# Scaling Low-Resource MT via Synthetic Data Generation with LLMs

**Ona de Gibert** **Joseph Attieh** **Teemu Vahtola** **Mikko Aulamo** **Zihao Li**
**Raúl Vázquez** **Tiancheng Hu** **Jörg Tiedemann**
University of Helsinki University of Cambridge
{first.last}@helsinki.fi, th656@cam.ac.uk

## Abstract

We investigate the potential of LLM-generated synthetic data for improving low-resource Machine Translation (MT). Focusing on seven diverse target languages, we construct a document-level synthetic corpus from English Europarl, and extend it via pivoting to 147 additional language pairs. Automatic and human evaluation confirm its overall high quality. We study its practical application by (i) identifying effective training regimes, (ii) comparing our data with the HPLT dataset, (iii) studying the effect of varying training data size, and (iiii) testing its utility beyond English-centric MT. Finally, we introduce SynOPUS, a public repository for synthetic parallel datasets. Our findings show that LLM-generated synthetic data, even when noisy, can substantially improve MT performance for low-resource languages.

## 1 Introduction

Machine translation (MT) has achieved remarkable success for high-resource languages, but its application to the vast majority of the world's languages remains severely hampered by the scarcity of high-quality parallel corpora. Traditional data augmentation techniques like back-translation (Sennrich et al., 2016) and pivoting (Costa-jussá et al., 2018; Cheng, 2019) preserve the human-written target and synthesize the other. The advent of Large Language Models (LLMs) presents a transformative opportunity, as reflected by the growing number of survey papers on the subject (Zhou et al., 2024; Ding et al., 2024; Wang et al., 2024; Nadas et al., 2025). LLM-based synthetic data generation, akin to sequence-level knowledge distillation (Kim and Rush, 2016), opens up the possibility of creating vast amounts of training data even where human-translated resources are virtually non-existent.

This raises the question: *Can an MT system trained on LLM-generated data benefit truly low-resource language pairs?* To date, there is little to no systematic investigation of (a) generating large-scale synthetic data using LLMs for low-resource languages, (b) evaluating its intrinsic quality, (c) quantifying its downstream impact when training or fine-tuning modern MT systems. This paper provides such a systematic investigation. We make the following contributions:

- We use GPT-4o[1] to generate a document-level synthetic parallel corpus by forward-translating English Europarl (Koehn, 2005) into seven diverse low-resource languages.
- We assess the corpus quality using both automatic metrics and human evaluation, finding the data to be generally of high quality.
- We comprehensively evaluate the utility of this synthetic data by demonstrating that:
  (1) Compact MT models trained from scratch solely on this data achieve strong baseline performance (e.g., 49.49 ChrF for English-Georgian, compared to NLLB's 48.31).
  (2) Fine-tuning pretrained state-of-the-art (SOTA) systems (OPUS-MT, NLLB-200-1.3B, Llama-3B) consistently yields substantial improvements (e.g., average gains of +2.95 ChrF for NLLB and +20.63 ChrF for Llama-3B).
  (3) Our synthetic data is complementary to existing corpora like HPLT, e.g., leading to further ChrF increases of up to +2.79 when combined (for English-Icelandic).
  (4) Fine-tuned models that are 10-20 times smaller than SOTA models perform similar or better than their large counterparts.
- We study the effect of training data size to investigate the scalability of our approach.
- We extend our dataset into a multi-way parallel corpus via pivoting. As a case study, we demonstrate that Finnish-Somali translation im-

[1] We use the `gpt-4o-2024-08-06` model.



*Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27663–27681
November 4-9, 2025 

proves by +14.78 ChrF and +21.64 ChrF when fine-tuning OPUS-MT.

- To promote reproducibility and future research, we introduce SynOPUS, a public repository of synthetic parallel corpora. We also publicly release our dataset, which we publish under a new version of Europarl (v8syn),[2] code,[3] and baseline models.[4]

Our results show that LLM-generated synthetic data, even when noisy, can train competitive MT models from scratch and consistently improves pretrained systems, especially for the least resourced languages in the resource spectrum. Our work demonstrates a clear path towards open high-quality MT for underrepresented languages, by harnessing widely available high-resource monolingual corpora and powerful LLMs.

## 2 Related Work

**Low-resource MT.** Low-resource MT targets language pairs with little to no parallel data available (Haddow et al., 2022). To mitigate the data scarcity problems, two main lines of research have emerged: (a) transfer learning (Zoph et al., 2016) and multilingual training (Johnson et al., 2017), and (b) data augmentation (Xia et al., 2019). First, transfer learning involves using a model trained on a high-resource language as a starting point for training the low-resource language, while multilingual training proposes to train jointly on multiple language pairs to compensate for the lack of text in a specific language. Second, data augmentation proposes to generate synthetic samples to train on, by perturbing, translating or otherwise modifying existing sentences (Fadaee et al., 2017). Below, we focus on data augmentation and recent work that uses LLMs to generate such data.

**Classical data augmentation.** The most popular approach for low-resource languages is back-translation, which involves translating the monolingual target-language data into the source language (Ko et al., 2021; Khenglawt et al., 2024). The reverse process, forward translation of source-side monolingual sentences, has also been explored, and while less common in MT, proved valuable for LLM pretraining. For example, Wang et al. (2025) used NLLB to forward-translate monolingual corpora in nine languages and demonstrated its value for LLM pretraining. This process also relates closely to sequence-level knowledge distillation (Kim and Rush, 2016), where compressing a large model involves training a small *student* model on synthetic data constructed by forward-translating it with the *teacher* model (Gordon and Duh, 2019).

**LLM-based data augmentation.** LLMs have opened new avenues for synthetic data generation, driven by their strong performance in low-resource language settings. Several studies assess the translation performance of LLMs: Claude on Yoruba-English (Enis and Hopkins, 2024), Claude on 13 low-resource languages of Mali (Dembele et al., 2025), and GPT-4 on 3 languages (Jiao et al., 2023). These efforts encouraged researchers to use LLMs for synthetic data generation. For instance, Oh et al. (2023) explore different prompting strategies to generate synthetic data for German-Korean translation with ChatGPT. Our work is most similar to Yang and Nicolai (2023), where they exploit data generation for MT between German and Galician with ChatGPT. However, the authors generate source synthetic sentences that are later translated, while we use original English sentences as source data, and experiment on more languages.

**Gap addressed in this work.** Despite the above advances, there is still no systematic study that produces a fully synthetic multi-way parallel corpus with SOTA LLMs for low-resource languages and evaluates that corpus both intrinsically and on downstream MT. We close this gap by extending Europarl, a multilingual resource with alignments across all the official EU languages, into seven low-resource languages and evaluating its quality and usefulness.

## 3 Dataset Construction

Our goal is to study real-world cases instead of selecting common language pairs in an artificially constructed low-resource scenario. We conduct a preliminary experiment to help us select the languages to prioritize (Section 3.1). We then forward-translate the English Europarl corpus (Section 3.2), and, in a final post-processing step, we filter out noise to ensure high-quality translations (Section 3.3). Finally, we expand our dataset via pivoting to all languages of Europarl (Section 3.4).

[2] https://opus.nlpl.eu/synthetic/Europarl.php The data is subject to the terms and conditions defined by the usage policies of OpenAI.

[3] https://github.com/Helsinki-NLP/low-res-lmt

[4] Helsinki-NLP/scaling-low-res-mt-via-synthetic-data-generation-with-llms

| | eu | gd | ka | is | mk | so | uk |
|---|---|---|---|---|---|---|---|
| n. after segmentation | 2 167 164 | 2 192 082 | 2 504 071 | 2 370 036 | 2 054 167 | 2 373 145 | 2 359 720 |
| n. after lang. id. | 2 160 061 | 2 182 553 | 2 481 357 | 2 362 411 | 2 044 219 | 2 364 985 | 2 351 562 |
| n. aligned sentences | 2 138 713 | 2 164 999 | 2 317 070 | 2 348 030 | 2 027 406 | 2 353 915 | 2 341 706 |

Table 1: Statistics of the different post-processing steps described in Section 3.3 for each language.

