

NLP-IIT Patna at WMT 2026: Corpus-Driven Adaptation of Multilingual MT Models for Low-Resource Arabic–Asian Machine Translation

Pushkar Sharma*, Mo Ahtasam*, Kshetrimayum Boynao Singh

Deepak Kumar and Asif Ekbal

Department of Computer Science and Engineering

Indian Institute of Technology Patna, India

{pushkarrokhel, ahtisham5736, boynfrancis, deepakkumar1538, asif.ekbal}@gmail.com

Abstract

We investigate low-resource Arabic-Asian machine translation across six Arabic-centric translation directions: Arabic↔English, Arabic↔Hindi, and Arabic↔Urdu. Fine-tuned MADLAD-400 ranks first in four of the six directions on the official primary leaderboard and places within the top three in the remaining two. We fine-tune three pre-trained multilingual models MADLAD-400, NLLB-200, and GemmaX2 and compare them with their zero-shot counterparts using BLEU, chrF2++, COMET-22, and TER. Across all language pairs, supervised fine-tuning consistently outperforms zero-shot inference. Among the evaluated models, fine-tuned MADLAD-400 achieves the strongest overall performance, NLLB-200 delivers competitive results, and GemmaX2-28-9B exhibits the largest relative improvement over its zero-shot baseline. These findings demonstrate the effectiveness of supervised adaptation for low-resource Arabic-Asian machine translation and highlight MADLAD-400 as the most robust model across the evaluated translation directions.

1 Introduction

This work describes our submission to the WMT 2026 Shared Task on Low-Resource Arabic-Asian Machine Translation (Laskar et al., 2026). Machine translation (MT) is essential for multilingual communication, particularly for low-resource language pairs where parallel data remain scarce. Arabic-South Asian MT is an important yet under-explored setting despite its practical significance. Arabic-speaking countries in the Gulf Cooperation Council (GCC) host approximately 35 million foreign workers, many from South Asia, who

rely on translation for healthcare, legal, employment, and public services (Chughtai, 2026). However, Arabic-Hindi and Arabic-Urdu translation remain under-resourced due to the limited availability of parallel corpora (Poria and Huang, 2025). The WMT 2026 Low-Resource Arabic-Asian-Machine Translation Shared Task (Laskar et al., 2026) provides a benchmark for six translation directions: Arabic↔English, Arabic↔Hindi, and Arabic↔Urdu. These language pairs present additional challenges arising from Arabic’s rich morphology, orthographic ambiguity (Alayba, 2025), and flexible word order, as well as the distinct writing systems used across Arabic, Hindi, and Urdu.

Prior work on Arabic MT has documented the broader linguistic and technical challenges posed by Arabic, while our analysis of the shared-task corpus shows that these challenges are accompanied by pronounced lexical and distributional differences: Arabic has a larger vocabulary, greater lexical diversity, higher evaluation OOV rates, and weaker train-evaluation lexical overlap than the other languages in the corpus. In addition, the limited overlap among the Perso-Arabic, Devanagari, and Latin scripts reduce cross-lingual lexical sharing in multilingual models. Together, these observations identify differences in the data that may affect translation adaptation, and they guide the construction and preprocessing of our training data; model-specific optimization strategies are then selected according to the architecture and scale of each pretrained model. We refer to this approach as corpus-driven adaptation.

We fine-tune three pretrained multilingual models: NLLB-200 (NLLB Team et al., 2022), MADLAD-400 (Kudugunta et al., 2023), and GemmaX2-28-9B (Cui et al., 2025). We compare supervised fine-tuning with zero-shot inference across all six translation directions using standard automatic evaluation metrics. Fine-tuned models consistently outperform their zero-shot counter-

*Equal contribution.

Code and resources are available at github.com/pushkar-7/Low-Resource-Arabic-Asian-MT.

Language Pair	Training	Validation	Testing
Arabic-English	21,000	1,000	1,000
Arabic-Hindi	20,100	1,000	1,000
Arabic-Urdu	20,300	1,000	1,000

Table 1: Corpus scale and alignment statistics for the Arabic-centric bidirectional parallel datasets. All language pairs exhibit strict source-target alignment, with comparable train, development, and test splits.

parts, with MADLAD-400 achieving the best overall performance by ranking first in four translation directions and within the top three in the remaining two.

