

Efficient Multilingual Neural Machine Translation via Corpus-Driven Vocabulary Pruning: An English-Arabic Case Study

Ahmed Amine Aliane^{*1,3}, Nasredine Semmar², and Hassina Aliane³

¹Arabic Institute for Translation, Algiers, Algeria

²CEA LIST, Paris-Saclay, France

³CERIST, Algeria

Abstract

The adoption of large pre-trained multilingual models for neural machine translation (MNMT) faces a major challenge: excessive memory and computational consumption due to overly large vocabularies and embedding layers. Although existing compression methods like pruning, quantization and knowledge distillation reduce parameter redundancy, they mainly preserve the structure of the original vocabulary, leaving a major source of inefficiency unresolved. We propose in this paper a general optimization framework combining a vocabulary pruning method with a targeted fine-tuning protocol for MNMT models. We evaluate the proposed framework using three models (M2M100, NLLB-200, mBART-50) on the English-Arabic language pair. Our approach reduces the vocabulary size by 82-87% depending on the architecture, without any loss in performance. Results show that optimized multilingual models can match or exceed the performance of dedicated bilingual baselines. In particular, the pruned and fine-tuned M2M100 model achieves a competitive BLEU score of 42.04 (against 44.59 for the OPUS-MT-en-ar bilingual model) while it significantly outperforms it on the COMET metric (0.8730 vs 0.7911) revealing superior semantic adequacy and fluency.

1 Introduction

Large-scale Multilingual Neural Machine Translation models such as Facebook’s M2M100 and NLLB-200 have transformed cross-lingual interaction by facilitating translation among hundreds of languages through a single architecture. These models have achieved state-of-the-art performance and significantly advanced zero-shot translation capabilities. However, to achieve high performance, they rely on massive parameter counts and require very large vocabularies (often exceeding 128,000

subword tokens), resulting in prohibitive memory footprints and inference latency, although most vocabulary entries remain inactive for any given language pair. While existing compression techniques such as quantization, and weight pruning have been extensively studied, they primarily focus on reducing model parameters rather than addressing the structural inefficiency introduced by large multilingual vocabularies. Moreover, most adaptation methods preserve the original vocabulary, thereby maintaining a high memory footprint even after compression. In practice, very few studies explicitly target the vocabulary size issue in large-scale MNMT systems, leaving an important gap in efficient multilingual model design.

In this paper, we propose a novel general optimization framework combining vocabulary pruning and targeted fine-tuning for MNMT: pruning retains only tokens relevant to a specific language pair, while fine-tuning allows the model to re-learn the optimal parameters configuration within the new vocabulary space.

We evaluate the proposed framework using three state-of-the-art multilingual models (M2M100, mBART-50, NLLB-200) for English-Arabic machine translation and the dedicated model OPUS-MT-en-ar as a baseline. Furthermore, we trained our models on both formal and hybrid (formal and informal) corpora, to explore the impact of data composition on the stability of pruned models. Our contributions include: (1) a minimalist architecture-agnostic framework for MNMT that is founded on vocabulary pruning, (2) extensive validation on English-Arabic translation using M2M100, mBART-50 and NLLB-200, obtaining scores comparable to or exceeding those obtained with a dedicated bilingual model, while reducing embedding-layer parameters by $\sim 82-87\%$, (3) an analysis of the synergy between vocabulary reduction and fine-tuning, showing that hybrid strategies combining formal and informal text are crucial for

^{*}Corresponding author: aaliane781@gmail.com

maintaining model robustness after compression, outperforming models trained on single-register corpora.

2 Related work

2.1 Large-scale Multilingual Neural Machine Translation Models

Multilingual Neural Machine Translation (MNMT) builds on the NMT successes. MNMT’s key innovation rests on leveraging a single shared model capable of translating between multiple languages instead of training a specific model for each targeted language pair (Firat et al., 2016; Johnson et al., 2017). The adoption of subword tokenization (Kudo, 2018), which provided an effective means of addressing the rare word problem in neural machine translation (Luong et al., 2015), has subsequently become an essential component of multilingual models by enabling a shared vocabulary across diverse languages and writing systems. The introduction of mBART (Liu et al., 2020) demonstrated the effectiveness of multilingual denoising pretraining for sequence-to-sequence translation for 25 languages. mBART-50 (Tang et al., 2020) extended this capability to a many-to-many setting, handling 50 languages without relying on English as an interlingua. M2M100 (Fan et al., 2021) scaled multilingual translation to 100 languages via a natively trained many-to-many architecture while NLLB-200 (Costa-Jussa et al., 2022) expanded this to 200 languages, achieving state-of-the-art results across diverse language families. These gains in quality and language coverage come at the cost of prohibitive memory requirements and computational overhead.

2.2 Model compression and domain adaptation

Extensive research has explored the use of compression techniques and adaptation strategies to improve the efficiency of large-scale multilingual neural translation models. Common approaches, including quantization (Zafrir et al., 2019; Shen et al., 2020; Mohammadshahi et al., 2022; Prato et al., 2020), weight pruning (See et al., 2016; Voita et al., 2019; Behnke and Heafield, 2020; Khan et al., 2021), layer pruning (Fan et al., 2019; Peer et al., 2022; Moslem et al., 2025) and knowledge distillation (Hinton et al., 2015; Kim and Rush, 2016; Tan et al., 2019) have been successfully applied to monolingual and multilingual models.

Compressing multilingual models is not language-neutral, pruning and quantization can cause uneven degradation, particularly for low-resource languages (Mohammadshahi et al., 2022; Gaido et al., 2025; Palomino, 2026). Crucially, these methods leave embedding layers largely unchanged, which constitutes a structural bottleneck for multilingual models where shared vocabulary size often exceeds 128,000 tokens, contributing substantially to the memory footprint.