### 3.1 Language Selection

We start by selecting a small set of low-resource languages for which GPT-4o can produce usable translations. To do so, we begin with a list of 204 European minority languages[5] and retain only the 39 languages that are supported by the FLORES+ benchmark (Goyal et al., 2022). For each of the 39 languages, we prompt GPT-4o to produce translations of (i) 100 random samples from the FLORES+ development set and (ii) 20 five-sentence chunks to simulate paragraph-level translation. We specify the script of the target language in the prompt.[6]

To contextualize GPT-4o's performance against existing well-performing translation models, we translate the same datasets with EMMA-500 (Ji et al., 2024), using both zero-shot and 3-shot settings, by selecting 3 unused examples from FLORES+. Additionally, we compare the results to the best available OPUS-MT model (Tiedemann et al., 2024) per language, selected from the OPUS-MT Dashboard (Tiedemann and De Gibert, 2023). We compare the performance of the three systems using ChrF (Popović, 2015).[7] Appendix A (Table 6) presents the results of the pilot evaluation. We proceed to select seven languages: Basque (*eu*), Scottish Gaelic (*gd*), Icelandic (*is*), Georgian (*ka*), Macedonian (*mk*), Somali (*so*), and Ukrainian (*uk*). We select these languages based on the linguistic diversity, low-resource coverage, model performance, and our practical interest.

### 3.2 Synthetic Data Generation

We use the English Europarl[8] (Koehn, 2005) as the source for generating the synthetic dataset. Europarl, which is derived from the proceedings of the European Parliament, offers well-defined document boundaries and is multi-way parallel across 21 European languages. We leverage the metadata within the Europarl corpus to segment the data into paragraphs in a way that each generated translation can be matched back to its exact English source, preserving the multi-way parallel structure. Paragraphs are sent in bulk to the OpenAI's Batch API[9]. We instruct the model to generate translations for source-target language pairs using the following prompt:

> *This is an English to* `TARGET` *translation, please provide the* `TARGET` *translation to this sentence in* `SCRIPT` *script. Do not provide any explanation or text apart from the translation.*

For instance, in the English-Ukrainian direction, we set the target language (`TARGET`) to *Ukrainian*, and the script information (`SCRIPT`) to *Cyrl*. We use script identifiers from the FLORES+ language codes, such as *ukr_Cyrl*.

### 3.3 Data Post-Processing

After translating the data with GPT-4o, we align the translated sequences with the original English sentences to produce parallel datasets. To produce aligned sentence pairs, we must first segment the paragraphs into individual sentences. For every language except Georgian, we use a sentence splitter from the Moses package (Koehn et al., 2007), selecting the language-specific system whenever it exists. Otherwise, we rely on the settings for the closest available language. For Somali we use the fallback to English, which seems to perform reasonably well. For Georgian we apply WtP (Minixhofer et al., 2023) with the `sat-3l-sm` model for sentence segmentation.

Because of the inherent noise in the translation process, and because of their tendencies to produce hallucinations, the LLMs may make errors in translation. To filter such cases, we apply language identification using `heliport`[10], which is based on the HeLI-OTS language identification models (Jauhiainen et al., 2022), to every generated segment

[5] The list is derived from https://en.wikipedia.org/wiki/Regional_and_minority_languages_in_Europe.

[6] Our initial experiments suggested that GPT-4o occasionally produces translations using a script different from that used in the FLORES+ dataset. This issue was the most prominent in Serbian, which uses both Cyrillic and Latin scripts.

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

[8] To the best of our knowledge, the Europarl corpus is not subject to any copyright restrictions.

[9] https://platform.openai.com/docs/guides/batch

[10] https://pypi.org/project/heliport/

after sentence segmentation, and discard those that are not classified correctly. On average, this step removes only about 0.45 % of sentences in each language.

Lastly, we align the cleaned target sentences with their English counterparts using the Yasa alignment tool (Lamraoui and Langlais, 2013), while preserving document boundaries so each sentence can be matched to its document and sentence identifiers. The resulting corpus contains 2–2.3 million aligned sentences per language pair. The data statistics are presented in Table 1.

### 3.4 MultiEuroparl: a Multi-Way Parallel Document-Level Corpus

An important design decision in our experiments was the focus on an inherently multi-way parallel dataset. Since all languages are aligned through English, we can use it as a pivot to project synthetic translations onto existing alignments.

In order to achieve that, we preserve sentence, paragraph, and document IDs from the original dataset during translation. Then, translated paragraphs are sentence-aligned to their input paragraphs and the alignment of English sentences to other existing languages in the original Europarl corpus are retrieved from OPUS. A minor complication is that sentence alignment is not one-to-one in all cases. We expand alignments by including neighboring sentence pairs until we get a match. In the worst case, this would cover the entire paragraph but, luckily, the data is quite well-behaved and aligns rather nicely also across language pairs.

Using the procedures above, we are able to create 147 new language pairs added to the original Europarl corpus, while keeping document information. Europarl has multi-way parallel corpora originally available in 21 languages. We add 7 new languages to the existing 21, yielding additional $21 \times 7 = 147$ language pairs when pairing each new language with all the existing ones. All of the language pairs are now available as training data for non-English-centric MT, a valuable source that comes for "free" due to the multilinguality and metadata of the source data. We study the usefulness of this data in Section 5.5.

## 4 Dataset Quality Analysis

In this section, we delve into the quality of the generated low-resource data. We first conduct a quantitative analysis (Section 4.1) producing numerical scores for each sentence pair and then proceed to ask native speakers of the target languages to rate a subsample of the dataset (Section 4.2). Finally, we compute inter-annotator agreement scores and correlation metrics.

### 4.1 Quantitative Analysis

To evaluate the quality of the generated parallel dataset, we compute two neural metrics at the segment level: Bicleaner-AI[11] (Zaragoza-Bernabeu et al., 2022) and COMETKiwi (Rei et al., 2022). These two metrics are optimized for different tasks and therefore behave differently: Bicleaner-AI is a binary classifier trained to determine whether two sentences are valid translations of each other. In contrast, COMETKiwi is a reference-free Quality Estimation (QE) metric based on COMET (Rei et al., 2020), trained to predict human judgment scores (on a 0–100 scale, normalized to 0-1) for machine-translated sentences.

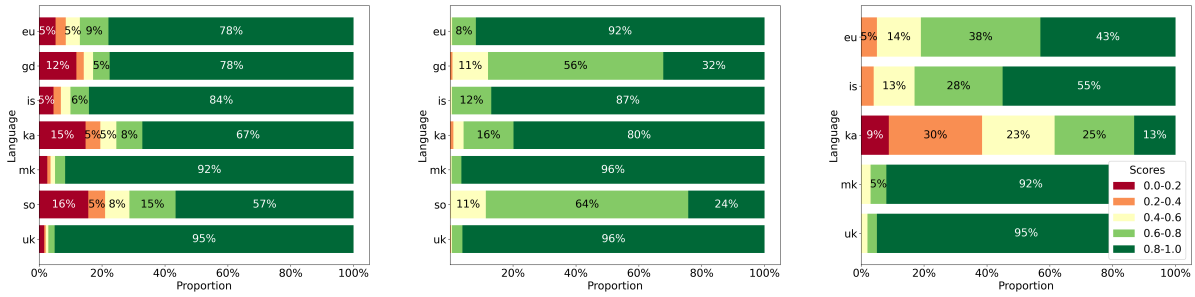
Figures 1a and 1b show the distribution of Bicleaner-AI scores and COMETKiwi per language pair. Looking at the Bicleaner-AI scores, we observe that over 92% of the sentences in Ukrainian and Macedonian fall in the highest bin and over 12% of the sentences for Somali, Georgian and Scottish Gaelic fall in the lowest bin. Although the general trend of COMETKiwi is similar to Bicleaner-AI, the results are interpreted differently as COMETKiwi reveals the actual quality of the sentences in the dataset generated. We can see that more than 85% of the sentences per language are in the top quality bins, noticing that the sentences with lower quality sentences are in Scottish Gaelic, Somali, and Georgian. However, COMETKiwi has not been explicitly trained on any of the languages in our dataset, even though its underlying model, XLM-R (Conneau et al., 2020), includes them. The fine-tuning for QE was conducted using data from the WMT General Shared Tasks (2017–2020). As such, these results are zero-shot and should be interpreted with caution.

### 4.2 Human Evaluation

To further assess the quality of our synthetic dataset, we conduct a human evaluation for five languages.[12] For each of language pair, we randomly sample 100 sentences and ask native speakers to

[11] We use the `bitextor/bicleaner-ai-full-large-en-xx` model.

