luso-mt: Compact Machine Translation for English, Chichewa and Tumbuka

Multilingual transfer learning and knowledge distillation for on-device translation

Byamasu Patrick Paul
Rexplore Research Labs, Lilongwe, Malawi patrick@rexplore.ai

Abstract

Machine translation for Chichewa and Tumbuka is constrained by scarce parallel data, few models licensed for commercial use and the computational cost of existing multilingual systems. We present luso-mt, a family of compact English–Chichewa and English–Tumbuka translation models built for on-device use. We adapt MADLAD-400 3B to Malawian data and use sequence-level knowledge distillation to train two encoder-decoder students from scratch: luso-mt-tiny (19.9M parameters, 40.5 MB in int8) and luso-mt-small (87.7M, 153.5 MB). On held-out NTREX-128 and SMOL test sets, luso-mt-small reaches 47.6, 46.9, 35.8 and 30.0 chrF++ (en→ny, ny→en, en→tum, tum→en), significantly above the comparably sized OPUS-MT models in every direction they support, and the two students are the only compact models we evaluated that translate into Tumbuka. luso-mt-tiny keeps 84–93% of its teacher's chrF++ at 1/148 of its size, and int8 quantisation changes chrF++ by at most 0.3. A larger model, luso-mt-3b, exceeds NLLB-200 distilled 600M in three of four directions. Across 18 student runs with three seeds, a control matched in training rows per epoch shows that most of the gain from distillation comes from additional training exposure; the teacher's translations add a further 1.5 chrF++ to luso-mt-small and a non-significant 0.7 to luso-mt-tiny. We release all three models under Apache-2.0 with all benchmark outputs; the training corpus is not released.

Technical report · November 2026

Code Models Benchmark

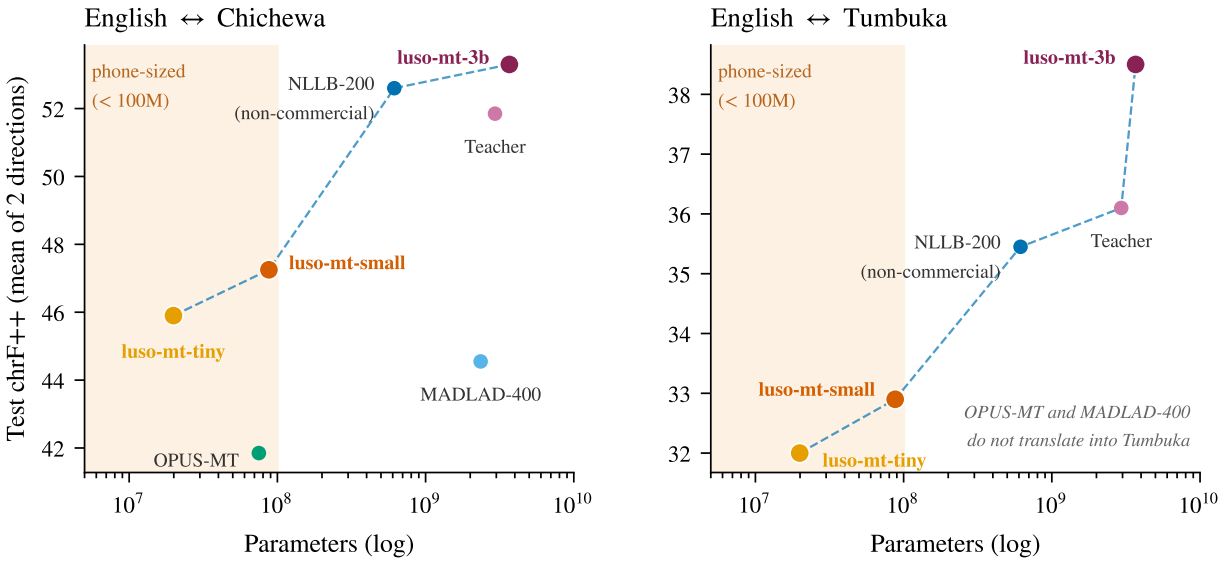


Figure 1. Translation quality against model size on the held-out test sets (chrF++, mean of the two directions of each language pair). The shaded region marks models small enough for phones (under 100M parameters), and the dashed line connects the best score at each size. Among the evaluated models, luso-mt-tiny and luso-mt-small score highest under 100M parameters for both language pairs and are the only compact models that translate into Tumbuka; luso-mt-3b scores highest overall. NLLB-200 is licensed for non-commercial use only.

1 Introduction

Chichewa (ISO 639-3 `nya`) is the national language of Malawi and is spoken as a first or second language by most Malawians; closely related varieties are spoken as Nyanja in Zambia, Mozambique and Zimbabwe. Tumbuka (`tum`) is the main language of Malawi's Northern Region and is also spoken in eastern Zambia. Both are Bantu languages with noun-class systems and agglutinative verbs: the single Chichewa word *sindinamuone* ("I did not see him") carries subject, tense, negation and object. Most public, health and agricultural information, and most of the web, is written in English.

Among the open translation models we evaluated (Section 4.1), those licensed for commercial use reach 41–43 chrF++ for Chichewa and 22–24 for Tumbuka into English, and none translates into Tumbuka. NLLB-200 [1] is stronger but is licensed for non-commercial use only, and cloud services require a network connection and charge per request.

This report studies how much translation quality can be obtained from models small enough for on-device use. We set three design constraints: the smallest model should fit in 50 MB and the default model in 200 MB after quantisation; one model should serve all four directions; and every component should be releasable under a permissive licence. Within these constraints we ask:

RQ1 How do compact multilingual models compare with existing models of similar size for English–Chichewa and English–Tumbuka translation?

RQ2 What trade-offs exist between translation quality, model size and inference efficiency when compressing a large multilingual translation model for these languages?

RQ3 How do sequence-level knowledge distillation, training exposure and student capacity affect translation quality in this setting?

We fine-tune MADLAD-400 3B [2] on Malawian data to obtain a teacher, translate the training sources with it, and train deep-encoder, shallow-decoder students from scratch on the original pairs and the teacher's translations [3]. The contributions are:

1. **luso-mt-tiny** (19.9M parameters) and **luso-mt-small** (87.7M), released under Apache-2.0, which translate in all four directions with one network each, including into Tumbuka.
2. **luso-mt-3b**, the teacher continued on a larger corpus and released under Apache-2.0, which has higher chrF++ than NLLB-200 distilled 600M in three of four directions.
3. A three-seed ablation of distillation and student size with a control matched in training rows per epoch, which attributes most of the apparent distillation gain to additional training exposure, and a documented failed training run.
4. An evaluation with confidence intervals, paired significance tests, metric signatures, tokeniser fertility, language identification, training-to-development gaps and error analysis, with every system's outputs released.

2 Related work

Multilingual translation A single encoder-decoder with a shared vocabulary can serve many directions when a token at the start of the input names the target language [4]. Scaled to hundreds of languages, this approach produced NLLB-200 [1, 5] and the translation models of MADLAD-400 [2]. Both cover Chichewa and NLLB-200 also covers Tumbuka, but NLLB-200 is released for non-commercial use only, while MADLAD-400 is released under Apache-2.0 without a Tumbuka target. OPUS-MT [6] offers small single-pair models, including English–Chichewa and Tumbuka–English.

African languages Masakhane built translation benchmarks for more than 30 African languages through community participation [7], and MAFAND-MT showed that a few thousand high-quality pairs suffice to adapt a large pretrained model to a new African language [8]. The WMT 2022 shared task on African languages included English–Chichewa [9], and web-mined Chichewa–English corpora have been released

[10]. We found no published system dedicated to Tumbuka; it appears only inside large multilingual resources. InkubaLM [11] is a compact model for African languages but covers neither Chichewa nor Tumbuka.

Compact translation Knowledge distillation trains a small student to imitate a larger teacher [12]; its sequence-level form trains the student on the teacher's beam-search outputs and substantially improves students much smaller than their teachers [3]. The Bergamot project used it to build CPU students for on-device translation [13, 14]. Because the decoder runs once per output token, a deep encoder with a shallow decoder decodes substantially faster than a balanced model with little loss in quality [15]. We use both techniques; our teacher must first be adapted to the target languages.

Synthetic data Back-translation creates pairs by translating target-language text into the source language [16], and tagging the synthetic sources lets the model distinguish them from natural input [17]. Most of our English text is produced this way. SMOL [18] provides professionally translated sentences and documents for many low-resource languages, including Chichewa and Tumbuka.

Evaluation practice A meta-evaluation of 769 MT papers found that most report BLEU alone and few test significance [19], and a large comparison with human judgements found chrF to agree with human system rankings more often than BLEU [20]. We therefore report chrF++ as the primary metric, test differences with paired bootstrap resampling [21], publish metric signatures [22] and release all system outputs.

3 Data and methodology

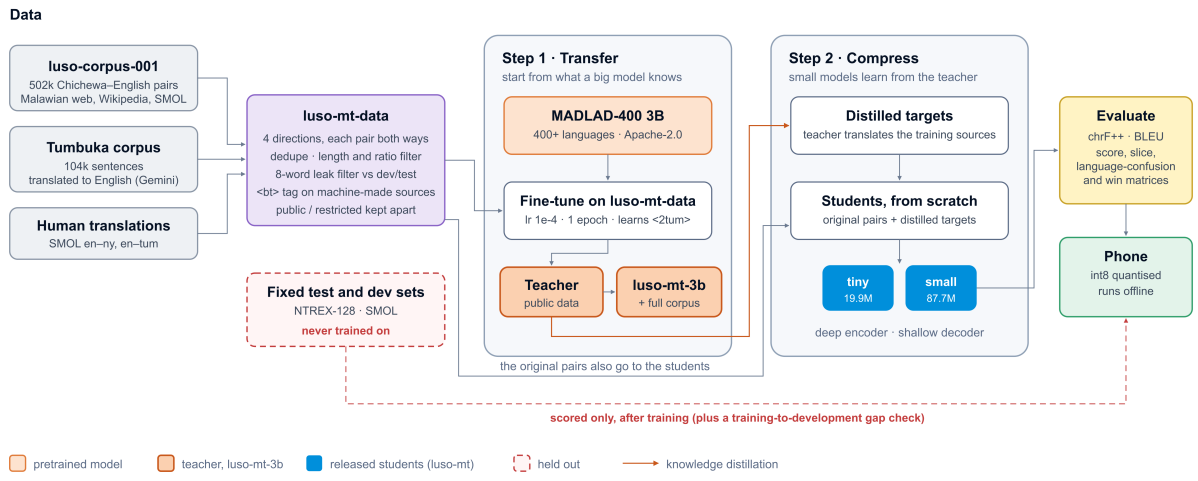


Figure 2. Training pipeline. Step 1 fine-tunes MADLAD-400 3B into a teacher, which is continued on the full corpus to give luso-mt-3b. Step 2 trains the students on the original pairs and the teacher's translations. Held-out sets are used only for scoring.