Domain adaptation via targeted fine-tuning is a well-established strategy for specializing multilingual models to specific language pairs, yet it typically preserves the full multilingual vocabulary. Vocabulary pruning, the direct reduction of the model’s vocabulary by removing tokens irrelevant to a target language pair, is a less explored avenue in MNMT literature. L’Hostis et al. (2016) explored in earlier work vocabulary selection and pruning strategies for efficient NMT, in bilingual settings, but did not address the challenges introduced by large-scale multilingual models and shared cross-lingual vocabularies. Existing work has mainly investigated vocabulary trimming (Ushio et al., 2023; Dorkin et al., 2024) and the impact of BPE vocabulary trimming on translation quality (Cognetta et al., 2024). More recently (Jiang et al., 2026) proposed re-tokenizing a pruned vocabulary with embedding relearning, improving the performance on low-resource and non-latin scripts languages. Zhang et al. (2026) proposed VocabTailor, a dynamic vocabulary selection method that filters input and language-specific tokens at inference time rather than through a corpus-driven keep-set, achieving aggressive vocabulary reduction while matching or exceeding static pruning baselines on several tasks including machine translation. However, despite these efforts, combining domain adaptation with vocabulary reduction for pretrained multilingual translation models remains underexplored. Other parameter-efficient alternatives such as (Chronopoulou et al., 2023), address a related but distinct axis of efficiency, leaving the vocabulary itself untouched.

3 Methodology

Unlike common compression methods, our framework employs a deterministic pruning strategy that preserves weight alignment for non-embedding layers while removing only tokens irrelevant to the target language pair.



Figure 1: Overview of the proposed optimization framework: corpus-driven vocabulary pruning followed by targeted fine-tuning.

Our approach is motivated by two main observations. First, conventional compression methods target attention and feed-forward weights, leaving the vocabulary-dependent layer unchanged. Thus, substantial memory overhead remains in the input embedding and output projection matrices, whose size grows linearly with the multilingual vocabulary (Fan et al., 2021). However, for language-pair specific translation tasks, only a small portion of this shared vocabulary is actively utilized while the remaining embedding parameters continue to occupy memory throughout both fine-tuning and inference. This argues for vocabulary pruning as an orthogonal direction for compressing multilingual translation models. Second, large multilingual models jointly pretrained on hundreds of languages learn language-agnostic representations that provide stronger initialization for bilingual fine-tuning than training from scratch (Arivazhagan et al., 2019), suggesting that most multilingual knowledge resides in shared parameters rather than the exhaustive joint vocabulary.

Building on these observations, we propose an optimization framework for multilingual NMT articulated around two complementary sequential phases (Figure 1):

1. A corpus-driven vocabulary pruning phase, which reshapes the model’s embedding space by retaining only tokens relevant to a given language pair, thus reducing the necessary memory without training from scratch.
2. A targeted fine-tuning protocol that adapts the pruned model to the language pair at hand, by learning an optimal new parameter configuration with the reduced vocabulary space.

This framework allows full exploitation of cross-lingual transfer inherited from massive pretraining, while overcoming hardware limitations. We validate our approach on three model families: M2M100, mBART-50 and NLLB-200, covering different model sizes and tokenization strategies.

3.1 Corpus-driven vocabulary pruning

Our approach performs structural compaction of the pretrained model’s embedding space by retaining only parameters associated with tokens observed in the target training data. Unlike approaches relying on sparse masks or internal layer modifications, we implement a five-stage pipeline that executes exact row-wise slicing operations. This process transforms the original vocabulary dimension V into a task-relevant subset $|\mathcal{K}|$, starting from corpus-driven token identification to in-place tensor updates and runtime tokenizer remapping.

Stage 1: Corpus-driven token collection. We randomly sample N sentence pairs from the training corpus (default $N = 150,000$), sized to ensure vocabulary coverage for the target language pair. Both the source and target texts are tokenized using the original pretrained tokenizer. The union of all token IDs appearing in at least one tokenized sentence is collected into a set $\mathcal{C}$:

$$\mathcal{C} = \bigcup_{i=1}^N (\text{tok}(S_i) \cup \text{tok}(T_i)),$$

where S_i and T_i denote the source and target sentences of pair i , respectively. This empirical harvest ensures that the retained vocabulary reflects actual usage patterns rather than predetermined selection thresholds. We validate this empirically in Section 4.5: across 83,000 held-out sentences that the trimmer never saw, at least 99.6% of Arabic tokens survive the pruned vocabulary, and even word types entirely unseen in the pruning corpus, including dialectal ones, decompose into retained productive subwords with at least 98.5% full coverage.

Stage 2: Mandatory token preservation. Special tokens that must survive pruning, including padding (PAD), end-of-sentence (EOS), beginning-of-sentence (BOS), and unknown (UNK)—are preserved independently of corpus frequency. Additionally, all language-tag tokens required for proper model routing (e.g. `__ar__`, `arb_Arab`, `ar_AR`) are collected into a set $\mathcal{M}$. The final keep-set $\mathcal{K}$ and its sorted ID mapping are defined as:

$$\mathcal{K} = \mathcal{C} \cup \mathcal{M}, \quad \phi : \{0, \dots, |\mathcal{K}| - 1\} \xrightarrow{\sim} \mathcal{K}.$$

This guarantees functional integrity after aggressive vocabulary reduction.

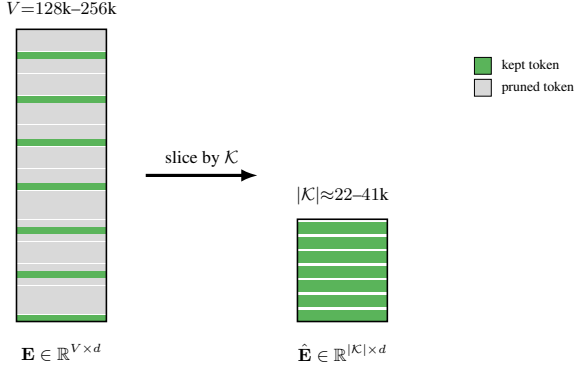


Figure 2: Row-wise slicing of the embedding matrix: only rows corresponding to the corpus-driven keep-set $\mathcal{K}$ (Stages 1–2) are retained when constructing the compacted embedding $\hat{\mathbf{E}}$ (Stages 3–4). The kept set amounts to 13–18% of the original vocabulary ($V = 128,104$ for M2M100, 250,054 for mBART-50, 256,206 for NLLB-200 600M; $|\mathcal{K}| = 22,535, 31,969$ and 40,952 respectively).

Stage 3: Weight extraction before resizing. The embedding matrix $\mathbf{E} \in \mathbb{R}^{V \times d}$ and the language-model head $\mathbf{W} \in \mathbb{R}^{V \times d}$ are indexed by $\mathcal{K}$ while they are still at their original size V , producing sliced copies

$$\hat{\mathbf{E}}, \hat{\mathbf{W}} \in \mathbb{R}^{|\mathcal{K}| \times d}.$$

Extracting weights before resizing is critical: resizing first would invalidate any index $\geq |\mathcal{K}|$, leading to incorrect weight extraction or misaligned embeddings.

