Cao, Mia Xu Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, and 5 others. 2022. Building Machine Translation Systems for the Next Thousand Languages. *Preprint*, arXiv:2205.03983.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 Conference on Machine Translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Loubna Ben Allal, Anton Lozhkov, Guilherme Penedo, Thomas Wolf, and Leandro von Werra. 2024. Cosmopedia.

Nicola Bertoldi and Marcello Federico. 2009. Domain Adaptation for Statistical Machine Translation with Monolingual Resources. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 182–189, Athens, Greece. Association for Computational Linguistics.

Nikolay Bogoychev and Rico Sennrich. 2020. Domain, Translationese and Noise in Synthetic Data for Neural Machine Translation. *Preprint*, arXiv:1911.03362.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Ondřej Bojar and Aleš Tamchyna. 2011. Improving Translation Model by Monolingual Data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 330–336, Edinburgh, Scotland. Association for Computational Linguistics.

Eleftheria Briakou, Jiaming Luo, Colin Cherry, and Markus Freitag. 2024. Translating Step-by-Step: Decomposing the Translation Process for Improved Translation Quality of Long-Form Texts. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1301–1317, Miami, Florida, USA. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Franck Burlot and François Yvon. 2018. Using Monolingual Data in Neural Machine Translation: a Systematic Study. In *Proceedings of the Third Conference on Machine Translation: Research Papers*,

pages 144–155, Brussels, Belgium. Association for Computational Linguistics.

Happy Buzaaba, Alexander Wettig, David Ifeoluwa Adelani, and Christiane Fellbaum. 2025. Lugha-Llama: Adapting Large Language Models for African Languages. *Preprint*, arXiv:2504.06536.

Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024. LLMs Are Few-Shot In-Context Low-Resource Language Learners. 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 405–433, Mexico City, Mexico. Association for Computational Linguistics.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged Back-Translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63, Florence, Italy. 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, and Solo Farabado Cissé. 2025. SMOL: Professionally translated parallel data for 115 under-represented languages. *Preprint*, arXiv:2502.12301.

Sahil Chaudhary. 2023. Code Alpaca: An Instruction-following LLaMA model for code generation. https://github.com/sahil280114/codealpaca.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. 2024. Self-Play Fine-Tuning Converts Weak Language Models to Strong Language Models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 6621–6642. PMLR.

Team Cohere, Aakanksha, Arash Ahmadian, Marwan Ahmed, Jay Alammar, Yazeed Alnumay, Sophia Althammer, Arkady Arkhangorodsky, Viraat Aryabumi, Dennis Aumiller, Raphaël Avalos, Zahara Aviv, Sammie Bae, Saurabh Baji, Alexandre Barbet, Max Bartolo, Björn Bebensee, Neeral Beladia, Walter Beller-Morales, and 207 others. 2025. Command A: An Enterprise-Ready Large Language Model. *Preprint*, arXiv:2504.00698.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Y. Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, and 19 others. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. *CoRR*, abs/2207.04672.

Yiming Cui, Ziqing Yang, and Xin Yao. 2024. Efficient and Effective Text Encoding for Chinese LLaMA and Alpaca. *Preprint*, arXiv:2304.08177.

Dan Dan Friedman and Adji Bousso Dieng. 2023. The vendi score: A diversity evaluation metric for machine learning. *Transactions on machine learning research*.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya Expanse: Combining Research Breakthroughs for a New Multilingual Frontier. *Preprint*, arXiv:2412.04261.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 514 others. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. 2023. SONAR: Sentence-Level Multimodal and Language-Agnostic Representations. *Preprint*, arXiv:2308.11466.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A Massive Collection of Cross-Lingual Web-Document Pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969, Online. Association for Computational Linguistics.

Maxim Enis and Mark Hopkins. 2024. From LLM to NMT: Advancing Low-Resource Machine Translation with Claude. *Preprint*, arXiv:2404.13813.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, and 1 others. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – News Test References for MT Evaluation of 128 Languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

Markus Freitag and Yaser Al-Onaizan. 2017. Beam Search Strategies for Neural Machine Translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 56–60, Vancouver. Association for Computational Linguistics.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 178 others. 2024. Gemma 2: Improving Open Language Models at a Practical Size. *Preprint*, arXiv:2408.00118.

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.

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent Machine Translation Evaluation through Fine-grained Error Detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks Are All You Need. *Preprint*, arXiv:2306.11644.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual Learning for Machine Translation. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2024. Exploring Human-Like Translation Strategy with Large Language Models. *Transactions of the Association for Computational Linguistics*, 11:229–246.

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. *Preprint*, arXiv:2302.09210.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Alex Jones, Isaac Caswell, Ishank Saxena, and Orhan Firat. 2023. Bilex Rx: Lexical Data Augmentation for Massively Multilingual Machine Translation. *Preprint*, arXiv:2303.15265.

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.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google Submission to the WMT 2024 Metrics Shared Task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical Significance Tests for Machine Translation Evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Jianshang Kou, Benfeng Xu, Chiwei Zhu, and Zhendong Mao. 2024. KNN-Instruct: Automatic Instruction Construction with K Nearest Neighbor Deduction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10337–10350, Miami, Florida, USA. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Wen Lai, Mohsen Mesgar, and Alexander Fraser. 2024. LLMs Beyond English: Scaling the Multilingual Capability of LLMs with Cross-Lingual Feedback. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8186–8213, Bangkok, Thailand. Association for Computational Linguistics.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-Based & Neural Unsupervised Machine Translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium. Association for Computational Linguistics.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023a. Bactrian-X: Multilingual Replicable Instruction-Following Models with Low-Rank Adaptation. *Preprint*, arXiv:2305.15011.

Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Omer Levy, Luke Zettlemoyer, Jason E Weston, and Mike Lewis. 2024. Self-Alignment with Instruction Backtranslation. In *The Twelfth International Conference on Learning Representations*.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023b. Textbooks Are All You Need II: phi-1.5 technical report. *Preprint*, arXiv:2309.05463.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, Yansong Tang, and Dongmei Zhang. 2025. WizardMath: Empowering Mathematical Reasoning for Large Language Models via Reinforced Evol-Instruct. In *The Thirteenth International Conference on Learning Representations*.

Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. 2024. WizardCoder: Empowering Code Large Language Models with Evol-Instruct. In *The Twelfth International Conference on Learning Representations*.

Xing Han Lù. 2024. BM25S: Orders of magnitude faster lexical search via eager sparse scoring. *Preprint*, arXiv:2407.03618.

Benjamin Marie and Atsushi Fujita. 2021. Synthesizing Monolingual Data for Neural Machine Translation. *Preprint*, arXiv:2101.12462.

Benjamin Marie, Raphael Rubino, and Atsushi Fujita. 2020. Tagged Back-translation Revisited: Why Does It Really Work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5990–5997, Online. Association for Computational Linguistics.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive Machine Translation with Large Language Models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland. European Association for Machine Translation.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual Generalization through Multitask Finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive Learning from Complex Explanation Traces of GPT-4. *Preprint*, arXiv:2306.02707.

OpenAI. 2024. Hello GPT-4o. `https://openai.com/index/hello-gpt-4o/`.

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 on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Martin Jaggi, Leandro von Werra, and Thomas Wolf. 2024. FineWeb2: A sparkling update with 1000s of languages.

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.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, 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.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Stephen Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. 1995. Okapi at TREC-3. In *Overview of the Third Text REtrieval Conference (TREC-3)*, pages 109–126. Gaithersburg, MD: NIST.

Holger Schwenk. 2008. Investigations on large-scale lightly-supervised training for statistical machine translation. In *Proceedings of the 5th International Workshop on Spoken Language Translation: Papers*, pages 182–189, Waikiki, Hawaii.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021a. WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021b. CCMatrix: Mining Billions of High-Quality Parallel Sentences on the Web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, and 14 others. 2024. Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567, Bangkok, Thailand. Association for Computational Linguistics.

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2024. A Benchmark for Learning to Translate a New Language from One Grammar Book. In *The Twelfth International Conference on Learning Representations*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An Instruction-following LLaMA model. `https://github.com/tatsu-lab/stanford_alpaca`.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 Technical Report. *Preprint*, arXiv:2503.19786.