3.1 Data

Most sentence pairs start from text written in Chichewa or Tumbuka, with the English side produced by machine translation (Table 1), so the Chichewa and Tumbuka side is natural text; the synthetic English is marked with a `<bt>` tag when it is used as a source (Section 3.2). Professionally translated pairs come from SMOL [18]. The corpus has two partitions. The public-source partition is drawn from openly distributed sources (Table 2) and is used for luso-mt-tiny, luso-mt-small and their teacher; the training corpus itself is not released. A restricted partition holds text whose terms do not permit redistribution; it is used only to train luso-mt-3b.

Table 1. Training data after cleaning, in sentence pairs (each pair is used in both directions). Word counts are whitespace tokens.

Source	Partition	English side	Pairs	English words	ny/tum words
Chichewa web text (FineWeb-2, HPLT, Wikipedia)	public-source	Gemini 3.8 Flash	85,063	1,903,573	1,463,772
Chichewa web text, long passages	public-source	Gemma 3 27B	1,692	149,793	122,778
SMOL sentences and documents (en-ny)	public-source	professional	3,686	63,166	53,464
Tumbuka Wikipedia and web text	public-source	Gemini 3.8 Flash	50,938	901,164	921,212
SMOL sentences and documents (en-tum)	public-source	professional	1,597	23,867	20,490
Public-source total			**142,976**	**3,041,563**	**2,581,716**
Additional Chichewa text	restricted	Gemini 3.8 Flash	403,327	7,522,007	5,903,308
Additional Tumbuka text	restricted	Gemini 3.8 Flash	51,057	837,718	697,429
All data			**597,360**	**11,401,288**	**9,182,453**

Table 2. Data provenance. We list the terms under which each source is distributed; this is not a legal review.

Source	Use	Terms
FineWeb-2 (`nya_Latn`, `tum_Latn`)	training text	ODC-By 1.0 for the collection; text remains subject to the terms of the original sites
HPLT 2.0 (Chichewa)	training text	HPLT dataset terms; text remains subject to the terms of the original sites
Wikipedia (Chichewa, Tumbuka)	training text	CC-BY-SA
SMOL (en-ny, en-tum)	training, dev and test	CC-BY-4.0
NTREX-128 (Chichewa)	dev and test only	CC-BY-SA-4.0
Gemini 3.8 Flash outputs	English side	generated through the provider's API under its terms of service
Gemma 3 27B outputs	English side (1,692 pairs)	from an existing release of Gemma 3 27B translations; Gemma terms of use
Restricted partition	luso-mt-3b only	terms do not permit redistribution

Synthetic English The English side of the web text was generated with Gemini 3.8 Flash in November 2026. Before use, it scored 37.0 (ny→en) and 43.9 (tum→en) chrF++, the two directions used for generation, on a 100-sentence SMOL sample per language pair that is separate from our test set. The prompt is given in Appendix B. Sentences were sent in batches with a fixed system prompt and temperature 0.1, and responses were accepted only if they returned exactly one translation per input. All machine-translated pairs were then checked by human reviewers. The 1,692 long passages translated with Gemma 3 27B were taken from an existing release rather than generated by us.

Cleaning Every sentence was identified with GlotLID [23] and, for Chichewa, with MillID; text below a confidence of 0.7 and 0.5 respectively was removed, and Tumbuka text had to score at least 0.5 for `tum_Latn`. Zambian and Mozambican varieties were removed with lexical and orthographic markers. Wikipedia bot stubs were capped at three per template. Pairs were deduplicated after normalisation and kept only with 2–150 words per side and a length ratio between 1:3 and 3:1. Chichewa sentences average 15.3 words and Tumbuka sentences 15.8; Chichewa uses 0.78 words per English word and Tumbuka 0.93.

Held-out sets and contamination control Test and development sets were fixed before training (Table 3). For Chichewa we use NTREX-128 [24], which is excluded from training entirely. For Tumbuka, which NTREX-128 does not cover, we use SMOL `smolsent` en-tum (863 sentences): after a seeded shuffle, the first 500 form the test set and the next 300 the development set. Any training pair sharing an exact match or

an 8-word window with a held-out sentence, on either side, was removed (1,544 pairs); sentences shorter than eight words are compared as a whole. A check after training found no exact match and no shared 8-word window between any of the 1,194,720 training rows, public-source or restricted, and the held-out sets. This does not rule out paraphrased overlap or contamination inherited from the pretrained teacher.

Table 3. Held-out sets; the same sentences serve both directions of a pair.

Set	Directions	Source	Sentences
test	en↔ny	NTREX-128, sentences 0–499 (news)	500 per direction
test	en↔tum	SMOL smolsent en-tum, seeded shuffle, first 500	500 per direction
dev	en↔ny	NTREX-128, sentences 500–999	500 per direction
dev	en↔tum	SMOL smolsent en-tum, next 300	300 per direction

3.2 Tokenisation

Each student uses one unigram SentencePiece vocabulary [25] shared by the three languages: 16,000 pieces for luso-mt-tiny and 32,000 for luso-mt-small, with character coverage 0.9999. The vocabulary is trained for each run on up to 3 million sentences of that run's training data, so distilled and non-distilled students have slightly different vocabularies. Four reserved symbols are added: the target tags `<2en>`, `<2ny>`, `<2tum>` and the `<bt>` tag for machine-made sources [17]. Fertility, the average number of subword pieces per word [26], determines sequence length and therefore memory and decoding time (Table 4). With an eighth of the vocabulary, the luso-mt-small tokeniser splits Chichewa and Tumbuka less than the 256,000-piece tokenisers of NLLB-200 and MADLAD-400 (1.53 and 1.65 against 1.79–2.02 tokens per word), consistent with the tokenisation premium of general-purpose tokenisers for low-resource languages [27]. The OPUS-MT en-ny tokeniser is included for reference only: it is the English source-side tokeniser of that model and was not designed for Chichewa or Tumbuka text.

Table 4. Tokeniser fertility (subword tokens per word) on the test sets (17,295 English, 9,409 Chichewa and 7,115 Tumbuka words) and share of words split into two or more pieces. Lower is better.

Tokeniser (vocabulary)	English	Chichewa	Tumbuka	Split ny	Split tum
luso-mt-tiny (16k)	1.56	1.69	1.81	41%	45%
luso-mt-small (32k)	1.41	1.53	1.65	34%	40%
NLLB-200 (256k)	1.44	1.79	1.86	49%	51%
MADLAD-400 (256k)	1.48	1.87	2.02	55%	60%
OPUS-MT en-ny (source)	1.32	3.39	3.54	87%	86%

3.3 Model

The students are encoder-decoder Transformers [28] in the T5 v1.1 form: pre-norm RMSNorm [29], gated-GELU feed-forward layers [30] and relative position biases with logarithmic buckets [31]. Unlike T5 v1.1, input and output embeddings are tied [32], which saves 5.1M (tiny) and 16.4M (small) parameters. A target tag selects the output language, so one file serves all four directions [4], and most depth is placed in the encoder, which runs once per sentence, rather than in the decoder, which runs once per output token [15]. For a 20-token output, luso-mt-small executes $12 + 4 \times 20 = 92$ layer passes against 252 for a balanced 12 + 12 model; layer passes are not equivalent to measured computation or latency. These architectural choices are adopted from prior work and were not ablated. Figure 3 shows the network and Table 5 the sizes.

3.4 Teacher and luso-mt-3b

MADLAD-400 3B is released under Apache-2.0 but writes Chichewa poorly and has no Tumbuka tag (Section 4.1). We add a `<2tum>` token and fine-tune the full model on all four directions of the public-

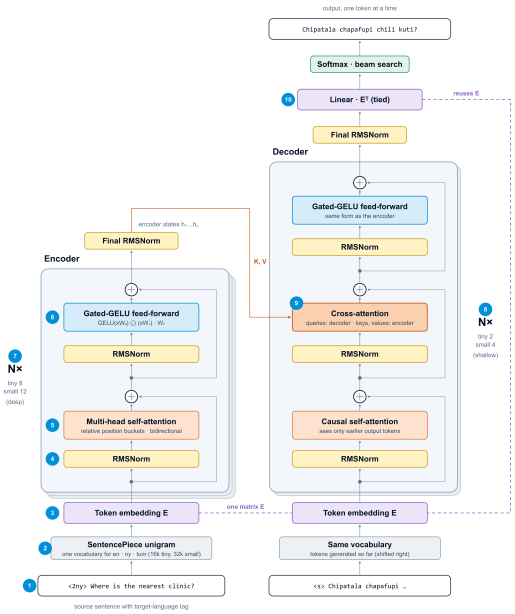


Figure 3. luso-mt student architecture: (1) target tag, (2) shared vocabulary, (3) tied embedding, (4) pre-norm RMSNorm, (5) self-attention with relative position buckets, (6) gated-GELU feed-forward, (7) deep encoder, (8) shallow decoder, (9) cross-attention, (10) tied output projection. The encoder runs once per sentence and the decoder once per output token, so most layers are placed in the encoder.