Stage 4: In-place resize and weight injection. The model’s embedding matrices are resized to the new vocabulary cardinality $|\mathcal{K}|$ via `resize_token_embeddings( $|\mathcal{K}|$ )`. The extracted matrices $\hat{\mathbf{E}}$ and $\hat{\mathbf{W}}$ are then written back to their corresponding positions. Model configuration parameters (`vocab_size`, `pad_token_id`, `eos_token_id`) are updated to reflect the reduced vocabulary, and any weight tying between encoder and decoder embeddings is re-established.

Stage 5: Remapped tokenizer wrapper. Since the on-disk SentencePiece vocabulary retains its original cardinality V , we construct a lightweight `RemappedTokenizer` wrapper that intercepts the primary tokenization and detokenization entry points (encoding and decoding). The forward map ϕ^{-1} converts external token IDs to internal indices during encoding, while the reverse map ϕ performs the inverse conversion during decoding. This allows the trimmed model to be used through

the standard tokenizer interface for most inference code, though batched tensor-level remapping in performance-sensitive paths accesses the wrapper’s internal lookup tensor directly.

The resulting trimmed models retain only tokens observed in the English–Arabic corpus, reducing the embedding layer by approximately 82-87% with a corresponding decrease in GPU memory footprint.

3.2 Targeted fine-tuning protocol and adaptation operations

Vocabulary pruning alters the token-ID embedding space by removing discarded token representations. While retained embeddings preserve their local semantics, the language model head and decoder must adapt to the remapped, lower-dimensional categorical distribution. Without adaptation, this induces systemic errors: softmax probabilities become skewed by unchanged weight magnitudes on remaining tokens (logits calibration), and training dynamics shift given the altered gradient flow through smaller embedding matrices (cross-entropy convergence). To address these issues and maintain steady generation outputs, we introduce an immediate targeted fine-tuning phase within our framework. The fine-tuning protocol uses supervised learning on task-specific parallel corpora, enabling the model to re-optimize its parameter configuration within the constrained lexical space while preserving beneficial cross-lingual representations inherited from pretraining.

4 Experiments and Discussion

In this section, we validate our proposed optimization framework on the English-Arabic machine translation task. We have chosen this language pair for three key reasons:

1. Arabic morphological richness makes it a stringent test case for compression methods.
2. Both languages are well represented in multilingual pretraining corpora, enabling fair zero-shot comparisons.
3. Existing bilingual baselines provide strong reference points for evaluating pruned model performance.

Below, we detail our experimental configuration, evaluation protocols and results across three state-of-the-art multilingual architectures.

4.1 Experimental Setup

Data. Two training corpus configurations are evaluated: 200,000 parallel sentence pairs from MultiUN (Eisele and Chen, 2010), representing formal diplomatic discourse and a second configuration augmenting this with heterogeneous OPUS-100 text (Tiedemann, 2012), to expose the pruned vocabulary to higher lexical entropy and informal register. Both configurations pass through a deterministic filtering pipeline removing anomalous pairs, alignment corruptions, and over-length sequences. Evaluation uses a held-out partition of 1,500 MultiUN test segments.

Training configuration. All models are fine-tuned for 5 epochs with a linear learning rate decay (500 warm-up steps, peak $\eta = 2 \times 10^{-5}$), an effective batch size of 512 (16 samples/device $\times$ 8 gradient-accumulation steps $\times$ 4 GPUs via Hugging Face Accelerate), and FP16 mixed precision. Gradient checkpointing and an 8-bit quantized AdamW optimizer (Dettmers et al., 2022) keep training within workstation-grade memory constraints, consistent with the ~ 82 -87% embedding-layer reduction from pruning.

Baselines. We compare against three reference points:

- (a) The original multilingual checkpoints evaluated zero-shot, establishing the pre-fine-tuning ceiling for each architecture.
- (b) The trimmed-only variants of each architecture, which serves to isolate the baseline impact of vocabulary reduction before any training occurs.
- (c) *OPUS-MT-en-ar*, a dedicated bilingual model serving as a baseline trained exclusively English–Arabic.

4.2 Evaluation framework and main translation results

4.2.1 Evaluation framework

Performance is measured on a strictly held-out partition (1,500 segments from the MultiUN test set) using four complementary metrics: **BLEU** (Papineni et al., 2002), **chrF++** (Popović, 2015), **TER** (Snover et al., 2006), and **COMET** (Rei et al., 2020). Using both surface-level (BLEU, chrF++) and meaning-based (COMET) evaluations allows us to distinguish between superficial token matching and genuine translation quality, an essential

Table 1: Translation results for English $\rightarrow$ Arabic. All fine-tuned and trimmed models use the vocabulary pruning framework described in Section 3.1. COMET scores: range 0–1. TER: lower is better ($\downarrow$).

Condition	Training Data	System	BLEU	chrF++	TER $\downarrow$	COMET
FINE-TUNED	MultiUN	M2M100	42.04	58.81	51.55	0.8730
		mBART-50	39.84	57.47	52.88	0.8742
		NLLB-200 600M	37.70	55.79	55.04	0.8640
	MultiUN + OPUS-100	M2M100	39.55	57.23	53.27	0.8680
		mBART-50	36.73	55.09	55.02	0.8663
		NLLB-200 600M	35.51	54.26	56.82	0.8589
	OPUS-100	M2M100	36.98	55.34	55.25	0.8609
BASELINE	—	NLLB-200 600M	26.43	48.07	64.03	0.8418
		M2M100 418M	24.95	46.64	65.09	0.8371
		mBART-50	13.65	31.70	80.77	0.7477
BASELINE	—	NLLB-200 Distilled 600M	26.82	48.44	62.91	0.8452
		M2M100 418M (Original)	24.86	46.62	65.16	0.8372
		mBART-50 (Original)	13.67	31.44	80.98	0.7452
Comparison	Bilingual	<i>OPUS-MT-en-ar</i>	44.59	58.77	58.59	0.7911

consideration for morphologically rich targets like Arabic, where BLEU may penalize fluent but lexically divergent outputs.