Qwen Team. 2024. Qwen2.5: A Party of Foundation Models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *Preprint*, arXiv:2307.09288.

Kosei Uemura, Mahe Chen, Alex Pejovic, Chika Maduabuchi, Yifei Sun, and En-Shiun Annie Lee. 2024. AfriInstruct: Instruction Tuning of African Languages for Diverse Tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13571–13585, Miami, Florida, USA. Association for Computational Linguistics.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-Instruct: Aligning Language Models with Self-Generated Instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, Tianxiang Hu, Shangjie Li, Binyuan Hui, Bowen Yu, Dayiheng Liu, Baosong Yang, Fei Huang, and Jun Xie. 2023. PolyLM: An Open Source Polyglot Large Language Model. *Preprint*, arXiv:2307.06018.

Yuxiang Wei, Federico Cassano, Jiawei Liu, Yifeng Ding, Naman Jain, Zachary Mueller, Harm de Vries, Leandro Von Werra, Arjun Guha, and LINGMING ZHANG. 2024a. SelfCodeAlign: Self-Alignment for Code Generation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Yuxiang Wei, Zhe Wang, Jiawei Liu, Yifeng Ding, and LINGMING ZHANG. 2024b. Magicoder: Empowering Code Generation with OSS-Instruct. In *Forty-first International Conference on Machine Learning*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *Preprint*, arXiv:1910.03771.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. 2024a. WizardLM: Empowering Large Pre-Trained Language Models to Follow Complex Instructions. In *The Twelfth International Conference on Learning Representations*.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024b. A Paradigm Shift in Machine Translation: Boosting Translation Performance of Large Language Models. In *The Twelfth International Conference on Learning Representations*.

Haoran Xu, Kenton Murray, Philipp Koehn, Hieu Hoang, Akiko Eriguchi, and Huda Khayrallah. 2025. X-ALMA: Plug & Play Modules and Adaptive Rejection for Quality Translation at Scale. In *The Thirteenth International Conference on Learning Representations*.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024c. Contrastive Preference Optimization: Pushing the Boundaries of LLM Performance in Machine Translation. In *Forty-first International Conference on Machine Learning*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024. Qwen2 Technical Report. *arXiv preprint arXiv:2407.10671*.

Wen Yang, Chong Li, Jiajun Zhang, and Chengqing Zong. 2023. BigTranslate: Augmenting Large Language Models with Multilingual Translation Capability over 100 Languages. *Preprint*, arXiv:2305.18098.

Armel Zebaze, Benoît Sagot, and Rachel Bawden. 2025a. Compositional Translation: A Novel LLM-based Approach for Low-resource Machine Translation. *Preprint*, arXiv:2503.04554.

Armel Randy Zebaze, Benoît Sagot, and Rachel Bawden. 2025b. In-Context Example Selection via Similarity Search Improves Low-Resource Machine Translation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1222–1252, Albuquerque, New Mexico. Association for Computational Linguistics.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. STaR: Bootstrapping Reasoning With Reasoning. In *Advances in Neural Information Processing Systems*, volume 35, pages 15476–15488. Curran Associates, Inc.

Weihao Zeng, Can Xu, Yingxiu Zhao, Jian-Guang Lou, and Weizhu Chen. 2024. Automatic Instruction Evolving for Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6998–7018, Miami, Florida, USA. Association for Computational Linguistics.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

Tarek Ziadé. 2023. MediaWiki Categories Model. https://huggingface.co/datasets/tarekziade/wikipedia-topics.

A Reproducibility Details

A.1 Models, Datasets and Tools

In Table 7, we list the links to the relevant resources used for the experiments.

A.2 Implementation Details

We use HuggingFace's Transformers library (Wolf et al., 2020). For training the unidirectional models, we use a per-device batch size of 4 and apply gradient accumulation over 4 steps. We use stacking with a maximum length of 512 tokens. Training is conducted over 5,000 steps using a learning rate of 1e-5, with 500 warmup steps. The learning rate decays to zero following a cosine schedule. We also apply a weight decay of 0.01. For the multidirectional models, which are trained across the 10 target languages, we use the same hyperparameters but extend training to 100,000 steps. We adopt the prompt template introduced by Xu et al. (2024b), and compute the loss only on the target sentence. In all experiments, we fine-tune in both directions (i.e., source-to-target and target-to-source), so each parallel sentence pair contributes two samples to the training set. On smaller datasets such as FLORES and SMOLSENT, we reduce the training steps to 1,000. In this setting, the batch size per device is 2, we use 100 warmup steps and we do not perform gradient accumulation. All other hyperparameters remain unchanged.

```
Translate this from English to Hausa:
English: "We now have 4-month-old mice
    that are non-diabetic that used to
    be diabetic," he added.
Hausa:
```

B Additional Experiments

B.1 More baselines

Table 8 presents a comparison of our fine-tuned student models on FLORES-200 against addi-

Datasets	
FLORES-200	https://huggingface.co/datasets/facebook/flores
NTREX HF	hhttps://huggingface.co/datasets/mteb/NTREX
TICO-19	https://huggingface.co/datasets/gmnlp/tico19
Models evaluated	
Aya-expanse-8B	https://huggingface.co/CohereLabs/aya-expanse-8b
Aya-expanse-32B	https://huggingface.co/CohereLabs/aya-expanse-32b
Command-R7B	https://huggingface.co/CohereLabs/c4ai-command-r7b-12-2024
Qwen2.5-7B-Instruct	https://huggingface.co/Qwen/Qwen2.5-7B-Instruct
Qwen2.5-32B-Instruct	https://huggingface.co/Qwen/Qwen2.5-32B-Instruct
LLaMAX2-7B-Alpaca	https://huggingface.co/LLaMAX/LLaMAX2-7B-Alpaca
LLaMAX3-8B-Alpaca	https://huggingface.co/LLaMAX/LLaMAX3-8B-Alpaca
LLaMA-2-7B	https://huggingface.co/meta-llama/Llama-2-7b-hf
LLaMA-3-8B	https://huggingface.co/meta-llama/Meta-Llama-3-8B
Gemma-2-9B-It	https://huggingface.co/google/gemma-2-9b-it
Gemma-2-27B-It	https://huggingface.co/google/gemma-2-27b-it
Gemma-3-4B-It	https://huggingface.co/google/gemma-3-4b-it
Gemma-3-27B-It	https://huggingface.co/google/gemma-3-27b-it
Gemma-3-4B-Pt	https://huggingface.co/google/gemma-3-4b-pt
Gemma-3-27B-Pt	https://huggingface.co/google/gemma-3-27b-pt
LLaMA-3.1-8B-It	https://huggingface.co/meta-llama/Meta-Llama-3.1-8B-Instruct
LLaMA-3.1-70B-It	https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct
NLLB-200-3.3B	https://huggingface.co/facebook/nllb-200-3.3B
Other resources	
MetricX24-XXL	https://huggingface.co/google/metricx-24-hybrid-xxl-v2p6
XCOMET-XL	https://huggingface.co/Unbabel/XCOMET-XL
BM25s	https://github.com/xhluca/bm25s
FastText	https://huggingface.co/facebook/fasttext-language-identification
Wikipedia topics	https://huggingface.co/datasets/tarekziade/wikipedia-topics
vLLM (Kwon et al., 2023)	https://github.com/vllm-project/vllm

Table 7: Links to datasets, benchmarks and models.

tional baselines. They surpass all evaluated 7B models, and notably, unidirectional models based on LLaMA-3-8B with beam search outperform LLaMA-3.3-70B It.

B.2 Translation into English

In Table 9, we present the performance of the fine-tuned student models on FLORES-200 for translation from low-resource languages into English. The results mirror our observations in the reverse direction: unidirectional fine-tuning consistently outperforms multidirectional fine-tuning, and using a strong base model (LLaMA-3-8B) leads to significantly better performance—comparable to models that are up to three times larger.