Table 5. Student sizes. Parameters include the tied embedding; attention heads have dimension 64.

Model	d_{model}	Heads	d_{ff}	Encoder layers	Decoder layers	Vocabulary	Parameters
luso-mt-tiny	320	5	1,024	8	2	16,000	19.9M
luso-mt-small	512	8	2,048	12	4	32,000	87.7M

source partition with label-smoothed cross-entropy,

$$\mathcal{L}_{\text{CE}}(x, y; \theta) = -\sum_{t} \left[(1-\varepsilon) \log p_\theta(y_t \mid y_{<t}, x) + \frac{\varepsilon}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \log p_\theta(v \mid y_{<t}, x) \right], \quad \varepsilon = 0.1, \tag{1}$$

repeating Tumbuka pairs three times and human-translated pairs twice. luso-mt-3b continues this teacher for one epoch on the full corpus (Table 6, Algorithm 2 in Appendix A); development loss kept falling when the larger corpus was added (Figure 4).

Table 6. Teacher and luso-mt-3b settings (Adafactor, cosine schedule, 3% warm-up, weight decay 0.01, label smoothing 0.1, bf16, 192 tokens per side, one NVIDIA H200).

	Teacher	luso-mt-3b
Initialisation	`google/madlad400-3b-mt` (2.94B)	the teacher
Training data	public-source partition	full corpus
Tumbuka repetition	3×	2×
Batch (sentences)	32×2 accumulation	32×4 accumulation
Learning rate	10^{-4}	10^{-4}
Optimiser steps (epochs)	7,917 (1)	11,000 (1)
Wall clock	2.2 h training; 6.9 h with distillation	5.5 h including evaluation

3.5 Distillation into students

Sequence-level distillation trains the student on the teacher's most likely translation rather than matching its token distributions [3, 12]. The teacher translates every public-source training source once with beam size 4, giving $\hat{y} \approx \arg\max_y q(y \mid x)$; of 285,952 translations, 815 whose length ratio to the reference falls outside 0.5–2 are discarded. Each student is trained from random initialisation on the original pairs $\mathcal{D}$ and

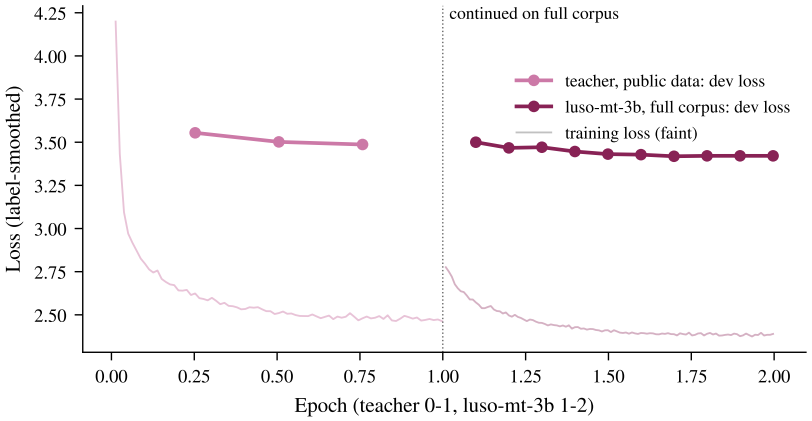


Figure 4. Training and development loss of the teacher (epoch 0–1, public-source partition) and of luso-mt-3b (epoch 1–2, continued on the full corpus), on the same development set.

the distilled pairs $\mathcal{K}$:

$$\mathcal{L}(\theta) = \sum_{(x,y)\in\mathcal{D}} \mathcal{L}_{\mathrm{CE}}(x, y; \theta) + \sum_{(x,\hat{y})\in\mathcal{K}} \mathcal{L}_{\mathrm{CE}}(x, \hat{y}; \theta). \tag{2}$$

During training, segmentations are sampled with subword regularisation [33] from the $\ell = 64$ best candidates with probability proportional to $P(\mathrm{s})^{\alpha}$, $\alpha = 0.5$, so most samples equal the standard segmentation. Table 7 lists the settings and Algorithm 3 the procedure.

Table 7. Student training settings. Students without distillation use the same settings with an upper bound of 20 epochs and no teacher translations; the matched-exposure control is described in Section 4.3.

Setting	Value
Optimiser	AdamW, learning rate 5×10^{-4}, inverse square root schedule, 4,000 warm-up steps, weight decay 0.01, gradient clipping 1.0
Batch and precision	256 sentence pairs, bf16; 192 tokens per side
Regularisation	dropout 0.1; label smoothing 0.1; subword sampling ($\alpha = 0.5$, 64 best segmentations)
Data balance	Tumbuka pairs repeated 3 times; every pair used in both directions
Stopping	up to 30 epochs; development chrF++ every 2,000 steps (greedy); patience 5 evaluations
Reference runs (1 × H200)	luso-mt-tiny: 29.4 epochs, 2.8 h; luso-mt-small: 27.4 epochs, 6.1 h; both ended by early stopping

4 Experimental setup

4.1 Baselines

We evaluated the open checkpoints we found that list Chichewa or Tumbuka (Table 8), all with the same pipeline; Hy-MT, PLaMo translate, TranslateGemma and SMaLL-100 cover neither language. Recent Gemma-based translation checkpoints reported to cover Chichewa and Tumbuka were not evaluated. MADLAD-400 reads Chichewa well (51.3 ny→en) but writes it poorly (37.8 en→ny), often with long, repetitive output, which motivated fine-tuning it as a teacher.

Table 8. Existing open models on our test sets (chrF++ / BLEU, beam 4). *n/s*: direction not supported.

Model	Params	en→ny	ny→en	en→tum	tum→en	Licence
NLLB-200 distilled	615M	50.0 / 20.1	55.2 / 34.2	38.2 / 13.0	32.7 / 12.0	CC-BY-NC-4.0
MADLAD-400 3B	2.9B	37.8 / 11.5	51.3 / 30.7	*n/s*	26.8 / 5.9	Apache-2.0
OPUS-MT en-ny / ny-en / tum-en	75–76M	41.0 / 12.0	42.7 / 18.6	*n/s*	24.1 / 5.4	CC-BY-4.0
OPUS-MT en-mul / mul-en	77–78M	41.4 / 13.0	41.8 / 18.9	*n/s*	22.1 / 3.4	CC-BY-4.0

4.2 Evaluation protocol

All systems translate the test sets of Section 3.1 with beam size 4. The primary metric is chrF++ [34], an F-score over character n-grams ($n \leq 6$) and word n-grams ($n \leq 2$), $\mathrm{chrF}_\beta = (1+\beta^2)PR/(\beta^2 P + R)$ with $\beta = 2$, which emphasises recall over precision. Character n-grams credit a correct stem with a wrong affix, which suits agglutinative languages, and chrF agrees with human system rankings more often than BLEU [20]. BLEU [35] is reported as a secondary metric. Both are computed with sacreBLEU 2.6.0 [22]; signatures are in Appendix D. Differences between systems are tested with paired bootstrap resampling over test sentences [21] with 1,000 resamples (Algorithm 1), which gives 95% confidence intervals and p-values. These intervals capture variation over test sentences, not over training runs; Section 4.3 addresses the latter. Language identity of outputs is measured with GlotLID.

4.3 Training runs and seeds

Each student is trained in three conditions: without distillation (20-epoch limit), with distillation (30-epoch limit), and a matched-exposure control that repeats the original pairs twice under the 30-epoch limit. The control sees as many rows per epoch as a distilled run (992,184 against 990,729 after Tumbuka up-sampling) and therefore the same number of optimiser steps per epoch, but no teacher translations. Early stopping applies to all conditions, so the steps actually taken differ between runs (Appendix C). Every condition is run with seeds 13, 14 and 15 (18 runs); seed 13 gives the released models. Conditions are compared with Welch's t-test over the three seed means. Runs are configured with Hydra and OmegaConf and logged in Weights & Biases; the configurations are released with the code.

5 Results

5.1 Main results

Table 9 gives test results for all systems. luso-mt-small is significantly better than OPUS-MT in the three directions both support (+6.6 en→ny, +4.2 ny→en, +5.9 tum→en), and the two students are the only compact models in the comparison that translate into Tumbuka. luso-mt-tiny, with 4.4 times fewer parameters, is 0.9–1.4 chrF++ behind luso-mt-small. NLLB-200, seven times larger than luso-mt-small, leads it by 2.4 chrF++ into Chichewa and into Tumbuka, 2.7 out of Tumbuka and 8.3 out of Chichewa. luso-mt-small reaches 86–96% of its teacher's chrF++ and luso-mt-tiny 84–93%.

5.2 Teacher and luso-mt-3b

One epoch of fine-tuning raised MADLAD-400 from 37.8 to 49.9 chrF++ into Chichewa, from 51.3 to 53.8 out of Chichewa and from 26.8 to 34.7 out of Tumbuka; the new `<2tum>` tag reached 37.5 into Tumbuka. Continuing on the full corpus (luso-mt-3b) improved every direction by a further 1.1–3.2 points ($p = 0.001$). luso-mt-3b has higher chrF++ than NLLB-200 distilled 600M into Chichewa (+1.7, $p = 0.002$), into Tumbuka (+2.5) and out of Tumbuka (+3.6, both $p = 0.001$); out of Chichewa it is 0.3 lower ($p = 0.19$). Its BLEU into Tumbuka is lower than NLLB-200's (12.6 against 13.0), and it is almost five times larger (Figure 5).

5.3 Distillation, training exposure and student size

On the seed-13 runs, adding the teacher's translations raises test chrF++ in every direction, by 2.9–4.3 for luso-mt-tiny and 3.5–5.1 for luso-mt-small (paired bootstrap, $p = 0.001$ in each direction). Because the distilled runs also see twice as many rows per epoch under a higher epoch limit, this comparison confounds the teacher's translations with training exposure. Table 10 and Figure 6 report all 18 runs, and Table 11 separates the two effects.