2 Data and Corpus Analysis

We analyze the official WMT 2026 Low-Resource Arabic-Asian MT dataset (Laskar et al., 2026), comprising three bidirectional language pairs: Arabic $\leftrightarrow$ English, Arabic $\leftrightarrow$ Hindi, and Arabic $\leftrightarrow$ Urdu. As shown in Table 1, each pair contains approximately 20-21K training sentence pairs and with 1K validation and 1K test pairs, and strict sentence-level alignment across all splits. Since the shared task is Arabic-centric, the same Arabic sentences are aligned with multiple target languages, resulting in an aggregated Arabic corpus with 65.66% cross-pair duplication, while duplication within individual corpora remains negligible (0-0.05%). We distinguish the official blind-test leaderboard results 4 from our devtest analysis 5. The former determines our shared-task ranking, whereas the latter is used to compare zero-shot and fine-tuned systems under a common evaluation set.

Corpus statistics (Table 2) further show that Arabic is the strongest lexical and distributional outlier among the languages represented in the shared-task data. Compared with Hindi and Urdu, Arabic has a shorter average sentence length (26.6 vs. 38.3 and 41.4 tokens, respectively), the largest vocabulary (98,062 unique tokens), the highest observed Type-Token Ratio (TTR = 0.1757), and the highest OOV rate (10.9% vs. $\sim$ 2%). We define OOV rate as the proportion of whitespace-delimited evaluation tokens that are absent from the corresponding training corpus vocabulary. Arabic also exhibits lower cross-split lexical overlap, with trigram overlap of only 9-12%, indicating weaker lexical overlap between training and evaluation data. All statistics are computed using whitespace tokenization for language-agnostic comparison.

3 System Description

We evaluate three architecturally distinct multilingual paradigms on the bidirectional Arabic-Asian translation task, selected under two constraints: each system must officially support all four languages, and each must be small enough to train under our compute budget. NLLB-200 (NLLB Team et al., 2022) is a dense encoder-decoder built specifically for many-to-many translation across 200 languages, with strong coverage of low-resource and non-Latin scripts; we use the 3.3B dense variant. MADLAD-400 (Kudugunta et al., 2023) is also an encoder-decoder, but represents the web-scale end of that family: a T5-style model trained on 250 billion tokens spanning more than 450 languages, drawn from a manually audited document-level corpus. GemmaX2-28-9B (Cui et al., 2025) is a decoder-only LLM, continually pretrained from Gemma2-9B on monolingual and parallel data in 28 languages and then instruction-tuned for translation, so the task is framed as instruction following rather than sequence-to-sequence transduction. Because all three declare support for Arabic, English, Hindi, and Urdu, no direction we evaluate falls outside a model’s stated coverage. Each system is evaluated under both zero-shot and direction-specific fine-tuned configurations across all six translation pathways (En $\leftrightarrow$ Ar, Hi $\leftrightarrow$ Ar, Ur $\leftrightarrow$ Ar). We note that GemmaX2’s pre-training draws partly on MADLAD-400, so the three systems are not fully independent in their data.

3.1 Models and Checkpoints

We fine-tune three pretrained multilingual models: NLLB-200 3.3B, MADLAD-400 10B, and GemmaX2-28-9B. Each model is fine-tuned independently for every translation direction. Due to differences in model scale, architecture, and computational requirements, we use model-specific post-training adaptation strategies. NLLB is fully fine-tuned, updating all model parameters during task-specific adaptation, whereas MADLAD-400 and GemmaX2 are adapted using parameter-efficient fine-tuning with Low-Rank Adaptation (LoRA)(Hu et al., 2022) and Quantized Low-Rank

<https://huggingface.co/facebook/nllb-200-3.3b>

<https://huggingface.co/google/madlad400-10b-mt>

<https://huggingface.co/ModelSpace/GemmaX2-28-9B-v0.1>

Language	Sentence Pairs	Mean Length	Unique Token	Type-Token Ratio	OOV (%)	Duplication (%)
Arabic	61,400	26.58	98,062	0.1757	10.9	0.069
English	21,000	23.45	45,234	0.0868	4.6	0.030
Hindi	20,100	38.33	58,233	0.0460	2.0	0.040
Urdu	20,300	41.42	32,960	0.0392	2.1	0.030

Table 2: Corpus-level lexical statistics highlight Arabic as the strongest distributional outlier. Arabic exhibits a higher unique-token vocabulary than English, elevated Out-of-Vocabulary (OOV) rates (10.9%), and a Type-Token Ratio (TTR=0.1757). Hindi and Urdu show comparable, low OOV rates ($\approx 2.0\%$), indicating higher vocabulary coverage relative to the training corpus size. TTR = Type-Token Ratio; OOV = Out-of-Vocabulary rate.

