

IndicMT at WMT 2026: Full Fine-Tuning of NLLB-200-600M with Round-Trip Filtering and Forward Self-Training for Arabic-Asian Translation

Pramit Jana^{1,*}, Arindam Adhikari¹, Priyobroto Acharya^{2,*}, Dipankar Das³

¹Department of Construction Engineering, Jadavpur University, Kolkata, India

²Department of Power Engineering, Jadavpur University, Kolkata, India

³Department of Computer Science and Engineering, Jadavpur University, Kolkata, India

{pramitj.const.ug, dipankar.das}@jadavpuruniversity.in

{arindamadhikari2004, priyobrotoacharya98}@gmail.com

* Equal contribution.

Abstract

We describe the IndicMT submission to the WMT 2026 Shared Task on Low-Resource Arabic-Asian Language Translation. Our system covers all ten translation directions between Arabic and Bangla, English, Hindi, Indonesian, and Urdu. For each direction, we fully fine-tune a separate NLLB-200-distilled-600M model using two NVIDIA T4 graphics processing units. The training pipeline combines Unicode normalization and length filtering with direction-specific round-trip consistency filtering of the available parallel data. We additionally augment each direction with 500 source-side sentences whose target translations are generated by the pre-trained NLLB model, a procedure corresponding to forward self-training rather than conventional back-translation. On the official test set, our strongest relative results are obtained for Arabic→Indonesian and Indonesian→Arabic. Across all five language pairs, translation from Arabic obtains higher BLEU scores than translation into Arabic. Our results demonstrate that full adaptation of a compact multilingual translation model, combined with model-based data selection and synthetic-target augmentation, provides a practical baseline under limited computational resources.

1 Introduction

Neural machine translation (NMT) has achieved strong performance for language pairs supported by large parallel corpora, particularly with attention-based encoder-decoder models and the Transformer architecture (Vaswani et al., 2017). Performance remains substantially less reliable in low-resource settings, where limited parallel data restricts supervised adaptation and increases sensitivity to noisy or domain-mismatched examples. Arabic-centric translation introduces additional challenges associated with rich morphology, cliticization, orthographic variation, and optional diacritics, especially when Arabic is paired with

typologically and script-diverse Asian languages. Massively multilingual machine translation models can partially address data scarcity through transfer across languages. No Language Left Behind (NLLB-200) supports direct translation among a broad set of languages, including all languages considered in this work (team et al., 2022). Its distilled 600M-parameter checkpoint is particularly suitable for adaptation in computationally constrained settings. Nevertheless, effective low-resource adaptation depends not only on model selection but also on the quality and coverage of the available training data. The WMT 2026 Shared Task on Low-Resource Arabic–Asian Language Translation evaluates five directions into Arabic and five directions from Arabic, pairing Arabic with Bangla, English, Hindi, Indonesian, and Urdu (Laskar et al., 2026). We participate in all ten directions. For each direction, we fully fine-tune a separate NLLB-200-distilled-600M model. To improve the training set, we use the pretrained model in two ways. First, we apply direction-specific round-trip consistency filtering to retain parallel examples whose source sentence can be reconstructed with sufficient character-level similarity. Second, we generate synthetic targets for 500 additional source-side sentences. The latter procedure is forward self-training, rather than conventional back-translation, because the generated side is the target side. Our contributions are threefold. First, we develop ten direction-specific systems covering all shared-task directions through full-parameter adaptation under a two-T4-GPU budget. Second, we combine round-trip selection of authentic bi-text with synthetic-target augmentation in a single NLLB-based pipeline. Third, we analyze the official results across lexical, character-level, edit-based, and semantic metrics, identifying a consistent BLEU asymmetry between translation from and into Arabic and a substantial development-to-test performance gap.

