

NITM AI Lab at WMT 2026 Shared Task: Low-Resource Arabic-Asian Language Translation

Nabin Koirala¹ and Rhythm Bhetwal² and Nurul Amin Choudhury²

¹Vignan's Foundation for Science Technology and Research

²National Institute of Technology Meghalaya

nabinkoirala53@gmail.com

rhythmbhetwal77@gmail.com

nurul.choudhury@nitm.ac.in

Abstract

This paper describes the NITM AI Lab submissions to the WMT26 Low-Resource Arabic-Asian Translation shared task. We participate in two translation directions: EnglishtoArabic (Sub-Task 1A) and ArabictoHindi (Sub-Task 2B). To address the data scarcity inherent to these language pairs, we compare three neural architectures: a domain-specialized bilingual model (OPUS-MT-TC-Big), a distilled massively multilingual model (NLLB-200-Distilled-600M), and a generalist multilingual sequence-to-sequence transformer (mBART-50-Many-to-Many). Our systems contrast full-parameter fine-tuning with parameter-efficient fine-tuning (PEFT) via Low-Rank Adaptation (LoRA), and we measure the effect of rule-based orthographic and numerical normalization together with targeted data augmentation from OPUS. On the official blind test set, the domain-specialized OPUS-MT model is the strongest system for EnglishtoArabic, while LoRA-tuned NLLB-200 performs best for ArabictoHindi, with OPUS augmentation giving a small additional gain in the latter direction.

1 Introduction

Machine translation between Arabic and other regional languages is difficult in low-resource settings because of limited parallel data, rich Arabic morphology, and inconsistent orthography across sources. Pretrained multilingual models such as mBART-50 (Tang et al., 2020) and NLLB-200 (NLLB Team et al., 2022) offer broad cross-lingual coverage, while bilingual models such as OPUS-MT (Tiedemann and Thottingal, 2020) offer a strong prior for a specific language pair at the cost of generality. Parameter-efficient fine-tuning methods such as LoRA (Hu et al., 2022) have been proposed to adapt large multilingual models to new domains without the overfitting risk of full fine-tuning on small datasets.

This paper describes our submissions to the WMT26 Low-Resource Arabic-Asian Translation shared task (Laskar et al., 2026), covering Sub-Task 1A (EnglishtoArabic) and Sub-Task 2B (ArabictoHindi). For each direction, we compare a domain-specialized bilingual model against two multilingual backbones under full fine-tuning and LoRA, and we quantify the effect of script-level normalization and OPUS-based data augmentation. Three design questions guide our systems: (1) whether a translation-pair-specialized pretrained model outperforms a generalist multilingual model under data scarcity, (2) whether full fine-tuning or parameter-efficient adaptation better mitigates overfitting, and (3) how much script-level and morphological normalization contributes to translation quality for Arabic.

PART I: Sub-Task 1A — English to Arabic

2 System Overview

For EnglishtoArabic, we submitted three systems spanning different pretraining designs and fine-tuning paradigms:

- **Primary System:** Full fine-tuning of OPUS-MT-TC-Big on the official 21,000-pair dataset.
- **Contrastive System 1:** LoRA adaptation of NLLB-200-Distilled-600M, augmented with 20,000 additional parallel sentences from OPUS.
- **Contrastive System 2:** Full fine-tuning of mBART-50 with aggressive orthographic and numerical normalization.

3 Data and Preprocessing

The base training pool is the official 21,000 English–Arabic parallel sentences released by the

shared task organizers (Laskar et al., 2026). Contrastive System 1 expands this to 41,000 pairs using additional data from OPUS (Tiedemann, 2012).

English normalization. Source text was lower-cased, whitespace-collapsed, and normalized to Unicode NFC form.

Arabic normalization. We applied the following steps, with scope noted per system:

- *Diacritic removal* (all systems): short vowel marks and Tashkeel signs were removed using the character range `[\u064B-\u065F\u0640\u0670]`.
- *Alef standardization* (all systems): Alef Wasla (U+0671), Alef Madda (U+0622), Alef with Hamza Below (U+0625), and Alef with Hamza Above (U+0623) were mapped to bare Alef (U+0627).
- *Morphological reduction* (Contrastive 1 & 2): Ta Marbuta (U+0629) to Ha (U+0647); Alef Maqsura (U+0649) to Yaa (U+064A).
- *Numeral mapping* (Contrastive 2 only): Eastern Arabic-Indic numerals (U+0660–U+0669) were converted to Western digits (0–9) to prevent token fragmentation.