Adaptation (QLoRA)(Dettmers et al., 2023), respectively. This design is motivated by prior work showing that parameter-efficient adaptation can achieve competitive translation performance while updating substantially fewer parameters in both multilingual NMT (Bapna and Firat, 2019; Üstün and Stickland, 2022; Singh et al., 2025b) and LLM-based machine translation (Alves et al., 2023; Singh et al., 2026b, 2025a, 2026a; Ahtasam et al., 2026a,b). Prior studies further demonstrate that the effectiveness and computational benefits of parameter-efficient adaptation depend on model scale and the available parameter budget (Üstün and Stickland, 2022). Accordingly, full-parameter fine-tuning is used for NLLB within our available training setup, while parameter-efficient adaptation is used for the larger MADLAD-400 and GemmaX2 models to reduce training memory and parameter update costs. The adaptation procedures are therefore not identical; all systems nevertheless use the same task formulation, training data, and evaluation protocol, with the optimization strategy selected according to each model configuration.

Each translation direction is trained independently with a fixed random seed of 42, gradient checkpointing, a linear learning-rate schedule with warmup, and an effective batch size of 32. Early stopping with a patience of two evaluation rounds is applied using the development set. Checkpoint selection follows the development objective appropriate to each model: COMET-22 for NLLB and development loss for MADLAD-400 and GemmaX2. GemmaX2-28-9B is trained for three epochs because its development performance plateaued after the third epoch, with no measurable improvement from additional training. NLLB is fine-tuned using full-parameter optimization in bf16 precision with the 8-bit AdamW optimizer provided by bitsandbytes. MADLAD-400 is loaded in 8-bit precision with its backbone weights frozen, and LoRA adapters are inserted into the

q , k , v , o , wi_0 , wi_1 , and wo projection layers. GemmaX2 is loaded using 4-bit NF4 quantization with double quantization and bf16 computation, with LoRA adapters applied to the q_proj , k_proj , v_proj , o_proj , $gate_proj$, up_proj , and $down_proj$ layers. Training is performed using the trl SFTTrainer with right padding, and cross-entropy loss is computed over the complete prompt-source-target sequence. All experiments are conducted on a single NVIDIA RTX 6000 Ada GPU with 49 GB of GPU memory. Table 3 summarizes the model checkpoints, adaptation strategies, and training configurations.

3.2 Preprocessing and Tokenization

No heuristic normalisation was applied prior to tokenisation. Conventional Arabic normalisation rules unifying the $\aleph$ (*alef*), $\aleph^a$ (*alef hamza above*), $\aleph_b$ (*alef hamza below*), and $\aleph^m$ (*alef madda*), merging ya (*ya*) with maqsura (*alef maqsura*), and stripping diacritics (Habash, 2010) are designed for Arabic and unsafe for Urdu. Urdu distinguishes several characters that these rules would collapse, including choti ye (*choti ye*), bari ye (*bari ye*), and gol he (*gol he*), along with the retroflex consonants tte (*tte*), ddal (*ddal*), and rre (*rre*); merging them changes the word. A second reason is practical: each model’s subword vocabulary was learned over unnormalised text(Kudo and Richardson, 2018), so normalising the input first would split words into subwords the model never saw during pretraining. After decoding, we converted both the system outputs and the references to Unicode NFC. This ensures that two strings which look identical but use different underlying codepoints are not counted as errors, since neither the character-level nor the token-level metric normalises its inputs by default.