4.2.2 Main Translation Results

Table 1 reports the translation performance of all evaluated configurations across four metrics. We group results into four conditions: fine-tuned models (pruned then fine-tuned), trimmed-only models, the original unmodified checkpoints as zero-shot baselines, and the dedicated bilingual OPUS-MT-en-ar model.

Fine-tuned models. The fine-tuned M2M100 trained on MultiUN achieves the highest BLEU (42.04) and chrF++ (58.81) among all systems, demonstrating that the pruning–fine-tuning pipeline successfully recovers and improves upon the original multilingual capabilities for the English–Arabic pair. The mBART-50 fine-tuned variant achieves the highest COMET score (0.8742), despite a slightly lower BLEU. The hybrid MultiUN + OPUS-100 training data consistently underperforms the MultiUN-only variant on the formal test set, which is expected given that the evaluation domain (MultiUN) matches the formal training data. However, the hybrid strategy is expected to offer better robustness on out-of-domain text, a hypothesis explored further in the error analysis below. NLLB-200 fine-tuned achieves a competitive 37.70 BLEU despite a far larger original vocabulary (256,000 tokens), suggesting pruning is effective across different model scales and tokenizer designs.

Trimmed-only models. Trimming alone without subsequent fine-tuning degrades performance sub-

stantially across all architectures. Notably, NLLB-200 trimmed (BLEU 26.43, COMET 0.8418) and M2M100 trimmed (BLEU 24.95, COMET 0.8371) both remain competitive with the zero-shot baseline (BLEU 26.82, COMET 0.8452), demonstrating that vocabulary pruning alone does not catastrophically destroy translation capability. The mBART-50 trimmed variant is the exception, collapsing to BLEU 13.65 and COMET 0.7477, which we attribute to the architecture’s stronger coupling between the decoder embedding and the language-model head, a coupling more severely disrupted by the remapping of token IDs.

Results from Table 1 confirm that the combination of vocabulary pruning and fine-tuning is essential. The pruning step reduces embedding-layer size by approximately 82-87% while the fine-tuning step recovers and improves translation quality, with the best system matching or exceeding dedicated bilingual baselines at a fraction of the memory cost. We additionally compare against a few-shot instruction-tuned LLM baseline (Aya-8B); see Appendix D.

Table 2: Out-of-domain translation results for English → Arabic on FLORES-200 devtest ($n = 1,012$ segments). All fine-tuned and trimmed models use the vocabulary pruning framework described in Section 3.1. TER: lower is better (↓). COMET not computed.

Condition	Training Data	System	BLEU	chrF++	TER ↓
FINE-TUNED	MultiUN	M2M100	24.79	44.62	68.12
		mBART-50	22.09	42.87	67.90
		NLLB-200 600M	27.83	48.25	64.11
	MultiUN + OPUS-100	M2M100	26.02	46.27	65.58
TRIMMED	—	NLLB-200 600M	29.47	49.81	60.19
		mBART-50	21.19	42.25	67.01

Out-of-domain robustness. To assess whether the pruning–fine-tuning framework generalizes beyond MultiUN formal register, we re-evaluate a subset of our strongest checkpoints on FLORES-200 devtest (Goyal et al., 2022; Costa-Jussa et al., 2022), a Wikipedia-domain benchmark with no overlap with MultiUN. Table 2 reports these results, read for relative ordering rather than direct comparison with Table 1. As an external reference, the NLLB-200 paper reports English → Arabic spBLEU scores in the low-to-mid 30s on this benchmark for their full-scale, non-pruned production model, averaged across Arabic dialect direc-

tions (Costa-Jussa et al., 2022). Our trimmed-only NLLB-200 600M checkpoint (29.47 BLEU) falls within this range despite being smaller with no domain fine-tuning, suggesting that pruning alone preserves most of general-domain competence. The mBART-50 models, our weakest systems (21.19–22.09 BLEU), sit roughly ten points below this reference, indicating greater competence loss for this architecture. The MultiUN + OPUS-100 fine-tuned M2M100 also outperforms its MultiUN-only counterpart here (26.02 vs 24.79 BLEU, 46.27 vs 44.62 chrF++), reversing the in-domain ranking and supporting our hypothesis that informal data exposure improves out-of-domain robustness at an in-domain cost. More strikingly, the trimmed-only NLLB-200 model (no fine-tuning) achieves the best FLORES-200 score overall (29.47 BLEU) surpassing its own fine-tuned counterpart (27.83 BLEU). This suggests that narrow-domain fine-tuning trades cross-domain generalization for in-domain specialization, while pruning alone better preserves pretrained multilingual representations. mBART-50 shows the opposite pattern, with fine-tuning still improving over the trimmed variant (22.09 vs 21.19 BLEU), consistent with its greater sensitivity to vocabulary remapping already noted in Table 1. We emphasize that this out-of-domain evaluation covers a subset of the configurations in Table 1, focused on the strongest MultiUN fine-tuned and trimmed-only systems; a full out-of-domain sweep across all training-data compositions and architectures is left for future work.

4.3 Confidence Intervals

Table 3 reports 95% bootstrap confidence intervals for the best fine-tuned system and the dedicated bilingual baseline. Intervals were produced using 1,000 bootstrap resamples over 1,500 test segments. Because BLEU is a non-linear corpus-level metric, bootstrap resampling systematically underestimates its true corpus-level value (Koehn, 2004). We therefore derive interval half-widths from the bootstrap distribution and re-centre them on the corpus-level point estimate. chrF++ and TER, being approximately linear at the segment level, require no such correction: their bootstrap means coincide with the point estimates to within rounding error.

The confidence intervals reveal two important findings. First, the BLEU gap between OPUS-MT-en-ar (44.59) and M2M100 (MultiUN) (42.04) is real but modest: the intervals do not overlap, con-

Table 3: 95% bootstrap confidence intervals (1,000 resamples, $n = 1,500$). Bounds are centred on the corpus-level point estimate; half-widths are taken from the bootstrap distribution. TER: lower is better ($\downarrow$).