4 Model Architectures and Training

Primary: **OPUS-MT-TC-Big (full fine-tuning).** We fine-tuned Helsinki-NLP/opus-mt-tc-big-en-ar (Tiedemann and Thottingal, 2020), a Transformer encoder-decoder pretrained specifically for English–Arabic translation, for 10 epochs with AdamW (Loshchilov and Hutter, 2019) and a cosine learning-rate schedule (peak LR 5×10^{-5} , 10% warmup). Effective batch size was 32 (per-device 8, 4 accumulation steps), in FP16.

Contrastive 1: NLLB-200-Distilled-600M with LoRA. To exploit the cross-lingual generalization of a massively multilingual model while limiting catastrophic forgetting, we applied LoRA (Hu et al., 2022) to the attention projections (q_proj, k_proj, v_proj, out_proj) of facebook/nllb-200-distilled-600M (NLLB Team et al., 2022). LoRA rank was $r = 16$, scaling factor $\alpha = 32$, dropout 0.05, updating only 0.76% of parameters ($\sim 4.7\text{M}$). Training ran for 12 epochs at learning rate 1.5×10^{-4} on the augmented 41,000-pair corpus.

Contrastive 2: mBART-50 (full fine-tuning). We fully fine-tuned facebook/mbart-large-50-many-to-many-mmt (Tang et al., 2020) for 10 epochs. Gradient checkpointing was enabled with per-device batch size 1 and 16 accumulation steps (effective batch 16), learning rate 2×10^{-5} , and early stopping (patience 3).

PART II: Sub-Task 2B — Arabic to Hindi

5 System Overview

For ArabictoHindi, we submitted three systems built around NLLB-200 and mBART-50, contrasting PEFT against full fine-tuning:

- **Primary System:** NLLB-200-Distilled-600M adapted via LoRA on the official constrained dataset.
- **Contrastive System 1:** The same NLLB-200+LoRA configuration, trained with an additional 10,000 OPUS Arabic–Hindi pairs.
- **Contrastive System 2:** Full fine-tuning of mBART-50-Many-to-Many on the official constrained dataset.

6 Data and Preprocessing

Arabic normalization is identical to Sub-Task 1A: diacritic and Tatweel removal, Alef variant standardization, Ta MarbutatoHa, Alef Maqsura-toYaa, and Eastern-to-Western numeral conversion (mBART system only). Target Hindi text underwent whitespace compression and Unicode NFC normalization for consistent tokenization.

The Primary and Contrastive 2 systems used only the official parallel data (Laskar et al., 2026), comprising 20,100 pairs after filtering empty lines introduced by normalization. Contrastive System 1 was augmented with 10,000 Arabic–Hindi pairs from OPUS (Tiedemann, 2012), yielding a training corpus of 30,078 pairs.

7 Model Architectures and Training

Primary: NLLB-200-Distilled-600M with LoRA (official data). LoRA (Hu et al., 2022) was applied to the attention projections (q_proj, k_proj, v_proj, out_proj) of facebook/nllb-200-distilled-600M (NLLB Team et al., 2022): rank $r = 32$, $\alpha = 32$, dropout 0.05. Training used AdamW at peak LR 1×10^{-4}

with a cosine schedule and 10% warmup, for 10 epochs in FP16, effective batch size 32, on the official 20,100-pair corpus.

Contrastive 1: NLLB-200-Distilled-600M with LoRA (official + OPUS). This system uses a separate LoRA configuration on the same attention projections: rank $r = 16$, $\alpha = 32$, dropout 0.05. Training used AdamW at peak LR 2×10^{-4} with a cosine schedule and 10% warmup, for 10 epochs in FP16, effective batch size 32, on the expanded 30,078-pair corpus.

Contrastive 2: mBART-50 (full fine-tuning). Full fine-tuning of facebook/mbart-large-50-many-to-many-mmt for 10 epochs, with gradient checkpointing, effective batch size 16 (per-device 1, accumulation 16), learning rate 2×10^{-5} , and early stopping (patience 3).

8 Experimental Configuration Summary

Table 1 summarizes all six systems across both sub-tasks. Forced BOS tokens used at inference are given in Section 9.

