Omnilingual MT: Machine Translation for 1,600 Languages

Anonymous ACL submission

Abstract

We introduce **Omnilingual MT**, a family of multilingual translation systems covering more than 1,600 languages—the broadest benchmarked coverage of any MT system to date. Omnilingual MT explores two complementary architectures: OMT-LLAMA, a decoder-only model (1B/3B/8B) built on LLAMA3 with multilingual continual pretraining, and OMT-NLLB, a 3B encoder-decoder model with a novel three-stage training strategy exploiting non-parallel data via auto-encoding. Our data strategy combines large-scale multilingual corpora with backtranslation for 2,000+ languages and MeDLEy, a grammatically-diverse seed dataset for 108 low-resource languages. For evaluation, we extend the BOUQuET benchmark to 275 languages and develop Met-BOUQuET with human annotations of machine quality in 161 directions and BLASER 3, a reference-free quality estimation metric. Omnilingual MT doubles the number of languages with "passable" quality from ∼200 to over 440 and translates in zero-shot mode into ∼1,200 languages versus 300–400 for baselines. Our specialized 1B–8B models match or exceed 70B general-purpose LLMs, demonstrating that specialization is a more reliable path to multilingual translation than scale alone.

1 Introduction

The success of No Language Left Behind (NLLB; NLLB Team et al., 2024) demonstrated that high-quality machine translation (MT) could extend to 200 languages, setting a new standard for linguistic inclusion. Despite its name, the project left the vast majority of the world's ∼7,000 languages without any direct support or mechanism of including new languages into an existing model. Early explorations toward 1,000-language MT (Siddhant et al., 2022) demonstrated its feasibility but did not yield sustained or extensible systems. Meanwhile, large language models (LLMs) have improved MT for high-resource languages (Alves et al., 2024; Dang et al., 2024; Gemma Team et al., 2025), but their multilingual gains remain dominated by well-resourced languages, and they are inefficient—often requiring tens of billions of parameters to match much smaller specialized models.

Since NLLB's FLORES-200 benchmark (introduced in Goyal et al. (2022) and later extended), there has been no attempt to radically increase the language coverage of MT evaluation. But even FLORES-200 exposed a structural asymmetry: modern models can often *interpret* languages beyond their declared coverage through cross-lingual transfer, but they struggle to *generate* them reliably.

We argue that foundational models should aspire to provide at least a base level of support for *every* language, rather than concentrating capacity on a privileged subset. In this work, we introduce **Omnilingual MT**, a family of multilingual translation systems that extend support to more than 1,600 languages, the broadest benchmarked coverage to date. Our contributions span data, modeling, and evaluation (see Table 1). In order to produce them, we explored several research directions:

Monolingual data: how can it be leveraged to boost MT performance for low-resourced languages? We validate its utility with traditional methods (backtranslation, bitext mining, use for language modeling continual pretraining) as well as a more innovative mid-training of a decoder using an autoencoding objective.

Seed data construction: how does one optimize the diversity of a newly collected multiway parallel dataset, given tight size constraints? We construct MeDLEy, a manually curated, grammatically-diverse multidomain seed dataset for 108 low-resource languages.

Architectures: are encoder-decoder Transformers still a viable option for NMT? We show that it is the case with the OMT-NLLB model that employs a cross-lingually aligned sentence encoder

Table 1: **Summary of Omnilingual MT:** Each row lists the artifact, its type, and its primary claim.

Name	Type	Description	Claim
OMT-LLaMA	Model	LLaMA3-based decoder-only MT models (1B, 3B, 8B)	LLM for MT, understanding and generating over 1,000 languages
OMT-NLLB	Model	Encoder–decoder MT built on OMNISONAR	1,600-to-250 language MT system
BLASER 3	Model	Reference-free MT quality estimation	Highest multilingual coverage (1,600+ source languages)
OmniTOX	Model	Toxicity detection	Largest language coverage (1,600 languages), ROC AUC 0.86
MeDLEy	Dataset	MT training data	108 extremely low-resource languages, domain and grammatical diversity
BOUQuET	Dataset	Extension of an MT evaluation benchmark	275 languages, 56 language families, 8 domains
Met-BOUQuET	Dataset	Human judgments of MT quality	161 language directions; largest quality estimation dataset by language coverage

(Omnilingual SONAR Team et al., 2026) in an encoder-decoder architecture with a novel three-stage training strategy exploiting non-parallel data.

Model size: can small translation-specific LLMs compete with larger general-purpose language models? We develop OMT-LLAMA, a series LLAMA3-based decoder-only models (1B/3B/8B) with a multilingual continual pretraining recipe, post-training via SFT and RL. All three outperform their much larger non-specialized counterpart, LLAMA3 70B.

Adaptability: what are the ways of boosting LLM-based MT quality for barely seen languages? We show that both fine-tuning and retrieval augmented generation work and their effects add up but cannot compensate for a weaker base model.

MT metrics: how to scale the language coverage of scarce training data for quality estimation metrics? With BLASER 3, we combine human annotations with multiple protocols, synthetic data, and a strong base encoder to develop an omnilingual quality estimation model.

With these strategies combined, we were able to extend the number of languages that a multilingual model can "understand well enough for practical use" from about 200 to over 400, and the number of languages with better-than-random generation ability, from 300-400 to about 1200. Despite conventional wisdom, specialised small models (<8B) match or exceed larger models. We hope that our work opens the way to *omnilinguality* in machine translation: models and resources aiming to provide better support to every language.

2 Related Work

Scaling multilingual MT. NLLB (NLLB Team et al., 2024) demonstrated high-quality MT for 200 languages using encoder-decoder Transformers with large-scale data curation and established FLoRes+ as the de facto evaluation standard. Google's massively multilingual translation work (Siddhant et al., 2022) explored models covering up to 1,000 languages but did not produce sustained, extensible systems. Recent specialized systems such as MADLAD (Kudugunta et al., 2023) and TranslateGemma (Finkelstein et al., 2026) have improved coverage beyond the NLLB-200 language set. More recently, Omnilingual SONAR Team et al. (2026) built a sentence encoder with unlimited language support evaluated with translation from 1,560 languages, but stuck to the NLLB-200 language set on the target side.

LLMs for MT. Large decoder-only models exhibit surprising zero-shot and few-shot translation abilities (Zhu et al., 2024), and specialized variants such as TowerLLM (Alves et al., 2024) validate LLM-based MT recipes. However, LLMs remain inefficient for MT: even 70B-parameter models lag behind smaller specialized systems on low-resource pairs. Recent work on MT-specific RL (Feng et al., 2025) and reasoning-based translation (Wang et al., 2025) has shown promise for high-resource pairs.

Data for long-tail languages. Creating training data for thousands of languages requires diverse strategies. Backtranslation (Sennrich et al., 2016; NLLB Team et al., 2024) and bitext mining (Schwenk et al., 2021) have been extended to massively multilingual settings. Manually curated seed

Type	Source	# Texts	# Langs
Mono.	CC-2000-WEB	20M	∼2000
	CC-2000-PDF	5M	∼1700
Parallel	Bible	600M	∼1600
	Panlex	2B	∼1000
	CC-NLLB-200	500M	∼200
	Other sources	90M	∼500
Synthetic	Backtranslation	135M	∼2000
	Mined paragraphs	100K	∼60

Table 2: CPT data sources.

	En-YY				XX-En			
	H	M	L	VL	H	M	L	VL
No BT	44.7	24.0	17.0	20.0	57.0	42.3	35.7	38.2
+BT (LQ)	46.1	26.1	20.3	23.6	58.0	43.8	38.1	40.9
+BT (MQ)	46.3	**26.6**	**20.8**	**24.3**	58.2	**44.0**	**38.3**	**41.3**
+BT (HQ)	**46.8**	26.1	20.7	24.1	**58.5**	43.6	37.7	40.9

Table 3: Effect of backtranslation on ChrF++ (FLoRes+). MQ = medium quality filtering. H/M/L/VL = resource levels.

datasets—NLLB Seed (NLLB Team et al., 2024), SMOL (Caswell et al., 2025), and Gatitos (Jones et al., 2023)—provide high-quality examples that bootstrap synthetic data pipelines.

Evaluation at scale. MT evaluation beyond 200 languages remains limited by the availability of the data and reliablity of the metrics. The most widely used multilingual MT benchmark, FLORES-200 (Goyal et al., 2022; NLLB Team et al., 2024), has been extended to FLORES+ (Burchell et al., 2024; Dale et al., 2025) covering, as of today, 228 language varieties.[1] Few benchmarks go beyond that, with a notable exception of ChiKhaPo (Chang and Bafna, 2025) focused on word translation evaluation with thousands of languages.