[12] We were, unfortunately, unable to find annotators for Scottish Gaelic and Somali.




(a) Bicleaner-AI scores. (b) COMET-Kiwi scores. (c) Human evaluation scores.

Figure 1: Distribution of scores of the dataset quality analysis. Each bar represents the proportion of sentence pairs falling within different score intervals. We normalize human evaluation scores to use the same scale across plots.

evaluate them. Each language pair was scored by 1–3 annotators. Following the Direct Assessment (DA) protocol (Graham et al., 2013) from the WMT2017 Shared Task (Bojar et al., 2017). Annotators were shown the source sentence and its translation, and were asked to assign a score on a 0–100 scale, using the guidelines provided (see Appendix B). All ratings were collected through a custom web interface built with Gradio (Abid et al., 2019).

We measure Inter-Annotator Agreement (IAA) using Krippendorff's alpha (interval level) and compute the z-scored version by normalizing each annotator's scores to account for differences in scoring behavior. Table 2 shows the results of the IAA, which indicates moderate consistency among annotators.

Figure 1c presents the results of our human evaluation. Consistent with the findings in the previous section, Macedonian and Ukrainian stand out with high-quality outputs, where human annotators rated over 92% of the data within the 80–100 score range. Icelandic and Basque exhibited more variability, with approximately 40% of data rated as good quality (scores between 60 and 100) and around 30% considered acceptable (scores between 40 and 60). In contrast, Georgian data was considerably lower, with about 40% judged by annotators as being of unacceptable quality (scores below 40).

**Correlation scores** We report Spearman's rank correlation ($\rho_s$) between Bicleaner-AI and COMETKiwi scores, and human judgments in Table 2. In general, the correlations with BicleanerAI are weak across most language pairs, suggesting limited alignment. Georgian shows a relatively higher correlation ($\rho_s = 0.39$), likely due to the greater variance in human scores for this language. We observe that Bicleaner-AI tends to assign lower scores to samples that received very high ratings from human annotators, indicating a potential underestimation of high-quality translations. In contrast, correlations are consistently higher for COMETKiwi. This is expected, as COMETKiwi is designed to evaluate translation quality (a task more closely aligned with human judgments).

| Pair | Ann. Count | z-IAA | $\rho_s$(BicleanerAI) | $\rho_s$ (COMETKiwi) |
|---|---|---|---|---|
| en-eu | 3 | 0.49 | 0.25 | 0.43 |
| en-is | 2 | 0.53 | 0.27 | 0.43 |
| en-ka | 1 | - | 0.39 | 0.64 |
| en-mk | 1 | - | 0.21 | 0.21 |
| en-uk | 2 | 0.39 | 0.15 | 0.22 |

Table 2: Annotator count (Ann. Count), Inter-Annotator Agreement (IAA), as measured by z-score normalized Krippendorff's Alpha, and Spearman correlation ($\rho_s$) of the human judgements with the Bicleaner-AI and COMETKiwi scores per language pair.

Overall, we can conclude that the generated dataset is of fairly good quality, with both automatic and human metrics indicating that most sentence pairs are of good to excellent quality, particularly for Ukrainian and Macedonian. Lower performance is observed for Georgian, Somali, and Scottish Gaelic. This is coherent with the GPT-4o pilot evaluation that we conducted on FLORES+, as GPT-4o performs the best in terms of translation accuracy on Ukrainian and Macedonian, and worse for the rest of the languages (see Appendix A, Table 6).

## 5 Leveraging our Synthetic Data for MT

We evaluate the quality of our synthetic data by analyzing model performance both before and after fine-tuning across multiple architectures (Sections 5.1 and 5.2), serving as a proxy for data quality. Furthermore, we compare our dataset to a web-

| Model | Language Pair | | | | | | | # Params |
|---|---|---|---|---|---|---|---|---|
| | en-eu | en-gd | en-is | en-ka | en-mk | en-so | en-uk | |
| Synthetic | 53.00 | 51.10 | 49.91 | 49.49 | 57.72 | 45.10 | 51.71 | 60.6M |
| OPUS-MT | 54.99 | 41.60 | 51.97 | 42.69 | **64.45** | 44.20 | **60.14** | 191.6M |
| OPUS-MT-ft | 55.68 | **52.07** | **53.80** | 50.16 | 61.99 | **46.35** | 56.98 | |
| Δ | +0.69* | +10.47* | +1.83* | +7.47* | −2.46* | +2.15* | −3.16* | |
| NLLB | 52.05 | 49.94 | 47.98 | 48.31 | 60.13 | 45.90 | 54.44 | 1.3B |
| NLLB-ft | **56.32** | 51.81 | 52.93 | **52.75** | 62.32 | 46.14 | 57.13 | |
| Δ | +4.27* | +1.87* | +4.95* | +4.44* | +2.19* | +0.24 | +2.69* | |
| Llama | 29.25 | 26.56 | 22.66 | 13.17 | 26.58 | 22.76 | 30.24 | 3B |
| Llama-ft | 49.85 | 47.01 | 46.12 | 25.06 | 55.60 | 42.31 | 49.68 | |
| Δ | +20.59* | +20.45* | +23.46* | +11.89* | +29.02* | +19.55* | +19.44* | |
| GPT-4o | 57.10 | 53.24 | 55.94 | 51.84 | 64.45 | 46.82 | 60.88 | |

Table 3: ChrF scores (%) on seven translation tasks. For each architecture (Synthetic, OPUS-MT, NLLB, Llama), we report the raw ChrF of the base and fine-tuned (ft) models when available, along with the absolute improvement (Δ). The rightmost column shows model size. "*" indicates a significant difference ($p < 0.05$) between base and fine-tuned models, based on paired *t*-test and bootstrap resampling (5,000 iterations).

crawled SOTA corpus (Section 5.3), investigate the effect of variable training data size (Section 5.4) and study the usefulness of MultiEuroparl (Section 5.5).

### 5.1 Experimental Setup

**Data** We focus on the translation direction from English into the low-resource language, as this is typically the more challenging scenario. For all experiments we use the synthetic data as training set, and the FLORES+ (Goyal et al., 2022) development and test sets for model selection and evaluation, respectively.

**Models** Since the languages under consideration are not linguistically similar, we train individual bilingual models for each target language and leave multilingual studies for future work. We experiment using the following models (more details are provided in Appendix C):

**Synthetic**: a transformer-base model (Vaswani et al., 2017) trained on the synthetic data with MarianNMT (Junczys-Dowmunt et al., 2018).

**OPUS-MT**: the best OPUS-MT model per language pair, based on the OPUS-MT Dashboard scores (Tiedemann and De Gibert, 2023). The full list of the selected models is provided in Appendix D. Each model is fine-tuned without modifying its original tokenizer.

**NLLB-200-distilled-1.3B**: the distilled 1.3B parameter NLLB-200 model (Meta AI, 2022). For fine-tuning NLLB, we used DeepSpeed (Rasley et al., 2020).

**Llama-3.2-3B-Instruct**: the 3B parameter Llama-3.2 Instruct model (Dubey et al., 2024). For the fine-tuned version, we adapt LoRA (Hu et al., 2022) using Unsloth (Han et al., 2023).

All models are run on four 32 GB NVIDIA Volta V100 GPUs and take less than 9 hours to train.

**Evaluation** We evaluate all models before and after fine-tuning. We report ChrF (Popović, 2015) as our main automatic metric, as it has been the standard metric for low-resource MT and it is shown to correlate more closely with human judgments than BLEU (Papineni et al., 2002). We report COMET[13] (Rei et al., 2020) for all our experiments in Appendix E.

We also evaluate GPT-4o on the full FLORES+ test set (in the pilot evaluation in Appendix A, we used only 100 samples from the development set) and include it as a reference in our results.

### 5.2 Overall Results and Analysis

Table 3 summarizes the ChrF scores across three experimental conditions: off-the-shelf inference, fine-tuned training, and their performance differentials. We assessed the statistical significance of all the differences using paired Student's *t*-tests and paired bootstrap resampling (5000 iterations at 95% confidence).

[13]We use the `Unbabel/wmt22-comet-da` model.

**Effectiveness for Training from Scratch** The 60M parameter baseline, trained exclusively on our dataset, surpasses the out-of-the-box performance of billion-parameter models like NLLB and Llama for Basque, Scottish Gaelic, Icelandic and Georgian, while nearly matching them for Somali. This shows that our corpus is rich enough to train functional MT systems without any external pretraining or multilingual transfer.