B.3 ChrF++ and XCOMET scores

We consider the XCOMET metric (Guerreiro et al., 2024), specifically its reference-based version: XCOMET-XL (which supports the same 100 languages as XLM RoBERTa (Conneau et al., 2020)). XCOMET scores range from 0 and 1, which we rescale to 0 to 100 (higher scores are better). We evaluate the translations produced on FLORES-200 by models fine-tuned on the TOPXGEN dataset and report the results in Table 10. The results are consistent with BLEU and MetricX.

B.4 Results on NTREX-128 and TICO-19

In this section, we evaluate the models on 2 additional benchmarks:

- **NTREX 128** (Barrault et al., 2019; Federmann et al., 2022) is an MT benchmark derived from WMT19 news data translated by professional human translators. It contains 1997 parallel sentences and is recommended for the evaluation of from-English translation directions. We use the first 1000 sentence pairs for evaluation, and the last 997 sentence pairs as the selection pool.

- **TICO-19** (Anastasopoulos et al., 2020) is

Models	Basque		Hausa		Igbo		Kinyarwanda		Nepali	
	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX
Other Baselines										
`LLaMA-3.3-70B It`	25.02	5.56	18.15	6.09	14.36	9.05	8.11	14.05	23.28	4.60
`LLaMA-3.1-8B-It`	15.47	11.15	9.61	11.75	7.75	17.39	3.73	20.71	12.21	7.66
`Aya-expanse-8B`	4.53	19.98	3.33	16.53	3.66	22.41	2.25	23.11	5.28	6.64
`Qwen-2.5-7B-It`	4.31	22.04	4.39	19.37	4.13	22.72	1.94	24.28	5.22	12.50
`LLaMAX2-7B Alpaca`	10.56	15.38	17.17	6.75	9.34	14.59	4.57	18.85	12.07	14.97
Our Models										
`LLaMA-2-7B` uni.	13.00	14.76	13.11	8.57	12.30	10.97	7.30	15.87	15.11	7.98
`LLaMA-2-7B` uni. *beam size=5*	14.93	12.55	13.77	8.42	12.90	10.04	7.61	14.73	16.08	6.31
`LLaMA-3-8B` uni.	23.70	6.25	19.65	5.20	16.28	7.31	11.76	9.91	21.88	4.21
`LLaMA-3-8B` uni. *beam size=5*	**25.64**	**5.27**	**20.52**	**5.07**	**17.02**	**6.54**	**13.60**	**8.51**	**23.24**	**3.77**
`LLaMA-3-8B` multi.	21.77	6.95	18.22	6.05	15.54	7.96	9.76	12.96	20.84	4.52
`LLaMA-3-8B` multi. *beam size=5*	24.07	5.68	19.36	5.76	16.09	7.13	11.62	11.40	22.37	3.94

Models	Somali		Sundanese		Swahili		Urdu		Xhosa	
	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX
Other Baselines										
`LLaMA-3.3-70B It`	9.35	11.14	17.27	5.46	33.89	4.75	26.31	3.88	6.21	16.14
`LLaMA-3.1-8B-It`	4.20	19.63	10.30	10.71	20.44	9.21	19.24	5.93	3.41	23.41
`Aya-expanse-8B`	4.46	18.08	9.66	3.42	4.82	21.81	6.60	9.03	3.60	23.99
`Qwen-2.5-7B-It`	3.50	21.19	5.90	16.69	4.94	21.92	5.10	13.56	2.55	24.28
`LLaMAX2-7B Alpaca`	9.00	10.72	8.60	12.13	21.62	7.34	9.17	17.59	12.03	10.31
Our Models										
`LLaMA-2-7B` uni.	8.89	11.73	14.70	7.19	19.19	11.42	14.89	8.43	9.25	15.24
`LLaMA-2-7B` uni. *beam size=5*	8.20	10.63	15.42	5.75	20.96	9.52	16.33	6.63	7.67	13.09
`LLaMA-3-8B` uni.	13.20	7.00	16.84	4.88	30.99	5.23	22.43	4.16	12.39	9.33
`LLaMA-3-8B` uni. *beam size=5*	**13.70**	**6.17**	**18.16**	**4.26**	**33.49**	**4.51**	**23.51**	**3.78**	**13.93**	**7.77**
`LLaMA-3-8B` multi.	12.03	7.94	16.24	5.66	28.53	6.03	21.65	4.54	11.74	10.56
`LLaMA-3-8B` multi. *beam size=5*	12.71	7.05	17.32	4.83	31.11	5.07	22.92	3.96	13.15	8.61

Table 8: BLEU and MetricX scores for 10 English → X directions from FLORES 200. Best results after fine-tuning are highlighted in bold.

an MT benchmark comprising texts on the COVID-19 pandemic covering 35 languages. Its validation and test sets consist of 971 (used as a selection pool) and 2100 samples respectively.

We focus on translating from English. We report the results obtained on NTREX-19 in Table 11 and those obtained on TICO-19 in Table 12. On both benchmarks, the best fine-tuned model usually ranks third, behind `NLLB-200-3.3B` and `Gemma-3-27B-It`.

B.5 Fine-tuning the generator and the back-translator

We conduct additional experiments to assess the impact of fine-tuning both the back-translator, `NLLB-200-3.3B`, and the generator's base model, `Gemma-3-27B-PT`, on the TOPXGEN dataset. For NLLB, we retain the same training hyperparameters as before, modifying only the maximum sequence length to 128. In contrast, for Gemma, we apply LoRA (Hu et al., 2022), fine-tuning the `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj` modules with rank $r = 32$, scaling factor $\alpha = 64$, and dropout rate 0.01. We use the same training hyperparameters as in previous runs and fine-tune on 4×H100 80GB GPUs. We compute BLEU and MetricX scores every 200 training steps and present the results in Figure 5. For languages like Basque, Nepali, and Urdu—where `Gemma-3-27B-IT` outperforms NLLB (see Table 2)—we observe that fine-tuning NLLB on the TOPXGEN dataset yields substantial improvements. Specifically, NLLB gains up to 3 BLEU points in Basque and Nepali, and 1 point in Urdu, with corresponding MetricX gains of approximately 1 point, and up to 3 in Nepali. In cases where NLLB and `Gemma-3-27B-IT` perform comparably (e.g., Sundanese and Swahili), BLEU remains largely unchanged, while MetricX shows modest improvements. Conversely, when `Gemma-3-27B-IT` underperforms NLLB, further fine-tuning NLLB on TOPXGEN leads to a decline in performance—more pronounced in BLEU, but also noticeable in MetricX. Fine-tuning `Gemma-3-27B-PT` on generations from `Gemma-3-27B-IT` yields significant gains in translation quality. However, we observe diminishing returns as training progresses across all directions. We hypothesize that the model does not learn trans-

Methods	Basque		Hausa		Igbo		Kinyarwanda		Nepali	
	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX
NLLB-200-3.3B	36.21	2.46	36.98	4.27	32.88	4.92	35.17	3.96	44.98	2.93
Gemma-3-27B-It	35.59	1.98	31.20	4.28	25.30	5.74	28.59	4.45	40.73	2.19
Gemma-2-27B-It	35.07	2.18	28.76	4.83	21.23	7.14	21.86	6.67	39.14	2.63
LLaMA-3.1-70B It	36.53	2.24	31.80	4.75	26.39	6.25	23.19	7.09	40.74	2.84
Gemma-2-9B-It	32.99	2.82	27.04	5.65	18.59	8.50	19.45	8.28	37.09	2.92
LLaMA-3.1-8B-It	26.78	4.12	19.91	8.69	15.27	10.89	12.54	12.17	28.03	4.97
Command-R7B	12.21	11.63	5.72	16.84	5.94	18.60	6.46	17.84	23.25	6.49
Aya-expanse-32B	24.41	5.16	9.56	15.39	8.49	15.20	10.79	13.99	28.33	4.49
Aya-expanse-8B	12.13	10.34	6.68	16.88	6.37	16.23	6.84	17.29	18.22	6.75
Qwen-2.5-32B-It	23.55	6.48	12.10	14.63	10.15	14.38	10.09	15.74	33.27	4.37
Qwen-2.5-7B-It	13.27	11.06	8.09	17.37	7.31	17.73	7.53	18.99	21.85	7.06
LLaMAX2-7B Alpaca	15.36	8.73	20.16	6.77	12.15	12.39	10.77	11.19	15.28	12.19
LLaMAX3-8B Alpaca	29.27	3.09	27.84	4.56	21.72	6.60	17.58	7.78	34.02	3.21
LLaMA-2-7B 5-SHOT BM25	8.22	13.54	6.68	17.13	5.83	17.39	5.98	17.53	10.32	11.24
LLaMA-3-8B 5-SHOT BM25	31.82	2.74	26.93	5.61	22.69	7.43	16.71	9.06	33.43	3.76
LLaMA-2-7B uni.	21.00	8.41	20.82	9.79	16.27	12.21	16.80	11.20	24.29	7.43
LLaMA-2-7B uni. *beam size=5*	21.78	7.75	20.81	9.26	15.38	11.59	17.09	10.77	26.22	6.75
LLaMA-3-8B uni.	34.71	3.21	32.42	6.07	26.22	7.84	27.42	6.70	38.69	3.79
LLaMA-3-8B uni. *beam size=5*	**35.43**	**3.07**	**33.07**	**5.81**	**27.70**	**7.35**	**28.35**	**6.33**	**40.00**	**3.62**
LLaMA-3-8B multi.	34.51	3.19	31.75	6.33	26.27	8.04	25.38	8.17	38.81	3.98
LLaMA-3-8B multi. *beam size=5*	35.07	3.05	32.67	6.04	27.56	7.60	26.34	7.74	39.41	3.76