2 Related Work

Early Arabic machine translation (MT) was dominated by rule-based and phrase-based statistical approaches, including systems developed with Moses (Koehn, 2005). Progress on Arabic–Asian language pairs remained limited, however, largely because of scarce parallel data and the linguistic complexity of Arabic. Its rich morphology, productive cliticization, and optional diacritics increase lexical variation and complicate alignment and segmentation, particularly when paired with typologically diverse Asian languages. Attention-based encoder–decoder models and Transformers (Vaswani et al., 2017) improved source–target modeling, but their effectiveness in low-resource settings remains strongly dependent on the availability and quality of parallel supervision.

Massively multilingual MT models mitigate data scarcity through cross-lingual transfer. NLLB-200 supports translation across 200 languages using large-scale multilingual training and mined parallel data (team et al., 2022); its distilled 600M-parameter variant offers a practical backbone for adaptation under limited computational resources. Parameter-efficient methods provide a less resource-intensive alternative to full fine-tuning: LoRA introduces trainable low-rank updates to frozen pretrained weights (Hu et al., 2021), whereas DoRA decomposes weights into magnitude and directional components (Liu et al., 2024). However, their effectiveness relative to full fine-tuning is less established for low-resource multilingual MT (Su et al., 2024). Prior work on Indic translation has explored LoRA- and DoRA-based adaptation of NLLB-200 and IndicTrans2, while also demonstrating the importance of linguistically informed preprocessing (Acharya et al., 2025).

The WMT 2026 Shared Task on Low-Resource Arabic–Asian Language Translation establishes a common benchmark for bidirectional translation between Arabic and Bengali, English, Hindi, Indonesian, and Urdu (Laskar et al., 2026). Within this setting, we study full fine-tuning of the distilled NLLB-200-600M model combined with back-translation-based data augmentation. In contrast to prior work centered on parameter-efficient adaptation, our approach evaluates whether full-model optimization and synthetic parallel data can jointly improve Arabic–Asian translation under strict dual-GPU constraints.

3 Methodology and Implementation Details

System overview. We initialize all systems from facebook/nllb-200-distilled-600M and fine-tune the model independently for each of the ten translation directions. This produces ten direction-specific checkpoints that share the same architecture and training configuration, rather than a single jointly trained multidirectional model. All experiments are conducted on two NVIDIA T4 GPUs with 15 GB memory per device. The same pretrained checkpoint is also used to filter the parallel data and generate synthetic training examples, as described in Section 3.2.

NLLB-200 is a multilingual encoder–decoder Transformer developed to support translation across 200 languages, with an emphasis on low-resource translation (team et al., 2022). We select its distilled 600M-parameter dense checkpoint because it supports Arabic, Bengali, English, Hindi, Indonesian, and Urdu through dedicated language-script identifiers and remains feasible for full-parameter fine-tuning under our hardware budget. This coverage permits direct translation in all task directions without an intermediate pivot language. We also considered IndicTrans2 (Gala et al., 2023) and Sarvam AI translation models. However, these systems primarily cover English and Indian languages and do not natively support Arabic as a translation endpoint. Our selection is therefore based on language coverage and computational suitability, rather than a controlled comparison of translation quality.

3.1 Preprocessing

We apply a uniform preprocessing pipeline to every language pair. Each sentence is first normalized using Unicode Normalization Form KC (NFKC) (Unicode Consortium, 2025). We then collapse consecutive whitespace and remove leading and trailing whitespace. A sentence pair is discarded if either side exceeds 100 tokens, with filtering performed jointly to preserve source–target alignment. Table 1 reports the number of examples retained after preprocessing.