Metrics are also a bottleneck. Surface-level metrics (ChrF++, BLEU) fail to capture meaning adequacy; neural metrics such as xCOMET (Guerreiro et al., 2024) and MetricX-24 (Juraska et al., 2024) are not guaranteed to work with target languages not included in their fine-tuning data. BLASER 2 (Dale and Costa-jussà, 2024) demonstrates cross-lingual generalization, but is limited to the 200 languages of the underlying sentence encoder.

3 Data

3.1 Continual Pretraining Data

Our data strategy prioritizes both breadth (number of languages) and depth (domain diversity). Table 2 summarizes the CPT data sources.

Monolingual data. We collect multilingual monolingual corpora based on web documents from Common Crawl. After quality filtering and language identification with GlotLID (Kargaran et al., 2023) (inspired by FineWeb2, Penedo et al. (2025)), we create CC-2000-WEB and CC-2000-PDF covering 2,000+ languages.

Parallel data. Bible texts (∼1,600 languages)[2] provide the broadest coverage. Panlex dictionaries (Kamholz et al. (2014), ∼1,000 languages) provide word-level alignment. CC-NLLB-200 reproduces the NLLB-200 composition of primary data sources. We further include Tatoeba (Tiedemann, 2012), SMOL (Caswell et al., 2025), Gatitos (Jones et al., 2023), and data contributed via a partnership program[3] as well as a collection of over 50 miscellaneous publicly available datasets.

3.2 Backtranslation

For languages with monolingual text but little or no parallel data, we generate synthetic bitext via backtranslation: (1) segment monolingual text into sentences using a fine-tuned SAT-12L-SM model (Frohmann et al., 2024); (2) filter via GlotLID with language-specific thresholds calibrated using Bible texts; (3) translate using NLLB-200 or LLAMA3; and (4) estimate quality using OMNISONAR sentence embedding similarity and model-free signals (length ratio, unique character ratio to exclude repetitive hallucinations, edit distance to exclude excessive source copying).

We define quality filtering presets based on embedding similarity. A sentence pair is classified as "medium-quality" (MQ) if its similarity exceeds a reference threshold calibrated to the mean similarity of Bible texts for that target language. High-quality (HQ) and low-quality (LQ) presets adjust this threshold by adding or subtracting one standard deviation, respectively. We find that medium-quality filtering provides the best balance across resource levels (Table 3). The final dataset contains ∼230M sentences across ∼2,000 languages.

[1] By "language variety" we denote the combination of language (e.g. Mandarin Chinese), writing system (e.g. Cyrillic script), and, in some cases, dialect (e.g. the Aranese dialect of the Occitan language included in FLORES+).

[2] With the majority of texts downloaded using the eBible tool: https://github.com/BibleNLP/ebible.

[3] More details will be disclosed after the anonymized review.

3.3 Bitext mining

We adopt global mining using the OMNISONAR multilingual embedding space (Omnilingual SONAR Team et al., 2026) to discover parallel paragraphs in monolingual corpora. Source data includes CC-2000-Web (non-English text) and FineWeb-Edu(Lozhkov et al., 2024) (English text). We encode all data into OMNISONAR embeddings and apply product quantization using FAISS (Johnson et al., 2019) to enable efficient billion-scale similarity search. Quantization is applied to batches of up to 50 million embeddings per language. We then perform KNN search with $k = 3$ nearest neighbours using the stopes library (Andrews et al., 2022), retrieving English candidates for each non-English sentence and vice versa. Candidate pairs are scored using margin-based cosine similarity in the OMNISONAR space, and pairs above language-specific thresholds are retained as mined bitexts.

3.4 MeDLEy: Seed Data for Low-Resource Languages

Existing multilingual datasets exhibit systematic gaps in domain diversity and coverage of marginalized languages. We create MeDLEy (**M**ulticentric, **D**omain-diverse, **L**inguistically-diverse, **E**asy-to-translate seed dataset), a manually curated parallel corpus for 108 extremely low-resource languages. MeDLEy starts from 605 paragraphs (∼2,200 sentences) in 5 source languages, designed to cover 61 cross-linguistic grammatical features across 5 domains. These are multiway-parallelized across 8 pivot languages and professionally translated into 108 low-resourced languages. By targeting grammatical *functions* (e.g., case, aspect, evidentiality) rather than language-specific forms, MeDLEy achieves broad grammatical coverage in any target language. MeDLEy covers 92 languages not in prior datasets such as SMOL (Caswell et al., 2025) or NLLB Seed. In controlled experiments, MeDLEy matches or outperforms existing seed datasets and shows particular gains for into-English translation; it also proved essential for the post-training recipe (§4.2).

3.5 Evaluation Data

BOUQuET. We use BOUQuET originally introduced by Omnilingual MT Team et al. (2025) as a manually created benchmark covering 55 languages across 8 domains with diverse registers. In this work, we extend it to 275 languages through additional professional translation, making it the largest multidomain MT evaluation benchmark. The extended benchmark and the associated leaderboard will be made publicly available. BOUQuET is organized into paragraphs, but in this paper, we report the results of sentence-level translation.

	En-YY					XX-En				
	H	**M**	**L**	**VL**	**All**	**H**	**M**	**L**	**VL**	**All**
128K	39.0	17.0	12.2	12.7	17.8	54.1	35.3	31.0	31.5	35.9
256K	**41.2**	**23.2**	**16.7**	**15.5**	**22.5**	**54.8**	**39.1**	**33.6**	**33.1**	**38.7**

Table 4: Tokenizer ablation (ChrF++ on FLoRes+). Extending from 128K to 256K tokens yields 26% relative improvement for En-YY.

Bible benchmark. We use the Gospel of John as an evaluation set, covering 1,560 languages. While Bible evaluation has limitations (narrow domain, formal register), it provides the broadest available reference-based evaluation coverage.

FLORES+. We report MT results on the FLORES+ dataset (version 2.1) with 213 language varieties in its `devtest` split.

Resource levels. When reporting MT evaluation results, we sometimes group them by language resource levels: "high" **H** (over 50M parallel sentences in the training data) "mid" **M** (over 1M), "low" **L** (over 40K), and "very low" **VL** (less then 40K parallel sentences).

Directions. With each benchmark dataset, we evaluate translation from all language varieties into English (XX-En) and from English into each language variety (En-YY).

4 Modeling

4.1 Vocabulary Extension

A critical prerequisite for omnilingual translation is adequate vocabulary coverage. The original LLAMA3 tokenizer produces high fertility for underserved scripts. We address this by: (1) adjusting pre-tokenization to better handle rare writing systems; and (2) extending from 128K to 256K tokens via continued BPE merging on balanced multilingual data. New token embeddings are initialized by averaging their sub-token decompositions (Gee et al., 2022).

Table 4 shows controlled ablation results on LLAMA3 1B Instruct (30K steps of CPT, identical

	BOUQuET		FLoRes+		Bible	
	XX-En	En-YY	XX-En	En-YY	XX-En	En-YY
CPT	.725	**.621**	.735	.523	.683	.609
+SFT	.732	.611	.735	.539	.685	.656
+RL	**.741**	.616	**.747**	**.541**	**.689**	**.661**

Table 5: Post-training results (xCOMET). SFT and RL yield cumulative improvements across benchmarks.

	FLoRes+			Bible		
	ChrF	xCOM	MtrX	ChrF	xCOM	MtrX
OMNISONAR	47.5	.65	5.66	40.6	.52	8.79
+Stage 1	49.7	.65	5.48	45.9	.53	8.59
(w/o autoencoding)	49.6	.65	5.51	44.8	.52	8.61
+Stage 2	50.1	.65	5.36	46.4	.51	8.58
OMT-NLLB	**50.9**	**.66**	**5.32**	**46.3**	.50	8.61

Table 6: OMT-NLLB 3-stage training results. Each stage improves over OMNISONAR.

data). The extended tokenizer achieves 26% relative ChrF++ improvement for En-YY (17.8→22.5) and 7% for XX-En.

4.2 OMT-LLaMA: Decoder-Only Models

Continual pretraining (CPT). OMT-LLAMA extends LLAMA3 3.1 8B (Grattafiori et al., 2024) with the 256K tokenizer via CPT on language modeling and translation tasks for 50K steps on 256 A100 GPUs, preceded by a 10K-step embedding warmup. We apply the same recipe at 1B and 3B scales.

Post-training (SFT + RL). We apply supervised fine-tuning (SFT) on a mixture of ∼600K diverse instruction-tuning examples (covering 10 languages) and translation data from SMOL and MeDLEy. SFT recovers instruction-following capabilities and biases the model towards high-quality translations, as seen in Table 5.