Methods	Somali		Sundanese		Swahili		Urdu		Xhosa	
	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX
NLLB-200-3.3B	31.02	5.50	42.04	3.33	45.70	2.88	39.73	3.08	40.06	3.71
Gemma-3-27B-It	28.76	4.57	37.78	2.92	46.09	2.19	36.90	2.14	31.96	5.14
Gemma-2-27B-It	22.13	6.85	32.60	3.74	46.23	2.26	37.04	2.54	28.17	5.76
LLaMA-3.1-70B It	21.32	8.06	36.14	3.64	46.89	2.41	40.01	2.37	22.71	8.68
Gemma-2-9B-It	20.44	7.95	31.41	4.08	43.47	2.66	34.82	2.98	26.05	6.66
LLaMA-3.1-8B-It	11.38	14.42	25.36	5.88	34.37	4.40	29.05	3.96	12.94	13.18
Command-R7B	5.52	17.66	17.75	8.86	16.51	10.98	15.92	8.86	8.72	16.50
Aya-expanse-32B	12.01	13.60	28.25	4.64	28.22	5.89	28.70	4.00	14.90	11.56
Aya-expanse-8B	8.44	16.54	18.89	7.01	12.75	11.65	17.97	7.55	8.66	15.76
Qwen-2.5-32B-It	10.53	15.31	28.98	5.40	28.98	6.67	31.00	3.93	15.93	12.47
Qwen-2.5-7B-It	8.05	18.09	21.69	7.74	14.79	12.09	22.43	6.37	9.54	16.22
LLaMAX2-7B Alpaca	12.35	11.21	21.17	8.13	27.67	5.96	12.71	13.89	18.75	7.29
LLaMAX3-8B Alpaca	24.03	5.57	31.00	4.02	38.65	3.03	32.03	3.30	28.42	5.06
LLaMA-2-7B 5-SHOT BM25	6.02	18.73	16.30	8.76	10.78	12.84	10.93	11.21	8.17	16.63
LLaMA-3-8B 5-SHOT BM25	14.09	11.42	33.77	3.83	39.46	3.41	33.30	3.41	17.08	10.34
LLaMA-2-7B uni.	15.94	11.93	30.78	6.44	27.55	7.99	21.30	8.17	20.73	11.28
LLaMA-2-7B uni. *beam size=5*	16.33	11.51	31.30	6.07	28.77	7.44	23.33	7.29	19.68	10.47
LLaMA-3-8B uni.	24.69	8.32	38.46	4.19	42.77	3.78	35.61	3.74	31.38	6.98
LLaMA-3-8B uni. *beam size=5*	**25.75**	**7.87**	**39.15**	**3.98**	**44.04**	**3.63**	**36.62**	**3.51**	**32.14**	**6.71**
LLaMA-3-8B multi.	23.29	9.05	37.55	4.49	42.51	3.95	35.44	3.91	29.34	7.83
LLaMA-3-8B multi. *beam size=5*	24.53	8.63	38.53	4.34	43.86	3.71	36.33	3.69	30.62	7.39

Table 9: BLEU and MetricX scores for 10 X → English directions from FLORES 200. Best results after fine-tuning are highlighted in bold.

lation per se through this process; rather, early training steps help it infer the structure of the translation task. Once this template is internalized, the model relies on its pretrained knowledge for translation. Further fine-tuning on its own generations appears counterproductive. As shown in Table 13, the fine-tuned models do not surpass the performance of Gemma-3-27B-PT in the 5-SHOT BM25 setting, suggesting that self-generated data does not enhance the model's understanding of the target languages.

B.6 Analysis of the TOPXGEN dataset

For each language, we compute the average MetricX-24 quality estimation (QE) scores over the first 20K sentences. We also report the average number of words and tokens per sentence for both the source (English) and target sides. As shown in Table 14, source sentences have a relatively consistent average word count across languages. However, in terms of tokens, sentences in low-resource languages (LRLs) typically require twice as many tokens as their English counterparts. For language pairs involving Hausa, Nepali, Somali, and Urdu, TOPXGEN achieves higher QE scores than both FLORES and SMOL, suggesting