3.2 Round-Trip Filtering and Forward Self-Training

Synthetic parallel data can supplement authentic bitext when human translations are limited. In conventional back-translation, target-language mono-

Direction	Raw	Retained	Training	Held-out
Arabic→Bengali	15,000	10,960	10,958	575
Bengali→Arabic	15,000	10,960	10,957	575
Arabic→English	17,000	11,583	11,563	1,287
English→Arabic	17,000	12,434	12,378	1,381
Arabic→Hindi	15,000	11,507	11,440	603
Hindi→Arabic	15,000	11,643	11,584	612
Arabic→Indonesian	15,000	10,184	10,180	536
Indonesian→Arabic	15,000	11,712	11,704	616
Arabic→Urdu	15,000	10,675	10,646	561
Urdu→Arabic	15,000	11,966	11,875	626

Table 1: Numbers of parallel sentence pairs before and after preprocessing, together with the training and held-out set sizes.

lingual sentences are translated into the source language, producing synthetic sources paired with authentic targets (Sennrich et al., 2016). In contrast, forward self-training translates source-side monolingual sentences into the target language and uses the generated hypotheses as pseudo-targets (Abdumumin et al., 2021; Jiao et al., 2021). Our pipeline combines round-trip filtering of authentic parallel data with the latter form of synthetic-target augmentation.

Round-trip consistency filtering. For a parallel pair (x, y) in a particular source-to-target direction, we translate the source sentence x into the target language and then translate the resulting hypothesis back into the source language:

$$\hat{y} = M_{x \rightarrow y}(x), \quad \hat{x} = M_{y \rightarrow x}(\hat{y}).$$

Both translations are generated using the pretrained facebook/nllb-200-distilled-600M checkpoint before direction-specific fine-tuning. We compute the similarity between x and $\hat{x}$ using Python’s character-level `difflib.SequenceMatcher` ratio and retain the authentic pair (x, y) if the ratio is at least 0.3. Round-trip similarity has previously been used as a quality signal for selecting pseudo-parallel data (Imankulova et al., 2019; Khatri and Bhattacharyya, 2020); our use differs in that the signal is applied to the available authentic bitext. Filtering is performed independently for every translation direction. Consequently, reversing a language pair may produce a different retained subset because reconstruction is evaluated in a different source language. The threshold of 0.3 follows prior use of round-trip similarity as a pseudo-parallel quality signal (Imankulova et al., 2019; Khatri and Bhattacharyya, 2020) and was fixed across all ten directions without per-direction tuning, given the two-GPU computational budget.

We did not conduct a threshold sweep; Section 5 lists this as an avenue for future ablation work.

Synthetic-target generation. For each direction, we sample 500 additional source-language sentences from the dataset. We remove examples overlapping with the official training and development sets using exact string matching against the normalized source sentences in the official training and development sets. The pretrained source-to-target NLLB model translates each sentence into the target language, and each source sentence is paired with its generated hypothesis:

$$\mathcal{D}_{\text{syn}} = \{(x_i, \hat{y}_i)\}_{i=1}^{500}.$$

No additional filtering of the synthetic pairs was applied beyond the generation step described above. Synthetic and authentic examples receive the same training weight. Because the generated side is the target side, we refer to this procedure as forward self-training.

3.3 Model Training

For each direction, we fully fine-tune all parameters of facebook/nllb-200-distilled-600M (team et al., 2022). We use Adafactor, which reduces optimizer-state memory by maintaining factored second-moment estimates for matrix-valued parameters (Shazeer and Stern, 2018). Training is conducted for three epochs with a learning rate of 3×10^{-5} , 150 warm-up steps, and cosine learning-rate decay. The per-device batch size is 2, and gradients are accumulated over four steps, giving an effective batch size of 16 across two GPUs. Input and target sequences are truncated to 128 subword tokens. We disable FP16 training and perform optimization in FP32. Checkpoints are selected using BLEU on the development set. Table 2 summarizes the complete configuration.

3.4 Inference

At inference, we use beam search with a beam size of 8, a length penalty of 1.2, and a maximum generation length of 128 tokens. We additionally block repeated trigrams to reduce repetitive generation. We apply the same decoding configuration to all translation directions to maintain a controlled evaluation setting.