9 Inference and Decoding

For Sub-Task 1A, OPUS-MT (Primary) used beam width 2 and length penalty 1.0; NLLB (Contrastive 1) used beam width 5, length penalty 1.0, and forced BOS token `arb_Arab`; mBART (Contrastive 2) used beam width 5, length penalty 1.2 to encourage complete generation, forced BOS token `ar_AR`, and the same Arabic normalization pipeline applied to model output.

For Sub-Task 2B, both NLLB systems used beam width 5, length penalty 1.0, max generation length 128 tokens, and forced BOS token `hin_Deva`; mBART used beam width 5, length penalty 1.2, max generation length 128 tokens, and forced BOS token `hi_IN`.

10 Results

Sub-Task 1A. The domain-specialized OPUS-MT model achieves the highest BLEU (Papineni et al., 2002) (23.59), outperforming both generalist multilingual models on n-gram precision. This is consistent with expectations: OPUS-MT-C-Big was pretrained specifically on large-scale English–Arabic bitexts, giving it a strong prior for this pair even before task fine-tuning. The NLLB-200+LoRA system reaches a competitive

chrF (Popović, 2015) of 51.93, indicating strong character-level morphological coverage. Contrastive System 2 (mBART-50) achieves the highest chrF (53.07) despite a lower BLEU, suggesting better morphological alignment at the subword level, likely due to the aggressive numeral and orthographic normalization applied only to this system.

Sub-Task 2B. NLLB-200 with LoRA fine-tuning is the strongest architecture for this direction, with the Primary and Contrastive 1 systems clearly ahead of mBART-50. Adding the 10,000 OPUS Arabic–Hindi pairs in Contrastive System 1 gives the highest dev-test BLEU (25.04) against Primary’s 24.60, a modest but consistent gain from augmentation. Primary retains a small edge in chrF (49.59 vs. 49.51), suggesting the official-only model is marginally more conservative at the character level while the augmented model captures slightly more of the correct n-gram content. The mBART-50 full fine-tuning system reaches a BLEU of 23.93, a reasonable result given its generalist pre-training and the absence of Arabic–Hindi domain-specific data.

10.1 Official Blind Test Results

Based on the official evaluation, our systems achieved competitive rankings across both translation directions. For English to Arabic (Subtask 1A), our Primary system ranked 8th among all participating systems, while Contrastive System 2 and Contrastive System 1 ranked 11th and 12th, respectively. For Arabic to Hindi (Subtask 2B), Contrastive System 1 achieved 3rd place and Contrastive System 2 achieved 4th place, while the Primary system ranked 9th. These rankings demonstrate the effectiveness of our approaches, particularly the OPUS-augmented NLLB model for Arabic to Hindi translation. Table 3 reports the official scores released by the organizers (Laskar et al., 2026).

Sub-Task 1A. On the blind test set, the Primary system outperforms both contrastive systems across every metric, confirming the development-set trend that domain-specialized pretraining is decisive for this pair. The gap between Primary and the contrastive systems is wider on the blind test than on the dev-test set, which we attribute to a distribution shift between the organizer-provided dev-test split and the held-out blind test set. Between the two contrastive systems, mBART-50 edges out NLLB+LoRA on the official test set across all

Task	System	Base Model	Strategy	Train Pairs	Batch	Beam LP
1A (En→Ar)	Primary	OPUS-MT-TC-Big	Full FT	21,000	32	2 1.0
	Contrastive 1	NLLB-200-600M	PEFT (LoRA)	41,000	32	5 1.0
	Contrastive 2	mBART-50-Large	Full FT	21,000	16	5 1.2
2B (Ar→Hi)	Primary	NLLB-200-600M	PEFT (LoRA)	20,100	32	5 1.0
	Contrastive 1	NLLB-200-600M	PEFT (LoRA)	30,078	32	5 1.0
	Contrastive 2	mBART-50-Large	Full FT	20,100	16	5 1.2

Table 1: System configuration summary for Sub-Tasks 1A and 2B.

Task	System	BLEU	chrF
1A	Primary	23.59	52.32
	Contr. 1	18.83	51.93
	Contr. 2	18.55	53.07
2B	Primary	24.60	49.59
	Contr. 1	25.04	49.51
	Contr. 2	23.93	49.56

Table 2: Development set (dev-test) evaluation results. 1A: Primary = OPUS-MT-TC-Big, Contr. 1 = NLLB-200+LoRA, Contr. 2 = mBART-50. 2B: Primary = NLLB-200+LoRA, Contr. 1 = NLLB-200+LoRA+OPUS, Contr. 2 = mBART-50 Full FT.

five metrics, whereas on the dev-test set the two were close on BLEU with NLLB narrowly ahead. This reversal suggests that the aggressive orthographic and numerical normalization applied to the mBART pipeline generalizes better to unseen data than to our dev-test set.