We further apply reinforcement learning using DAPO (Yu et al., 2025) with a combined ChrF++/MetricX reward signal and KL regularization. Table 5 shows that SFT and RL yield cumulative improvements, with RL providing additional gains on hard subsets without degrading any direction.

Retrieval-augmented translation (RAG). At inference time, we optionally retrieve parallel examples from our training data using a combination of full-text search and OMNISONAR embedding similarity, inserting them as few-shot examples. RAG provides complementary gains to fine-tuning, particularly for XX-En translation (§6.5).

4.3 OMT-NLLB: Encoder-Decoder Model

We train OMT-NLLB, a compact 3B-parameter encoder-decoder model that translates from 1,600 source languages into 250 target languages. Our key innovation is a three-stage training strategy that exploits non-parallel data through a cross-lingually aligned encoder.

Stage 1: Decoder enhancement with auto-encoding. We initialize from OMNISONAR, which maps 1,600+ languages into a shared sentence embedding space. Keeping the encoder frozen, we train the decoder on parallel MT data and monolingual data via auto-encoding—the decoder reconstructs monolingual input from the OMNISONAR sentence representations. This substantially increases training data for low-resource languages and boosts the generation fluency.

Stage 2: Removing the sentence-level bottleneck. The original OMNISONAR architecture uses pooled sentence-level representations for cross-attention, limiting fine-grained information transfer. We remove this bottleneck and introduce token-level encoder-decoder attention. A decoder warm-up phase (encoder frozen) allows adaptation to full-sequence attention without training instabilities.

Stage 3: End-to-end fine-tuning. Both encoder and decoder are jointly updated on parallel data, enabling full optimization of token-level interactions.

Table 6 shows progressive improvements at each stage over the OMNISONAR baseline.

5 Proposed Evaluation

XSTS+R+P human evaluation protocol. We extend the established XSTS semantic similarity protocol (Licht et al., 2022) with two additions designed to address systematic biases in previous MT evaluation protocols. The register (R) component adds scoring criteria that account for pragmatic discrepancies such as formality mismatches between source and translation. The paragraph-level (P) component adds consistency checks that penalize translations with sentence-level semantic equivalence but paragraph-level inconsistencies, such as inconsistent grammatical gender assignment to referents, shifting formality registers within a paragraph, or contradictory temporal framing across

sentences. XSTS+R+P uses a 5-point scale: 1 (no useful information preserved), 2 (some information preserved but translation is largely unusable), 3 (core meaning conveyed, moderate errors), 4 (good translation with minor issues), and 5 (perfect translation). XSTS+R+P achieves Krippendorff's α of 0.80, indicating substantially higher inter-annotator agreement than XSTS (0.57) and RSQM (0.44).

Met-BOUQuET. To train and evaluate automatic metrics, we collect the largest human annotation dataset for MT quality: Met-BOUQuET, comprising annotations across 161 language directions and 119 unique languages using the XSTS+R+P protocol. Round 1 covers 104 directions with diverse system outputs; Round 2 compares OMT-LLAMA against the strongest baseline per direction across 57 directions.

BLASER 3: Reference-Free Quality Estimation Existing reference-free QE metrics (xCOMET, MetricX) use encoders with limited cross-lingual coverage. BLASER 3 is a small dense model (inspired by Dale and Costa-jussà (2024)) built on top of frozen OMNISONAR embeddings (1,600 + languages) and is trained multi-task on 6 evaluation protocols (such as DA and MQM) using human annotations from the dev split of Met-BOUQuET, annotated data from NLLB Team et al. (2024) and WMT evaluation campaigns, and synthetic data scored with reference-based MetricX.

On Met-BOUQuET, BLASER 3 achieves Spearman correlation of 0.55 with human judgments, compared to 0.47 for MetricX and 0.43 for xCOMET, making it our primary reference-free metric.

OmniTOX OmniTOX extends MuTox (Costa-jussà et al., 2024) using OMNISONAR embeddings, providing binary toxicity classification for 1,600 languages—an eight-fold increase in language coverage over its predecessor, which supported approximately 200 languages. The model applies a classification head on top of OMNISONAR sentence embeddings, trained on toxicity annotations spanning multiple languages and domains. OmniTOX achieves a mean per-language ROC AUC of 0.86 across 30 diverse evaluation languages, outperforming MuTox by +0.06 mean per-language ROC AUC.

	XX-En				En-YY			
	H	M	L	VL	H	M	L	VL
OMT-LLAMA 8B	**65.1**	53.5	**38.7**	27.6	60.7	45.8	**30.8**	**18.8**
OMT-NLLB	64.0	**56.1**	37.4	**28.4**	**61.1**	**47.6**	27.3	17.5
OMT-LLAMA 3B	64.4	54.6	38.8	28.0	59.4	45.0	28.7	16.8
GPT-OSS 120B	64.6	49.9	33.6	25.4	63.6	43.7	26.0	16.0
Gemma 3	65.3	51.3	34.8	27.4	62.6	40.9	24.0	14.8
NLLB-200	61.5	52.4	32.1	25.5	62.9	46.8	24.6	17.5
LLAMA 3 70B	65.0	47.6	33.3	26.2	60.2	37.2	23.7	16.2
OMT-LLAMA 1B	63.0	51.5	34.7	24.1	56.5	42.2	25.1	14.9
Tower+ 72B	66.5	43.3	31.7	25.2	58.9	26.9	15.2	11.3
LLAMA 3 8B	61.0	39.8	27.5	21.6	52.7	26.0	13.1	9.5

Table 7: BOUQuET results (ChrF++) by resource level. OMT models lead across nearly all categories. Selected baselines; full results in supplementary.

6 Results

We evaluate on BOUQuET (275 languages), FLoRes+ (~200 languages), and Bible (1,560 languages), using ChrF++ for reference-based and BLASER 3 for reference-free evaluation, complemented by xCOMET (Guerreiro et al., 2024) and MetricX-24 (Juraska et al., 2024). Baselines include NLLB-200 (NLLB Team et al., 2024), LLAMA3 (Grattafiori et al., 2024), Gemma 3 (Gemma Team et al., 2025), GPT-OSS 20B/120B (OpenAI, 2025), MADLAD (Kudugunta et al., 2023), Aya-101 (Üstün et al., 2024), Tower+ (Alves et al., 2024), TranslateGemma (Finkelstein et al., 2026), and Hunyuan-MT (Zheng et al., 2025).

6.1 Standard Benchmarks

Table 7 reports ChrF++ on BOUQuET by resource level. For XX-En, both OMT-LLAMA and OMT-NLLB outperform all baselines at every level. For En-YY, OMT-NLLB excels on high/mid-resource languages while OMT-LLAMA is stronger on low/very-low-resource languages. Evaluation with FLoRes+ (Table 10 in the appendix) provides qualitatively similar results.

6.2 Long-Tail Understanding and Generation

We evaluate on the Bible benchmark (1,560 languages) to assess performance on the long tail of languages.

Understanding (XX-En). Figure 1 shows that OMT-LLAMA and OMT-NLLB consistently outperform all baselines across 1,560 languages. OMT-LLAMA 8B strictly outperforms all baselines in 1,045 languages. OMT-NLLB outperforms all baselines in XX-En across all 1,600 eval-

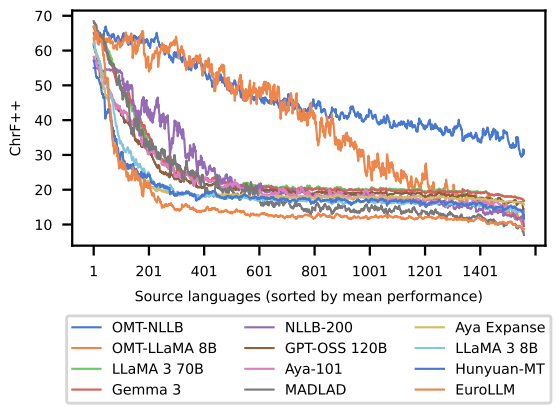

Figure 1: Long-tail **understanding** (XX-En, ChrF++ on Bible). OMT models (blue/orange) outperform baselines across 1,560 languages. Languages sorted by average performance; curves smoothed.