**Impact on Fine-Tuning Pretrained Models** Fine-tuning consistently improves NLLB and Llama, confirming that the synthetic data is well-suited for adaptation. OPUS-MT also benefits from fine-tuning in five of seven cases; however, performance drops for Macedonian and Ukrainian, the two highest-resource low-resource pairs in our set. This suggests that when the model is trained on enough real parallel data, it ends up fitting too closely to the synthetic examples.

**Quality versus Usefulness** Based on the results from the previous section, we observe high quality for Ukranian and Macedonian, medium quality for Basque and Icelandic, and noticeably lower quality for the rest. Yet, noisy does not mean useless. In fact, Table 3 shows that the languages with the noisiest synthetic corpora also result in the largest downstream gains (Scottish Gaelic, Georgian and Somali). When the alternative is no data at all, quantity is better than quality. However, for mid-resource languages such as Macedonian and Ukrainian, cleaner text is already available beforehand and multilingual pretrained models benefit from large quantities of data from closely related languages. Therefore, additional synthetic data offers diminishing returns and mainly hurts the performance of these systems. The lower the resource level, the more tolerant MT training is to noise.

**Challenges with General Purpose LLMs** Llama initially struggles (13-30 ChrF), reflecting its lack of inherent translation capability for low-resource languages. While adapter training yields substantial improvements (+11-29 ChrF), the model still underperforms compared to smaller, translation-specific models. This indicates that while synthetic data allows for adaptation, it cannot fully compensate for mismatches between the pretraining objective and the translation task itself. In Llama's case, the 3B parameter scale appears unnecessarily large for this specific MT task, and leads to unnecessarily large fine-tuning times.

**Competitiveness with GPT-4o** GPT-4o[14] is the best performing system for almost all language directions, however fine-tuned models like OPUS-MT-ft and NLLB-ft still offer competitive results, despite being much smaller. OPUS-MT-ft is far behind by 1-2 ChrF points in most languages and NLLB-ft even outperforms GPT-4o for Georgian. This indicates that while GPT-4o is a challenging system to beat, smaller and more efficient models can achieve comparable results with far fewer parameters. These models close the performance gap using only our moderately-sized synthetic corpus (≈2M sentences), a mere fraction of the vast data required to train a frontier model like GPT-4o. This underscores a key finding: a targeted strategy of generating high-quality data provides a powerful and practical pathway to SOTA performance. Our approach significantly lowers the computational and financial barriers, making high-quality MT for low-resource languages much more accessible.

**A practical recipe for exploiting fully–synthetic low-resource data** These experiments point to a clear best practice:

1. **Generate in bulk for truly low resource languages.** Prioritize volume over perfection, as even noisy data drives significant gains when no alternatives exist.
2. **Fine-tune MT multilingual models.** NLLB-200 benefits consistently across all language directions. Our findings are consistent with previous research that finds that fine-tuning NLLB is among the best approaches for low-resource MT (Iyer et al., 2024; Zhu et al., 2024; Scalvini et al., 2025; Tapo et al., 2025; de Gibert et al., 2025).
3. **Avoid general-purpose LLMs for low-resource MT.** Despite Llama's large gains, its inefficient computational costs and inferior translation performance confirm that translation-specific encoder-decoder models leverage synthetic data more effectively.

### 5.3 Comparison with HPLT v2

To further assess our dataset's utility, we conduct comparative experiments against HPLT v2 (de Gibert et al., 2024; Burchell et al., 2025), a "real" parallel corpus derived from web sources (Internet

[14]Since GPT-4o's training data is not public, FLORES+ may be included, making evaluation unfair.

| Training Data | Language Pair | | |
|---|---|---|---|
| | en-eu | en-is | en-mk |
| Synthetic | 53.00 | 49.91 | 57.72 |
| HPLT | 54.63 | 50.60 | 62.09 |
| Δ | +1.63* | +0.69* | +4.37* |
| HPLT | 54.63 | 50.60 | 62.09 |
| HPLT + Synthetic | **56.20** | **53.39** | **62.92** |
| Δ | +1.57* | +2.79* | +0.83* |

Table 4: ChrF scores for the comparison of our data with HPLT.

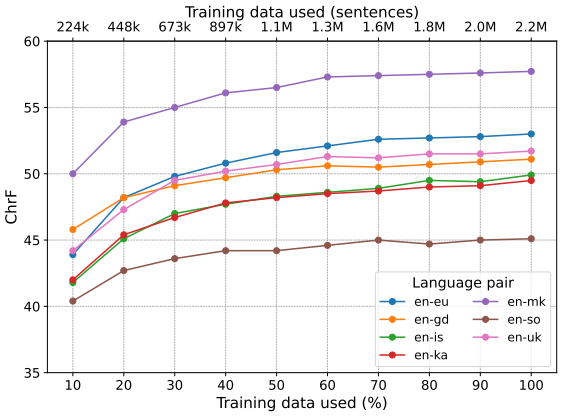



Figure 2: Learning curves with different data sizes.

Archive[15] and Common Crawl).[16]

We train systems for three out of the four overlapping language pairs (English paired with Basque, Icelandic and Macedonian). This decision was motivated by the significant variation in HPLT v2 data sizes (see Appendix F, Table 10), with Ukrainian having approximately ten times more data than the others; therefore, we leave it out. We train three models: (1) the same synthetic baseline as described earlier (Synthetic), (2) a model trained on HPLT dataset (HPLT), and (3) a model trained on the concatenation of our synthetic dataset and the HPLT (HPLT + Synthetic). All models follow the same architecture (transformer-base) and hyperparameters. All models are evaluated using ChrF on the same test set described previously. Table 4 reports the detailed ChrF scores.

**Comparable Performance** The models trained on our synthetic data alone perform on the same ballpark as the ones trained on the HPLT dataset, with an average difference of 2.23 ChrF points. The largest performance difference is observed for Macedonian, following a similar pattern as our experiments in the previous section. These results demonstrate that our synthetic dataset is of sufficiently high quality to challenge real-world parallel corpora, even when trained from scratch.

**Complementary when Combined** Adding our corpus to HPLT yields the best overall performance across all language pairs, with significant improvements. This proves the effectiveness of our synthetic data in low-resource MT. The consistent increases suggest that our data introduces useful diversity and complements the HPLT dataset, as it represents previously unseen material.

[15] https://archive.org/
[16] https://commoncrawl.org/

**Combined beats Transfer Learning** If we compare these results with Table 3, we can observe how training on the combined HPLT and synthetic datasets not only matches the performance of the fine-tuned NLLB for Basque, but surpasses it for both Icelandic and Macedonian, even though the Synthetic model is 21.6 times smaller. This highlights the power of data augmentation: enriching real-world corpora with high-quality synthetic data can outperform SOTA transfer learning approaches in low-resource settings.

### 5.4 Effect of Training Data Size

To assess the efficiency and scalability of our synthetic data approach, we train models with increasing fractions of the available synthetic data (10–100%) by creating cumulative subsets. For each subset, we trained a Marian NMT model using the same tokenizer and identical hyperparameters to the baselines presented in Section 5.1 to ensure comparability across runs, where data is the only changing variable.

**Effective Scaling with Synthetic Data** Figure 2 shows that performance improves consistently as more data is used, confirming that additional synthetic data is beneficial across all language directions. However, the gains decrease after 50–60% has been used. For example, English-Basque improves by +8.7 chrF from 10% to 50%, but only +1.4 additional points from 50% to 100%. This means that larger synthetic corpora brings gains but substantial improvements can already be achieved with a fraction of the full dataset. This suggests that future research on synthetic data generation should prioritize data quality and diversity for greater benefits than further scaling alone.

| Model | Language Pair | | | |
|---|---|---|---|---|
| | fi-so | so-fi | fi-uk | uk-fi |
| Synthetic | 38.72 | 35.87 | 42.01 | 46.01 |
| OPUS-MT | 26.31 | 15.86 | **50.56** | **55.24** |
| OPUS-MT-ft | 41.09 | 37.50 | 49.36 | 53.30 |
| Δ | +14.78* | +21.64* | -1.21* | -1.94* |
| NLLB | 40.33 | 39.66 | 45.59 | 47.78 |
| NLLB-ft | **42.21** | **42.31** | 47.61 | 51.62 |
| Δ | +1.88* | +2.65* | +2.02* | +3.84* |

Table 5: ChrF scores for Finnish-Somali and Finnish-Ukrainian translation.