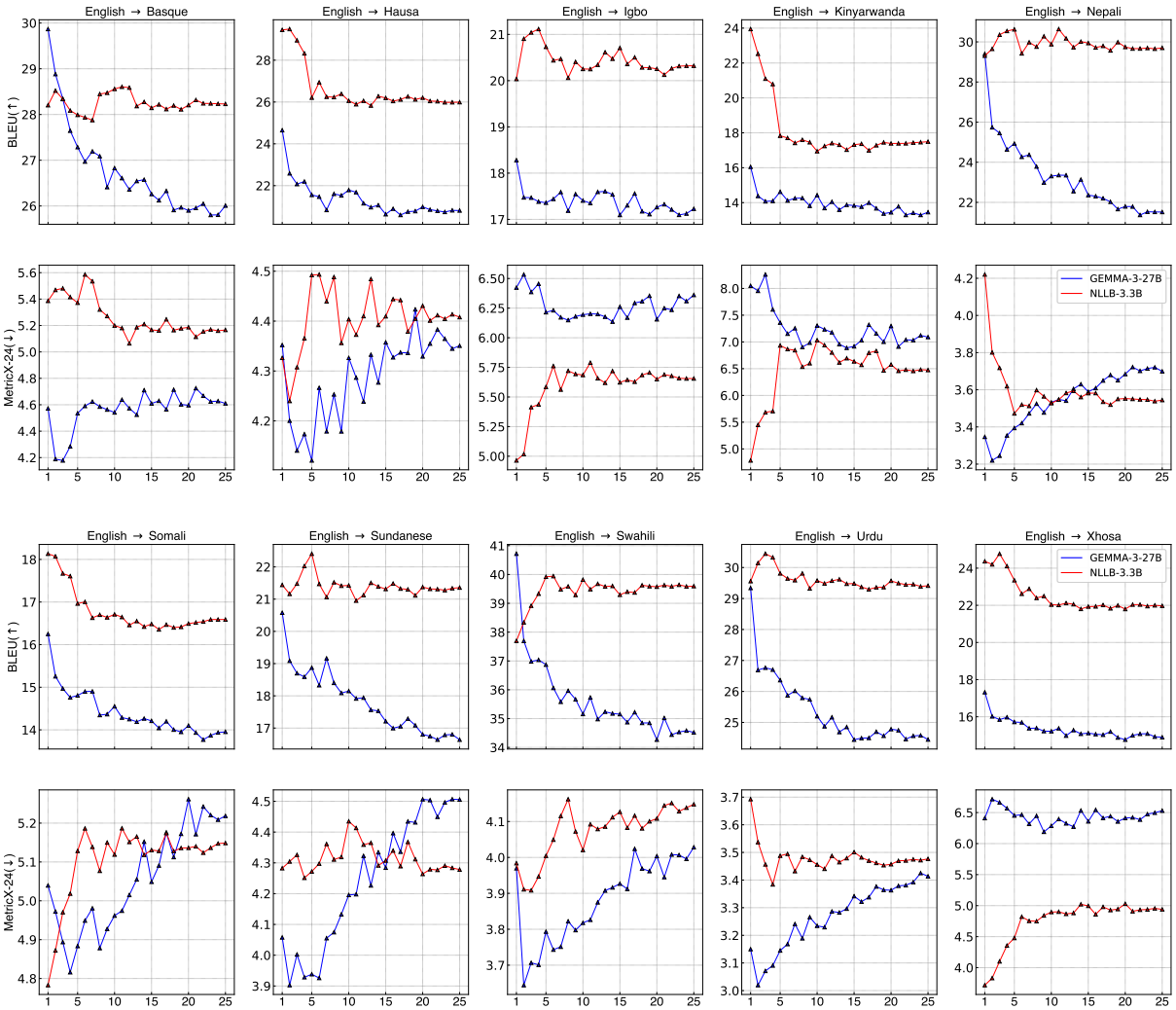

Figure 5: BLEU and MetricX results for 10 English→X directions from FLORES 200. We fine-tune `NLLB-200-3.3B` and `Gemma-3-27B-PT`. We consider 1 model per direction and report the scores (greedy decoding) every 200 steps.

Methods	Basque		Hausa		Igbo		Kinyarwanda		Nepali	
	chrF++	XCOMET	chrF++	XCOMET	chrF++	XCOMET	chrF++	XCOMET	chrF++	XCOMET
NLLB-200-3.3B	46.31	76.88	51.58	65.65	40.56	19.04	46.90	26.93	44.59	75.82
Gemma-3-27B-It	49.71	78.97	44.94	58.57	36.41	18.40	37.18	23.12	45.76	83.12
Gemma-2-27B-It	46.69	73.43	42.23	58.72	30.65	18.95	28.31	24.61	43.07	79.71
LLaMA-3.1-70B It	49.28	77.61	43.40	55.46	36.13	18.74	29.38	24.69	45.74	78.20
Gemma-2-9B-It	41.00	60.10	38.41	50.13	25.82	19.76	21.99	24.25	39.87	74.51
LLaMA-3.1-8B-It	39.97	51.19	31.30	31.23	22.76	19.77	19.25	23.33	32.68	55.28
Command-R7B	18.76	54.79	11.24	27.13	9.60	23.63	12.54	22.83	23.29	50.46
Aya-expanse-32B	32.58	30.74	23.35	21.84	19.43	19.58	22.10	24.56	30.64	51.95
Aya-expanse-8B	25.53	26.96	19.29	22.00	17.22	19.02	19.13	23.35	22.01	62.55
Qwen-2.5-32B-It	30.99	27.15	20.99	22.32	20.45	20.23	20.03	24.86	30.76	45.73
Qwen-2.5-7B-It	26.31	21.13	20.02	20.42	15.90	19.98	15.56	21.85	23.11	31.12
LLaMAX2-7B Alpaca	31.16	34.21	40.75	36.32	30.27	18.74	20.11	24.34	36.55	39.17
LLaMAX3-8B Alpaca	34.38	56.20	41.54	55.12	33.60	19.42	19.07	25.07	41.82	69.42
LLaMA-2-7B 5-SHOT BM25	16.90	22.75	11.64	25.21	11.25	21.95	13.92	23.18	15.40	26.13
LLaMA-3-8B 5-SHOT BM25	41.36	62.92	34.05	39.59	23.20	20.60	18.37	23.94	37.47	61.10
LLaMA-2-7B uni.	36.44	39.97	37.46	41.80	31.02	18.47	26.53	22.55	35.84	51.37
LLaMA-2-7B uni. *beam size=5*	38.59	48.68	37.62	46.40	32.41	17.56	27.47	22.87	38.38	61.84
LLaMA-3-8B uni.	46.83	74.21	44.73	61.11	36.65	18.53	34.51	22.25	42.99	78.17
LLaMA-3-8B uni. *beam size=5*	**48.28**	**80.10**	**45.21**	**64.56**	**37.73**	**18.60**	**36.32**	**23.22**	**44.72**	**82.32**
LLaMA-3-8B multi.	45.36	71.05	42.80	57.24	35.65	18.35	30.30	21.93	42.39	75.78
LLaMA-3-8B multi. *beam size=5*	47.02	77.89	43.76	61.57	36.61	18.24	32.23	23.08	44.42	81.35

Methods	Somali		Sundanese		Swahili		Urdu		Xhosa	
	chrF++	XCOMET	chrF++	XCOMET	chrF++	XCOMET	chrF++	XCOMET	chrF++	XCOMET
NLLB-200-3.3B	41.58	63.09	45.28	66.86	58.60	78.97	47.17	70.44	46.78	51.58
Gemma-3-27B-It	39.02	57.19	42.95	70.17	58.37	81.34	44.61	77.74	38.44	39.86
Gemma-2-27B-It	32.03	35.44	39.54	61.69	58.13	82.22	32.32	33.64	41.68	71.66
LLaMA-3.1-70B It	33.33	36.81	42.48	67.08	57.52	78.73	46.05	73.17	28.32	29.61
Gemma-2-9B-It	27.81	27.45	37.63	58.40	53.64	75.32	38.20	62.87	27.10	28.04
LLaMA-3.1-8B-It	22.56	21.65	33.22	48.63	45.53	56.04	38.60	55.91	18.93	23.06
Command-R7B	12.23	27.42	30.79	73.38	26.97	25.13	19.81	31.53	14.80	24.35
Aya-expanse-32B	28.18	26.07	33.84	52.53	33.71	27.80	31.86	44.64	23.70	24.36
Aya-expanse-8B	24.87	22.92	33.46	82.88	25.24	21.24	27.02	37.71	21.26	22.46
Qwen-2.5-32B-It	23.56	21.52	31.17	35.60	34.26	26.47	30.87	39.42	23.01	24.18
Qwen-2.5-7B-It	21.24	20.82	26.60	34.35	25.25	21.48	24.03	26.01	18.34	21.66
LLaMAX2-7B Alpaca	33.31	30.17	31.55	41.34	49.18	51.16	32.08	27.12	34.42	28.34
LLaMAX3-8B Alpaca	35.14	50.75	34.94	56.56	50.83	70.09	38.90	57.27	33.91	36.14
LLaMA-2-7B 5-SHOT BM25	13.76	22.46	24.55	36.89	15.59	23.45	15.02	24.66	12.44	23.61
LLaMA-3-8B 5-SHOT BM25	20.97	24.82	38.27	52.82	46.86	60.37	35.91	54.45	16.69	24.76
LLaMA-2-7B uni.	32.56	32.35	39.99	60.48	44.42	46.61	34.98	40.43	31.05	29.19
LLaMA-2-7B uni. *beam size=5*	33.33	35.75	41.90	67.52	46.66	55.64	37.23	51.21	32.80	33.36
LLaMA-3-8B uni.	38.34	52.08	42.90	70.50	55.21	75.45	42.37	69.20	36.84	38.35
LLaMA-3-8B uni. *beam size=5*	**39.08**	**56.78**	**44.18**	**74.00**	**57.01**	**80.62**	**43.36**	**73.92**	**38.52**	**44.37**
LLaMA-3-8B multi.	36.82	47.22	42.00	68.25	53.11	72.29	41.38	66.97	35.77	36.21
LLaMA-3-8B multi. *beam size=5*	37.66	52.61	43.27	72.62	54.94	78.22	42.77	72.59	37.37	42.73

Table 10: ChrF++ and XCOMET-XL scores for 10 English → X directions from FLORES 200. Best fine-tuning results are highlighted in bold.

that its topic-guided generation produces natural and coherent text in LRLs, accurately translated by the back-translation model.