Exposure accounts for most of the gain Averaged over directions and seeds, distillation raises chrF++ by 3.3 (luso-mt-tiny) and 4.7 (luso-mt-small) over the 20-epoch baseline. The control recovers 2.6 and 3.2 of these points (79% and 68%), leaving 0.7 and 1.5 attributable to the teacher's translations. For luso-mt-small the remainder is consistent: every distilled seed scores above every control seed in all four directions

Table 9. Test results: chrF++ with 95% bootstrap confidence interval / BLEU, 500 sentences per direction, beam size 4. Bold: best in each group. Every difference between luso-mt-small and any other system is significant (paired bootstrap, 1,000 resamples, $p \leq 0.007$). luso-mt-3b is significantly better than NLLB-200 in en→ny, en→tum and tum→en ($p \leq 0.002$); the difference in ny→en is not significant ($p = 0.19$). *n/s*: not supported.

Model	Params	en→ny	ny→en	en→tum	tum→en	Licence
Phone-sized (under 100M parameters)						
luso-mt-small	88M	**47.6** ±1.2 / 20.1	**46.9** ±1.2 / 23.6	**35.8** ±1.0 / 8.2	**30.0** ±0.9 / 8.8	Apache-2.0
luso-mt-tiny	20M	46.3 ±1.2 / 19.1	45.5 ±1.2 / 22.0	34.9 ±1.0 / 8.1	29.1 ±0.9 / 7.9	Apache-2.0
OPUS-MT (per pair)	75M	41.0 ±1.0 / 12.0	42.7 ±1.1 / 18.6	*n/s*	24.1 ±0.8 / 5.4	CC-BY-4.0
Large						
luso-mt-3b	2.9B	**51.7** ±1.3 / 23.1	54.9 ±1.3 / 32.9	**40.7** ±1.3 / 12.6	**36.3** ±1.1 / 14.2	Apache-2.0
Teacher (public data)	2.9B	49.9 ±1.3 / 21.9	53.8 ±1.2 / 31.4	37.5 ±1.3 / 9.7	34.7 ±1.1 / 12.7	not released
NLLB-200 distilled	615M	50.0 ±1.4 / 20.1	**55.2** ±1.3 / 34.2	38.2 ±1.3 / 13.0	32.7 ±1.1 / 12.0	CC-BY-NC-4.0
MADLAD-400 3B	2.9B	37.8 ±1.5 / 11.5	51.3 ±1.2 / 30.7	*n/s*	26.8 ±1.1 / 5.9	Apache-2.0

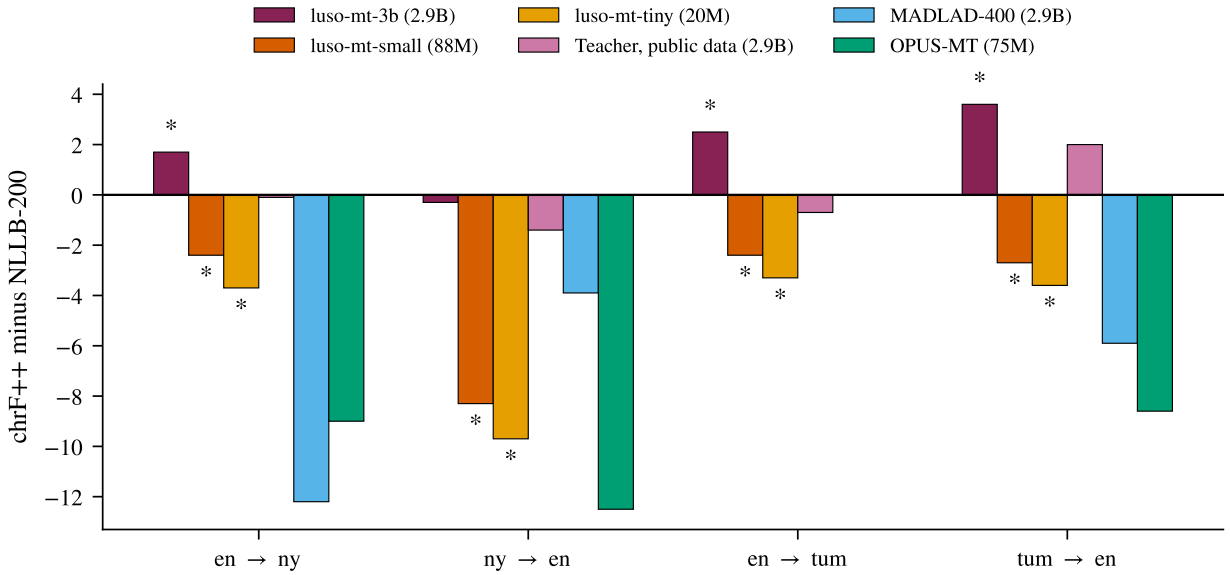


Figure 5. Test chrF++ of each system minus that of NLLB-200. Asterisks mark luso-mt models whose difference from NLLB-200 is significant (paired bootstrap, $p < 0.05$).

Table 10. Students across seeds: test chrF++, mean ± standard deviation over seeds 13–15. The matched-exposure control repeats the original pairs twice so that it sees as many rows per epoch as the distilled run.

Training condition	Seeds	en→ny	ny→en	en→tum	tum→en	Mean
tiny, no distillation (20 epochs)	3	41.9 ±0.1	42.6 ±0.1	30.9 ±0.2	26.3 ±0.2	35.4 ±0.1
tiny, no distillation, pairs ×2 (30 epochs)	3	45.1 ±0.5	44.8 ±0.5	33.9 ±0.3	28.2 ±0.3	38.0 ±0.3
tiny, distilled (30 epochs)	3	46.1 ±0.2	45.3 ±0.2	34.6 ±0.3	28.7 ±0.4	38.7 ±0.3
small, no distillation (20 epochs)	3	41.7 ±0.8	41.5 ±0.4	31.3 ±0.6	26.1 ±0.3	35.2 ±0.5
small, no distillation, pairs ×2 (30 epochs)	3	45.6 ±0.9	44.9 ±0.6	34.2 ±0.4	28.6 ±0.4	38.3 ±0.5
small, distilled (30 epochs)	3	47.4 ±0.3	46.5 ±0.6	35.5 ±0.2	29.8 ±0.2	39.8 ±0.3

($p = 0.02$). For luso-mt-tiny it is not significant ($p = 0.053$), although each distilled seed has a higher mean than each control seed. Without distillation the two student sizes perform alike (difference −0.3 at 20 epochs, +0.3 in the control); with distillation luso-mt-small leads by 1.1.

What the control matches The 20-epoch baselines reached their best development score at the last or second-to-last evaluation (36,000–38,000 of 38,760 steps), so the epoch limit ended them while they were still improving. The controls and distilled runs were ended by early stopping, not by the 30-epoch limit,

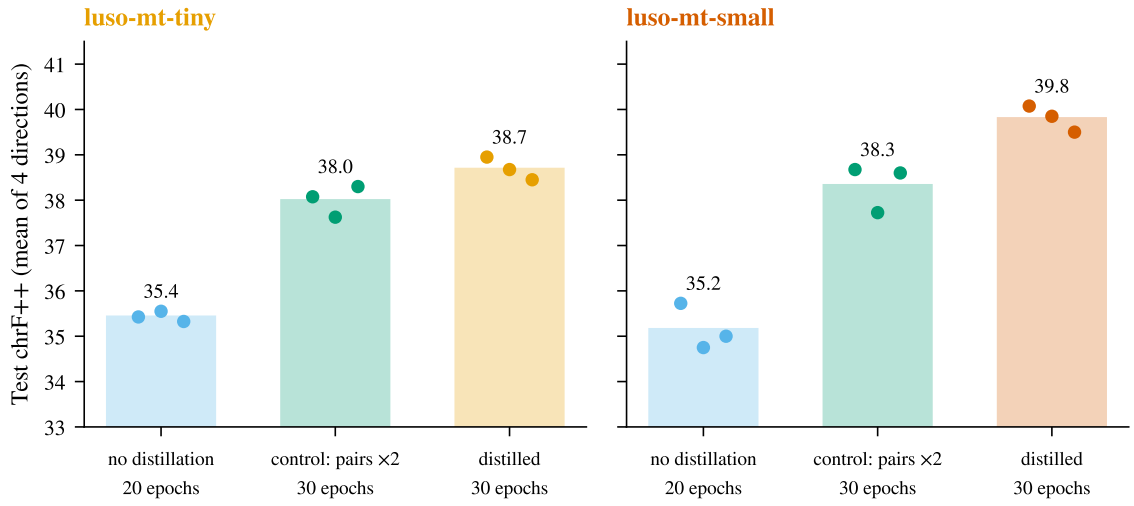


Figure 6. All 18 student runs: test chrF++ (mean of four directions) for each seed (dots) and condition (bars).

Table 11. Effect of each change on test chrF++ (mean of four directions), as a difference of means over three seeds with a Welch 95% confidence interval.

Change	luso-mt-tiny	luso-mt-small
Teacher's translations (distilled − control)	+0.7 [0.0, +1.4]	+1.5 [+0.4, +2.6]
Training exposure (control − 20-epoch baseline)	+2.6 [+1.8, +3.3]	+3.2 [+2.0, +4.3]
Total (distilled − 20-epoch baseline)	+3.3 [+2.7, +3.8]	+4.7 [+3.6, +5.7]
Student size (luso-mt-small − luso-mt-tiny)		
no distillation, 20 epochs	−0.3 [−1.5, +0.9]	
control, 30 epochs	+0.3 [−0.7, +1.4]	
distilled, 30 epochs	+1.1 [+0.5, +1.7]	

but at different points: the selected control checkpoints come from 40,000–86,000 optimiser steps and the distilled ones from 76,000–104,000 (Table 18). The control therefore matches the data seen per epoch and the training budget, not the number of updates. The controls stopped because their development score had not improved for five evaluations, so the remaining gap is not explained by a budget limit; whether the teacher's translations help by improving the targets or by allowing useful training to continue longer cannot be separated with this design.