System	BLEU [95% CI]	chrF++ [95% CI]	TER $\downarrow$ [95% CI]
M2M100 (MultiUN)	42.04 [40.87–43.22]	58.81 [57.91–59.71]	51.55 [50.30–52.80]
OPUS-MT en-ar (baseline)	44.59 [43.20–45.98]	58.77 [57.79–59.75]	58.59 [57.11–60.07]

firming a statistically significant difference on this metric. Second, the chrF++ scores are statistically indistinguishable (58.81 vs 58.77, with overlapping CIs), and M2M100 (MultiUN) substantially outperforms OPUS-MT-en-ar on TER (51.55 vs 58.59, non-overlapping intervals). Combined with the COMET advantage of M2M100 (MultiUN) (0.8730 vs 0.7911), this suggests that the bilingual model’s BLEU lead is driven primarily by surface n-gram alignment with the MultiUN reference style, while the pruned multilingual model produces semantically superior and more fluent translations as captured by the meaning-based metrics.

4.4 Segment-Level Error Analysis (M2M100 Models)

To complement the corpus-level metrics reported above, we conduct a fine-grained segment-level analysis on the two M2M100 fine-tuned variants: M2M100 (MultiUN) and M2M100 (MultiUN + OPUS-100). Both models were re-run on 500 held-out test segments and evaluated at the sentence level using chrF++, which we bucket into three quality tiers: *good* (≥ 50), *medium* (25–50), and *poor* (< 25). We additionally flag empty outputs and token-level repetition, both of which are known failure modes for Arabic NMT. Analyses comparing the two fine-tuned variants use this 500-segment subset throughout; control analyses involving only M2M100 (MultiUN) and the untrimmed baseline use the full 1,500-segment test set.

Score distribution. Table 4 summarises the quality distribution for both models. M2M100 (MultiUN) achieves a higher proportion of good segments (69.2% vs 66.4%) and an identical poor rate (3.6%) compared to the hybrid variant. The average segment chrF++ is 59.27 vs 57.56, consistent with the corpus-level gap in Table 1. Critically, neither model produces any empty or repetitive outputs across 500 segments, confirming that the vocabulary pruning procedure does not introduce pathological decoding failures.

Table 4: Segment-level quality distribution for M2M100 fine-tuned variants (500 test segments, chrF++ buckets: good ≥ 50 , medium 25–50, poor < 25).

Model	Avg chrF++	Good (%)	Medium (%)	Poor (%)	Empty	Repetitive
M2M100 (MultiUN)	59.27	69.2	27.2	3.6	0	0
M2M100 (MultiUN + OPUS)	57.56	66.4	30.0	3.6	0	0

Length bucket analysis. Table 5 breaks down performance by source sentence length. Across both models a consistent and counter-intuitive pattern emerges: *long* sentences (≥ 35 tokens) attain a lower poor rate than short sentences (≤ 15 tokens). This reflects the nature of MultiUN: long diplomatic sentences follow predictable, formula-heavy syntax well covered by training data, while short segments often consist of citations, proper nouns, or abbreviations underrepresented in the vocabulary.

Table 5: Segment performance by source length bucket. Short: ≤ 15 tokens; medium: 16–34; long: ≥ 35 .

Model	Length bucket	Count	Avg chrF++	Poor (%)
M2M100 (MultiUN)	Short (≤ 15)	125	60.97	5.6
	Medium (16–34)	234	58.57	3.8
	Long (≥ 35)	141	58.94	1.4
M2M100 (MultiUN + OPUS)	Short (≤ 15)	125	58.77	7.2
	Medium (16–34)	234	56.77	3.0
	Long (≥ 35)	141	57.80	1.4

The hybrid (MultiUN + OPUS) model shows an elevated poor rate for short segments (7.2% vs 5.6%) but a lower poor rate for medium segments (3.0% vs 3.8%), suggesting that exposure to the stylistically diverse OPUS-100 data slightly destabilises performance on the short formal fragments characteristic of the MultiUN domain while marginally improving fluency on moderately-length sentences.

Pruning versus corpus effects. To disentangle pruning from corpus properties, we evaluate the untrimmed M2M100-418M baseline (full 128k vocabulary, no fine-tuning) on the full 1,500-segment test set. It exhibits the same length-dependent pattern, and more strongly: poor rate falls from 13.9% (short) to 3.4% (long) versus 6.0% to 2.9% for the pruned and fine-tuned model. Since the untrimmed model has zero pruning exposure, this reflects contextual redundancy rather than vocabulary removal: source-side type–token ratio falls from 0.97 (short) to 0.78 (long), showing that longer sentences are lexically more repetitive and thus easier to translate correctly, independently of vocabulary size.

Rare-form parity test. We further test whether the length effect masks a bias against rare Ara-

bic morphology. Tagging test segments whose reference contains at least one word type from the bottom 5% frequency tail of the fine-tuning corpus (27% of segments overall, rising to 42% of long segments), we find that rare-form segments in the long bucket are not poorer than other long segments (3.5% vs 2.5% poor rate, z -test $p = 0.58$). Rare-form segments do show elevated poor rates in the short (14.3% vs 4.4%, $p = 0.003$) and medium (5.9% vs 1.7%, $p = 0.003$) buckets, but the untrimmed baseline shows a comparable medium-bucket penalty (7.1% vs 3.2%, $p = 0.03$), indicating that this penalty reflects inherent task difficulty (consistent with the citation and proper-noun cases identified in our qualitative error taxonomy below) rather than an artefact of vocabulary pruning. In other words, given sufficient context, rare morphology is not a distinct failure mode for either the pruned or the full-vocabulary model.

Output length ratio and head-to-head comparison. Neither model exhibits systematic over or under-generation (average ratios of 0.97 and 0.96 with under 1% of segments showing severe length mismatch). Aligning the 500 matched segments across both variants, M2M100(MultiUN) scores higher on 246 segments (49.2%) versus 174 (34.8%) for the hybrid variant, with 80 ties. The formal model’s largest wins ($\Delta\text{chrF++} > 10$ on 55 segments) concentrate on long institutional passages with direct MultiUN coverage, while the hybrid variant’s advantages (23 segments) concentrate on medium-length, informally phrased segments, consistent with OPUS-100 exposure.