Our approach differs from adapter-based low-resource MT systems in two respects. First, we

Hyperparameter	Value
Optimizer	Adafactor
Learning rate	3×10^{-5}
Learning-rate schedule	Cosine decay
Warm-up steps	150
Batch size	2 per device
Gradient accumulation	4 steps
Effective batch size	16
Training epochs	3
Maximum sequence length	128 subword tokens
Numerical precision	FP32
Checkpoint criterion	Development-set BLEU

Table 2: Hyperparameters used for all direction-specific systems.

fully fine-tune a compact multilingual model rather than applying LoRA or DoRA to a substantially larger backbone. Second, we reuse NLLB-200 for data filtering and pseudo-label generation, avoiding an additional multilingual sentence-embedding model. The result is a self-contained pipeline for all ten Arabic–Asian directions that operates within a two-GPU computational budget.

4 Results and Discussion

Evaluation metrics. We evaluate development outputs using corpus-level BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002), computed with SacreBLEU (Post, 2018). The shared-task organizers additionally report ChrF2, a character n-gram F-score with recall weight $\beta = 2$ (Popović, 2017); Translation Edit Rate (TER) (Snover et al., 2006); COMET, a learned neural metric for translation evaluation (Rei et al., 2020); and an mBERT-based score (Zhang et al., 2020). Higher BLEU, ChrF2, COMET, and mBERT-based scores indicate better performance, whereas lower TER is better. Table 3 reports development-set BLEU, while Table 4 presents the official WMT 2026 shared-task results for the IndicMT team. Because BLEU emphasizes surface-form overlap, we interpret it together with the character-level and semantic evaluation metrics reported on the official test set.

On the development set, Arabic→English obtains the highest BLEU score (38.87), followed by Arabic→Indonesian (36.34) and Arabic→Hindi (33.14). Bengali→Arabic yields the lowest score (18.74). More generally, translation from Arabic outperforms translation into Arabic for all five paired directions. The directional differences are especially large for Indonesian (16.63 BLEU), Hindi (12.22), and English (14.33), but smaller

Direction	BLEU
Arabic→English	38.87
Arabic→Indonesian	36.34
Arabic→Hindi	33.14
Arabic→Urdu	29.54
English→Arabic	24.54
Arabic→Bengali	21.01
Hindi→Arabic	20.92
Indonesian→Arabic	19.71
Urdu→Arabic	18.77
Bengali→Arabic	18.74

Table 3: Development-set BLEU scores after three epochs of direction-specific fine-tuning.

for Urdu (10.77) and Bengali (2.27). This pattern indicates a systematic directional asymmetry in the present systems. Possible contributing factors include differences in target-side generation difficulty, pretraining coverage, and the lexical and morphological variation of Arabic. Establishing the individual contribution of these factors, however, would require controlled linguistic and data-distribution analyses.

The official test results exhibit a similar directional pattern. Among the Arabic-target systems, English→Arabic achieves the highest BLEU (16.05), followed by Indonesian→Arabic (13.58). Among the Arabic-source systems, Arabic→English obtains the highest BLEU (25.49), closely followed by Arabic→Indonesian (24.67) and Arabic→Hindi (24.11). Arabic→Indonesian is the strongest system in relative leaderboard performance, ranking second, while Indonesian→Arabic ranks third. These results suggest that the proposed training procedure is particularly effective for the Arabic–Indonesian pair, although system rankings should be interpreted in relation to the number and strength of submissions in each subtask.