### 5.5 Beyond English-Centric MT: The Finnish Use Case

To explore the multilingual potential of our expanded dataset via pivoting (Section 3.4) and move beyond English-centric translation, we train MT models for two additional language pairs: Finnish–Somali and Finnish-Ukrainian. The data consists of 1 876 672 sentences for Finnish-Ukranian and 1 882 712 for Finnish-Somali.

These languages were selected due to their prominence in Finland's linguistic landscape, where Ukrainian and Somali are among the most widely spoken foreign languages, accounting for approximately 0.7% and 0.5% of the population, respectively (Official Statistics of Finland (OSF), 2024). Developing high-quality models for these pairs is therefore both practical and relevant.

Our setup for this experiment is similar to the ones above. We first evaluate the out-of-the-box capabilities of OPUS-MT and NLLB. Next, we train a synthetic baseline (transformer-base), and finally, we fine-tune OPUS-MT and NLLB. We exclude Llama fine-tuning from this stage, as previous results have shown that it consistently underperforms. Table 5 reports the results.

**Usefulness of Pivoted Data** It is important to note that our synthetic baselines are weaker here than in previous experiments, providing greater headroom for fine-tuning improvements. OPUS-MT obtains clear gains from fine-tuning on synthetic data for the low-resource Finnish–Somali pair (+14.78 and +21.64 ChrF). However, for Finnish-Ukranian, fine-tuning does not improve. NLLB, which already exhibits a strong baseline, sees consistent gains across all directions. Overall, these results highlight the utility of synthetic data, particularly for low-resource language pairs.

## 6 SynOPUS: a New Synthetic Parallel Corpus Repository

The increasing adoption of LLMs in generating synthetic data underscores the growing need to systematically organize synthetic datasets. Although it is well known that many parallel datasets already contain MT content (Thompson et al., 2024), when synthetic data is intentionally produced, especially when involving significant financial or computational resources, proper archiving are paramount for promoting reuse, ensuring transparency, and maximizing resource utility. Therefore, with the release of our dataset, we introduce SynOPUS,[17] a new repository for parallel synthetic datasets, i.e., data that has been (partially) generated by translating text into other languages using MT systems or LLMs. We invite the community to contribute with their own datasets.

## 7 Conclusions

In this work, we thoroughly studied the quality and usefulness of LLM-generated synthetic data for low-resource MT. We presented a new synthetic corpus at document-level by forward translating Europarl, a parliamentary corpus, with GPT-4o. Then, we evaluated the resulting dataset both quantitatively and through human evaluation. Furthermore, we investigated the usefulness of this dataset for low-resource MT by: (i) identifying the most effective strategy for training, (ii) comparing our dataset with the public HPLT dataset, (iii) extending our analysis beyond English-centric MT by generating a multi-way parallel corpus via pivoting through alignments to English, and (iiii) studying the effect of varying training data size.

Our study highlights a crucial and often overlooked opportunity: the ability to create valuable parallel resources for low-resource MT by leveraging widely available high-resource monolingual data. This challenges the traditional reliance on scarce real target-language data for data augmentation approaches, and opens new directions for scalable MT development.

For future work, we aim to explore optimal methods for combining real and synthetic data, as well as extending our experiments to the document-level and investigating the use of synthetic data for monolingual LLM pretraining.

[17] https://opus.nlpl.eu/synthetic/

## Limitations

**Domain Bias** First and foremost, because of the origin of our source data, which is Europarl, a corpus compiled from parliamentary proceedings, the presented dataset belongs to a very specific domain. This implies that our models may suffer from domain bias and that any system trained on this data may not generalize well to informal, conversational, or domain-specific language; where linguistic style, vocabulary, and discourse structure differ significantly.

**Language Coverage** While we focus on seven diverse languages (varied language families, linguistic typologies, written scripts), our approach relies on GPT-4o's ability to produce a certain language reasonably well. Even though we rely on the results of our pilot study (shown in Table 6), our method may not translate well to other languages. Determining where the threshold lies, that is to say, how well a language must be supported for GPT-generated data to be viable; remains an open question.

**Human Evaluation Scope** We aim to provide enough pointers to evaluate the quality of our dataset both numerically and qualitatively. However, our human evaluation is limited to a 100 samples per language pair due to a lack of resources. Furthermore, we use Direct Assessment (DA), a widely accepted but increasingly outdated method. More recent evaluation approaches, such as Error Span Annotation (ESA) (Kocmi et al., 2024), offer more fine-grained insights into translation errors, but were beyond our reach for this study.

**Data Contamination of the Test Set** Due to the the closed-source nature of GPT-4o, there is a risk of data contamination, since the model may have already seen our test set (FLORES+) during pretraining. Recent studies (Mansurov et al., 2025) have shown that distilled data may inherit intrinsic biases from the teacher model and this may have an impact on benchmark results. We note, however, that there is a strong domain mismatch between our source data (formal Europarl proceedings) and the FLORES+ benchmark (general encyclopedic text), which reduces the likelihood of simple memorization affecting results. While this risk is inherent to most widely used test sets and cannot be fully controlled, we acknowledge it here for transparency.

**Data Contamination of the Source** GPT-4o pretraining data also likely includes the Europarl corpus. This means our experiments could be affected by data contamination, in the sense that the model may have had indirect prior exposure to the underlying content and domain, even if not in our low-resource target languages. Because our setup requires cross-lingual generation into languages that are not present in Europarl, direct memorization is unlikely. Still, there remains a risk of overestimating translation quality, which should be kept in mind when interpreting our results.

## Ethical Considerations

**Hallucinated Content** LLMs are known to generate hallucinated content, outputs that are fluent and well-formed but factually incorrect or irrelevant (Vazquez et al., 2025). This phenomenon is a risk in iself, as it can introduce noise and propagate misinformation in downstream MT models. In our case, we observed that in some cases, the model disregards the input and instead generates a response similar to, "You have been trained on data up until October 2023" in the target language. This issue is most prevalent in Georgian, with around 9,000 cases, and Ukrainian, with approximately 4,000 cases. For these languages, we removed each line containing the string "2023". While such hallucinations appear to be an intrinsic limitation of current LLMs, they highlight the need for careful post-processing and validation when using synthetic data.

**Reproducibility** Since our synthetic data is generated using a closed-source LLM, the exact reproduction of our work is not possible. To mitigate this, we publicly release the generated dataset along with all preprocessing scripts and training code.

**Cost-Benefit Trade-off** Our empirical results demonstrate that augmenting training data with high-quality LLM-generated translations improve translation performance for low-resource languages, outperforming existing baselines. This benefit is valuable in contexts where existing parallel training data is scarce or even unavailable. However, the cost of generating such translations with LLMs is significant both in terms of compute and financial expense, and may therefore be unreasonable for many groups. For example, generating the synthetic data used in our work costed

approximately $5 000. While the improvements in translation quality may justify the use of LLMs for data generation, future work should explore more cost-efficient methods for synthetic data generation. Promising directions could include e.g., distilling larger models into smaller ones and selective data augmentation to reduce the volume of unnecessary synthetic data while preserving improvements in performance.

ChatGPT was used to assist code development for this project.

## Acknowledgments

First and foremost, we would like to sincerely thank our annotators for their valuable contributions. Ander González Docasal, Harritxu Gete Ugarte, and Nerea Mandiola Solozabal for the Basque annotation; Helga Hilmisdóttir and Reynir Eggertsson for Icelandic; Artur Voit-Antal and Yana Matyash for Ukrainian; Biljana Stojanovska for Macedonian; and Elene Kavteladze for Georgian.

We acknowledge OpenAI for their generous provision of $5,000 in API credits, which were instrumental in generating the translations that comprise the datasets used in this work.

This project has received funding from the European Union's Horizon Europe programme (GA No 101070350) and from UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (GA No 10052546). This work was also supported by the Digital Europe Programme under grant agreement No 101195233. and by the GreenNLP project, which is funded by the Research Council of Finland. Tiancheng Hu is supported by Gates Cambridge Trust (grant OPP1144 from the Bill & Melinda Gates Foundation).

## References

Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. 2019. Gradio: Hassle-Free Sharing and Testing of ML Models in the Wild. *arXiv preprint arXiv:1906.02569*.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, et al. 2017. Findings of the 2017 Conference on Machine Translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214.