Qualitative error taxonomy. Inspecting worst-performing segments reveals three recurring error types. First, *bibliographic leakage*: segments consisting of partial citations (e.g. “(3) of the commentary to article 12, p. 126”) are frequently transliterated rather than rendered in Arabic script, yielding near-zero chrF++ scores. These segments expose a gap between the multilingual formal register of MultiUN prose and the parenthetical citation style it occasionally embeds. Second, *reference mismatch*: certain references point to text that is substantially different in the Arabic side (e.g. the source discusses one topic while the reference Arabic sentence discusses a completely different subject), inflating the apparent error count for otherwise fluent translations. Third, *domain shift in short proper nouns*: place names, personal names, and institution acronyms (e.g. “Vanuatu”,

“N.P.D.”) tend to be transliterated in the hypothesis but transcribed differently in the reference, penalising chrF++ despite producing pragmatically correct output. These findings reinforce two practical recommendations: (i) complement formal-register fine-tuning data with citation- and entity-specific augmentation to close the short-segment gap, and (ii) give COMET significant weight alongside BLEU for Arabic NMT evaluation, as it is less sensitive to the reference-mismatch artefacts described above.

4.5 Pruning validity on held-out data

Section 4.4 shows that vocabulary pruning does not, by itself, explain the length-dependent and rare-form error patterns observed on the MultiUN test set. To directly address whether corpus-driven pruning introduces a bias against productive but infrequent Arabic forms, and to explicitly validate the pruning strategy on rare morphology, clitic combinations, dialectal variation, named entities and domain-specific terminology, we conduct a dedicated coverage and challenge-set audit on data that was never used for pruning or fine-tuning.

Held-out corpora. We use three held-out sources, none overlapping with pruning or fine-tuning data. The first is the FLORES-200 devtest set ($n = 1,012$, Section 4.2). The second is a purpose-built 2,000 challenge set spanning eight linguistic categories: *rare_morph* (bottom 5% frequency tail), *clitic* (stacked proclitics/enclitics, regex identified), *ner* (spaCy-tagged named entities), *domain_term* (log-odds distinctive terms vs. MultiUN), and four AraBench-sampled dialects. The third source is the full AraBench test split ($n = 80,413$), used only for the coverage audit, full construction details, including decontamination and filtering criteria are given in Appendix A.

Vocabulary coverage audit. For each corpus, we tokenize the Arabic references with the original pretrained tokenizer and measure token/type representability after pruning. Coverage is detailed further in Appendix B (Table 6). Retained-token rates reach 99.9%, 99.8% and 99.6% for FLORES, the challenge set and AraBench respectively, with $\sim 96\%$ of sentences per corpus tokenizing with zero pruned tokens. Of the 105,569 removed token IDs, all were hapax or entirely unseen in the pruning corpus, accounting for only 138 occurrences across

FLORES and challenge-set sentences combined, and 3,763 across AraBench.

Frequency-stratified bias test. Table 6 (Appendix B) bins held-out Arabic word types by pruning corpus frequency and confirms coverage is essentially flat: every bin with non-zero frequency reaches 100% coverage, and even entirely unseen word types (bin 0) are covered at 98.5–98.8%, including under the maximal distributional shift of dialectal AraBench (57,469 of 70,433 held-out types unseen by construction). This is consistent with the subword solution to the rare-word problem (Luo et al., 2015): rare and unseen word types remain representable because they decompose into retained stems, clitics and affixes, rather than being individually preserved.

Challenge-set translation results. Table 7 (Appendix C) reports BLEU, chrF++, TER and COMET for the pruned and fine-tuned M2M100 (MultiUN) model on each challenge category. The three mined formal categories (clitics, domain-term, ner), all meet or exceed the FLORES out-of-domain reference (24.79 BLEU); with COMET ≥ 0.78 , indicating no category collapses. The *rare_morph* slice is harder (16.59 BLEU) -expected given its bottom-tail composition- but since these forms remain fully representable in $\mathcal{K}$ (coverage audit above), the gap reflects content unfamiliarity rather than vocabulary removal. The dialectal slices score lowest (BLEU 4.7–8.3), consistent with the model being fine-tuned to produce MSA while the references are dialectal, the same surface-level mismatch any MSA-trained system exhibits against dialectal references. This aligns with human evaluation findings from the WMT 2025 general translation shared task (Kocmi et al., 2025) which reported persistent difficulties in producing Egyptian dialect, with several systems defaulting to MSA-like outputs instead. This BLEU/COMET divergence echoes a broader finding from the WMT 2025 on automated translation evaluation shared task (Lavie et al., 2025) which found automatic metrics particularly unreliable across linguistically diverse targets, reinforcing our reliance on COMET alongside BLEU for the dialectal categories. COMET nonetheless stays at 0.68–0.74, and the AraBench audit (98.8% coverage of unseen word types) confirms that the pruned vocabulary itself can represent dialectal morphology. Dialect handling therefore appears to be a register mismatch orthogonal to pruning (see Limi-

tations). Taken together, these results confirm that pruning does not systematically penalize infrequent Arabic forms: the two-lowest scoring categories reflect content unfamiliarity and register mismatch, not vocabulary size.

5 Conclusion

We have presented in this paper a minimalist, yet powerful, two-stage framework for efficient multilingual NMT that combines corpus-driven vocabulary pruning with targeted fine-tuning. Applied to M2M100, mBART-50, and NLLB-200 on the English–Arabic translation task, our approach achieves approximately 82–87% reduction in embedding-layer size while matching or exceeding the performance of the uncompressed baselines. The best pruned M2M100 fine-tuned on MultiUN, outperforms the zero-shot baseline by more than 15 BLEU points and approaches dedicated bilingual model at a substantially lower memory cost. A dedicated coverage and challenge-set audit further shows that pruning preserves near-total vocabulary coverage across frequency bins, including on rare morphology, clitic combinations, named entities, domain terminology and dialectal Arabic never seen during pruning or fine-tuning. Segment-level analysis shows robust long-sentence performance and no empty or repetitive outputs on M2M100 variant. The main remaining challenge lies in short segments containing bibliographic references, proper nouns, and domain-specific abbreviations, which are failure modes independent of the pruning procedure and call for targeted data augmentation.

Future work will explore extending the pruning framework to larger model variants applying the approach to additional language pairs with non-Latin scripts, and integrating structured knowledge distillation to further reduce the parameter count of the attention layers beyond what vocabulary pruning alone achieves. We also plan to investigate entity-preserving strategies, such as NER-guided vocabulary retention, to mitigate failure modes on proper nouns, citations and domain-specific abbreviations identified in our error analysis.