Sub-Task 2B. Consistent with the dev-test trend, on the blind test set Contrastive 1 achieves the highest BLEU (23.2485) and chrF2 (48.0121), narrowly ahead of Primary. This confirms that the additional 10,000 OPUS Arabic–Hindi pairs provide a small but reproducible generalization benefit on both dev-test and genuinely unseen data. Primary retains a slim edge on COMET (Rei et al., 2020) and mBERT, suggesting the two systems trade off differently depending on whether the metric emphasizes surface-level n-gram overlap (BLEU, chrF2) or embedding-based semantic adequacy (COMET, mBERT). Contrastive 2 (mBART-50) remains the weakest system across all five blind-test metrics, consistent with our dev-test findings and with the expectation that a fully fine-tuned generalist model trained on comparatively little constrained data cannot match the parameter-efficient multilingual adaptation of NLLB-200.

11 Discussion and Conclusion

Our submissions to Sub-Tasks 1A and 2B reveal complementary findings about architecture choice in low-resource machine translation. In English–

Arabic, the domain-specialized OPUS-MT-TC-Big model outperforms both generalist multilingual architectures on BLEU, confirming that task-specific pretraining provides a strong inductive bias that generic multilingual representations struggle to match under tight data constraints. The generalist mBART-50, in contrast, achieves the highest chrF when combined with aggressive normalization, indicating that character-level morphological accuracy can benefit from broader cross-lingual transfer when output surface forms are carefully regularized.

In Arabic–Hindi, LoRA fine-tuning of NLLB-200 is the most effective strategy. The shared multilingual pretraining coverage of Arabic and Hindi appears to be well captured by NLLB’s representations, and LoRA’s parameter efficiency limits overfitting on the small official dataset. The modest gain from adding OPUS augmentation data was consistent across both the dev-test set and the blind test, suggesting that in this direction the augmentation corpus was well-aligned with the target domain – in contrast to Sub-Task 1A, where the additional data instead lowered BLEU relative to the domain-specialized OPUS-MT model.

Across both tasks, rule-based orthographic normalization – particularly diacritic removal and Alef standardization – is confirmed as an important pre-processing step for Arabic NMT, consistently reducing vocabulary sparsity and stabilizing training regardless of architecture. Comparing dev-test and official blind test results, overall system rankings are broadly consistent between the two evaluation sets for Sub-Task 2B, while Sub-Task 1A shows a wider gap between Primary and the contrastive systems on the blind test than on dev-test. This underscores the risk of drawing firm conclusions from development-set evaluation alone under low-resource conditions, where dev-test and blind test distributions may diverge even when drawn from related sources.

Task	System	BLEU	chrF2	TER	COMET	mBERT
1A	Primary	14.0772	49.1256	75.7463	82.2053	83.1776
	Contrastive 1	5.7452	40.4801	96.9625	76.8072	79.4595
	Contrastive 2	7.0269	40.9526	85.8900	78.2673	79.6415
2B	Primary	22.5518	47.6432	68.2783	68.2350	83.5656
	Contrastive 1	23.2485	48.0121	68.2865	68.0563	83.4639
	Contrastive 2	22.2600	47.5587	69.2859	67.9737	83.3759

Table 3: Official blind test results for Sub-Tasks 1A and 2B. 1A: Primary = OPUS-MT-TC-Big, Contrastive 1 = NLLB-200+LoRA, Contrastive 2 = mBART-50. 2B: Primary = NLLB-200+LoRA (official data only), Contrastive 1 = NLLB-200+LoRA+OPUS, Contrastive 2 = mBART-50 Full FT.

Limitations

Our study is limited to the models, preprocessing strategies, and data augmentation methods explored in this work. In addition, our evaluation relies solely on the automatic metrics provided by the shared task, without complementary human evaluation or detailed error analysis.

Acknowledgments

We thank the WMT26 Low-Resource Arabic-Asian Translation shared task organizers for the official datasets and for organizing the shared task.

References

- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.
- 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.
- Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the*

40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

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.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, 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 (LREC’12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Jörg Tiedemann and Santhosh Thottingal. 2020. **OPUS-MT – building open translation services for the world**. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.