The Vendi Score (Dan Friedman and Dieng, 2023), calculated using SONAR embeddings, quantifies the diversity of a text—higher values indicate greater diversity. The results are summarized in Table 14. On the target side, TOPXGEN generally achieves higher Vendi scores than FLORES (e.g., 1.123 vs. 1.096 in Somali), suggesting more diverse generations. Source-side sentences are also more diverse in TOPXGEN, though the difference is less pronounced. Notably, SMOLSENT, despite its smaller size, exhibits high diversity and occasionally surpasses TOPXGEN—particularly in languages like Hausa on the target side. However, as shown in Figure 3, this increased diversity does not consistently lead to better translation quality than that achieved by TOPXGEN.

We also evaluate how well the paragraphs generated by TOPXGEN align with their intended topics. To do this, we generate 1,000 paragraphs per language using `Gemma-3-27B-It`, and ask both `Gemma-3-27B-It` and `Llama-4-Scout-17B-16E-Instruct` to assess whether each paragraph accurately addresses its assigned topic. We consider two settings: the paragraph written in the low-resource language, and its English translation obtained via sentence-by-sentence translation using `NLLB-3.3B`. Results are presented in Table 15. According to `Gemma-3-27B-It`, 97% of the paragraphs it generates are on-topic, though this rate decreases to 90% when the same

Methods	Basque		Hausa		Igbo		Kinyarwanda		Nepali	
	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX
NLLB-200-3.3B	18.29	8.29	25.43	4.51	22.43	5.15	21.05	4.76	22.12	5.32
Gemma-3-27B-It	21.62	5.79	19.54	4.97	16.64	7.76	14.98	7.99	18.38	4.23
Gemma-2-27B-It	18.31	7.36	17.37	5.69	14.07	10.44	8.06	14.85	17.39	4.81
LLaMA-3.1-70B It	21.70	6.61	18.58	6.31	18.44	8.76	8.95	14.04	18.42	5.15
Gemma-2-9B-It	13.40	10.14	14.69	6.94	11.76	14.20	5.03	19.88	14.64	5.62
LLaMA-3.1-8B-It	13.23	12.47	10.46	11.72	10.07	17.09	4.55	20.45	9.60	9.09
Command-R7B	4.02	13.21	2.56	18.52	2.98	20.01	2.92	21.52	4.35	9.79
Aya-expanse-32B	8.79	16.72	6.42	16.62	5.65	20.08	5.00	20.84	8.02	9.08
Aya-expanse-8B	5.04	19.93	4.86	16.32	4.27	21.52	3.29	22.82	4.33	7.39
Qwen-2.5-32B-It	6.91	19.06	6.25	16.65	6.68	19.15	4.23	22.44	7.97	10.57
Qwen-2.5-7B-It	4.69	21.65	5.12	18.91	5.40	21.84	2.97	23.83	4.00	13.34
LLaMAX2-7B Alpaca	10.29	15.88	18.45	6.98	10.10	13.79	5.52	18.36	9.82	14.73
LLaMAX3-8B Alpaca	11.37	11.33	17.37	6.41	15.10	9.80	5.20	19.01	16.97	6.48
LLaMA-2-7B 5-SHOT BM25	4.22	22.20	1.96	21.23	3.57	21.84	3.51	22.34	2.79	19.29
LLaMA-3-8B 5-SHOT BM25	14.61	10.07	12.56	10.41	14.03	15.19	5.45	20.42	12.07	8.17
LLaMA-2-7B uni.	10.95	16.01	13.27	9.32	13.01	11.64	7.42	16.93	10.51	9.51
LLaMA-2-7B uni. *beam size=5*	12.01	14.17	13.45	9.30	14.35	10.70	8.06	15.56	11.62	7.96
LLaMA-3-8B uni.	19.10	7.86	19.44	5.84	17.88	8.19	11.89	11.18	15.89	5.36
LLaMA-3-8B uni. *beam size=5*	**20.77**	**6.78**	**20.02**	**5.84**	**19.00**	**7.38**	**13.16**	**9.66**	**17.22**	**4.81**
LLaMA-3-8B multi.	18.11	8.44	17.91	6.76	16.60	8.55	9.38	13.98	15.35	5.79
LLaMA-3-8B multi. *beam size=5*	18.79	7.13	18.54	6.45	18.18	8.04	10.64	11.97	16.10	5.00

Methods	Somali		Sundanese		Swahili		Urdu		Xhosa	
	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX
NLLB-200-3.3B	19.82	4.74	-	-	38.63	5.38	32.16	5.11	18.73	4.61
Gemma-3-27B-It	14.82	5.48	-	-	36.97	4.23	28.62	3.71	12.41	8.58
Gemma-2-27B-It	11.04	10.83	-	-	35.99	4.53	26.89	4.49	9.38	12.98
LLaMA-3.1-70B It	12.03	10.82	-	-	34.83	4.96	30.50	4.24	7.50	16.12
Gemma-2-9B-It	8.78	14.21	-	-	32.52	5.93	21.08	5.90	7.61	18.14
LLaMA-3.1-8B-It	5.79	19.23	-	-	22.25	9.92	21.19	6.70	3.49	23.06
Command-R7B	2.88	19.62	-	-	6.90	19.57	3.55	13.77	2.44	21.81
Aya-expanse-32B	8.53	14.47	-	-	12.15	16.73	11.18	8.44	5.24	2.39
Aya-expanse-8B	7.08	17.18	-	-	6.81	21.40	7.11	10.08	4.03	23.73
Qwen-2.5-32B-It	5.84	18.67	-	-	12.29	17.44	11.70	10.30	4.98	22.21
Qwen-2.5-7B-It	5.20	20.75	-	-	6.86	21.38	5.56	14.21	3.00	24.10
LLaMAX2-7B Alpaca	11.38	9.57	-	-	24.80	7.91	11.74	15.73	11.10	11.21
LLaMAX3-8B Alpaca	13.66	7.30	-	-	29.45	7.19	22.58	6.60	10.51	11.33
LLaMA-2-7B 5-SHOT BM25	2.78	21.07	-	-	4.02	21.96	3.62	19.31	2.68	23.02
LLaMA-3-8B 5-SHOT BM25	6.03	17.96	-	-	26.37	8.55	19.42	7.57	3.25	21.78
LLaMA-2-7B uni.	10.92	12.75	-	-	19.89	13.11	14.05	9.78	9.00	16.50
LLaMA-2-7B uni. *beam size=5*	11.18	11.31	-	-	21.88	11.21	15.42	8.20	7.75	14.29
LLaMA-3-8B uni.	15.30	7.77	-	-	32.06	6.42	22.34	5.25	11.53	11.41
LLaMA-3-8B uni. *beam size=5*	**16.15**	**6.83**	-	-	**34.31**	**5.53**	**24.26**	**4.62**	**12.18**	**9.66**
LLaMA-3-8B multi.	14.04	8.74	-	-	30.25	7.10	21.80	5.69	10.65	12.16
LLaMA-3-8B multi. *beam size=5*	14.79	7.80	-	-	33.00	5.99	23.51	4.97	11.42	10.31

Table 11: BLEU and MetricX scores for 9 English → X directions from NTREX 128 (Federmann et al., 2022).

paragraphs are translated into English. Similarly, `Llama-4-Scout-17B-16E-Instruct` finds that 93% of the original paragraphs align with their topics, dropping to 83–85% after translation. In summary, the generated paragraphs are generally well-aligned with the provided topics. Even in cases where strict topical alignment is not achieved, the content remains relevant for machine translation training, where the primary requirement is having semantically equivalent sentences across languages.