3.3 Inference and Evaluation Protocol

To ensure fair comparison, all systems were decoded using deterministic beam search, repetition

Hyperparameter	NLLB-200	MADLAD-400	GemmaX2-28-9B
Architecture	Encoder-Decoder	Encoder-Decoder (T5)	Decoder-only LLM
Size	3.3B	10B	9B
Adaptation	Full FT	LoRA	QLoRA
Precision / Quantization	bf16	8-bit base	4-bit NF4
Epochs	5	5	3
Per-device batch size	16	8	4
Gradient accumulation	2	4	8
Effective batch size	32	32	32
Learning rate	5×10^{-5}	5×10^{-5}	2×10^{-4}
Warmup ratio	0.10	0.10	0.05
Weight decay	0.01	0.01	0.01
LoRA (r/α /dropout)	-	16 / 32 / 0.05	16 / 32 / 0.05
Optimizer	8-bit AdamW	AdamW	AdamW

Table 3: Training hyperparameters used for the three multilingual translation systems

control, and length penalties to reduce hallucinations and excessive output expansion.

Translation quality was assessed using four complementary automatic metrics: BLEU (Papineni et al., 2002), COMET-22 (Rei et al., 2022), ChrF2++ (Popović, 2017), and Translation Edit Rate (TER) (Snover et al., 2006). While BLEU measures n -gram precision, it penalises valid morphological variants and is known to underperform for morphologically rich target languages such as Arabic (Beseiso et al., 2022). ChrF2++ computes character- and word-level n -gram F-scores and is widely adopted as the standard lightweight metric for such languages precisely because it captures inflectional variation without the high-resource dependency of neural metrics (Popović, 2017). COMET-22 provides neural semantic adequacy scores, while TER measures the post-editing effort required to correct the system output. This four-metric suite enables cross-directional comparison that is robust to both Arabic’s morphological complexity and Urdu and Hindi’s agglutinative character-level behaviour. This unified evaluation framework enables systematic analysis of how corpus characteristics and architectural design jointly influence Arabic-centric multilingual translation performance under both zero-shot and supervised adaptation settings

4 Experimental Results

4.1 Main Results

We participated in six of the ten sub-tasks of the shared task, covering the Arabic-English, Arabic-Hindi, and Arabic-Urdu directions (Sub-Tasks 1A, 1B, 1E, 2A, 2B, and 2E). We submitted MADLAD-400 as our primary system and NLLB-200 and GemmaX2-28-9B as contrastive systems; the or-

ganizers evaluate primary and contrastive submissions on separate leaderboards. All systems were scored by the organizers on the official blind challenge test set using automatic metrics (BLEU, ChrF, TER, COMET, and BERT-based measures); the complete per-system evaluation matrices are published on the task page. Table 4 reports our official per-direction rankings. On the primary leaderboard, our MADLAD-400 system placed first in four of the six directions (En→Ar, Ar→Hi, Hi→Ar, Ur→Ar), second in Ar→En, and third in Ar→Ur. On the contrastive leaderboard, NLLB-200 placed first for Ar→Hi and Ar→Ur, and GemmaX2 placed second in five of the six directions.

4.2 Zero-shot vs. Fine-tuning

Table 5 compares zero-shot and fine-tuned performance on the devtest split. Fine-tuning improves every system on every metric in every direction. The gains are moderate for the two encoder-decoder models, which are already usable zero-shot: on macro-average, MADLAD gains +9.3 BLEU, +7.2 ChrF2++, +0.023 COMET-22, and −10.6 TER, with NLLB slightly smaller. For GemmaX2-28-9B the gains are of a different kind: zero-shot output is adequate only for Ar→En, while the remaining five directions produce near-zero ChrF2++ with TER above 100, indicating largely off-target output; fine-tuning recovers all five, lifting the macro-average by +21.4 BLEU and +0.251 COMET-22.