Translation into Arabic remains more difficult according to BLEU for every paired direction. The test-set differences are 9.44 BLEU for English, 11.47 for Hindi, 2.47 for Bengali, 11.10 for Indonesian, and 8.59 for Urdu. The smaller Bengali difference indicates that the directional asymmetry is not uniform across languages. Moreover, the metrics do not always induce the same system-level interpretation. For example, Arabic→Hindi records a lower COMET score than Arabic→English and Arabic→Indonesian despite achieving a similar BLEU score. This divergence highlights the importance of reporting both overlap-based and model-based metrics, particularly for morphologically rich target languages with multiple valid surface real-

Task	Direction	Rank	BLEU	ChrF2	TER	COMET	mBERT
1A	English→Arabic	5th	16.0545	49.3012	76.4907	81.9755	83.4821
1B	Hindi→Arabic	6th	12.6454	46.2765	80.9425	80.3644	82.2278
1C	Bengali→Arabic	5th	11.5256	45.1398	82.3128	79.3155	81.7783
1D	Indonesian→Arabic	3rd	13.5751	46.6783	79.8812	80.8181	82.4175
1E	Urdu→Arabic	5th	12.5320	45.8115	81.2936	79.9090	82.0278
2A	Arabic→English	8th	25.4940	51.9472	64.0603	79.7175	80.2948
2B	Arabic→Hindi	5th	24.1114	49.1763	67.0918	69.3231	84.1642
2C	Arabic→Bengali	4th	13.9948	47.7002	76.5946	76.9274	82.3213
2D	Arabic→Indonesian	2nd	24.6707	54.5822	63.9452	81.9856	84.7227
2E	Arabic→Urdu	4th	21.1249	46.3495	69.5899	73.8075	82.4280

Table 4: Official WMT 2026 test-set results for the IndicMT team. Higher BLEU, ChrF2, COMET, and mBERT scores are better, whereas lower TER is better.

izations.

Direction	Dev.	Test	Δ BLEU
Arabic→English	38.87	25.49	−13.38
Arabic→Indonesian	36.34	24.67	−11.67
Arabic→Hindi	33.14	24.11	−9.03
Arabic→Urdu	29.54	21.12	−8.42
English→Arabic	24.54	16.05	−8.49
Hindi→Arabic	20.92	12.65	−8.27
Bengali→Arabic	18.74	11.53	−7.21
Arabic→Bengali	21.01	13.99	−7.02
Indonesian→Arabic	19.71	13.58	−6.13
Urdu→Arabic	18.77	12.53	−6.24

Table 5: Difference between development- and official test-set BLEU. Negative values indicate lower performance on the official test set.

As shown in Table 5, test BLEU is lower than development BLEU for every direction, with decreases ranging from 6.13 to 13.38 points. The largest decreases occur for Arabic→English and Arabic→Indonesian, whereas the smallest occur for Indonesian→Arabic and Urdu→Arabic. These differences should not be interpreted as direct evidence of overfitting because the development and official test sets may differ in domain, sentence length, reference style, or data construction. The results instead indicate a distributional or evaluation mismatch between the two sets. Future analysis should compare their domains, length distributions, lexical coverage, and subword fragmentation rates, and should report results over multiple random seeds to separate systematic generalization differences from training variance.

5 Conclusion

We presented direction-specific Arabic–Asian MT systems based on full fine-tuning of the distilled NLLB-200-600M model under a two-GPU computational budget. The proposed pipeline combines

deterministic preprocessing, round-trip consistency filtering, and forward self-training with synthetic target sentences. On the official WMT 2026 test set, the systems achieved their strongest relative results for Arabic→Indonesian and Indonesian→Arabic, ranking second and third, respectively. Across all five paired languages, translation from Arabic obtained higher BLEU than translation into Arabic, although the magnitude of this asymmetry varied substantially by language.

The consistent development-to-test degradation further indicates that robustness to domain and evaluation-set variation remains an important challenge. Future work will evaluate the individual contributions of filtering and synthetic augmentation through controlled ablations, examine iterative back-translation using genuine target-side monolingual data, and compare full fine-tuning with parameter-efficient adaptation. Additional analysis using multiple random seeds, human evaluation, and language-specific error categories would provide a more reliable account of adequacy, fluency, and morphological correctness.