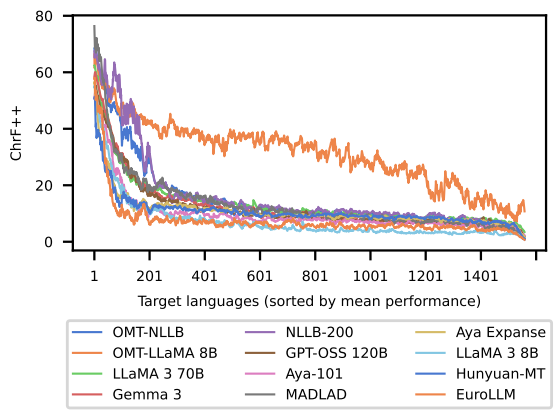

Figure 2: Long-tail **generation** (En-YY, ChrF++ on Bible). Baselines degrade to random at 300–400 languages; OMT-LLAMA (blue) sustains quality for ∼1,200 languages.

uated languages despite its smaller output vocabulary.

To understand how many languages our models understand "sufficiently well" for practical use, we define two quality thresholds: average XSTS+R+P score of 3.5 ("good") and 2.5 ("passable"). We use the BOUQuET test set to map these values to the more neutral MetricX metric, obtaining thresholds of around 4.0 and 6.7, respectively. With them, we estimate that OMT-LLAMA achieves "passable" quality for 440 languages and OMT-NLLB for 416—nearly double the 221 for NLLB-200.

Generation (En-YY). Figure 2 reveals a striking generation bottleneck: baseline models degrade to near-random quality at 300–400 languages, while OMT-LLAMA maintains non-trivial quality for ∼1,200. Determining the "passable" generation quality levels would require more human evaluation and is deferred to future work.

6.3 Model Size vs. Performance

Figure 3 shows the Pareto frontier of model size vs. quality on BOUQuET. The 1B OMT-LLAMA matches baselines 3×–8× larger, and the 8B OMT-LLAMA outperforms the 70B LLAMA3—demonstrating that *specialization* outperforms scale for multilingual MT.

6.4 Added toxicity

Using OmniTOX, we systematically benchmarked added toxicity—the increase in toxicity between source text and translation—across 207 languages for evaluated systems. We compute debiased toxicity logit differences between source and translation,

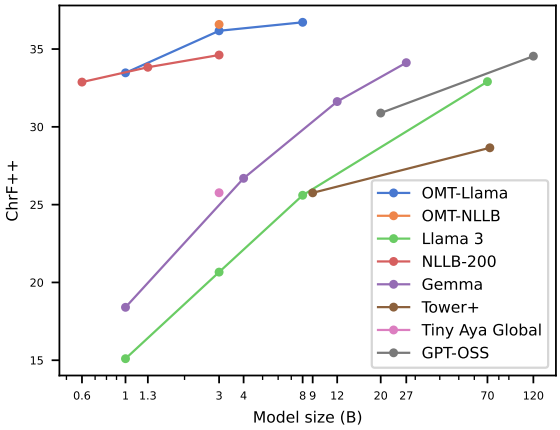

Figure 3: Model size vs. translation quality (ChrF++ on BOUQuET). OMT models define the Pareto frontier—specialized 1B–8B models match or exceed 70B general-purpose LLMs.

with per-language baselines established on non-toxic BOUQuET sentences and flagging thresholds calibrated at a 5% false positive rate. None of the evaluated systems exhibit meaningful added toxicity, with flagging rates below 1.5% across all model configurations and language directions.

6.5 Extensibility

An omnilingual model provides broad coverage, but targeted improvement for specific languages is often needed. We evaluate fine-tuning and RAG on 5 low-resource (LR) and 20 very low-resource (VLR) languages using available parallel data (11K–300K examples). As Table 8 shows, fine-tuning improves En-YY by +4.8 ChrF++ but slightly degrades XX-En. RAG improves XX-En

System	20 VLRL En-YY	20 VLRL XX-En	5 LRL En-YY	5 LRL XX-En
OMT-LLaMA + FT + RAG	**29.3**	**36.7**	46.7	**53.2**
OMT-LLaMA + FT	28.9	29.0	**47.5**	42.7
OMT-LLaMA + RAG	26.8	36.2	44.8	54.0
OMT-LLaMA	24.1	29.9	43.0	48.4
LLaMA-base + FT + RAG	20.1	27.9	33.2	40.9
LLaMA-base + FT	19.4	26.4	31.2	38.8
LLaMA-base + RAG	17.5	22.9	25.0	31.1
LLaMA-base	9.3	20.7	16.1	28.7

Table 8: Results of extending the OMT-LLAMA and LLaMA-base models (BOUQuET, dev split, ChrF++) with a sample of low-resourced and very low-resourced languages.

Direction group	Win%	OMT	Base	N
high → high	0.75	4.03	3.45	16
high → low	0.83	2.99	1.61	18
low → high	0.80	3.26	2.69	15
low → low	0.62	2.96	2.66	8
Total	**0.77**	**3.38**	**2.67**	**57**

Table 9: Human evaluation (Met-BOUQuET Round 2). Win rate = proportion of directions where OMT-LLAMA's mean XSTS+R+P score exceeds the baseline's. Scores are on a 1–5 scale.

by +6.3 ChrF++. Combining both yields the best results: +5.1 En-YY and +6.8 XX-En. Crucially, even with both techniques, the base LLAMA3 does not outperform unadapted OMT-LLAMA, confirming that OMT-LLAMA is the superior starting point for adaptation.

6.6 Human Evaluation

We conduct human evaluation with Met-BOUQuET Round 2, comparing OMT-LLAMA against the strongest baseline per direction across 57 directions and 80 languages (Table 9). OMT-LLAMA achieves a 77% win rate overall, with the largest gains for high→low directions (83%), where baselines fail to produce meaningful translations. The improvement from 2.67 to 3.38 mean XSTS+R+P score represents a jump from "minimally useful" to "useful."

7 Conclusion

Omnilingual MT demonstrates that scaling multilingual translation to 1,600+ languages requires coordinated advances across data, modeling, and evaluation. Our data strategy—combining massive public corpora with backtranslation for 2,000+ languages and the manually curated MeDLEy dataset—addresses the generation bottleneck for long-tail languages. Our two architectural approaches—OMT-LLAMA (decoder-only, 1B/3B/8B) and OMT-NLLB (encoder-decoder, 3B with novel three-stage training)—deliver complementary strengths: OMT-LLAMA covers more output languages while OMT-NLLB excels at cross-lingual understanding.

Large-scale evaluation reveals that Omnilingual MT doubles the number of "passable quality" languages from ∼200 to over 440, and maintains generation quality for ∼1,200 languages where baselines support 3-4 times less. This highlights that genuinely inclusive MT is a challenge that requires targeted data creation (such as MeDLEy and backtranslation) and innovative use of monolingual data (bottlenecked autoencoding) rather than simply scaling model parameters.

Our specialized 1B–8B models match or exceed the comparable 70B general-purpose LLM in massively multilingual MT, offering a clear Pareto advantage. Therefore, model specialization through continual pretraining and post-training provides a more efficient path to multilingual translation quality than "simply" increasing model size.

The human evaluation confirms a 77% win rate of OMT against strongest baselines across 57 diverse translation directions. Our evaluation suite, BOUQuET (275 languages) and Met-BOUQuET (161 directions), provides the community with tools to assess MT at unprecedented scale and develop even better translation systems and evaluation metrics.

Finally, we show that post-training techniques targeting multilingual extension of models do not compensate the lack of large improvements obtained by embracing massively multilingual training data and vocabulary in the earlier stages of training. This should motivate model developers interested in boosting multilingual tasks performance to develop models that are highly multilingual by design and are well equipped to extend their support to any additional language, if necessary.

Looking forward, we believe that foundational models should aim to provide at least a base level of support for *every* language—not only those with abundant data and rich markets. Omnilingual MT positions large-scale linguistic inclusion as an ongoing challenge demanding continued investment in data creation, architecture design, and evaluation methodology. The recipes presented here can in principle be reproduced on top of diverse base models, and we hope they inspire further research toward genuinely universal language technology.

Limitations and Ethical Considerations

Omnilingual MT has several limitations. OMT-NLLB generates only 250 target languages vs. ∼1,200 for OMT-LLAMA. Both models are based on the Llama 3 family, now two years old, which makes it stale by the LLM standards. Training data overrepresents formal registers and religious domains. We did not exhaustively optimize the prompts of the baseline models. The extensibility results cover only 25 languages.

Our evaluation is also limited. The long-tail evaluation relies on Bible text (domain-limited, formal register), and human evaluation covers only 57 directions. While we expand the number of languages with "passable" understanding from ∼200 to ∼440, this still covers only a fraction of the world's languages. And determining the languages for which our models meet the "passable" bar of generation quality would require either more human evaluation or extensive research into automated multilingual metrics of fluency evaluation.

A longer and more detailed technical report can be found at *(URL blinded for the anonymized review)* .

Human annotators were paid a fair rate.

References

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. *Preprint*, arXiv:2402.17733.