Gains are not uniform across directions. Surface-metric improvements concentrate in the into-Arabic directions—for the primary system, TER falls by 14.4 (En→Ar), 13.1 (Hi→Ar),

https://www2.statmt.org/wmt26/arabic-asian-mt-task.html#_evaluation

Languages Direction	System	BLEU $\uparrow$	ChrF2 $\uparrow$	TER $\downarrow$	COMET $\uparrow$	mBERT $\uparrow$	Rank
en $\rightarrow$ ar	MADLAD-400 (primary)	22.4455	55.1675	67.2558	84.7356	85.6693	1
	NLLB-200 (contr-1)	19.2107	52.9859	71.9966	83.7861	84.6091	4
	GemmaX2 (contr-2)	20.3824	53.6442	70.0626	84.1867	85.0212	3
hi $\rightarrow$ ar	MADLAD-400 (primary)	19.8538	53.1681	70.4659	84.2953	84.8184	1
	NLLB-200 (contr-1)	16.6289	50.5424	75.0421	83.0640	83.7877	3
	GemmaX2 (contr-2)	17.6224	51.3379	73.5956	83.6086	84.1502	2
ur $\rightarrow$ ar	MADLAD-400 (primary)	18.9970	52.2288	71.8622	83.8677	84.4310	1
	NLLB-200 (contr-1)	15.8666	49.6695	76.0974	82.7522	83.5203	3
	GemmaX2 (contr-2)	17.0747	50.7366	74.5466	83.4283	83.9132	2
ar $\rightarrow$ en	MADLAD-400 (primary)	33.6687	59.6708	57.3039	82.0053	83.1130	2
	NLLB-200 (contr-1)	30.3160	57.0117	59.7033	82.0480	82.0480	4
	GemmaX2 (contr-2)	32.0481	60.0195	60.1451	82.3304	82.3304	2
ar $\rightarrow$ hi	MADLAD-400 (primary)	29.2775	54.7254	62.5191	72.3411	85.4462	1
	NLLB-200 (contr-1)	27.8738	53.4990	63.3534	71.7645	85.1154	1
	GemmaX2 (contr-2)	26.3278	52.5436	68.2332	71.4596	84.3758	2
ar $\rightarrow$ ur	MADLAD-400 (primary)	22.4936	48.6969	68.1138	75.5240	82.7617	3
	NLLB-200 (contr-1)	25.6828	51.0409	65.1504	76.3313	83.5296	1
	GemmaX2 (contr-2)	24.6513	50.1675	69.5522	76.0019	82.9466	2

Table 4: Official WMT 2026 Challenge Test results across the six evaluated Arabic-Asian translation directions. BLEU, ChrF2, COMET, and mBERT are higher-is-better ($\uparrow$), while TER is lower-is-better ($\downarrow$). Rank denotes the official ranking of each system in the corresponding direction. **MADLAD-400** is our primary submission, while NLLB-200 and GemmaX2 are contrastive submissions.

and 12.2 (Ur $\rightarrow$ Ar), against 10.1 for Ar $\rightarrow$ En-consistent with supervised adaptation, reducing the surface-form mismatch identified for Arabic in Section 2. A weakness remains after fine-tuning: into-Arabic BLEU stays low (20.8-24.3 for the primary system) relative to the out-of-Arabic directions (28.1-36.4), even where COMET-22 is high, which we examine next.

4.3 Cross-model Comparison

The two encoder-decoder models and the decoder-only model differ most at the starting point, not the end point. Zero-shot, NLLB and MADLAD translate all six directions; GemmaX2-28-9B translates only Ar $\rightarrow$ En, despite its translation-oriented instruction tuning. After per-direction fine-tuning the three systems converge to within 0.008 COMET-22 on macro-average (0.835/0.829/0.827), so the blind-test separation between them is small and direction-dependent. MADLAD is the primary submission because it leads all four macro-averaged metrics after fine-tuning and is best or tied-best on most individual directions. NLLB remains a strong baseline: with a third of the parameters and full fine-tuning, it trails the primary system by 0.006 COMET-22 on average and wins its track in the

two directions where the primary system is weakest. GemmaX2-28-9B shows the largest relative improvement precisely. The pattern is consistent with a substantial mismatch in translation behavior or target-language control under zero-shot prompting, which is largely resolved by supervised adaptation. With 0.60% trainable parameters, it reaches within 0.008 COMET-22 of the primary system, suggesting that supervised adaptation substantially improves target-language control and translation behavior for this model. Adaptation efficiency is thus high for both parameter-efficient systems: under 1% of parameters recover or exceed parity with full fine-tuning of a smaller model.