Limitations

While our approach is applicable to any language pair, it has been validated on the single English–Arabic pair. We selected this pair due to the Arabic’s morphological complexity and non-latin

script, providing a rigorous testbed. Although the methodology is transferable, further evaluation on diverse language families is required to determine whether the reported gains generalize across model architectures and tokenization schemes. Furthermore, the corpus-driven pruning strategy can, in principle, lead to performance degradation on rare entities (e.g. proper nouns, citations) absent from the training data. Our coverage audit (Section 4.5) shows that pruning removes almost exclusively hapax or entirely unseen tokens, with 99.6–99.9% held-out token retention and flat coverage across frequency bins, including 98.8% coverage of 57,000 dialectal word types never seen in the training corpus. The residual degradation we do observe is confined to content that is genuinely absent from training (proper nouns, citations) and to register mismatch (dialectal targets against an MSA-trained model), both of which are orthogonal to vocabulary removal: the full-vocabulary baseline exhibits the same rare-form penalty on medium-length segments (Section 4.4), and the dialect slices of our challenge set show a comparable COMET range (0.68–0.74) to the model’s other out-of-domain results despite very low BLEU (4.7–8.3), consistent with an MSA/dialect surface mismatch rather than a vocabulary-coverage failure. Targeted entity-preserving strategies, such as NER-guided vocabulary retention, remain a promising mitigation, and dialect-aware fine-tuning is left to future work. Finally, we observe a trade-off between in-domain specialization and generalization: models fine-tuned exclusively on formal data (MultiUN) showed slight degradation on general domain text (FLORES-200) compared to hybrid (MultiUN and OPUS) variant while the latter exhibited reduced in-domain performance. This suggests that balancing training data composition is essential to maintain cross-domain adaptability alongside memory optimization.

References

- Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *arxiv preprint arxiv:1907.05019*.
- Maximiliana Behnke and Kenneth Heafield. 2020. Losing heads in the lottery: Pruning transformer attention in neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2664–2674.
- Alexandra Chronopoulou, Dario Stojanovski, and Alexander Fraser. 2023. Language-family adapters for low-resource multilingual neural machine translation. *arxiv preprint arxiv:2209.15236*.
- Marco Cignetta, Tatsuya Hiraoka, Rico Sennrich, Yuval Pinter, and Naoaki Okasaki. 2024. An analysis of bpe vocabulary trimming in neural machine translation. In *Proceedings of the Fifth Workshop on Insights from Negative Results in NLP*, pages 48–50.
- Martha R. Costa-Jussa, James Cross, Onur rlrbi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, and Elah an others Kalbassi. 2022. No language left behind: Scaling human-centered machine translation. *arxiv preprint arxiv:2207.04672*.
- Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 2022. 8-bit optimizers via block-wise quantization. In *arxiv preprint arxiv:2110.02861*.
- Alexis Dorkin, Taido Purason, and Kairit Sirts. 2024. Prune or retrain: optimizing the vocabulary of multilingual models for estonian. In *Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages*, pages 104–108.
- Andreas Eisele and Yu Chen. 2010. Multiun: A multilingual corpus from united nations documents. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, pages 2868–2872.
- Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Man-deep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.
- Angela Fan, Edouard Grave, and Armand Joulin. 2019. Reducing transformer depth on demand with structured dropout. *arxiv preprint arxiv: 1909.11556*.
- Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *arxiv preprint arxiv:1601.01073*.
- Marco Gaido, Thamme Gowda, Roman Grundkiewicz, and Matteo Negri. 2025. Findings of the wmt2025 shared task on model compression: early insights on compressing llms for machine translation. In *Proceedings of the Tenth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers*, pages 484–494.
- Naman Goyal, Cynthia Gao, Vishrav Chaudary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc’Aurelio Ranzato, Francis Guzman, and

- Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the association for computational linguistics* vol. 10, pages 522–538.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arxiv preprint arxiv:1503.02531*.
- Fan Jiang, Honglin Yu, Grace Chung, and Trevor Cohn. 2026. Tokenizer-aware cross-lingual adaptation of decoder-only llms through embedding relearning and swapping. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics Volume 1: Long Papers*, pages 7606–7636.
- Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Younghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google’s multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.
- Zeeshan Khan, Kartheek Akella, Vinay P. Namboodiri, and C. V. Jawahar. 2021. More parameters? no thanks! In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 96–102.
- Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327.
- Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakouga, Jessica Lundin, Christof Monz, Kenton Murray, and 10 others. 2025. Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*.
- Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395.
- Taku Kudo. 2018. 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)*, pages 66–75.
- Alon Lavie, Greg Hanneman, Sweta Agrawal, Diptesh Kanojia, Chi-Kiu Lo, Vilém Zouhar, Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, Archchana Sindhuja, Jiayi Wang, David Ifeoluwa Adelani, Brian Thompson, Tom Kocmi, Markus Freitag, and Daniel Deutsch. 2025. Findings of the WMT25 shared task on automated translation evaluation systems: Linguistic diversity is challenging and references still help. In *Proceedings of the Tenth Conference on Machine Translation*.
- Gurvan L’Hostis, David Grangier, and Michael Auli. 2016. Vocabulary selection strategies for neural machine translation. *arxiv preprint arxiv:1610.00072*.
- Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.
- Minh-Thang Luong, Ilya Sutskever, Quoc V. Le, Oriol Vinyals, and Wojciech Zaremba. 2015. Addressing the rare word problem in neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 11–19.
- Alireza Mohammadshahi, Vassilina Nikoulina, Alexandre Berard, Caroline Brun, James Henderson, and Laurent Besacier. 2022. What do compressed multilingual machine translation models forget? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4308–4329.
- Yasmin Moslem, Muhammad Hazim Al Farouq, and John D. Kelleher. 2025. Iterative layer pruning for efficient translation inference. pages 1022–1027.
- Alonso Palomino. 2026. Selected-layer codec compression for compact speech translation models: an iswslt english-chinese submission. pages 40–46.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.
- David Peer, Sebastian Stabiger, Stefan Engl, and Antonio Rodrigues-Sanchez. 2022. Greedy layer pruning: speeding up transformer models for natural language processing. *arxiv preprint arxiv:2105.14339*.
- Maja Popović. 2015. chrF: character n-gram f-score for machine translation evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.
- Gabriele Prato, Ella Charlaix, and Mehdi Rezagholizadeh. 2020. Fully quantized transformer for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1–14.
- Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. COMET: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2685–2702.

Abigail See, Minh-Thang Luong, and Christopher D. Manning. 2016. Compression of neural machine translation models via pruning. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 291–301.