Pierre Andrews, Guillaume Wenzek, Kevin Heffernan, Onur Çelebi, Anna Sun, Ammar Kamran, Yingzhe Guo, Alexandre Mourachko, Holger Schwenk, and Angela Fan. 2022. stopes-modular machine translation pipelines. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 258–265.

Laurie Burchell, Jean Maillard, Antonios Anastasopoulos, Christian Federmann, Philipp Koehn, and Skyler Wang. 2024. Findings of the WMT 2024 shared task of the open language data initiative. In *Proceedings of the Ninth Conference on Machine Translation*, pages 110–117, Miami, Florida, USA. Association for Computational Linguistics.

Isaac Caswell, Elizabeth Nielsen, Jiaming Luo, Colin Cherry, Geza Kovacs, Hadar Shemtov, Partha Talukdar, Dinesh Tewari, Baba Mamadi Diane, Koulako Moussa Doumbouya, Djibrila Diane, Solo Farabado Cissé, Edoardo Ferrante, Alessandro Guasoni, Mamadou K. Keita, Sudhamoy DebBarma, Ali Kuzhuget, David Anugraha, Muhammad Ravi Shulthan Habibi, and 3 others. 2025. SMOL: Professionally translated parallel data for 115 underrepresented languages. *Preprint*, arXiv:2502.12301.

Emily Chang and Niyati Bafna. 2025. Chikhapo: A large-scale multilingual benchmark for evaluating lexical comprehension and generation in large language models. *arXiv preprint arXiv:2510.16928*.

Marta R. Costa-jussà, Mariano Coria Meglioli, Pierre Andrews, David Dale, Prangthip Hansanti, Elahe Kalbassi, Alex Mourachko, Christophe Ropers, and Carleigh Wood. 2024. MuTox: Universal MUltilingual audio-based TOXicity dataset and zero-shot detector. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5725–5734, Bangkok, Thailand. Association for Computational Linguistics.

David Dale, Laurie Burchell, Jean Maillard, Idris Abdulmumin, Antonios Anastasopoulos, Isaac Caswell, and Philipp Koehn. 2025. Findings of the WMT 2025 shared task of the open language data initiative. In *Proceedings of the Tenth Conference on Machine Translation*, pages 495–502, Suzhou, China. Association for Computational Linguistics.

David Dale and Marta R. Costa-jussà. 2024. BLASER 2.0: a metric for evaluation and quality estimation of massively multilingual speech and text translation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16075–16085, Miami, Florida, USA. Association for Computational Linguistics.

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.

Zhaopeng Feng, Shaosheng Cao, Jiahan Ren, Jiayuan Su, Ruizhe Chen, Yan Zhang, Zhe Xu, Yao Hu, Jian Wu, and Zuozhu Liu. 2025. Mt-r1-zero: Advancing llm-based machine translation via r1-zero-like reinforcement learning. *Preprint*, arXiv:2504.10160.

Mara Finkelstein, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Parker Riley, Daniel Deutsch, Cole Dilanni, Colin Cherry, Eleftheria Briakou, and 1 others. 2026. Translategemma technical report. *arXiv preprint arXiv:2601.09012*.

Markus Frohmann, Igor Sterner, Ivan Vulić, Benjamin Minixhofer, and Markus Schedl. 2024. Segment any text: A universal approach for robust, efficient and adaptable sentence segmentation. *arXiv preprint arXiv:2406.16678*.

Leonidas Gee, Andrea Zugarini, Leonardo Rigutini, and Paolo Torroni. 2022. Fast vocabulary transfer for language model compression. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 409–416.

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.

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.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Alexander Jones, Isaac Caswell, Orhan Firat, and Ishank Saxena. 2023. GATITOS: Using a new multilingual lexicon for low-resource machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 371–405, Singapore. 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.

David Kamholz, Jonathan Pool, and Susan M Colowick. 2014. Panlex: Building a resource for panlingual lexical translation. In *LREC*, volume 14, pages 3145–3150.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. GlotLID: Language identification for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6155–6218, Singapore. Association for Computational Linguistics.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: A multilingual and document-level large audited dataset. *Advances in Neural Information Processing Systems*, 36:67284–67296.

Daniel Licht, Cynthia Gao, Janice Lam, Francisco Guzman, Mona Diab, and Philipp Koehn. 2022. Consistent human evaluation of machine translation across language pairs. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 309–321, Orlando, USA. Association for Machine Translation in the Americas.

Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. 2024. Fineweb-edu: the finest collection of educational content.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Omnilingual MT Team, Pierre Andrews, Mikel Artetxe, Mariano Coria Meglioli, Marta R. Costa-jussà, Joe Chuang, David Dale, Cynthia Gao, Jean Maillard, Alex Mourachko, Christophe Ropers, Safiyyah Saleem, Eduardo Sánchez, Ioannis Tsiamas, Arina Turkatenko, Albert Ventayol-Boada, and Shireen Yates. 2025. Bouquet: dataset, benchmark and open initiative for universal quality evaluation in translation. *Preprint*, arXiv:2502.04314.

Omnilingual SONAR Team, João Maria Janeiro, Pere-Lluís Huguet Cabot, Ioannis Tsiamas, Yen Meng, Vivek Iyer, Guillem Ramírez, Loic Barrault, Belen Alastruey, Yu-An Chung, Marta R. Costa-Jussa, David Dale, Kevin Heffernan, Jaehyeong Jo, Artyom Kozhevnikov, Alexandre Mourachko, Christophe Ropers, Holger Schwenk, and Paul-Ambroise Duquenne. 2026. Omnilingual sonar: Cross-lingual and cross-modal sentence embeddings bridging massively multilingual text and speech. *Preprint*, arXiv:2603.16606.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. Fineweb2: One pipeline to scale them all – adapting pre-training data processing to every language. *Preprint*, arXiv:2506.20920.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. 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.

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.

Aditya Siddhant, Ankur Bapna, Orhan Firat, Yuan Cao, Mia Xu Chen, Isaac Caswell, and Xavier Garcia. 2022. Towards the next 1000 languages in multilingual machine translation: Exploring the synergy between supervised and self-supervised learning. *arXiv preprint arXiv:2201.03110*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Jiaan Wang, Fandong Meng, and Jie Zhou. 2025. Deeptrans: Deep reasoning translation via reinforcement learning. *Preprint*, arXiv:2504.10187.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, and 16 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *Preprint*, arXiv:2503.14476.

Mao Zheng, Zheng Li, Bingxin Qu, Mingyang Song, Yang Du, Mingrui Sun, and Di Wang. 2025. Hunyuan-mt technical report. *Preprint*, arXiv:2509.05209.

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.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin 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. *arXiv preprint arXiv:2402.07827*.

A Languages at a glance

The next Table lists the languages covered by the evaluation datasets that we contribute with this work (275 in BOUQuET, and 120 in Met-BOUQuET), our seed training dataset MeDLEy and our previous and ongoing evaluation community effort of FLoRes+.