B.7 Topic Modeling

We use BERTopic[8] (Grootendorst, 2022) to assess whether the most relevant words identified through clustering align with the intended topics. As shown in Table 16, we present results on six Basque paragraphs translated into English from the TOPXGEN dataset. We experiment with two setups: (1) using `gpt-4o-mini` to generate a topic label for each paragraph, and (2) extracting the top 10 words most relevant to each paragraph's context. We find that the GPT-generated topics often encompass the ground truth topics, which are typically more fine-grained. For instance, Theodore Shackley is iden-

[8] https://maartengr.github.io/BERTopic/index.html

Methods	Basque		Hausa		Igbo		Kinyarwanda		Nepali	
	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX
NLLB-200-3.3B	-	-	30.96	3.97	-	-	22.52	7.44	32.72	4.02
Gemma-3-27B-It	-	-	19.59	4.53	-	-	11.29	10.84	27.37	3.44
Gemma-2-27B-It	-	-	18.60	5.29	-	-	6.93	17.01	26.11	3.96
LLaMA-3.1-70B It	-	-	20.98	5.62	-	-	8.07	15.57	26.17	4.20
Gemma-2-9B-It	-	-	16.06	6.60	-	-	4.33	21.08	22.37	4.55
LLaMA-3.1-8B-It	-	-	11.83	11.38	-	-	3.53	21.47	14.75	7.29
Command-R7B	-	-	3.22	18.86	-	-	2.14	22.57	7.24	8.56
Aya-expanse-32B	-	-	6.80	16.78	-	-	3.73	22.19	11.79	7.53
Aya-expanse-8B	-	-	5.38	15.80	-	-	2.79	23.29	7.47	6.37
Qwen-2.5-32B-It	-	-	7.52	16.45	-	-	3.58	23.14	12.52	9.22
Qwen-2.5-7B-It	-	-	5.89	18.88	-	-	2.29	24.38	6.89	12.67
LLaMAX2-7B Alpaca	-	-	17.64	6.58	-	-	5.01	19.74	13.83	14.89
LLaMAX3-8B Alpaca	-	-	19.49	5.81	-	-	4.29	20.06	24.10	5.20
LLaMA-2-7B 5-SHOT BM25	-	-	2.62	19.59	-	-	2.98	21.64	6.63	17.19
LLaMA-3-8B 5-SHOT BM25	-	-	14.40	8.96	-	-	5.13	20.08	24.48	6.00
LLaMA-2-7B uni.	-	-	14.24	8.68	-	-	5.32	18.62	13.46	9.07
LLaMA-2-7B uni. *beam size=5*	-	-	14.95	8.27	-	-	6.16	17.27	14.80	7.67
LLaMA-3-8B uni.	-	-	20.37	5.50	-	-	9.10	13.05	22.79	4.68
LLaMA-3-8B uni. *beam size=5*	-	-	**21.31**	**5.17**	-	-	**10.09**	**11.64**	**24.10**	**4.40**
LLaMA-3-8B multi.	-	-	19.33	6.17	-	-	7.69	15.61	22.17	4.90
LLaMA-3-8B multi. *beam size=5*	-	-	20.04	5.88	-	-	8.66	14.04	23.35	4.33

Methods	Somali		Sundanese		Swahili		Urdu		Xhosa	
	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX
NLLB-200-3.3B	12.43	12.73	-	-	36.95	3.95	32.87	3.59	-	-
Gemma-3-27B-It	9.12	13.33	-	-	34.35	3.67	28.43	3.18	-	-
Gemma-2-27B-It	6.61	16.67	-	-	34.48	3.87	25.84	4.08	-	-
LLaMA-3.1-70B It	6.74	16.98	-	-	33.59	4.14	27.66	3.65	-	-
Gemma-2-9B-It	5.21	18.66	-	-	30.23	4.79	22.33	4.96	-	-
LLaMA-3.1-8B-It	3.58	21.88	-	-	20.39	9.34	20.41	5.75	-	-
Command-R7B	1.65	22.02	-	-	6.55	19.48	5.39	12.83	-	-
Aya-expanse-32B	4.76	18.98	-	-	12.02	16.93	12.00	7.59	-	-
Aya-expanse-8B	4.04	20.73	-	-	7.83	21.26	8.98	8.87	-	-
Qwen-2.5-32B-It	3.91	21.30	-	-	13.00	16.95	12.66	9.64	-	-
Qwen-2.5-7B-It	3.27	22.34	-	-	7.82	21.32	7.02	13.49	-	-
LLaMAX2-7B Alpaca	7.13	16.01	-	-	22.46	7.27	10.95	16.62	-	-
LLaMAX3-8B Alpaca	8.05	14.42	-	-	28.03	6.00	22.61	5.15	-	-
LLaMA-2-7B	1.35	22.57	-	-	5.29	20.11	5.02	17.49	-	-
LLaMA-3-8B	1.89	21.75	-	-	25.87	7.04	21.58	6.15	-	-
LLaMA-2-7B uni.	6.41	17.53	-	-	20.33	11.35	14.35	8.90	-	-
LLaMA-2-7B uni. *beam size=5*	6.85	16.72	-	-	22.51	9.36	16.35	7.13	-	-
LLaMA-3-8B uni.	9.16	14.42	-	-	31.28	5.42	23.50	4.31	-	-
LLaMA-3-8B uni. *beam size=5*	**9.59**	**13.90**	-	-	**32.61**	**4.76**	**24.39**	**3.94**	-	-
LLaMA-3-8B multi.	8.26	15.25	-	-	29.25	6.01	22.82	4.64	-	-
LLaMA-3-8B multi. *beam size=5*	8.80	14.47	-	-	30.84	5.33	24.07	4.20	-	-

Table 12: BLEU and MetricX scores for 6 English → X directions from TICO-19 (Anastasopoulos et al., 2020). Best results after fine-tuning are highlighted in bold.

tified as a CIA officer, Frank Shields as a tennis player, and Deng Xi as a Chinese philosopher-the relevant words reflect these identities accurately.

B.8 Qualitative Analysis

In Table 17, we observe that Gemma-3-27B-It's generations display lexical overlap—both at the character and word level—with Google Translate's translations of the corresponding English sentences, as identified by the back-translator. Furthermore, Google Translate consistently produces translations that are similar to those obtained from back-translating Gemma-3-27B-It's generations. This suggests that the model is not hallucinating content, but instead generating semantically faithful and plausible sentences in the target language, reinforcing the relevance of our data generation pipeline.

Methods	Basque		Hausa		Igbo		Kinyarwanda		Nepali	
	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX
Gemma-3-27B-PT 5-shot BM25	31.04	3.73	25.98	4.24	19.39	6.10	21.02	5.77	34.06	3.71
Gemma-3-4B-PT 5-shot BM25	19.78	8.29	15.41	8.33	10.15	13.73	6.98	17.32	25.48	5.18
Gemma-3-4B-It	9.39	15.31	6.95	13.04	5.81	18.72	4.15	21.22	16.99	6.56
Gemma-3-4B-PT uni.	22.57	6.44	19.22	5.29	15.88	7.64	11.35	10.70	21.25	4.02
Gemma-3-4B-PT multi.	21.94	6.93	18.12	6.05	14.96	8.45	9.57	13.42	22.31	4.13
Methods	**Somali**		**Sundanese**		**Swahili**		**Urdu**		**Xhosa**	
	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX
Gemma-3-27B-PT 5-shot BM25	16.99	4.96	24.54	3.88	42.17	3.91	29.72	3.33	19.10	5.41
Gemma-3-4B-PT 5-shot BM25	9.77	10.42	16.91	7.41	30.83	6.65	21.32	4.97	8.63	15.04
Gemma-3-4B-It	4.93	15.94	9.32	8.68	18.31	10.75	16.16	6.33	4.08	20.60
Gemma-3-4B-PT uni.	13.04	6.41	17.29	4.62	32.00	5.10	22.39	3.88	12.60	9.24
Gemma-3-4B-PT multi.	12.22	7.31	16.49	5.71	30.39	5.51	21.92	4.09	11.64	10.50

Table 13: BLEU and MetricX scores for 10 English → X directions from FLORES. Best results after fine-tuning are highlighted in bold.