Sheng Shen, Zhen Dong, Jiayu Ye, Linjian Ma, Zhewei Yao, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. 2020. Q-bert: Hessian based ultra low precision quantization of bert. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8815–8821.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.

Xu Tan, Yi Ren, Di He, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2019. Multilingual neural machine translation with knowledge distillation. In *arxiv preprint arxiv: 1902.10461*.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arxiv preprint arxiv:2008.00401*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pages 2214–2218.

Asahi Ushio, Yi Zhou, and Jose Camacho-Collados. 2023. An efficient multilingual language model compression through vocabulary trimming. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14725–14739.

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. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939.

Elena Voita, David Talbot, Fedor Moiseev, Rico Senrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808.

Ofir Zafrir, Guy Boudoukh, Peter Izsak, and Moshe Wasserblat. 2019. Q8bert: Quantized 8-bit bert. *arxiv preprint arxiv:1910.06188*.

Hanling Zhang, Yayu Zhou, Tongcheng Fang, Zhihang Yuan, Guohao Dai, Wanli Ouyang, and Wang Yu.

2026. Vocabtailor: Dynamic vocabulary selection for downstream in small language models. *arxiv preprint arxiv:2508.15229*.

A Held-Out Corpus Construction

This appendix details the construction of the three held-out sources used in the coverage and challenge-set audit (Section 4.5), none of which overlaps with the pruning or fine-tuning data.

The first source is the FLORES-200 devtest set ($n = 1,012$), already introduced in Section 4.2. The second is a purpose-built challenge set of 2,000 segments (250 per category) spanning eight categories chosen to probe specific linguistic phenomena: `rare_morph` (target sentences drawn from the bottom 5% frequency tail of the fine-tuning corpus), `clitic` (target sentences with stacked Arabic proclitics and enclitics, identified with a regex heuristic), `ner` (source sentences containing named entities, tagged with a spaCy NER model), `domain_term` (source sentences whose English terms are lexically distinctive relative to the MultiUN training corpus by a log-odds criterion), and four dialect categories (`dialect_egyptian`, `dialect_gulf`, `dialect_levantine`, `dialect_maghrebi`) sampled from the AraBench test split. The first four categories are mined from the FLORES dev/devtest and OPUS-100 test pools; the dialect categories are drawn from an 80,000-sentence AraBench pool. An exact-match decontamination check against the prepared MultiUN train, development and test partitions removed 15 overlapping candidates, and all mined sentences pass a minimum-length quality filter (at least four words on both sides).

The third source is the full AraBench test split ($n = 80,413$), used only for the coverage audit given its size.

Because the challenge set concentrates hard phenomena by construction, its overall score sits below the general out-of-domain FLORES result (21.54 vs 24.79 BLEU). It should therefore be read relatively, as a test of whether any category collapses below what general out-of-domain difficulty alone would predict, rather than as an absolute measure of translation quality.

B Vocabulary Coverage Details

For each held-out corpus (Appendix A), we tokenize the Arabic references with the original pre-trained tokenizer and measure token- and type-

level representability after pruning. Retained-token rates reach 99.9%, 99.8% and 99.6% for FLORES, the challenge set and AraBench respectively, with $\sim 96\%$ of sentences per corpus tokenizing with zero pruned tokens. Table 6 stratifies coverage by word-type frequency in the pruning corpus.

Table 6: Coverage of held-out Arabic word types by the pruned vocabulary, stratified by frequency in the pruning corpus (bin 0 = unseen in the pruning corpus). Coverage is the percentage of word types that decompose entirely into retained subwords.

Corpus-freq bin	FLORES		Challenge		AraBench	
	Types	% cov.	Types	% cov.	Types	% cov.
0 (unseen)	3108	98.71	5191	98.54	57469	98.82
1	617	100.00	907	100.00	2530	100.00
2–5	1048	100.00	1106	100.00	3057	100.00
6–20	1290	100.00	1247	100.00	2681	100.00
21–100	1629	100.00	1620	100.00	2527	100.00
>100	1937	100.00	2077	100.00	2169	100.00

C Challenge-set Translation Results

Table 7 reports BLEU, chrF++, TER and COMET for the pruned and fine-tuned M2M100 (MultiUN) model on each challenge category, together with the overall score on the full 2,000-segment set.

Table 7: Challenge-set translation results for M2M100 (MultiUN, pruned and fine-tuned), 250 segments per category except Overall (2,000). TER: lower is better ($\downarrow$). Reference points: MultiUN in-domain test BLEU 42.04 / COMET 0.8730 (Table 1); FLORES out-of-domain BLEU 24.79 (Table 2).

Category	BLEU	chrF++	TER $\downarrow$	COMET
Overall (2,000)	21.54	40.19	79.02	0.747
clitic	28.57	49.69	70.02	0.805
domain_term	28.69	47.89	65.14	0.792
ner	26.45	45.35	67.97	0.788
rare_morph	16.59	34.14	86.03	0.741
dialect_egyptian	8.18	24.96	94.15	0.744
dialect_gulf	8.29	26.07	101.34	0.732
dialect_levantine	4.73	20.13	113.52	0.694
dialect_maghrebi	5.79	20.04	103.44	0.680

D Comparison with a Few-Shot LLM Baseline

Table 8 reports results for Aya-8B (Üstün et al., 2024), an open-weight multilingual LLM, prompted with 5 fixed in-domain exemplars (seed 42, temperature 0, no fine-tuning or vocabulary pruning), on the same MultiUN test set and metrics as Table 1.

Table 8: Aya-8B (5-shot) versus pruned/fine-tuned and bilingual systems on the MultiUN test set ($n = 1,500$). TER: lower is better ($\downarrow$).

System	BLEU	chrF++	TER $\downarrow$	COMET
Aya-8B (5-shot)	36.77	55.14	56.57	0.8754
M2M100 (pruned + FT, MultiUN)	42.04	58.81	51.55	0.8730
NLLB-200 600M (pruned + FT, MultiUN)	37.70	55.79	55.04	0.8640
OPUS-MT-en-ar (bilingual baseline)	44.59	58.77	58.59	0.7911

M2M100 outperforms Aya-8B on BLEU, chrF++, and TER, with comparable COMET scores. NLLB-200 trails Aya-8B on all four metrics, while OPUS-MT-en-ar shows the same BLEU/COMET divergence noted in Section 4.3.