Laurie Burchell, Ona de Gibert, Nikolay Arefyev, Mikko Aulamo, Marta Bañón, Pinzhen Chen, Mariia Fedorova, Liane Guillou, Barry Haddow, Jan Hajič, Jindřich Helcl, Erik Henriksson, Mateusz Klimaszewski, Ville Komulainen, Andrey Kutuzov, Joona Kytöniemi, Veronika Laippala, Petter Mæhlum, Bhavitvya Malik, Farrokh Mehryary, Vladislav Mikhailov, Nikita Moghe, Amanda Myntti, Dayyán O'Brien, Stephan Oepen, Proyag Pal, Jousia Piha, Sampo Pyysalo, Gema Ramírez-Sánchez, David Samuel, Pavel Stepachev, Jörg Tiedemann, Dušan Variš, Tereza Vojtěchová, and Jaume Zaragoza-Bernabeu. 2025. An Expanded Massive Multilingual Dataset for High-Performance Language Technologies (HPLT). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17452–17485, Vienna, Austria. Association for Computational Linguistics.

Yong Cheng. 2019. Joint Training for Pivot-Based Neural Machine Translation. In *Joint training for neural machine translation*, pages 41–54. Springer.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard 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.

Marta R Costa-jussá, Noé Casas, and Maite Melero. 2018. English-Catalan Neural Machine Translation in the Biomedical Domain through the Cascade Approach. *arXiv preprint arXiv:1803.07139*.

Ona de Gibert, Graeme Nail, Nikolay Arefyev, Marta Bañón, Jelmer van der Linde, Shaoxiong Ji, Jaume Zaragoza-Bernabeu, Mikko Aulamo, Gema Ramírez-Sánchez, Andrey Kutuzov, Sampo Pyysalo, Stephan Oepen, and Jörg Tiedemann. 2024. A new massive multilingual dataset for high-performance language technologies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1116–1128, Torino, Italia. ELRA and ICCL.

Ona de Gibert, Robert Pugh, Ali Marashian, Raul Vazquez, Abteen Ebrahimi, Pavel Denisov, Enora Rice, Edward Gow-Smith, Juan Prieto, Melissa Robles, Rubén Manrique, Oscar Moreno, Angel Lino, Rolando Coto-Solano, Aldo Alvarez, Marvin Agüero-Torales, John E. Ortega, Luis Chiruzzo, Arturo Oncevay, Shruti Rijhwani, Katharina Von Der Wense, and Manuel Mager. 2025. Findings of the AmericasNLP 2025 Shared Tasks on Machine Translation, Creation of Educational Material, and Translation Metrics for Indigenous Languages of the Americas. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 134–152, Albuquerque, New Mexico. Association for Computational Linguistics.

Alou Dembele, Nouhoum Souleymane Coulibaly, and Michael Leventhal. 2025. The Serendipity of Claude AI: Case of the 13 Low-Resource National Languages of Mali. *arXiv preprint arXiv:2503.03380*.

Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh Tuan Luu, and Shafiq Joty. 2024. Data Augmentation using LLMs: Data perspectives, learning paradigms and challenges. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1679–1705, Bangkok, Thailand. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The Llama 3 Herd of Models.

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

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data Augmentation for Low-Resource Neural Machine Translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–573, Vancouver, Canada. Association for Computational Linguistics.

Mitchell A Gordon and Kevin Duh. 2019. Explaining Sequence-Level Knowledge Distillation as Data-Augmentation for Neural Machine Translation. *arXiv preprint arXiv:1912.03334*.

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.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous Measurement Scales in Human Evaluation of Machine Translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of Low-Resource Machine Translation. *Computational Linguistics*, 48(3):673–732.

Daniel Han, Michael Han, and Unsloth team. 2023. Unsloth.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. LoRA: Low-Rank Adaptation of Large Language Models. *ICLR*, 1(2):3.

Vivek Iyer, Bhavitvya Malik, Pavel Stepachev, Pinzhen Chen, Barry Haddow, and Alexandra Birch. 2024. Quality or quantity? on data scale and diversity in adapting large language models for low-resource translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1393–1409, Miami, Florida, USA. Association for Computational Linguistics.

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. 2022. HeLI-OTS, Off-the-Shelf Language Identifier for Text. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3912–3922, Marseille, France. European Language Resources Association.

Shaoxiong Ji, Zihao Li, Indraneil Paul, Jaakko Paavola, Peiqin Lin, Pinzhen Chen, Dayyán O'Brien, Hengyu Luo, Hinrich Schütze, Jörg Tiedemann, et al. 2024. Emma-500: Enhancing Massively Multilingual Adaptation of Large Language Models. *arXiv preprint arXiv:2409.17892*.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is ChatGPT a Good Translator? Yes with GPT-4 as the Engine. *arXiv preprint arXiv:2301.08745*.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Marcin Junczys-Dowmunt, Kenneth Heafield, Hieu Hoang, Roman Grundkiewicz, and Anthony Aue. 2018. Marian: Cost-effective High-Quality Neural Machine Translation in C++. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 129–135.

Vanlalmuansangi Khenglawt, Sahinur Rahman Laskar, Partha Pakray, and Ajoy Kumar Khan. 2024. Addressing Data Scarcity Issue for English–Mizo Neural Machine Translation Using Data Augmentation and Language Model. *Journal of Intelligent & Fuzzy Systems*, 46(3):6313–6323.

Yoon Kim and Alexander M Rush. 2016. Sequence-Level Knowledge Distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327.

Wei-Jen Ko, Ahmed El-Kishky, Adithya Renduchintala, Vishrav Chaudhary, Naman Goyal, Francisco Guzmán, Pascale Fung, Philipp Koehn, and Mona Diab. 2021. Adapting High-Resource NMT Models to Translate Low-Resource Related Languages Without Parallel Data. 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 802–812, Online. Association for Computational Linguistics.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja

Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024. Error Span Annotation: A Balanced Approach for Human Evaluation of Machine Translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of machine translation summit x: papers*, pages 79–86.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Fethi Lamraoui and Philippe Langlais. 2013. Yet Another Fast, Robust and Open Source Sentence Aligner. Time to Reconsider Sentence Alignment? In *Proceedings of Machine Translation Summit XIV: Papers*, Nice, France.

Jonibek Mansurov, Akhmed Sakip, and Alham Fikri Aji. 2025. Data Laundering: Artificially Boosting Benchmark Results through Knowledge Distillation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8332–8345, Vienna, Austria. Association for Computational Linguistics.

Meta AI. 2022. NLLB-200-distilled-1.3B: Hugging face model card. `https://huggingface.co/facebook/nllb-200-distilled-1.3B`.

Benjamin Minixhofer, Jonas Pfeiffer, and Ivan Vulić. 2023. Where's the Point? Self-Supervised Multilingual Punctuation-Agnostic Sentence Segmentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7215–7235, Toronto, Canada. Association for Computational Linguistics.

Mihai Nadas, Laura Diosan, and Andreea Tomescu. 2025. Synthetic Data Generation Using Large Language Models: Advances in Text and Code. *arXiv preprint arXiv:2503.14023*.

Official Statistics of Finland (OSF). 2024. Number of Foreign-Language Speakers Exceeded 600,000 during 2024. `https://stat.fi/en/publication/cm1jg8tr20lco07vwvoif9s6i`. Accessed: 2025-04-25.

Seokjin Oh, Su Ah Lee, and Woohwan Jung. 2023. Data Augmentation for Neural Machine Translation using Generative Language Model. *arXiv preprint arXiv:2307.16833*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. DeepSpeed: System Optimizations Enable Training Deep Learning Models with Over 100 Billion Parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 3505–3506, New York, NY, USA. Association for Computing Machinery.

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

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-Unbabel 2022 Submission for the Quality Estimation Shared Task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Barbara Scalvini, Iben Nyholm Debess, Annika Simonsen, and Hafsteinn Einarsson. 2025. Rethinking Low-Resource MT: The Surprising Effectiveness of Fine-Tuned Multilingual Models in the LLM Age. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 609–621, Tallinn, Estonia. University of Tartu Library.

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.

Allahsera Auguste Tapo, Kevin Assogba, Christopher M Homan, M. Mustafa Rafique, and Marcos Zampieri.