Limitations

This study has several limitations. First, we do not conduct controlled ablation experiments to isolate the contributions of preprocessing, round-trip consistency filtering, and synthetic-data augmentation. Consequently, the effect of each component cannot be quantified independently. Second, we do not compare the fine-tuned systems against the zero-shot NLLB-200-distilled-600M baseline, which limits our ability to attribute the observed performance to direction-specific fine-tuning. Third, BLEU is consistently lower on the official test set than on the development set (Table 5). This discrepancy may reflect domain, difficulty, or reference-

distribution differences between the two sets, as well as possible overfitting; however, the present experiments do not distinguish among these factors. Finally, our evaluation relies exclusively on automatic metrics. Human evaluation and targeted error analysis are required to assess adequacy, fluency, and language-specific phenomena such as Arabic morphological realization.

References

- Idris Abdulmumin, Bashir Shehu Galadanci, and Abubakar Isa. 2021. Enhanced back-translation for low resource neural machine translation using self-training. In *Information and Communication Technology and Applications*, pages 355–371, Cham. Springer International Publishing.
- Priyobroto Acharya, Haranath Mondal, Dipanjan Saha, Dipankar Das, and Sivaji Bandyopadhyay. 2025. [JUNLP: Improving low-resource Indic translation system with efficient LoRA-based adaptation](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 1201–1209, Suzhou, China. Association for Computational Linguistics.
- Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul Nawale, Anupama Sujatha, Ratish Pudupully, Vivek Raghavan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. [IndicTrans2: Towards High-Quality and Accessible Machine Translation Models for All 22 Scheduled Indian Languages](#). *Transactions on Machine Learning Research*.
- J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. [Lora: Low-rank adaptation of large language models](#). *ArXiv*, abs/2106.09685.
- Aizhan Imankulova, Takayuki Sato, and Mamoru Komachi. 2019. [Filtered pseudo-parallel corpus improves low-resource neural machine translation](#). *ACM Transactions on Asian and Low-Resource Language Information Processing*, 19:1–16.
- Wenxiang Jiao, Xing Wang, Zhaopeng Tu, Shuming Shi, Michael Lyu, and Irwin King. 2021. [Self-training sampling with monolingual data uncertainty for neural machine translation](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2840–2850, Online. Association for Computational Linguistics.
- Jyotsana Khatri and Pushpak Bhattacharyya. 2020. [Filtering back-translated data in unsupervised neural machine translation](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4334–4339, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit (MT Summit X)*, pages 79–86.
- Sahinur Rahman Laskar, Firoj Alam, Bishwaraj Paul, Irfan Ahmad, Maya Silvi Lydia, and Pankaj Kundan Dadure. 2026. Findings of the WMT 2026 shared task on low-resource arabic-asian language translation. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. [Dora: Weight-decomposed low-rank adaptation](#). *Preprint*, arXiv:2402.09353.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Maja Popović. 2017. [chrF++: words helping character n-grams](#). In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.
- Matt Post. 2018. [A call for clarity in reporting BLEU scores](#). In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.
- Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. [COMET: A neural framework for MT evaluation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [Improving neural machine translation models with monolingual data](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.
- Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 4596–4604. PMLR.
- 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, Cambridge, Massachusetts.

USA. Association for Machine Translation in the Americas.

Tong Su, Xin Peng, Sarubi Thillainathan, David Guzmán, Surangika Ranathunga, and En-Shiun Lee. 2024. [Unlocking parameter-efficient fine-tuning for low-resource language translation](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4217–4225, Mexico City, Mexico. Association for Computational Linguistics.

Nllb team, Marta Ruiz Costa-jussà, James Cross, Onur cCelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Alison Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. [No language left behind: Scaling human-centered machine translation](#). *ArXiv*, abs/2207.04672.

Unicode Consortium. 2025. Unicode standard annex #15: Unicode normalization forms. <https://www.unicode.org/reports/tr15/>. Accessed 2026-07-27.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with bert](#). *Preprint*, arXiv:1904.09675.