Variation across seeds Standard deviations across seeds are 0.1–0.9 chrF++ per direction and at most 0.5 for the mean over directions. The released seed-13 students score 0.3 chrF++ above the means of their conditions, within one standard deviation, so their scores in Table 9 are slightly favourable relative to a typical run.

5.4 Training dynamics and a failed first run

The first students used subword sampling with $\alpha = 0.1$ over all segmentations and a learning rate of 7×10^{-4}. luso-mt-tiny stopped at epoch 10.3 with development chrF++ of 9.2 (en→ny) and 11.2 (en→tum); its outputs were fluent but unrelated to the source (*Malawi a Malawi a Malawi …*). Decoded training batches showed words cut into near-character fragments that changed at every pass:

```
normal:   _<2ny> _Where _is _the _near est _clinic ?
sampled:  _<2ny> _Where _is _ t he _ ne a rest _ c l in i c ?
```

Training loss on this text (4.6) stayed far below development loss on normally segmented text (7.2). The corrected runs changed several settings at once (sampling with $\alpha = 0.5$ over the 64 best segmentations, learning rate 5×10^{-4}, more epochs and re-tying the embedding after checkpoint loading), so the failure is most plausibly explained by the segmentation but the cause is not isolated. Figure 7 shows the failed and

corrected runs.

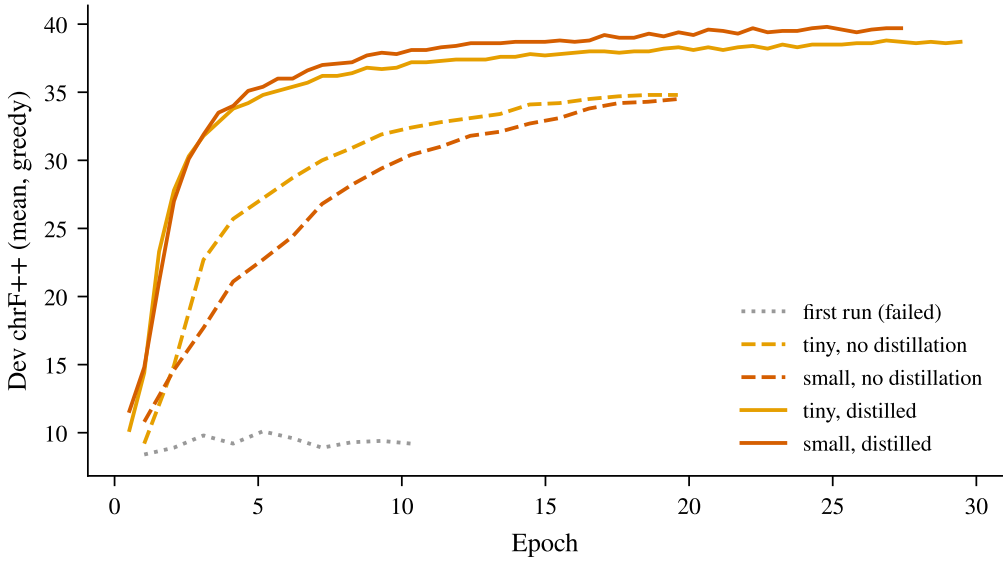


Figure 7. Development chrF++ during training (mean over directions, greedy decoding). Distilled students (solid) reach higher scores earlier than students trained without distillation (dashed); the failed run (dotted) never learnt to use the source.

5.5 Analysis

Sentence length All systems score lower on short sentences (Figure 8). Into Chichewa, the gap between luso-mt-small and NLLB-200 shrinks from 8.1 points on short sentences to 2.0 on long ones; out of Chichewa it remains large at every length.

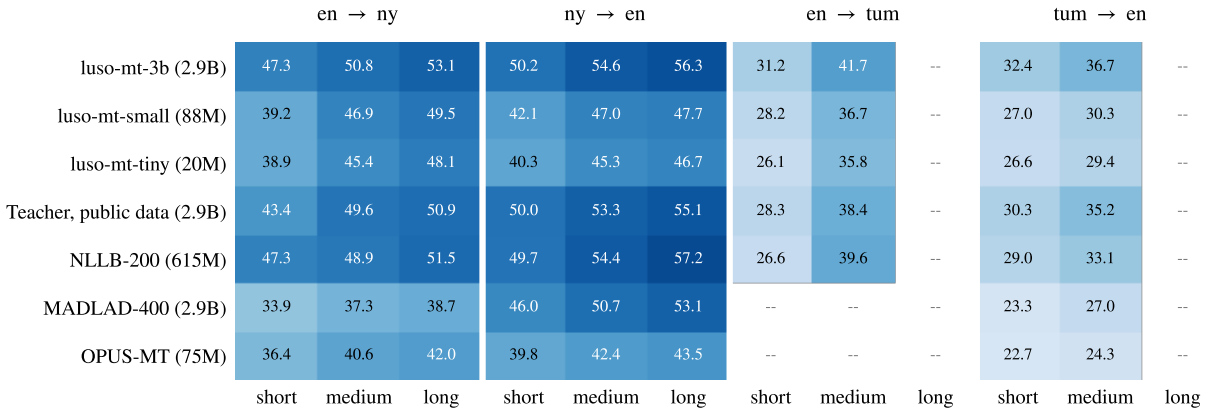


Figure 8. chrF++ by source length (short: under 10 words; long: over 25 words). A dash marks slices with fewer than 20 sentences.

Language identity All luso-mt models produce the requested language in 98–100% of outputs according to GlotLID (Figure 9); asked for Tumbuka, luso-mt-small never answered in Chichewa and luso-mt-3b did so in 0.4% of cases, against 2% for NLLB-200. Producing the right language does not imply a correct translation.

Sentence-level comparison Figure 10 gives, for every pair of systems, the share of test sentences on which one obtains a higher sentence-level chrF++ than the other. luso-mt-small scores higher than OPUS-MT on 72% of sentences and higher than NLLB-200 on 36%; luso-mt-3b scores higher than NLLB-200 on 56.5%. A higher sentence-level score is not the same as a better translation, which only human evaluation can establish (Section 8).

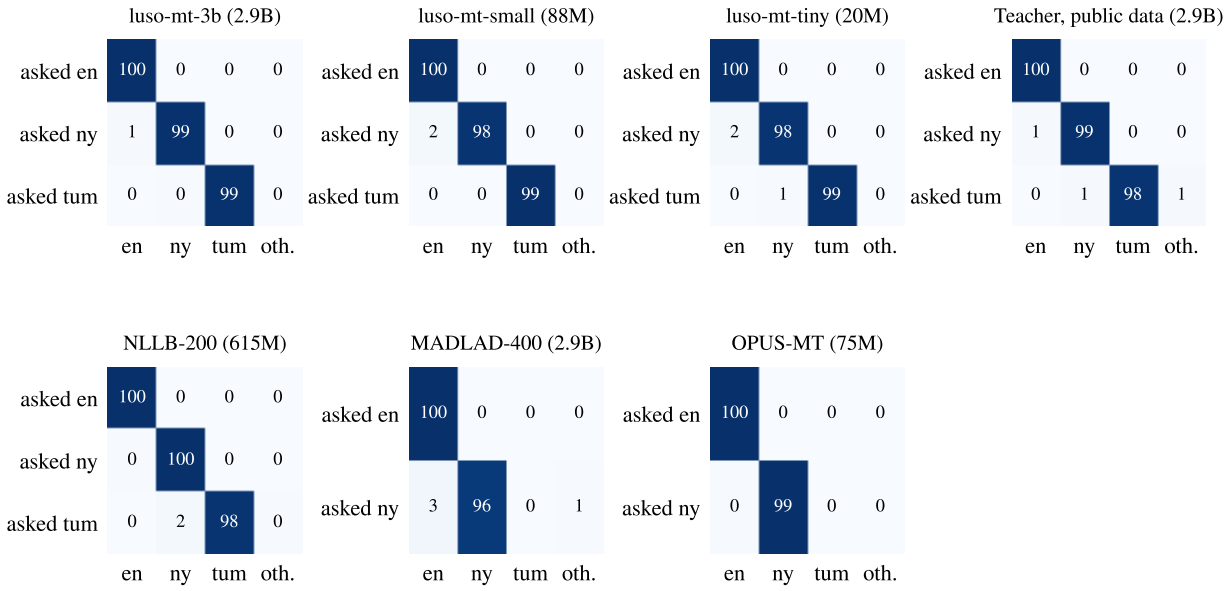


Figure 9. Language of the output (columns, GlotLID) for each requested language (rows), in percent of test outputs.