	Basque	Hausa	Igbo	Kinyarwanda	Nepali	Somali	Sundanese	Swahili	Urdu	Xhosa
Source mean number of words	22.18	23.47	21.10	22.80	18.21	22.80	22.15	23.02	19.04	23.03
Target mean number of words	17.49	26.72	23.05	19.12	16.16	25.50	19.69	22.96	24.30	15.41
Source mean Gemma-3-27B-It tokens	27.91	28.84	26.46	28.33	22.84	28.20	27.87	28.38	23.54	28.62
Target mean Gemma-3-27B-It tokens	46.11	50.44	54.32	57.20	34.32	57.44	43.06	48.89	35.08	56.02
MetricX24-XXL QE scores TopXGen	3.19	**3.58**	5.68	5.38	**2.86**	**4.34**	3.88	2.91	**2.62**	5.89
MetricX24-XXL QE scores FLORES	**2.65**	3.63	5.02	**3.43**	4.13	4.89	**3.66**	3.58	3.69	**3.82**
MetricX24-XXL QE scores SMOL	-	4.00	**4.62**	3.97	-	6.50	-	3.87	-	3.83
	Vendi scores (Dan Friedman and Dieng, 2023) based on SONAR embeddings (Duquenne et al., 2023)									
TopXGen Source	1.075	1.070	1.069	1.068	1.089	1.072	1.072	1.078	1.086	1.066
TopXGen Target	1.103	1.098	1.103	1.113	1.115	1.123	1.107	1.106	1.106	1.107
SmolSent Source	-	1.076	1.076	1.076	-	1.076	-	1.076	-	1.076
SmolSent Target	-	1.125	1.116	1.106	-	1.134	-	1.114	-	1.113
FLORES Source	1.051	1.051	1.051	1.051	1.051	1.051	1.051	1.051	1.051	1.051
FLORES Target	1.066	1.074	1.074	1.081	1.077	1.096	1.086	1.081	1.067	1.077

Table 14: Statistics of the TopXGen dataset in comparison to FLORES and SMOL.

	Basque	Hausa	Igbo	Kinyarwanda	Nepali	Somali	Sundanese	Swahili	Urdu	Xhosa
Gemma-3-27B-It										
In English										
Yes	972	973	982	965	980	973	965	977	977	969
No	28	27	18	35	20	27	35	23	23	31
In LRL										
Yes	946	924	903	878	938	926	930	943	913	891
No	54	76	97	122	62	74	70	57	87	109
Llama-4-Scout-17B-16E-Instruct										
In English										
Yes	904	935	903	919	956	934	918	902	959	919
No	59	53	77	77	23	63	82	67	38	70
In LRL										
Yes	896	854	837	832	903	881	881	894	871	805
No	100	143	159	163	96	112	114	103	127	192

Table 15: Repartition of 1000 paragraphs based on whether they discuss the topic they correspond to. We query Gemma-3-27B-It and Llama-4-Scout-17B-16E-Instruct (at Meta, 2025) in two settings: before and after translation by the back-translator (NLLB-200-3.3B).

Ground Truth Topics	gpt-4o-mini-2024-07-18 Labels	Most Representative Words
Theodore Shackley	CIA interventions in Peru	cia | the | united | of | states | in | trainingl , | war | was
Prasophyllum atratum	Biodiversity and conservation	species | the | australia | is | , | flowers | of | . | plant | it
Frank Shields	Development of Alternative Sports	tennis | sport | , | . | and | the | to | in | is | players
Guardians of the Galaxy Vol. 3 (soundtrack)	Influence of Music in Film	film | music | the | nosov | , | soundtrack | of | 's | a | .
Llanfawr Quarries	Local heritage and infrastructure	the | monuments | quarries | , | of | castle | and | . | these | to
Deng Xi	Influential Chinese Scholars	the | , | of | . | and | dynasty | chinese | qing | to | a

Table 16: Alignement between the ground truth topics and topics derived by BERTopic on six samples.

Basque	
Hauetako bakoitzak bere ezaugarriak ditu, batzuk produktu freskoen saldaritzan espezializatuz (okindegiak, frutariak) eta beste batzuk otoitz-zerbitzu ezarritik haratago doazen produktu eta zerbitzuak eskainiz (oinarrizko janari-elementuen salmenta nagusitzen duten tabako-dendetan, adibidez). Hauetako bakoitzak bere ezaugarriak ditu, batzuk produktu freskoen salmentan espezializatuta daude (baserriak, fruta-dendak) eta beste batzuk otoitz-zerbitzu finkatutik haratago doazen produktuak eta zerbitzuak eskaintzen dituzte (adibidez, oinarrizko elikagaien salmentan nagusi diren tabako-dendetan).	Each of these has its own characteristics, some specializing in the sale of fresh products (farms, fruit shops) and others offering products and services that go beyond the established prayer service (for example in tobacco shops that dominate the sale of basic food items). Each of these has its own characteristics, with some specializing in the sale of fresh produce (bakeries, greengrocers) and others offering products and services that go beyond the established prayer service (for example, tobacconists who mainly sell basic food items).
Hausa	
Wadannan matakan sun hada da sabbin kayayyakin safarar dukiya, da kuma inganta tsaron filin yayin da ake tafiya da komo. Wadannan matakan sun hada da sabbin hanyoyin jigilar kayayyaki, da kuma inganta tsaro a fagen sufuri.	These measures included new means of transporting goods, as well as improved field security during transportation. These measures include new transportation equipment, and improved field security during travel and return.
Kinyarwanda	
Impamvu Korea Times yifashishwa cyane, ni uko idashyira agahato ku makuru, kandi ngo ikunda kugaragaza ibintu bitandukanye na byinshi bisanzwe bimenyeshwa n'izindi nzego za Leta. Impamvu yo kwamamara muri Korea Times nuko idakurikirana amakuru, kandi ikunda kwerekana amakuru atandukanye nizindi nzego za leta.	The reason for the popularity of the Korea Times is that it does not censor information, and tends to present information that differs from most other government agencies. The reason the Korea Times is so widely used is that it does not impose restrictions on news, and it tends to present things that are different from what is usually reported by other government agencies.
Somali	
Qoyskiisu waxay ahaayeen kuwo qani ah oo leh xiriirro badan, taasoo ka caawisay inuu helaa fursado badan oo uu kaga shaqeeyo adeegga milatari. Qoyskiisu waxay ahaayeen kuwo hodan ah oo lahaa xidhiidho badan, taas oo ka caawisay inuu helo fursado badan oo uu kaga shaqeeyo adeegga milatariga.	His family was wealthy and had many connections, which helped him get many opportunities to work in the military service. His family was wealthy and well-connected, which helped him find many opportunities to serve in the military.
Sundanese	
Inskripsi ieu, ditulis dina basa Latin jeung basa Yunani, mangrupa conto anu saé kana kabijakan administrasi Romawi anu ngamimitahan panggunaan basa lokal pikeun mastikeun komunikasi anu efektif jeung populasi setempat. Prasasti, ditulis dina basa Latin sarta Yunani, mangrupakeun conto alus ngeunaan kawijakan administrasi Romawi nu wanti pamakéan basa lokal pikeun mastikeun komunikasi éféktif jeung populasi lokal.	The inscription, written in Latin and Greek, is a good example of a Roman administrative policy that encouraged the use of local languages to ensure effective communication with the local population. This inscription, written in Latin and Greek, is a good example of the Roman administrative policy of encouraging the use of local languages to ensure effective communication with the local population.
Swahili	
Pamoja na uzuri wake wa pekee, mlima huu pia umekuwa chanzo cha hadithi na misemo ya kitaifa kwa watu wa Wales kwa muda mrefu. Pamoja na uzuri wake wa kipekee, mlima huu pia kwa muda mrefu umekuwa chanzo cha hadithi na misemo ya kitaifa kwa watu wa Wales	Along with its unique beauty, this mountain has also long been the source of stories and national sayings for the people of Wales. Along with its unique beauty, this mountain has also long been the source of national legends and sayings for the Welsh people.
Xhosa	
Kukho imilinganiselo eyahlukeneyo esetyenziswa ukunje, ngokuphumela kwindlela yokuxabisa iimpahla, kodwa kwakungekho zibakala ezivela kunyaka. Imilinganiselo eyahlukeneyo iyasetyenziswa, ekhokelela kwinkqubo yokuxabisa, kodwa akuzange kubekho zibakala zonyaka.	Various measurements are used, resulting in a valuation system, but there were no facts from the year. There are different measures used today, resulting in a way of valuing goods, but there were no facts from the year.

Table 17: Examples of generations with their back-translator's translations. In red we provide Google Translate's translations (in the source language) of `NLLB-200-3.3B`'s translations in English. In blue we provide Google Translate's translations (in English) of the generator's (`Gemma-3-27B-It`) generations.