2025. Bayelemabaga: Creating resources for Bambara NLP. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12060–12070, Albuquerque, New Mexico. Association for Computational Linguistics.

Brian Thompson, Mehak Dhaliwal, Peter Frisch, Tobias Domhan, and Marcello Federico. 2024. A Shocking Amount of the Web is Machine Translated: Insights from Multi-Way Parallelism. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1763–1775, Bangkok, Thailand. Association for Computational Linguistics.

Jörg Tiedemann, Mikko Aulamo, Daria Bakshandaeva, Michele Boggia, Stig-Arne Grönroos, Tommi Nieminen, Alessandro Raganato, Yves Scherrer, Raúl Vázquez, and Sami Virpioja. 2024. Democratizing Neural Machine Translation with OPUS-MT. *Language Resources and Evaluation*, 58(2):713–755.

Jörg Tiedemann and Ona De Gibert. 2023. The OPUS-MT Dashboard–A Toolkit for a Systematic Evaluation of Open Machine Translation Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 315–327.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. *Advances in Neural Information Processing Systems*, 30.

Raul Vazquez, Timothee Mickus, Elaine Zosa, Teemu Vahtola, Jörg Tiedemann, Aman Sinha, Vincent Segonne, Fernando Sanchez Vega, Alessandro Raganato, Jindřich Libovický, Jussi Karlgren, Shaoxiong Ji, Jindřich Helcl, Liane Guillou, Ona De Gibert, Jaione Bengoetxea, Joseph Attieh, and Marianna Apidianaki. 2025. SemEval-2025 Task 3: Mu-SHROOM, the Multilingual Shared-task on Hallucinations and Related Observable Overgeneration Mistakes. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2472–2497, Vienna, Austria. Association for Computational Linguistics.

Jiayi Wang, Yao Lu, Maurice Weber, Max Ryabinin, David Adelani, Yihong Chen, Raphael Tang, and Pontus Stenetorp. 2025. Multilingual Language Model Pretraining using Machine-translated Data. *arXiv preprint arXiv:2502.13252*.

Ke Wang, Jiahui Zhu, Minjie Ren, Zeming Liu, Shiwei Li, Zongye Zhang, Chenkai Zhang, Xiaoyu Wu, Qiqi Zhan, Qingjie Liu, et al. 2024. A Survey on Data Synthesis and Augmentation for Large Language Models. *arXiv preprint arXiv:2410.12896*.

Mengzhou Xia, Xiang Kong, Antonios Anastasopoulos, and Graham Neubig. 2019. Generalized Data Augmentation for Low-Resource Translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5786–5796.

Wayne Yang and Garrett Nicolai. 2023. Neural Machine Translation Data Generation and Augmentation Using Chatgpt. *arXiv preprint arXiv:2307.05779*.

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

Yue Zhou, Chenlu Guo, Xu Wang, Yi Chang, and Yuan Wu. 2024. A Survey on Data Augmentation in Large Model Era. *arXiv preprint arXiv:2401.15422*.

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.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer Learning for Low-Resource Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575.

## A Pilot Evaluation Results

Table 6 reports the ChrF scores of GPT-4o and EMMA on low-resource MT for a 100 random sample of the FLORES+ development set. We compare performance at the sentence (Sent.) and chunk (Chunk) level. For EMMA, both zero-shot (Sent./0) and three-shot (Sent./3) settings are reported. The best OPUS-MT model is included as a reference. There is no OPUS-MT model for Aragonese and Aranese.

We do not evaluate our synthetic baseline models on these sample sentences, as the development set was used during training. Evaluating on this data risks overfitting and leads to optimistically biased performance estimates, which may not reflect true generalization.

| | | OPUS-MT | GPT-4o | | EMMA | |
|---|---|---|---|---|---|---|
| Language | Code | Sent. | Sent. | Chunk | Sent./3 | Sent./0 |
| Aragonese | arg_Latn | - | 43.9 | 47.5 | 38.1 | 39.8 |
| Aranese | arn_Latn | - | 41.2 | 45.4 | 16.7 | 27.6 |
| Armenian | hye_Armn | 47.5 | 56.3 | 56.5 | 45.7 | 48.7 |
| Asturian | ast_Latn | 59.6 | 59.6 | 62.8 | 52.7 | 53.3 |
| Bashkir | bak_Cyrl | 40.6 | 51.4 | 52.2 | 24.2 | 29.1 |
| Basque | eus_Latn | 55.0 | 57.3 | 60.4 | 41.5 | 44.0 |
| Belarusian | bel_Cyrl | 44.4 | 46.6 | 49.2 | 39.0 | 37.9 |
| Bosnian | bos_Latn | 58.8 | 63.9 | 65.4 | 51.6 | 53.3 |
| Catalan | cat_Latn | 67.4 | 67.9 | 69.8 | 56.7 | 56.5 |
| Crimean Tatar | crh_Latn | 35.9 | 37.6 | 40.0 | 13.7 | 23.2 |
| Croatian | hrv_Latn | 61.5 | 60.9 | 62.2 | 50.9 | 51.2 |
| Esperanto | epo_Latn | 59.7 | 63.2 | 65.6 | 60.8 | 60.5 |
| Friulian | fur_Latn | 49.9 | 45.7 | 50.0 | 45.4 | 43.8 |
| Galician | glg_Latn | 62.5 | 63.3 | 66.2 | 55.7 | 55.9 |
| Georgian | kat_Geor | 42.7 | 51.3 | 42.4 | 45.3 | 45.9 |
| Hebrew | heb_Hebr | 61.3 | 58.1 | 59.2 | 41.1 | 47.0 |
| Icelandic | isl_Latn | 53.0 | 55.6 | 59.3 | 41.1 | 41.6 |
| Irish | gle_Latn | 60.5 | 61.3 | 61.6 | 48.9 | 49.3 |
| Ligurian | lij_Latn | 43.9 | 36.2 | 40.1 | 34.5 | 35.3 |
| Limburgish | lim_Latn | 36.5 | 43.7 | 44.6 | 28.3 | 29.8 |
| Lombard | lmo_Latn | 34.1 | 35.3 | 38.8 | 25.3 | 28.4 |
| Luxembourgish | ltz_Latn | 55.5 | 59.4 | 60.9 | 51.5 | 49.9 |
| Macedonian | mkd_Cyrl | 64.5 | 65.3 | 64.2 | 55.2 | 56.8 |
| Northern Uzbek | uzn_Latn | 12.7 | 59.3 | 60.7 | 30.5 | 44.4 |
| Occitan | oci_Latn | 64.0 | 67.5 | 67.0 | 51.7 | 46.0 |
| Sardinian | srd_Latn | 50.1 | 43.4 | 46.2 | 40.0 | 48.8 |
| Scottish Gaelic | gla_Latn | 42.6 | 52.4 | 55.7 | 44.5 | 45.9 |
| Serbian | srp_Cyrl | 63.0 | 64.1 | 65.0 | 35.5 | 46.7 |
| Sicilian | scn_Latn | 39.9 | 45.0 | 48.7 | 43.4 | 43.8 |
| Somali | som_Latn | 44.2 | 47.6 | 53.0 | 42.7 | 38.6 |
| Tatar | tat_Cyrl | 43.0 | 53.6 | 54.9 | 21.1 | 33.9 |
| Tosk Albanian | als_Latn | 53.9 | 61.3 | 63.1 | 44.7 | 54.4 |
| Turkish | tur_Latn | 62.8 | 66.1 | 67.1 | 36.8 | 34.5 |
| Turkmen | tuk_Latn | 42.6 | 55.1 | 54.8 | 22.3 | 25.8 |
| Ukrainian | ukr_Cyrl | 60.1 | 60.1 | 62.1 | 48.7 | 49.4 |
| Uyghur | uig_Arab | 37.1 | 38.1 | 36.8 | 30.2 | 31.1 |
| Venetian | vec_Latn | 44.6 | 49.5 | 52.0 | 34.7 | 37.1 |
| Welsh | cym_Latn | 64.9 | 73.3 | 73.6 | 59.4 | 59.8 |
| Yiddish | ydd_Hebr | 0.0 | 40.5 | 42.6 | 41.4 | 51.5 |
| Average | | 49.2 | 53.9 | 55.6 | 40.8 | 43.6 |

Table 6: ChrF scores of the evaluation of GPT-4o and EMMA on low-resource MT. Highlighted rows correspond to the final set of selected languages for our study.