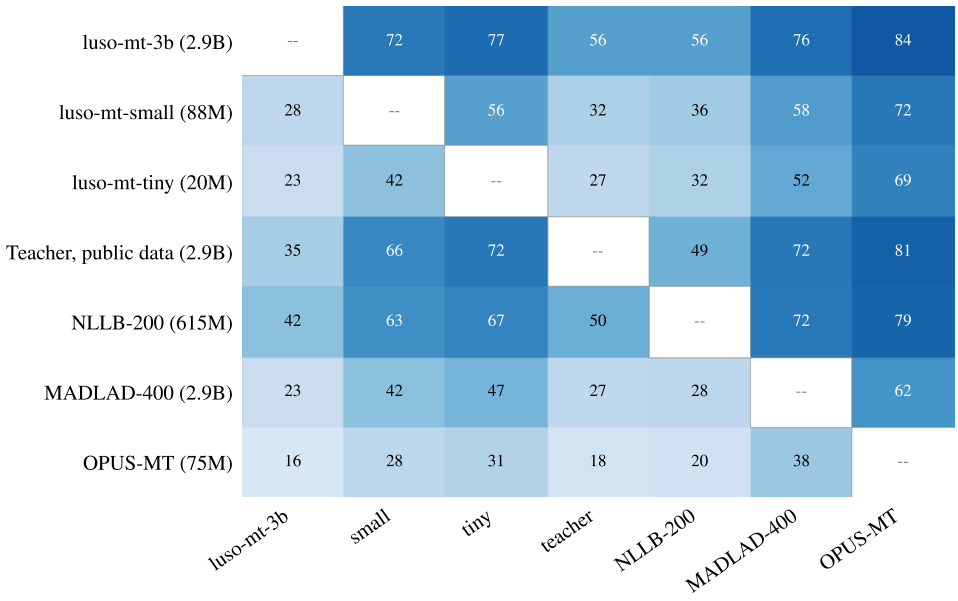


Figure 10. Share of test sentences (pooled over the directions both systems support) on which the row system's sentence-level chrF++ is higher than the column system's.

Training-to-development gap Table 12 compares chrF++ on training pairs with development chrF++: 500 human-translated training pairs per direction for the students, and 300 randomly sampled training pairs per direction, mostly with machine-translated targets, for luso-mt-3b. The gap is small for Chichewa and larger for Tumbuka, whose pairs are repeated in training, reaching 21.2 points for luso-mt-small tum→en and 23–35 points for luso-mt-3b. Such gaps can reflect overfitting, but also differences between training and development distributions; for luso-mt-3b the probe is drawn from synthetic-target pairs, so the two gaps are not directly comparable and neither establishes memorisation by itself.

Errors None of luso-mt-small's 2,000 test outputs copies its source or repeats itself heavily. Its errors are lexical and semantic, and some are serious (Table 13): a unit changed from inches to millimetres without changing the number, and the colloquial word for diabetes rendered as music. Such errors produce fluent output in the right language and are only partly visible in automatic metrics. Conversely, some low scores reflect the reference rather than the output: of the 200 outputs into English that score below 25 chrF++ on sources of 6–20 words, 174 are from Tumbuka, where several references are free translations.

Table 12. Training-to-development chrF++ gap per direction (students: 500 human-translated training pairs per direction; luso-mt-3b: 300 random training pairs per direction). Values above 15 in bold.

Model	en→ny	ny→en	en→tum	tum→en
luso-mt-small	6.4	9.6	14.8	**21.2**
luso-mt-tiny	4.1	4.1	11.4	12.7
small, no distillation	5.2	4.3	**18.0**	**15.9**
tiny, no distillation	3.5	1.3	10.1	9.9
luso-mt-3b	**22.9**	**26.0**	**28.7**	**35.3**

Table 13. luso-mt-small outputs on test sentences, with the professional reference.

Dir.	Source	Reference	luso-mt-small	Observation
ny→en	Amuna atatu, azaka za 33, 34 ndi 39, amangidwa galimoto itagundidwa mobwerezabwereza m'nyumba ina ku Londonderry.	Three men, aged 33, 34 and 39, have been arrested after a car was repeatedly rammed into a house in Londonderry.	Three men, aged 33, 34, and 39, have been arrested after a car was repeatedly attacked in a house in Londonderry.	Names and numbers correct; one weaker verb (*attacked*).
ny→en	Rosa idzabweretsa mvula yokwana mainchesi 10 m'malo ena a Mexico Lolemba.	Rosa will bring up to 10 inches of rain in parts of Mexico on Monday.	Rosa will bring up to 10 millimeters of rain in some parts of Mexico on Monday.	Unit error: *mainchesi* (inches) becomes millimetres.
ny→en	Madzulo ano abwenzi omwe amakhulupirira kuti akudziwa mayiyo afika pamalowa …	This afternoon friends who believe they know the woman have arrived at the scene …	Nowadays, friends who believe they know the woman have arrived at the scene …	Time expression (*madzulo ano*, this afternoon).
tum→en	Mtundu wachiwili wa suga wachuluka chomene mu vyaka ivi …	Type 2 diabetes has become so common in recent decades …	The second type of music is very abundant in these years …	Colloquial *suga* (diabetes) rendered as music.
tum→en	Longorani kuti muli na ndalama zakukwana zakugwiriskira ntchito paulendo.	Demonstrate possession of sufficient financial resources to cover travel expenses.	Make sure you have enough money to use on a trip.	Adequate meaning, different register (sentence chrF++ 11.7).
tum→en	Dr. Carter wakutemwa kusewera basketball, street hockey, flag football, golf na kuchimbira.	Dr. Carter enjoys playing basketball, street hockey, flag football, golf and running.	Dr. Carter enjoys playing basketball, street hockey, flag football, golf, and running.	Near-perfect, including borrowed sports names.

6 On-device evaluation

This section reports file size, server CPU speed and quality after quantisation; phone measurements are not yet available.

File size and server CPU speed Table 14 reports file size and throughput on an 8-vCPU server instance, translating 40 news sentences into Chichewa one at a time in a single run. Int8 refers to PyTorch dynamic quantisation of the linear layers; the embedding matrix stays in fp32, so both files can shrink further. Both students meet their storage budgets (40.5 MB against 50 MB and 153.5 MB against 200 MB). On this server, int8 weights raise throughput by 44% for luso-mt-tiny and by 146% for luso-mt-small, and beam search with four hypotheses runs at about the speed of greedy decoding.

Table 14. File size and server CPU throughput, in sentences per second (8 vCPUs, batch size 1).

		File size		Sentences per second		
Model	Parameters	fp32	int8	fp32, greedy	fp32, beam 4	int8, greedy
luso-mt-tiny	19.9M	79.5 MB	40.5 MB	8.32	8.21	11.96
luso-mt-small	87.7M	350.9 MB	153.5 MB	2.96	3.11	7.28

Quality after quantisation File size alone does not show whether translation quality survives quantisation. Table 15 compares the fp32 and int8 students on the full test sets.

Int8 quantisation changes chrF++ by at most 0.3 points in any direction for either student, so the quality reported for the fp32 models carries over to the quantised files. The fp32 scores computed on CPU differ from the GPU scores in Table 9 by at most 0.1 point, owing to batched decoding on a different device.

Table 15. Test chrF++ of the students in fp32 and after int8 dynamic quantisation (beam 4, CPU).

Model	en→ny	ny→en	en→tum	tum→en
luso-mt-tiny, fp32	46.3	45.5	34.8	29.0
luso-mt-tiny, int8	46.1	45.5	34.5	29.0
change	−0.2	0.0	−0.3	0.0
luso-mt-small, fp32	47.7	47.0	35.9	30.0
luso-mt-small, int8	47.7	46.8	35.6	30.0
change	0.0	−0.2	−0.3	0.0

Export for Android For the luso8 Android application, both students were exported to ONNX. The fp32 ONNX models reproduce the PyTorch outputs exactly on 200 test sentences per model, and int8 ONNX quantisation changes chrF++ by -0.1 to $+0.2$ relative to the fp32 ONNX models on the full test sets. In an initial manual test the int8 luso-mt-tiny translated offline inside the application; no latency, memory or battery measurements on phones have been made yet (Section 8).

7 Discussion

Compression Most of the teacher's automatic score survives a reduction to 1/148 of its parameters, and a four-fold larger student adds only about one chrF++ point. Under the size budgets of this work luso-mt-tiny is the more efficient model; whether the difference between the two students matters to readers, and how either behaves on phones, has not yet been measured.

Translation into English The students lose most relative to their teacher when translating into English: luso-mt-small is 6.9 chrF++ below the teacher out of Chichewa and 4.7 out of Tumbuka, against 2.3 into Chichewa and 1.7 into Tumbuka. In the public-source partition 95% of the English targets are Gemini translations, whereas the Chichewa and Tumbuka targets are natural text. A plausible explanation is that a small student trained from scratch fits synthetic English less well than a teacher pretrained on large amounts of English; this has not been tested.

Teacher quality luso-mt-3b, trained on more data, scores 1.1–3.2 points above the teacher used for distillation. Whether a stronger teacher yields stronger students is untested; testing it requires distilling the same students from luso-mt-3b.

Distillation and exposure The single-seed ablation suggested gains of up to 5.1 chrF++ from distillation; the control attributes most of this to longer training on more rows, because the 20-epoch baseline was still improving when it stopped. The remainder is consistent with Kim and Rush [3], who attribute the benefit of sequence-level distillation to targets that are easier for a small model to fit. That it is larger for the larger student suggests that luso-mt-small can exploit the teacher's outputs where luso-mt-tiny is closer to its capacity; this was not tested directly. Distillation gains reported without an exposure control can overstate the contribution of the teacher.

Tumbuka Tumbuka scores are lower than Chichewa scores. Contributing factors include less training data, a test set of everyday sentences whose references are sometimes free translations, and larger training-to-development gaps. The two test sets differ in domain, so scores are not comparable across the two languages.

Cost The training experiments used 70.6 GPU-hours, about US$351 at list price, of which the seed and control runs account for 39.7 hours (Appendix C). This excludes synthetic data generation, data collection, storage and human evaluation.

8 Conclusion and future work

RQ1. On the selected test sets, luso-mt-small is significantly better than the per-pair OPUS-MT models, the evaluated models closest in size, in every direction both support, and the two students are the only compact models evaluated that translate into Tumbuka. **RQ2.** luso-mt-tiny keeps 84–93% of the teacher's chrF++ at 1/148 of its size and 40.5 MB in int8, and int8 quantisation costs at most 0.3 chrF++. The gap to large models is largest from Chichewa into English (8.3 chrF++ below NLLB-200 for luso-mt-small). Inference speed has been measured on a server CPU only. **RQ3.** Over three seeds, most of the gain from adding teacher translations comes from additional training exposure; at matched rows per epoch the teacher's translations add 1.5 chrF++ to luso-mt-small, consistently across seeds, and 0.7 to luso-mt-tiny, which is not significant. Student size matters little without distillation and adds about one point with it.