ISO 639-3	ISO 15924	Language family	BOUQuET	MeDLEy	met-BOUQuET	FLoRes+
aar	Latn	Afro-Asiatic	✓	×	✓(R1)	×
abl	Latn	Austronesian	✓	×	✓(R2)	×
abz	Arab	Timor-Alor-Pantar	×	×	✓(R2)	×
ace	Arab	Austronesian	×	×	×	✓
ace	Latn	Austronesian	×	×	×	✓
acm	Arab	Afro-Asiatic	×	×	×	✓
acq	Arab	Afro-Asiatic	×	×	×	✓
aeb	Arab	Afro-Asiatic	×	×	×	✓
afr	Latn	Indo-European	✓	×	✓(R2)	✓
agr	Latn	Chicham	✓	×	✓(R1)	×
ahk	Thai	Sino-Tibetan	×	✓	×	×
aiq	Arab	Indo-European	✓	×	×	×
akb	Latn	Austronesian	×	✓	✓(R2)	×
als	Latn	Indo-European	✓	×	×	✓
amh	Ethi	Afro-Asiatic	✓	×	×	✓
ami	Latn	Austronesian	✓	×	✓(R1)	×
ane	Latn	Austronesian	✓	×	×	×
any	Latn	Atlantic-Congo	×	✓	✓(R2)	×
apc	Arab	Afro-Asiatic	✓	×	✓(R2)	✓
arb	Arab	Afro-Asiatic	×	✓	×	✓
arb	Latn	Afro-Asiatic	×	×	×	✓
arg	Latn	Indo-European	×	×	×	✓
arh	Latn	Chibchan	✓	×	×	×
arn	Latn	Araucanian	✓	×	×	×
ars	Arab	Afro-Asiatic	×	×	×	✓
ary	Arab	Afro-Asiatic	×	×	×	✓
arz	Arab	Afro-Asiatic	✓	×	✓(R1)	✓
arz	Latn	Afro-Asiatic	✓	×	✓(R1)	×
asm	Beng	Indo-European	✓	×	✓(R2)	✓
ast	Latn	Indo-European	×	×	×	✓
ati	Latn	Atlantic-Congo	×	✓	×	×
awa	Deva	Indo-European	×	×	✓(R2)	✓
ayr	Latn	Aymaran	✓	×	×	✓
ayz	Latn	Maybratic	✓	×	×	×
azb	Arab	Turkic	✓	✓	×	✓
azj	Latn	Turkic	✓	×	×	✓
azm	Latn	Otomanguean	✓	×	×	×
azz	Latn	Uto-Aztecan	✓	✓	×	×
bak	Cyrl	Turkic	✓	×	×	✓
bam	Latn	Mande	✓	✓	✓(R2)	✓
ban	Latn	Austronesian	×	×	×	✓
bas	Latn	Atlantic-Congo	✓	×	×	×
bba	Latn	Atlantic-Congo	✓	✓	×	×
bbc	Latn	Austronesian	×	✓	×	×
bcc	Arab	Indo-European	×	✓	✓(R2)	×
bel	Cyrl	Indo-European	✓	×	✓(R2)	✓
bem	Latn	Atlantic-Congo	×	×	✓(R2)	✓
ben	Beng	Indo-European	✓	×	✓(R1, R2)	✓
ben	Latn	Indo-European	✓	×	×	×
bfa	Latn	Nilotic	×	✓	×	×
bft	Arab	Sino-Tibetan	✓	×	×	×
bhb	Deva	Indo-European	✓	×	✓(R2)	×
bho	Deva	Indo-European	✓	×	×	✓
bib	Latn	Mande	×	✓	×	×
bjn	Arab	Austronesian	×	×	×	✓
bjn	Latn	Austronesian	×	×	×	✓
blt	Latn	Tai-Kadai	×	✓	×	×
bod	Tibt	Sino-Tibetan	✓	×	×	✓
bom	Latn	Atlantic-Congo	×	✓	×	×
bos	Latn	Indo-European	✓	×	×	✓
bre	Latn	Indo-European	✓	×	✓(R2)	×

brh	Arab	Dravidian	✓	×	×	×
brx	Deva	Sino-Tibetan	✓	×	×	✓
bsh	Arab	Indo-European	✓	×	×	×
bsk	Arab	Burushaski	✓	×	×	×
bsq	Latn	Kru	×	✓	✓(R2)	×
bug	Latn	Austronesian	×	×	×	✓
bul	Cyrl	Indo-European	✓	×	×	✓
cak	Latn	Mayan	✓	×	×	×
cat	Latn	Indo-European	✓	×	✓(R2)	✓
ceb	Latn	Austronesian	✓	×	✓(R2)	✓
ces	Latn	Indo-European	✓	×	✓(R1)	✓
che	Cyrl	Nakh-Daghestanian	✓	×	×	×
chr	Cher	Iroquoian	✓	×	×	×
chv	Cyrl	Turkic	✓	×	×	✓
cja	Arab	Austronesian	✓	×	×	×
cjk	Latn	Atlantic-Congo	✓	✓	×	✓
ckb	Arab	Indo-European	✓	×	✓(R2)	✓
ckl	Latn	Afro-Asiatic	✓	×	×	×
cmn	Hans	Sino-Tibetan	✓	×	✓(R1)	✓
cmn	Hant	Sino-Tibetan	✓	×	✓(R1)	✓
crh	Latn	Turkic	×	×	×	✓
crk	Cans	Algic	✓	×	✓(R1)	×
crk	Latn	Algic	✓	×	×	×
cux	Latn	Otomanguean	✓	×	×	×
cym	Latn	Indo-European	✓	×	×	✓
dan	Latn	Indo-European	✓	×	×	✓
daq	Deva	Dravidian	✓	×	×	×
dar	Cyrl	Nakh-Daghestanian	×	×	×	✓
deu	Latn	Indo-European	✓	×	✓(R1)	✓
dga	Latn	Atlantic-Congo	×	✓	✓(R2)	×
dgo	Deva	Indo-European	✓	×	×	✓
dik	Latn	Nilotic	✓	✓	✓(R2)	✓
diq	Latn	Indo-European	✓	×	×	×
div	Thaa	Indo-European	✓	×	×	×
djc	Latn	Dajuic	✓	×	×	×
dje	Latn	Songhay	✓	×	✓(R1)	×
dnj	Latn	Mande	×	✓	×	×
dtm	Latn	Dogon	✓	×	×	×
dts	Latn	Dogon	✓	×	×	×
dua	Latn	Atlantic-Congo	✓	×	×	×
dyu	Latn	Mande	×	✓	×	✓
dzo	Tibt	Sino-Tibetan	✓	✓	×	✓
ekk	Latn	Uralic	✓	×	✓(R2)	✓
ell	Grek	Indo-European	✓	×	✓(R1)	✓
enb	Latn	Nilotic	✓	×	×	×
eng	Latn	Indo-European	✓	✓	✓(R1, R2)	✓
enl	Latn	Lengua-Mascoy	✓	×	×	×
epo	Latn	Esperanto	×	×	×	✓
eto	Latn	Atlantic-Congo	✓	×	×	×
eus	Latn	Basque	✓	×	✓(R2)	✓
ewe	Latn	Atlantic-Congo	×	✓	×	✓
ewo	Latn	Atlantic-Congo	✓	×	×	×
fao	Latn	Indo-European	✓	×	×	✓
fia	Copt	Nubian	✓	×	×	×
fij	Latn	Austronesian	×	✓	×	✓
fil	Latn	Austronesian	×	×	×	✓
fin	Latn	Uralic	✓	×	×	✓
fon	Latn	Atlantic-Congo	×	✓	×	✓
fra	Latn	Indo-European	✓	✓	✓(R1, R2)	✓
fry	Latn	Indo-European	✓	×	✓(R2)	×
fuc	Latn	Atlantic-Congo	✓	×	×	×
fur	Latn	Indo-European	×	×	×	✓
fuv	Latn	Atlantic-Congo	✓	✓	✓(R2)	✓
fvr	Latn	Furan	✓	×	×	×
gax	Latn	Afro-Asiatic	✓	×	×	×
gaz	Latn	Afro-Asiatic	✓	×	✓(R1)	✓
gil	Latn	Austronesian	✓	×	✓(R1)	×
gkp	Latn	Mande	✓	×	×	×
gla	Latn	Indo-European	✓	×	×	✓
gle	Latn	Indo-European	✓	×	×	✓