Metric stability separates the targets more than the systems. For all three systems, into-Arabic directions score high on COMET-22 but low on BLEU with elevated TER, whereas out-of-Arabic directions show the reverse. This is the pattern predicted by the corpus analysis in Section 2: Arabic’s larger vocabulary, higher lexical diversity, and $\sim 11\%$ evaluation out-of-vocabulary rate mean that adequate translations are frequently realized with different surface forms, which n -gram metrics penalize and COMET-22 does not. The residual weakness of all systems is therefore

concentrated where the data analysis located it-on the Arabic-target side-rather than in any single architecture.

5 Conclusion

We described the NLP-IIT-Patna submissions to the WMT 2026 Low-Resource Arabic-Asian shared task (Laskar et al., 2026): three pretrained multilingual models-NLLB-200, MADLAD-400, and GemmaX2-28-9B-fine-tuned per direction on the official data under a shared pre-processing, decoding, and evaluation pipeline. Our primary MADLAD-400 system placed first on the primary leaderboard in four of the six directions we entered, and our contrastive systems led or placed second in their track across most directions. The empirical picture is consistent across all three architectures: supervised adaptation closes the behavioral gap of decoder-only models and specializes already-capable encoder-decoders; after adaptation, the systems converge, and the remaining difficulty is organized by the target language, concentrated on where the corpus analysis located Arabic’s vocabulary sparsity and train-evaluation shift. For low-resource Arabic-centric translation, careful adaptation of existing multilingual models, rather than new architectures, were sufficient for state-of-the-art results, provided evaluation accounts for Arabic’s surface-form sparsity when interpreting n -gram metrics. Extending the pipeline with monolingual Arabic data and to the remaining task languages is the natural next step.

Acknowledgment

The authors sincerely thank the COIL-D (Centre of Indian Language Data) Project under Bhashini, funded by the Ministry of Electronics and Information Technology (MeitY), Government of India, for providing the computational resources and infrastructure that enabled this research

References

- Mo Ahtasam, Kshetrimayum Boynao Singh, Jamaluddin, Deepak Kumar, and Asif Ekbal. 2026a. [Toward speaker-preserving arabic speech: A bilingual resource and comparative evaluation of cross-lingual voice cloning](#). In *The 6th Muslims in ML (MusIML) Workshop at ICML 2026*.
- Mo Ahtasam, Jamal uddin, and Mohammad Nadeem. 2026b. [Balancing linguistic intelligibility and speaker identity in zero-shot cross-lingual voice cloning](#). In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, pages 103–110, San Diego, USA (in-person and online). Association for Computational Linguistics.
- Abdulaziz M. Alayba. 2025. [Arabic natural language processing \(NLP\): A comprehensive review of challenges, techniques, and emerging trends](#). *Computers*, 14(11):497.
- Duarte M Alves, Nuno M Guerreiro, João Alves, José Pombal, Ricardo Rei, José GC de Souza, Pierre Colombo, and André FT Martins. 2023. [Steering large language models for machine translation with finetuning and in-context learning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11127–11148, Singapore. Association for Computational Linguistics.
- Ankur Bapna and Orhan Firat. 2019. [Simple, scalable adaptation for neural machine translation](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1538–1548, Hong Kong, China. Association for Computational Linguistics.
- Majdi Beseiso, Samiksha Tripathi, Bashar Al-Shboul, and Renad Aljadid. 2022. [Semantics based English-Arabic machine translation evaluation](#). *Indonesian Journal of Electrical Engineering and Computer Science*, 27(1):189–197.
- Alia Chughtai. 2026. [Where do the 35 million foreigners living in the GCC come from?](#) Al Jazeera. Published 9 March 2026.
- Menglong Cui, Pengzhi Gao, Wei Liu, Jian Luan, and Bin Wang. 2025. [Multilingual machine translation with open large language models at practical scale: An empirical study](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5420–5443, Albuquerque, New Mexico. Association for Computational Linguistics.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. [QLoRA: Efficient finetuning of quantized LLMs](#). In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 10088–10115.
- Nizar Y Habash. 2010. *Introduction to Arabic natural language processing*. Morgan & Claypool Publishers.
- 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 *The Tenth International Conference on Learning Representations (ICLR)*.