Future work These conclusions rest on automatic metrics and server measurements. Work under way will test and extend them:

- **Human evaluation.** A blind evaluation by bilingual evaluators, collected through a Model Lab in Corpus Studio and the luso8 Android application.
- **Benchmarks on phones.** Loading time, memory, latency, on-device quality and battery use on a low-end and a mid-range Android device.
- **More data.** Human Chichewa and Tumbuka translations from Corpus Studio, extending both the training data and the test sets.
- **Model improvements.** Longer training budgets, since exposure accounted for most of the distillation gain, and students distilled from luso-mt-3b, which would test whether teacher quality bounds student quality.

We invite feedback on the released models and this report.

9 Limitations and ethics

9.1 Limitations

- **Automatic metrics only.** chrF++ and BLEU differences are not human quality judgements; human evaluation is under way (Section 8).
- **Small, single-reference test sets** of 500 sentences per direction, from different domains for the two languages.
- **Seeds.** Student conditions use three seeds, which estimate run-to-run variation only roughly; the teacher and luso-mt-3b were trained once.
- **Exposure matched per epoch, not per update.** Early stopping ended control and distilled runs after different numbers of optimiser steps (Appendix C), and each run trains its own vocabulary on its own training data.
- **Synthetic English.** Most English training text was produced by Gemini and Gemma and then checked by human reviewers; it may still carry the style of those systems, and quality into English remains the weakest.
- **Training-to-development gaps** are large for Tumbuka and for luso-mt-3b.
- **Architecture not ablated.** The design choices of Section 3.3 follow prior work.
- **No systematic phone measurements** (Section 6).

9.2 Threats to validity

Checkpoints were selected on development sets with greedy decoding; test sets were never used for selection. Baselines were decoded with up to 256 new tokens and the students with up to 192, and a few outputs reach these limits. Language identity relies on GlotLID, whose error rate on short outputs we did

not measure. The overlap filter does not detect paraphrases or contamination inherited from the pretrained teacher, and the test domains differ from the health and agriculture text where we expect the models to be used.

9.3 Open gaps

Beyond SMOL, we are not aware of a Tumbuka MT test set, nor of public Chichewa or Tumbuka test sets for health, agriculture, education or government text. To our knowledge, no learned metric has been validated against human judgements in either language, and no evaluation set covers regional varieties of Nyanja or Tumbuka spelling variation.

9.4 Responsible use, data and release

The models are intended for general translation, assistive reading and writing, and as components of local applications. They are not suited to unreviewed medical, legal or safety-critical translation; names, numbers, units and specialised terms should be checked. Releasing a model under Apache-2.0 does not by itself establish that every training source and synthetic output is cleared for every commercial use; Table 2 lists the terms of each source. All three models are released under Apache-2.0 as weights only; the training corpus, including the restricted partition used for luso-mt-3b, is not released. NTREX-128 (CC-BY-SA-4.0) and SMOL (CC-BY-4.0) are used under their licences, and the released benchmark follows CC-BY-SA-4.0.

Table 16. Released artefacts.

Artefact	Location	Licence
luso-mt-small	https://huggingface.co/suru8-ai/luso-mt-small	Apache-2.0
luso-mt-tiny	https://huggingface.co/suru8-ai/luso-mt-tiny	Apache-2.0
luso-mt-3b	https://huggingface.co/suru8-ai/luso-mt-3b	Apache-2.0
Benchmark: all system outputs	https://huggingface.co/datasets/suru8-ai/luso-mt-benchmark	CC-BY-SA-4.0
Code	https://github.com/rexplore-ai/luso-mt	Apache-2.0

Acknowledgements

We thank the contributors who record and annotate Chichewa and Tumbuka with us in Corpus Studio, and the authors of the datasets, models and tools this work builds on, in particular SMOL, NTREX-128, FineWeb-2, HPLT, MADLAD-400, GlotLID, MillID, SentencePiece and sacreBLEU.

References

[1] NLLB Team, "No language left behind: Scaling human-centered machine translation," *arXiv preprint arXiv:2207.04672*, 2022. [Online]. Available: https://arxiv.org/abs/2207.04672

[2] S. Kudugunta, I. Caswell, B. Zhang, X. Garcia, D. Xin, A. Kusupati, R. Stella, A. Bapna, and O. Firat, "MADLAD-400: A multilingual and document-level large audited dataset," in *Advances in Neural Information Processing Systems 36, Datasets and Benchmarks Track*, 2023, pp. 67 284–67 296. [Online]. Available: https://arxiv.org/abs/2309.04662

[3] Y. Kim and A. M. Rush, "Sequence-level knowledge distillation," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 1317–1327. [Online]. Available: https://doi.org/10.18653/v1/D16-1139

[4] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, M. Hughes, and J. Dean, "Google's multilingual neural machine translation system:

Enabling zero-shot translation," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 339–351, 2017. [Online]. Available: https://doi.org/10.1162/tacl_a_00065

[5] NLLB Team, "Scaling neural machine translation to 200 languages," *Nature*, vol. 630, pp. 841–846, 2024. [Online]. Available: https://doi.org/10.1038/s41586-024-07335-x

[6] J. Tiedemann and S. Thottingal, "OPUS-MT – building open translation services for the world," in *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, 2020, pp. 479–480. [Online]. Available: https://aclanthology.org/2020.eamt-1.61

[7] W. Nekoto, V. Marivate, T. Matsila *et al.*, "Participatory research for low-resourced machine translation: A case study in African languages," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 2144–2160. [Online]. Available: https://doi.org/10.18653/v1/2020.findings-emnlp.195

[8] D. I. Adelani, J. O. Alabi, A. Fan *et al.*, "A few thousand translations go a long way! leveraging pre-trained models for African news translation," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022, pp. 3053–3070. [Online]. Available: https://doi.org/10.18653/v1/2022.naacl-main.223

[9] D. I. Adelani, M. M. I. Alam, A. Anastasopoulos *et al.*, "Findings of the WMT'22 shared task on large-scale machine translation evaluation for African languages," in *Proceedings of the Seventh Conference on Machine Translation*, 2022, pp. 773–800. [Online]. Available: https://doi.org/10.18653/v1/2022.wmt-1.72

[10] P. Vegi, S. J, B. Paul, A. Mishra, P. Banjare, P. K R, and C. Viswanathan, "WebCrawl African: A multilingual parallel corpora for African languages," in *Proceedings of the Seventh Conference on Machine Translation*, 2022, pp. 1076–1089. [Online]. Available: https://doi.org/10.18653/v1/2022.wmt-1.105

[11] A. L. Tonja, B. F. P. Dossou, J. Ojo, J. Rajab, F. Thior, E. P. Wairagala, A. Aremu, P. Moiloa, J. Abbott, V. Marivate, and B. Rosman, "InkubaLM: A small language model for low-resource African languages," *arXiv preprint arXiv:2408.17024*, 2024. [Online]. Available: https://arxiv.org/abs/2408.17024

[12] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015. [Online]. Available: https://arxiv.org/abs/1503.02531

[13] N. Bogoychev, R. Grundkiewicz, A. F. Aji, M. Behnke, K. Heafield, S. Kashyap, E.-I. Farsarakis, and M. Chudyk, "Edinburgh's submissions to the 2020 machine translation efficiency task," in *Proceedings of the Fourth Workshop on Neural Generation and Translation*, 2020, pp. 218–224. [Online]. Available: https://doi.org/10.18653/v1/2020.ngt-1.26

[14] N. Bogoychev, J. Van der Linde, and K. Heafield, "TranslateLocally: Blazing-fast translation running on the local CPU," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2021, pp. 168–174. [Online]. Available: https://doi.org/10.18653/v1/2021.emnlp-demo.20

[15] J. Kasai, N. Pappas, H. Peng, J. Cross, and N. A. Smith, "Deep encoder, shallow decoder: Reevaluating non-autoregressive machine translation," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://arxiv.org/abs/2006.10369

[16] R. Sennrich, B. Haddow, and A. Birch, "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)*, 2016, pp. 86–96. [Online]. Available: https://doi.org/10.18653/v1/P16-1009

[17] I. Caswell, C. Chelba, and D. Grangier, "Tagged back-translation," in *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, 2019, pp. 53–63. [Online]. Available: https://doi.org/10.18653/v1/W19-5206

[18] I. Caswell, E. Nielsen, J. Luo, C. Cherry *et al.*, "SMOL: Professionally translated parallel data for 115 under-represented languages," in *Proceedings of the Tenth Conference on Machine Translation*, 2025, pp. 1103–1123. [Online]. Available: https://doi.org/10.18653/v1/2025.wmt-1.85

[19] B. Marie, A. Fujita, and R. Rubino, "Scientific credibility of machine translation research: A meta-evaluation of 769 papers," 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)*, 2021, pp. 7297–7306. [Online]. Available: https://doi.org/10.18653/v1/2021.acl-long.566

[20] T. Kocmi, C. Federmann, R. Grundkiewicz, M. Junczys-Dowmunt, H. Matsushita, and A. Menezes, "To ship or not to ship: An extensive evaluation of automatic metrics for machine translation," in *Proceedings of the Sixth Conference on Machine Translation*, 2021, pp. 478–494. [Online]. Available: https://aclanthology.org/2021.wmt-1.57

[21] P. Koehn, "Statistical significance tests for machine translation evaluation," in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 2004, pp. 388–395. [Online]. Available: https://aclanthology.org/W04-3250

[22] M. Post, "A call for clarity in reporting BLEU scores," in *Proceedings of the Third Conference on Machine Translation: Research Papers*, 2018, pp. 186–191. [Online]. Available: https://doi.org/10.18653/v1/W18-6319