glg	Latn	Indo-European	✓	×	×	✓
gom	Deva	Indo-European	✓	×	×	✓
gor	Latn	Austronesian	×	✓	×	×
grt	Latn	Sino-Tibetan	×	✓	×	×
guc	Latn	Arawakan	✓	×	✓(R1)	×
gug	Latn	Tupian	✓	×	✓(R2)	✓
guj	Gujr	Indo-European	✓	×	×	✓
guz	Latn	Atlantic-Congo	✓	×	×	×
gxx	Latn	Kru	✓	×	×	×
hat	Latn	Indo-European	✓	×	✓(R2)	✓
hau	Latn	Afro-Asiatic	✓	×	✓(R2)	✓
heb	Hebr	Afro-Asiatic	✓	×	✓(R2)	✓
heh	Latn	Atlantic-Congo	✓	✓	×	×
hig	Latn	Afro-Asiatic	×	✓	×	×
hin	Deva	Indo-European	✓	✓	✓(R1, R2)	✓
hin	Latn	Indo-European	✓	×	✓(R1)	×
hne	Deva	Indo-European	✓	×	×	✓
hrv	Latn	Indo-European	✓	×	✓(R1)	✓
hun	Latn	Uralic	✓	×	✓(R1)	✓
hve	Latn	Huavean	✓	×	×	×
hye	Armn	Indo-European	✓	×	✓(R2)	✓
ibo	Latn	Atlantic-Congo	✓	×	✓(R2)	✓
ijc	Latn	Ijoid	✓	×	×	×
ilo	Latn	Austronesian	✓	×	×	✓
ind	Latn	Austronesian	✓	✓	✓(R1, R2)	✓
irk	Latn	Afro-Asiatic	✓	✓	×	×
isl	Latn	Indo-European	✓	×	×	✓
ita	Latn	Indo-European	✓	×	✓(R1, R2)	✓
jav	Latn	Austronesian	✓	×	✓(R1)	✓
jmc	Latn	Atlantic-Congo	✓	✓	×	×
jnj	Latn	Ta-Ne-Omotic	✓	×	×	×
jpn	Jpan	Japonic	✓	×	✓(R1)	✓
kaa	Cyrl	Turkic	✓	×	✓(R1)	✓
kab	Latn	Afro-Asiatic	×	×	×	✓
kac	Latn	Sino-Tibetan	✓	✓	×	✓
kai	Latn	Afro-Asiatic	✓	×	×	×
kal	Latn	Eskimo-Aleut	✓	×	✓(R1)	×
kam	Latn	Atlantic-Congo	✓	✓	×	✓
kan	Knda	Dravidian	✓	×	×	✓
kas	Arab	Indo-European	×	×	×	✓
kas	Deva	Indo-European	×	×	×	✓
kat	Geor	Kartvelian	✓	×	✓(R2)	✓
kaz	Cyrl	Turkic	✓	×	×	✓
kbp	Latn	Atlantic-Congo	×	✓	×	✓
kde	Latn	Atlantic-Congo	×	✓	×	×
kdj	Latn	Nilotic	×	✓	✓(R2)	×
kea	Latn	Indo-European	✓	×	×	✓
kek	Latn	Mayan	✓	✓	×	×
khk	Cyrl	Mongolic-Khitan	✓	×	×	✓
khm	Khmr	Austroasiatic	✓	×	✓(R1)	✓
khq	Latn	Songhay	✓	×	×	×
khw	Latn	Indo-European	✓	×	×	×
kik	Latn	Atlantic-Congo	×	×	×	✓
kin	Latn	Atlantic-Congo	✓	×	✓(R2)	✓
kir	Cyrl	Turkic	✓	×	×	✓
kls	Arab	Indo-European	✓	×	×	×
kmb	Latn	Atlantic-Congo	✓	✓	×	✓
kmr	Latn	Indo-European	✓	×	×	✓
knc	Arab	Saharan	✓	✓	✓(R2)	✓
knc	Latn	Saharan	×	×	×	✓
knw	Latn	Kxa	✓	×	×	×
kor	Kore	Koreanic	✓	×	✓(R1)	✓
krt	Latn	Saharan	✓	×	×	×
kru	Deva	Dravidian	✓	×	✓(R1)	×
ksf	Latn	Atlantic-Congo	✓	×	×	×
ktu	Latn	Atlantic-Congo	✓	✓	×	✓
kuj	Latn	Atlantic-Congo	✓	×	×	×
kus	Latn	Atlantic-Congo	×	✓	×	×
kwy	Latn	Atlantic-congo	✓	×	×	×
kxp	Arab	Indo-European	✓	×	×	×

lao	Laoo	Tai-Kadai	✓	×	×	✓
led	Latn	Central Sudanic	✓	×	×	×
lgg	Latn	Central Sudanic	✓	✓	×	×
lia	Latn	Atlantic-Congo	×	✓	×	×
lij	Latn	Indo-European	✓	×	✓(R1)	✓
lim	Latn	Indo-European	✓	×	×	✓
lin	Latn	Atlantic-Congo	✓	×	✓(R1, R2)	✓
lir	Latn	Pidgin	✓	×	✓(R2)	×
lit	Latn	Indo-European	✓	×	×	✓
lld	Latn	Indo-European	×	×	×	✓
lmo	Latn	Indo-European	×	×	×	✓
loa	Latn	North Halmahera	✓	×	×	×
loh	Latn	Surmic	✓	×	×	×
lon	Latn	Atlantic-Congo	×	✓	×	×
ltg	Latn	Indo-European	×	×	×	✓
ltz	Latn	Indo-European	×	×	×	✓
lua	Latn	Atlantic-Congo	×	✓	✓(R2)	✓
lug	Latn	Atlantic-Congo	✓	✓	×	✓
luo	Latn	Nilotic	✓	✓	✓(R2)	✓
lus	Latn	Sino-Tibetan	×	×	×	✓
lvs	Latn	Indo-European	✓	×	×	✓
mad	Latn	Austronesian	×	✓	✓(R2)	×
maf	Latn	Afro-Asiatic	✓	×	×	×
mag	Deva	Indo-European	×	×	×	✓
mah	Latn	Austronesian	×	✓	×	×
mai	Deva	Indo-European	✓	×	×	✓
mak	Latn	Austronesian	×	✓	×	×
mal	Mlym	Dravidian	✓	×	✓(R2)	✓
mam	Latn	Mayan	✓	✓	×	×
mar	Deva	Indo-European	✓	×	✓(R2)	✓
mas	Latn	Nilotic	✓	×	×	×
men	Latn	Mande	×	✓	×	×
mey	Latn	Afro-Asiatic	✓	×	×	×
mfe	Latn	Indo-European	×	×	×	✓
mhr	Cyrl	Uralic	×	×	×	✓
mie	Latn	Otomanguean	✓	×	×	×
mim	Latn	Otomanguean	×	✓	✓(R2)	×
min	Arab	Austronesian	✓	×	✓(R2)	✓
min	Latn	Austronesian	×	×	×	✓
mio	Latn	Otomanguean	×	✓	×	×
miq	Latn	Misumalpan	✓	✓	×	×
mkd	Cyrl	Indo-European	✓	×	×	✓
mlt	Latn	Afro-Asiatic	✓	×	×	✓
mni	Beng	Sino-Tibetan	×	×	×	✓
mni	Mtei	Sino-Tibetan	×	✓	✓(R2)	✓
mos	Latn	Atlantic-Congo	✓	✓	✓(R2)	✓
mri	Latn	Austronesian	✓	×	×	✓
mrw	Latn	Austronesian	×	✓	×	×
mtq	Latn	Austroasiatic	✓	×	×	×
mya	Mymr	Sino-Tibetan	✓	×	✓(R1)	✓
myv	Cyrl	Uralic	×	×	×	✓
myx	Latn	Atlantic-Congo	×	✓	×	×
mzl	Latn	Mixe-Zoque	✓	×	×	×
mzm	Latn	Atlantic-Congo	×	✓	×	×
naq	Latn	Khoe-Kwadi	✓	×	×	×
nga	Latn	Atlantic-Congo	×	✓	×	×
ngl	Latn	Atlantic-Congo	×	✓	×	×
ngu	Latn	Uto-Aztecan	×	✓	×	×
nhe	Latn	Uto-Aztecan	✓	×	×	×
nia	Latn	Austronesian	×	✓	×	×
nij	Latn	Austronesian	×	✓	×	×
nim	Latn	Atlantic-Congo	×	✓	×	×
nld	Latn	Indo-European	✓	×	✓(R1, R2)	✓
nlv	Latn	Uto-Aztecan	✓	×	✓(R2)	×
nno	Latn	Indo-European	✓	×	×	✓
nob	Latn	Indo-European	×	×	×	✓
npi	Deva	Indo-European	✓	×	✓(R2)	✓
nqo	Nkoo	N'Ko	×	×	×	✓
nso	Latn	Atlantic-Congo	✓	×	✓(R2)	✓
nuj	Latn	Atlantic-Congo	×	✓	×	×