- Taku Kudo and John Richardson. 2018. [Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.
- Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. [MADLAD-400: A multilingual and document-level large audited dataset](#). In *Advances in Neural Information Processing Systems*, volume 36.
- 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.
- 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. 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.
- Sampoorna Poria and Xiaolei Huang. 2025. [Bhaasha, bhāṣā, zaban: A survey for low-resourced languages in south asia – current stage and challenges](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1386–1406, Suzhou, China. Association for Computational Linguistics.
- Ricardo Rei, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT Martins. 2022. [Comet-22: Unbabel-ist 2022 submission for the metrics shared task](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Kshetrimayum Boynao Singh, Deepak Kumar, Ramakrishna Appicharla, Asif Ekbal, and Partha Pakray. 2026a. Leveraging high-fidelity llms for synthetic data generation: A scalable pivot strategy.
- Kshetrimayum Boynao Singh, Deepak Kumar, and Asif Ekbal. 2025a. [Evaluation of LLM for English to Hindi legal domain machine translation systems](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 823–833, Suzhou, China. Association for Computational Linguistics.
- Kshetrimayum Boynao Singh, Deepak Kumar, and Asif Ekbal. 2025b. Instruction-tuned english to bhojpuri neural machine translation using contrastive preference optimization. In *Proceedings of the Tenth Conference on Machine Translation*, pages 638–643.
- Kshetrimayum Boynao Singh, Saksham Singh, Partha Pakray, and Asif Ekbal. 2026b. Reasoning as supportive context for machine translation: A case study on hindi to bengali language pair. In *26th Annual Conference of the European Association for Machine Translation*, page 67.
- 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.
- Ahmet Üstün and Asa Cooper Stickland. 2022. [When does parameter-efficient transfer learning work for machine translation?](#) In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7919–7933, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Input Formatting

Each parallel corpus is read line by line; we verify that the source and target files contain an identical number of sentence pairs and discard any pair in which either side is empty after whitespace stripping. No further text normalization is applied prior to tokenization, and every model uses its native tokenizer together with its own language-signaling mechanism, as illustrated in Figure 1. For NLLB-200 we set the tokenizer source language, encode the target side with `text_target`, and constrain the decoder at generation time with `forced_bos_token_id`, using the identifiers `arb_Arab`, `eng_Latn`, `hin_Deva`, and `urd_Arab`. For MADLAD-400 each source sentence is prefixed with a target-language tag drawn from `<2ar>`, `<2en>`, `<2hi>`, and `<2ur>`, with identical formatting for zero-shot and fine-tuned inference. For GemmaX2-28-9B each input is cast as an instruction-following

NLLB-200

```
tokenizer.src_lang = {src}  
model.generate(..., forced_bos_token_id  
    = tok.convert_tokens_to_ids({tgt}))  
e.g. src=eng_Latn, tgt=urd_Arab
```

MADLAD-400

```
<2tgt> {source}  
e.g. <2ur> This is a test sentence.
```

GemmaX2-28-9B

```
Translate the following text from {src} to  
{tgt}:  
{source}  
Translation:  
(fine-tuning: {target}</eos> appended here)
```

Figure 1: Input formats used for each translation model. Placeholders in braces are replaced at run time; {src} and {tgt} denote the source and target language identifiers in each model’s own convention.

prompt; during fine-tuning the reference translation and the EOS token are appended after Translation:, and the loss is computed over the complete prompt-source-target sequence without masking the instruction tokens. All source sequences are truncated to 512 tokens. After decoding, outputs are normalized with Unicode NFC, and for GemmaX2-28-9B, only the first non-empty generated line is retained.