## B Human Evaluation Annotation Guidelines

In Figure 3, we provide an exact copy of the annotation guidelines given to the annotators.

**Introduction**

You will be asked to evaluate the quality of machine-translated (MT) sentences by comparing each one directly to its human-written original sentence (the source sentence). You will assign a score, based on how well the translation preserves meaning, fluency, and naturalness. This is what is known as Direct Assessment (DA, Graham et al., 2013). DA elicits human assessments of translation adequacy on an analogue rating scale (0–100), where human assessors are asked to rate how adequately the APE system output expresses the meaning of the human reference translation (Bojar et al., 2017). In this annotation project, you will be shown 100 samples of source-hypothesis pairs. Your task is to evaluate each translation pair through DA.

**Annotation Guidelines**

1. Carefully read the sentence pair. Try to understand the intended meaning of the source.
2. Evaluate whether the sentences are parallel or not. Compare the MT sentence with the source. Does the MT output preserve the key meaning of the source sentence?
3. Evaluate whether the target sentence contains fluency mistakes. Is the MT sentence grammatically correct? Are there any strange phrases, broken structure, or missing words?
4. Decide the score based on the scoring scale below.
5. Ensure that you double-check your annotations prior to moving to the next example. Re-read both source and translation. Does the score reflect meaning and fluency? Were you consistent with your previous scores? Adjust the score if needed to maintain fairness and consistency.

**Scoring scale**

Use the full range of the scale. Do not be afraid to give very low or very high scores when appropriate.

| Score | Interpretation |
|---|---|
| 100 | Perfect: grammatically flawless, fluent, and semantically identical to the source. |
| 85–99 | Excellent: small stylistic or fluency issues; all meaning preserved. |
| 70–84 | Good: mostly fluent; minor issues in grammar, wording, or slight meaning distortion. |
| 50–69 | Acceptable: understandable, but multiple issues with grammar, style, or partial meaning loss. |
| 30–49 | Poor: hard to understand, major meaning lost, broken grammar. |
| 1–29 | Very poor: barely comprehensible or mostly wrong meaning. |
| 0 | Incomprehensible: completely unrelated, meaningless, or unreadable. |

Figure 3: Annotation guidelines: Instructions.

## C Training Regimes

- **Synthetic**: We employ a shared 32k SentencePiece (Kudo and Richardson, 2018) vocabulary trained on the synthetic corpus; other settings follow the original Transformer-base recipe. Mini-batch fitting is enabled to optimize memory usage. Validation every 2500 updates checks perplexity. Early-stopping is employed on the development set, with a patience of 10.
- **OPUS-MT-ft**: We fine-tune each model without modifying its original tokenizer; appropriate language tags are prefixed at train and test time for multilingual models. Mini-batch fitting is enabled to optimize memory usage. Validation every 500 updates checks perplexity. Early-stopping is employed on the development set, with a patience of 20.
- **NLLB-200-distilled-1.3B-ft**: We fine-tune the NLLB-distilled-1.3B model with DeepSpeed on four V100 GPUs in FP16 mixed precision. Training uses a per-GPU batch size of 32 sentences, a maximum sequence length of 128 tokens, the Adam optimiser with a 1 $\times 10^{-4}$ learning rate, and runs for up to four epochs. DeepSpeed ZeRO-1 is used for basic tensor sharding; everything else is left on-GPU. Early-stopping is employed on the development set, with a patience of 5.
- **Llama-3.2-3B-Instruct-ft**: We adapted the model with LoRA using the Unsloth framework. We used the quantized 4-bit version of the model, applying LoRA adapters, and we used with prompts designed to mimic a professional translator's task using Unsloth's template system. Training was done using SFTTrainer with fp16 mixed precision, gradient accumulation, and 50k training steps with effective batch size of 16 utterances.

## D OPUS-MT Models selected for fine-tuning

We select the best available OPUS-MT model based on the OPUS-MT Dashboard (Tiedemann and De Gibert, 2023), by looking at the BLEU score on the FLORES+ dataset.

- en-eu: translate-en-eu-v1.0-hplt_opus
- en-gd: deu+eng+fra+por+spa-ine/tf-big
- en-is: translate-en-is-v1.0-hplt_opus
- en-mk: deu+eng+fra+por+spa-sla/tf-big
- en-so: deu+eng+fra+por+spa-afa/tf-big
- en-uk: eng-zle/tf-big
- en-ka: deu+eng+fra+por+spa-cau/tf-big
- fi-so: mul-mul/tf-big
- fi-uk: fin-zle/tf-big
- so-fi: afa-fiu/tf-base
- uk-fi: zle-fin/tf-big

## E COMET scores for MT training

We report the COMET scores for all our experiments to provide a more comprehensive evaluation. Table 7 shows the COMET scores equivalent to Table 3 for our fine-tuning experiments. Table 8 shows the COMET scores equivalent to Table 4 for our comparison with HPLT. Finally, Table 9 shows the COMET scores equivalent to Table 5 for our study on Finnish-centric translation.

| Model | Language Pair | | | | | | | # Params |
|---|---|---|---|---|---|---|---|---|
| | en-eu | en-gd | en-is | en-ka | en-mk | en-so | en-uk | |
| Synthetic | 81.51 | 78.04 | 80.16 | 80.72 | 82.24 | 78.15 | 78.89 | 60.6M |
| OPUS-MT | 83.27 | 71.30 | 79.69 | 69.09 | 87.34 | 77.06 | **89.02** | 191.6M |
| OPUS-MT-ft | 84.15 | 79.30 | 83.21 | 81.60 | 86.19 | 79.34 | 87.61 | |
| Δ | +0.88 | +8.00 | +3.52 | +12.51 | -1.15 | +2.28 | -1.41 | |
| NLLB | 84.55 | 78.73 | 82.06 | 80.49 | 87.45 | 80.06 | 87.21 | 1.3B |
| NLLB-ft | **86.84** | **79.43** | **85.36** | **86.94** | **88.74** | **80.90** | 88.14 | |
| Δ | +2.29 | +0.7 | +3.3 | +6.45 | +1.29 | +0.84 | +0.93 | |
| Llama | 40.94 | 44.61 | 36.96 | 33.68 | 42.50 | 43.79 | 50.86 | 3B |
| Llama-ft | 70.00 | 75.92 | 76.63 | 54.53 | 82.09 | 76.67 | 80.80 | |
| Δ | +29.06 | +31.31 | +39.67 | +20.85 | +39.59 | +32.88 | +29.94 | |
| GPT-4o | 86.65 | 80.11 | 86.70 | 85.76 | 90.04 | 80.67 | 91.13 | |

Table 7: COMET scores for our fine-tuning experiments.

| Training Data | Language Pair | | |
|---|---|---|---|
| | en-eu | en-is | en-mk |
| Synthetic | 81.51 | 80.16 | 82.24 |
| HPLT | 82.47 | 78.09 | 85.38 |
| Δ | +0.96 | -2.07 | + 3.14 |
| HPLT | 82.47 | 78.09 | 85.38 |
| HPLT + Synthetic | **84.53** | **82.82** | **86.96** |
| Δ | +2.06 | +4.73 | +1.58 |

Table 8: COMET scores for the comparison of our data with HPLT.

| Model | Language Pair | | | |
|---|---|---|---|---|
| | fi-so | so-fi | fi-uk | uk-fi |
| Synthetic | 75.09 | 66.55 | 76.89 | 77.36 |
| OPUS-MT | 54.06 | 34.33 | **89.07** | **88.01** |
| OPUS-MT-ft | 76.72 | 68.32 | 87.77 | 87.10 |
| Δ | +22.66 | +33.99 | -1.30 | -0.91 |
| NLLB | 77.35 | 76.37 | 85.17 | 84.40 |
| NLLB-ft | **78.60** | **78.62** | 87.09 | 87.04 |
| Δ | +1.25 | +2.25 | +1.92 | +2.64 |

Table 9: COMET scores for Finnish-Somali and Finnish-Ukrainian translation.

## F HPLT Data Sizes

We report the total amount of sentences of the HPLT v2 dataset in Table 10 for the overlapping language pairs with our selected languages.

| | en-eu | en-is | en-mk | en-uk |
|---|---|---|---|---|
| n. sentences | 1 491 873 | 2 694 541 | 3 991 617 | 25 125 019 |

Table 10: Data sizes in amount of sentences in the HPLT v2 dataset.