nus	Latn	Nilotic	✓	✓	✓(R2)	✓
nya	Latn	Atlantic-Congo	✓	×	×	✓
nyy	Latn	Atlantic-Congo	×	✓	×	×
oci	Latn	Indo-European	×	×	×	✓
ory	Orya	Indo-European	✓	×	×	✓
osi	Latn	Austronesian	×	×	✓(R2)	×
pag	Latn	Austronesian	×	×	×	✓
pam	Latn	Austronesian	×	✓	×	×
pan	Guru	Indo-European	✓	×	×	✓
pap	Latn	Indo-European	×	×	×	✓
pbs	Latn	Otomanguean	✓	×	×	×
pbt	Arab	Indo-European	✓	×	✓(R2)	✓
pcm	Latn	Indo-European	✓	✓	✓(R2)	×
pes	Arab	Indo-European	✓	×	✓(R1, R2)	✓
plt	Latn	Austronesian	✓	×	×	✓
pol	Latn	Indo-European	✓	×	✓(R1, R2)	✓
por	Latn	Indo-European	✓	×	✓(R1)	✓
prs	Arab	Indo-European	×	×	×	✓
quc	Latn	Mayan	✓	✓	×	×
quh	Latn	Quechuan	✓	×	×	×
quy	Latn	Quechuan	×	×	×	✓
quz	Latn	Quechuan	✓	×	×	×
rhg	Rohg	Indo-European	×	✓	×	×
rim	Latn	Atlantic-Congo	×	✓	×	×
rmy	Latn	Indo-European	×	✓	×	×
rob	Latn	Austronesian	✓	×	×	×
roh	Latn	Indo-European	✓	×	×	×
ron	Latn	Indo-European	✓	×	✓(R1)	✓
run	Latn	Atlantic-Congo	×	×	×	✓
rus	Cyrl	Indo-European	✓	✓	✓(R1, R2)	✓
sag	Latn	Atlantic-Congo	×	×	×	✓
san	Deva	Indo-European	×	×	×	✓
sat	Olck	Austroasiatic	✓	✓	×	✓
sba	Latn	Central Sudanic	✓	✓	✓(R1)	×
scn	Latn	Indo-European	✓	×	×	✓
sgc	Latn	Nilotic	✓	×	×	×
shk	Latn	Nilotic	×	✓	×	×
shn	Mymr	Tai-Kadai	✓	✓	×	✓
sif	Latn	Siamou	✓	×	×	×
sin	Sinh	Indo-European	✓	×	×	✓
skr	Arab	Indo-European	✓	×	×	×
slk	Latn	Indo-European	✓	×	×	✓
slv	Latn	Indo-European	✓	×	×	✓
sme	Latn	Uralic	✓	×	×	×
smo	Latn	Austronesian	×	×	×	✓
sna	Latn	Atlantic-congo	✓	×	×	✓
snd	Arab	Indo-European	✓	×	×	✓
snd	Deva	Indo-European	×	×	×	✓
som	Latn	Afro-Asiatic	✓	×	×	✓
sot	Latn	Atlantic-Congo	✓	×	✓(R2)	✓
spa	Latn	Indo-European	✓	✓	✓(R1, R2)	✓
sro	Latn	Indo-European	✓	×	×	×
srd	Latn	Indo-European	×	×	×	✓
srp	Cyrl	Indo-European	✓	×	×	✓
ssw	Latn	Atlantic-Congo	✓	×	✓(R2)	✓
sun	Latn	Austronesian	✓	×	×	✓
swe	Latn	Indo-European	✓	×	✓(R1)	✓
swh	Latn	Atlantic-Congo	✓	✓	✓(R1, R2)	✓
syl	Beng	Indo-European	×	✓	×	×
szl	Latn	Indo-European	✓	×	✓(R2)	✓
taj	Deva	Sino-Tibetan	×	✓	×	×
tam	Latn	Dravidian	✓	×	×	×
tam	Taml	Dravidian	✓	×	✓(R2)	✓
taq	Latn	Afro-Asiatic	✓	✓	×	✓
taq	Tfng	Afro-Asiatic	✓	✓	×	✓
tat	Cyrl	Turkic	✓	×	✓(R2)	✓
tda	Latn	Songhay	✓	×	✓(R2)	×
tel	Latn	Dravidian	✓	×	×	✓
tel	Telu	Dravidian	✓	×	×	×
tem	Latn	Atlantic-Congo	×	✓	×	×

teo	Latn	Nilotic	×	✓	×	×
tgk	Cyrl	Indo-European	✓	×	×	✓
tgl	Latn	Austronesian	✓	×	✓(R1, R2)	×
tha	Thai	Tai-Kadai	✓	×	✓(R1)	✓
tir	Ethi	Afro-Asiatic	✓	✓	✓(R1)	✓
toc	Latn	Totonacan	✓	×	×	×
tpi	Latn	Indo-European	✓	×	×	✓
tpl	Latn	Otomanguean	✓	×	×	×
tsg	Latn	Austronesian	✓	×	✓(R2)	×
tsn	Latn	Atlantic-Congo	✓	✓	✓(R2)	✓
tso	Latn	Atlantic-Congo	✓	×	✓(R2)	✓
tsz	Latn	Tarascan	✓	✓	×	✓
tui	Latn	Atlantic-Congo	✓	×	×	×
tuk	Latn	Turkic	×	×	×	✓
tum	Latn	Atlantic-Congo	×	✓	×	✓
tur	Latn	Turkic	✓	×	✓(R1)	✓
twi	Latn	Atlantic-Congo	✓	×	×	✓
tyv	Cyrl	Turkic	×	×	×	✓
tzh	Latn	Mayan	✓	✓	×	×
tzm	Tfng	Afro-Asiatic	✓	✓	✓(R2)	×
tzo	Latn	Mayan	×	✓	✓(R2)	×
uig	Arab	Turkic	✓	×	×	✓
ukr	Cyrl	Indo-European	✓	×	✓(R1)	✓
umb	Latn	Atlantic-Congo	✓	✓	×	✓
urd	Arab	Indo-European	✓	×	✓(R1)	✓
urd	Latn	Indo-European	✓	×	×	×
uzn	Latn	Turkic	✓	×	×	✓
uzs	Arab	Turkic	×	×	×	✓
vec	Latn	Indo-European	×	×	×	✓
ven	Latn	Atlantic-congo	✓	×	×	×
vie	Latn	Austroasiatic	✓	×	✓(R1)	✓
wlv	Latn	Mataguayan	✓	×	×	×
vmw	Latn	Atlantic-Congo	✓	✓	×	✓
war	Latn	Aystronesian	✓	×	✓(R2)	✓
wol	Latn	Atlantic-Congo	✓	✓	✓(R2)	✓
wsg	Deva	Dravidian	×	✓	×	×
wuu	Hans	Sino-Tibetan	✓	×	×	✓
xho	Latn	Atlantic-Congo	✓	×	✓(R2)	✓
xon	Latn	Atlantic-congo	×	✓	×	×
xsr	Deva	Sino-Tibetan	×	✓	×	×
xuu	Latn	Khoe-Kwadi	✓	×	×	×
yao	Latn	Atlantic-Congo	×	✓	×	×
ybb	Latn	Atlantic-Congo	×	✓	×	×
ydd	Hebr	Indo-European	✓	×	×	✓
ydg	Arab	Indo-European	✓	×	×	×
yor	Latn	Atlantic-Congo	✓	✓	✓(R1, R2)	×
yua	Latn	Mayan	✓	✓	×	×
yue	Hant	Sino-Tibetan	✓	×	×	×
zai	Latn	Otomanguean	✓	×	×	×
zgh	Tfng	Afro-Asiatic	×	×	×	✓
zsm	Latn	Austronesian	✓	×	✓(R1)	✓
zne	Latn	Atlantic-Congo	✓	✓	×	×
zul	Latn	Atlantic-Congo	✓	×	✓(R2)	✓

B Additional evaluation results

Resource level	XX-En high	XX-En mid	XX-En low	XX-En v. low	XX-En total	En-YY high	En-YY mid	En-YY low	En-YY v. low	En-YY total	All total
OMT-NLLB	65.0	58.3	54.6	52.3	57.3	57.6	46.0	41.0	34.6	44.6	50.9
OMT-LLaMA 8B	64.6	56.4	50.6	48.3	54.6	57.4	46.5	39.8	33.8	44.2	49.4
NLLB-200	63.5	55.3	50.9	49.3	54.2	57.2	44.3	39.4	32.5	43.0	48.6
Gemma 3	64.2	52.3	46.4	48.1	51.4	58.6	40.1	31.5	26.4	38.1	44.8
TranslateGemma	62.5	51.2	45.5	47.5	50.4	55.4	39.5	30.5	27.5	37.2	43.8
GPT-OSS 120B	62.5	49.5	42.9	44.7	48.5	58.2	40.9	31.7	29.2	38.8	43.6
LLaMA 3 70B	64.1	49.8	44.8	45.7	49.5	56.7	37.9	31.1	26.3	36.8	43.2
MADLAD	63.1	45.4	43.8	44.3	47.1	57.2	34.1	31.4	29.3	35.7	41.4
GPT-OSS 20B	61.2	47.1	39.9	40.7	45.8	54.8	33.9	25.0	24.1	32.6	39.2
Aya-101	58.2	48.5	42.1	43.5	47.1	48.5	32.3	22.9	20.3	29.9	38.5
Tower+	65.0	45.8	43.1	45.8	47.4	54.5	27.3	22.6	22.9	28.8	38.1
LLaMA 3 8B	61.0	44.1	38.9	39.7	44.1	51.4	27.5	21.7	20.6	27.9	36.0
Tiny Aya Global	61.3	39.9	34.6	35.6	40.4	53.6	27.6	19.5	19.7	27.4	33.9
Hunyuan-MT	57.3	38.1	35.5	37.7	39.7	45.6	22.2	20.5	23.4	24.9	32.3
Aya Expanse	59.0	34.8	34.5	38.2	38.3	48.0	22.3	20.1	21.8	24.9	31.6
EuroLLM	62.5	33.7	34.2	38.2	38.2	55.7	16.6	18.8	22.3	23.2	30.7

Table 10: Translation performance on FLoRes+ (ChrF++) by the non-English language resource level.