[23] A. H. Kargaran, A. Imani, F. Yvon, and H. Schuetze, "GlotLID: Language identification for low-resource languages," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 6155–6218. [Online]. Available: https://doi.org/10.18653/v1/2023.findings-emnlp.410

[24] C. Federmann, T. Kocmi, and Y. Xin, "NTREX-128 – news test references for MT evaluation of 128 languages," in *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, 2022, pp. 21–24. [Online]. Available: https://doi.org/10.18653/v1/2022.sumeval-1.4

[25] T. Kudo and J. Richardson, "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*, 2018, pp. 66–71. [Online]. Available: https://doi.org/10.18653/v1/D18-2012

[26] P. Rust, J. Pfeiffer, I. Vulić, S. Ruder, and I. Gurevych, "How good is your tokenizer? on the monolingual performance of multilingual language models," 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)*, 2021, pp. 3118–3135. [Online]. Available: https://doi.org/10.18653/v1/2021.acl-long.243

[27] A. Petrov, E. La Malfa, P. H. S. Torr, and A. Bibi, "Language model tokenizers introduce unfairness between languages," in *Advances in Neural Information Processing Systems 36*, 2023. [Online]. Available: https://arxiv.org/abs/2305.15425

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30*, 2017. [Online]. Available: https://arxiv.org/abs/1706.03762

[29] B. Zhang and R. Sennrich, "Root mean square layer normalization," in *Advances in Neural Information Processing Systems 32*, 2019. [Online]. Available: https://arxiv.org/abs/1910.07467

[30] N. Shazeer, "GLU variants improve transformer," *arXiv preprint arXiv:2002.05202*, 2020. [Online]. Available: https://arxiv.org/abs/2002.05202

[31] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020. [Online]. Available: https://jmlr.org/papers/v21/20-074.html

[32] O. Press and L. Wolf, "Using the output embedding to improve language models," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 2017, pp. 157–163. [Online]. Available: https://aclanthology.org/E17-2025

[33] T. Kudo, "Subword regularization: Improving neural network translation models with multiple subword candidates," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 66–75. [Online]. Available: https://doi.org/10.18653/v1/P18-1007

[34] M. Popović, "chrF++: words helping character n-grams," in *Proceedings of the Second Conference on Machine Translation*, 2017, pp. 612–618. [Online]. Available: https://doi.org/10.18653/v1/W17-4770

[35] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002, pp. 311–318. [Online]. Available: https://doi.org/10.3115/1073083.1073135

A Algorithms

Algorithm 1 Paired bootstrap test (as implemented in sacreBLEU)

Require: outputs A, B of two systems and references R on N test sentences; $K = 1{,}000$ resamples

$\delta \leftarrow M(A, R) - M(B, R)$ ▷ observed corpus-level difference for metric M
for $k = 1, \ldots, K$ **do**
    draw N sentence indices I_k uniformly with replacement
    $a_k \leftarrow M(A_{I_k}, R_{I_k})$; $b_k \leftarrow M(B_{I_k}, R_{I_k})$
end for
report each score with the 95% interval of $\{a_k\}$ and $\{b_k\}$
$p \leftarrow$ share of resamples whose difference, centred at zero, is at least as extreme as δ

Algorithm 2 Teacher fine-tuning and continuation

Require: MADLAD-400 3B M_0; public-source pairs $\mathcal{D}_{\text{pub}}$; full corpus $\mathcal{D}_{\text{all}}$

add the token `<2tum>` to M_0 and initialise its embedding randomly
$\mathcal{D}_1 \leftarrow \mathcal{D}_{\text{pub}}$ in both directions, Tumbuka $\times 3$, human translations $\times 2$
$T \leftarrow$ train M_0 on $\mathcal{D}_1$ for one epoch (Eq. 1); keep the checkpoint with the lowest dev loss
$\mathcal{D}_2 \leftarrow \mathcal{D}_{\text{all}}$ in both directions, Tumbuka $\times 2$, human translations $\times 2$
luso-mt-3b $\leftarrow$ continue T on $\mathcal{D}_2$ for one epoch with batch 128

Algorithm 3 Student training with sequence-level distillation

Require: teacher T; public-source pairs $\mathcal{D}$ in both directions, machine-made sources tagged `<bt>`

$\mathcal{K} \leftarrow \{(x, \mathrm{BeamSearch}(T, x, 4)) : x \in \mathcal{D}\}$, keeping length ratios in $[0.5, 2]$
train a unigram tokeniser on $\mathcal{D} \cup \mathcal{K}$; initialise a student S randomly (Table 5)
repeat
    sample a batch from $\mathcal{D} \cup \mathcal{K}$ (Tumbuka $\times 3$) and resample its segmentation
    update S on the loss of Eq. 2; every 2,000 steps compute development chrF++ (greedy)
until 5 evaluations without improvement or 30 epochs
keep the best checkpoint; score the test sets (beam 4) and the training-to-development gap

B Synthetic data generation

English translations of the Chichewa and Tumbuka web text were generated with Gemini 3.8 Flash through an OpenAI-compatible API, temperature 0.1, with segments sent in batches as a JSON array and the response accepted only when it contained exactly one translation per segment. The system prompt was:

> *You are a professional Malawian translator, fluent in Chichewa (Chinyanja), Tumbuka and English. Translate faithfully and naturally. Keep names, numbers and places unchanged. Do not add explanations.*

The user message was "Translate each item from Chichewa (Malawi) [or Tumbuka (Malawi)] to English. Return ONLY a JSON array with exactly *n* elements, in order." followed by the segments.

C Compute and training length

Table 17. Compute for all experiments in this report, including runs not used in the results. Wall clock of the training process as logged in Weights & Biases (evaluation: Hugging Face job records); cost at list price, H200 US$5 and L4 US$0.80 per hour.

Experiment	Hardware	Wall clock	Cost	Outcome
Existing-model evaluation	1 × L4	0.58 h	< US$1	baselines
Teacher fine-tuning and distillation	1 × H200	6.94 h	US$35	teacher, 285,952 translations
Students without distillation	1 × H200	3.25 h	US$16	ablation
luso-mt-tiny	1 × H200	2.79 h	US$14	released
luso-mt-small	1 × H200	6.11 h	US$31	released
luso-mt-3b	1 × H200	5.45 h	US$27	released
Exploratory runs (failed first students, 209M student)	1 × H200	5.78 h	US$29	not used
Subtotal (reference runs)		**30.9 h**	**US$152**	
luso-mt-tiny, seeds 14–15 (both conditions)	1 × H200	6.26 h	US$31	Table 10
luso-mt-tiny control, seeds 13–15	1 × H200	5.87 h	US$29	Table 10
luso-mt-small, seeds 14–15 (both conditions)	1 × H200	16.10 h	US$81	Table 10
luso-mt-small control, seeds 13–15	1 × H200	11.50 h	US$58	Table 10
Total		**70.6 h**	**US$351**	

Short attempts on A10G and L4 GPUs that ran out of memory or did not start, and smoke tests, are omitted (under US$1 together). The costs exclude synthetic data generation, data collection, storage and human evaluation. Table 18 gives the training length of every student run.

Table 18. Training length of every student run (Weights & Biases logs): epochs and optimiser steps completed, the step of the selected checkpoint and its development chrF++ (greedy, mean of four directions). Batch 256; one epoch is 1,938 steps without distillation and about 3,870 for the control and distilled runs.

Condition	Seed	Epochs	Steps	Selected step	Dev chrF++
luso-mt-tiny, no distillation	13	20.0	38,760	36,000	34.8
	14	20.0	38,760	36,000	34.6
	15	20.0	38,760	36,000	35.0
luso-mt-tiny, control	13	20.1	78,000	68,000	37.3
	14	17.0	66,000	56,000	37.1
	15	24.8	96,000	86,000	37.9
luso-mt-tiny, distilled	13	29.4	114,000	104,000	38.8
	14	23.8	92,000	82,000	38.8
	15	22.2	86,000	76,000	38.3
luso-mt-small, no distillation	13	20.0	38,760	38,000	34.5
	14	20.0	38,760	38,000	33.9
	15	20.0	38,760	38,000	34.1
luso-mt-small, control	13	19.1	74,000	64,000	38.0
	14	12.9	50,000	40,000	37.3
	15	19.1	74,000	64,000	37.9
luso-mt-small, distilled	13	27.4	106,000	96,000	39.8
	14	28.9	112,000	102,000	39.8
	15	22.7	88,000	78,000	39.4

D Metric signatures

```
chrF++ nrefs:1|case:mixed|eff:yes|nc:6|nw:2|space:no|version:2.6.0
BLEU   nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.6.0
```

E Reproducibility

The code (https://github.com/rexplore-ai/luso-mt) contains the data, training and evaluation scripts, each with its dependencies declared inline, and Hydra/OmegaConf configurations for every training run, so that a seed sweep is a single command (`launch.py -m experiment=tiny_kd seed=13,14,15`). `compute_results.py`, `make_tables.py` and `make_figures.py` regenerate every table and figure from the released benchmark. Pretrained checkpoints are loaded from the Hugging Face Hub by name; the training corpus is not public, so training runs cannot be reproduced exactly from public data.

Commands The 18 student runs of Table 10 and the tables and figures of this report are produced by:

```
uv run launch.py -m experiment=tiny_base,tiny_rep2,tiny_kd seed=13,14,15
uv run launch.py -m experiment=small_base,small_rep2,small_kd seed=13,14,15 hf.timeout=8h
python paper/collect_experiments.py # gather eval.json of every run
python paper/make_tables.py && python paper/make_figures.py
```

Each run trains on one H200 through Hugging Face Jobs, logs to Weights & Biases and writes `eval.json` with test and development scores; seed 13 of the distilled and 20-epoch conditions are the released reference runs.

F Citation

```
@techreport{luso-mt-2026,
title       = {luso-mt: Compact Machine Translation for English, Chichewa and Tumbuka},
author      = {Byamasu, Patrick Paul},
institution = {Rexplore Research Labs},
year        = {2026}
}
```