Direction	Zero-shot (devtest dataset)				Fine-tuned (devtest dataset)			
	COMET	ChrF2++	BLEU	TER↓	COMET	ChrF2++	BLEU	TER↓
<i>NLLB-200 (Contrastive-1)</i>								
Ar→En	0.829	47.5	21.9	66.7	0.842	53.9	29.2	60.3
En→Ar	0.864	45.8	15.8	82.6	0.876	49.0	21.1	71.8
Ar→Hi	0.709	46.1	29.5	67.9	0.735	50.9	35.2	63.9
Hi→Ar	0.847	42.8	13.9	84.3	0.859	46.0	17.8	75.6
Ar→Ur	0.781	43.6	25.1	70.6	0.800	48.6	32.2	65.2
Ur→Ar	0.854	42.5	14.4	81.5	0.865	45.8	17.9	75.7
Avg.	0.814	44.7	20.1	75.6	0.829	49.0	25.6	68.7
<i>MADLAD-400 (Primary)</i>								
Ar→En	0.830	46.5	20.8	67.6	0.848	56.5	32.3	57.5
En→Ar	0.867	46.0	15.7	81.7	0.883	51.1	24.3	67.3
Ar→Hi	0.704	42.2	25.6	70.0	0.742	51.7	36.4	63.3
Hi→Ar	0.855	44.7	14.6	83.6	0.869	48.7	21.6	70.5
Ar→Ur	0.752	35.8	16.9	75.2	0.792	46.3	28.1	68.0
Ur→Ar	0.863	44.1	14.5	84.1	0.876	47.9	20.8	71.9
Avg.	0.812	43.2	18.0	77.0	0.835	50.4	27.3	66.4
<i>GemmaX2-28-9B (Contrastive-2)</i>								
Ar→En	0.831	47.7	21.9	67.4	0.844	56.4	31.3	60.5
En→Ar	0.601	10.5	3.1	122.4	0.874	48.8	21.7	70.5
Ar→Hi	0.442	0.9	0.1	101.9	0.726	49.6	34.2	68.8
Hi→Ar	0.578	0.8	0.1	128.8	0.859	45.8	18.4	75.7
Ar→Ur	0.461	0.6	0.2	101.0	0.794	47.6	31.1	69.2
Ur→Ar	0.541	1.2	0.2	128.2	0.867	45.4	17.6	76.6
Avg.	0.576	10.3	4.3	108.3	0.827	48.9	25.7	70.2

Table 5: Zero-shot and fine-tuned results for all three systems on the devtest split. COMET denotes COMET-22; for TER lower is better. Avg. is the macro-average over the six directions. Official blind-test rankings are given in Table 4.

Direction / Source	GemmaX2	MADLAD-400	NLLB-200
ar→en مليار 86) 5-فيرياري دولار	5. Ferrari (\$86 billion) → Ferrari's value stands at \$86 billion.	5. Ferrari (\$86 billion) → Ferrari (USD 86 Billion)	Five Ferraris worth \$86 billion. → 5-Ferrari (\$86 billion)
en→ar A fleeting thought has entered my mind.	मेरे मन में एक क्षणिक विचार आया। → عابرة. فكرة لي خطرت	فكرة دخلت → عقلي دخلت عابرة فكرة ذهني. إلى عابرة	عابرة فكرة → ذهني دخلت عابرة فكرة ذهني. إلى دخلت
ar→hi هذه تكون أن وعسى الغزأوية. الأحران نهاية	May this be the end of the bitter sorrows. → उम्मीद है कि यह निराशाजनक दुखों का अंत होगा।	यह संभवतः गज़ा के दुखों का अंत होगा। → हो सकता है कि यह गाजा की पीड़ा का अंत हो।	और उम्मीद है कि यह आक्रमणकारी शोक का अंत होगा। → उम्मीद है कि यह आक्रमणकारी शोक का अंत होगा।
hi→ar *سؤدی سورتس سؤدت	*Quoted from Saudi Sports → السعودية الرياضية عن *نقلاً	→ السعودية سبورتس صحيفة من *مقتبس السعودية سبورتس عن *نقلاً	نقلاً * → السعودية الرياضة من *مقتبس السعودية "الرياضة عن
ar→ur الإمارات عن *نقلاً اليوم	*Quoted from Emirates Today → اليوم الإمارات ماخذ:	→ گیا کیا نقل سے ٹوڈ الإمارات □□□□□□□□ □□□□□□ ہے۔ گیا کیا نقل سے	آج حمل و نقل کی امارات عرب متحدہ سے امارات عرب متحدہ آج اقتباس۔
ur→ar ایک میں ذہن میرے آیا۔ خیال لمحہ لمحہ	An idea came to my mind in an instant. → تدریجاً. الفكرة جاءتني	واحدة فكرة جاءت → فکرت بلحظة لحظة ذهني. إلى لحظة كل في	ذهني في فكرة جاءت → فكرة. أتني فجأة، الأخرى. تلو لحظة

Table 6: Qualitative comparison of zero-shot (orange) and fine-tuned (green) outputs across the three systems. Source segments are shown beneath each translation direction.