

# Findings of the WMT 2026 Shared Task on Low-Resource Arabic-Asian Language Translation

Sahinur Rahman Laskar<sup>1\*</sup>, Firoj Alam<sup>2</sup>, Bishwaraj Paul<sup>3</sup>, Irfan Ahmad<sup>4</sup>  
Maya Silvi Lydia<sup>5</sup>, Pankaj Dadure<sup>1</sup>

<sup>1</sup>UPES, Dehradun, India, <sup>2</sup>Qatar Computing Research Institute, Doha, Qatar,

<sup>3</sup>Bahash-AI, Silchar, India, <sup>4</sup>King Fahd University of Petroleum and Minerals, Saudi Arabia,

<sup>5</sup>Universitas Sumatera Utara, Medan, Indonesia

sahinurlaskar.nits@gmail.com

## Abstract

We present the findings of the WMT 2026 Shared Task on Low-Resource Arabic-Asian Language Translation, which evaluated bidirectional translation between Arabic and English, Hindi, Bangla, Indonesian, and Urdu. We release *AraAsian1.0*, a news-domain parallel corpus with 20.1K-21.0K training pairs per language pair, and evaluate submissions on ten translation directions using BLEU, ChrF2, TER, COMET, and BERT-based metrics. Of 23 registered teams, 13 submitted systems, producing 70 primary and 65 contrastive runs. Fine-tuned multilingual MT models remained the strongest overall: MADLAD-based systems won four primary directions, while target-aware pivoting, joint multilingual adaptation, and MBR-based ensembling led the remaining tracks. The best contrastive run exceeded the best primary run in only three of ten directions, and each gain was below 0.2 BLEU, indicating that additional data, prompting, or multi-pass refinement did not provide consistent improvements. Bangla had the lowest winning BLEU in both directions, despite comparable training-set size. BLEU and COMET selected the same winner in nine of ten primary results but only six of ten contrastive results, with the largest disagreements occurring for LLM-based and reranked outputs. These results emphasize domain-matched adaptation, target-aware transfer, and multi-metric reporting in low-resource Arabic-Asian translation. We made the *AraAsian1.0* dataset available for the community.<sup>1</sup>

## 1 Introduction

Machine translation (MT) has advanced rapidly through multilingual pretraining, transfer learning, and instruction-tuned language models (Team et al., 2022; Hendy et al., 2023). These advances, however, remain unevenly distributed: systems are

typically strongest for language pairs with abundant parallel data and well-developed evaluation resources (Joshi et al., 2020; Haddow et al., 2022). When direct bitext is limited, translation quality becomes more sensitive to pretrained language coverage, domain mismatch, tokenization, and the quality of synthetic or mined data (Koehn and Knowles, 2017).

These challenges are particularly relevant for Arabic paired with South and Southeast Asian languages. Arabic has rich inflectional morphology and a right-to-left script (Farghaly and Shaalan, 2009; Darwish et al., 2020), while Hindi, Bangla, Urdu, and Indonesian span different language families and writing systems (Kakwani et al., 2020; Aji et al., 2022). Although large multilingual models cover many of these languages, direct Arabic-Asian parallel corpora and standardized bidirectional evaluations remain limited. Existing resources and benchmarks often center English as one side of the pair (Ramesh et al., 2022; Goyal et al., 2022), leaving less evidence about how modern MT strategies transfer to Arabic-centric settings.

The WMT 2026 Shared Task on Low-Resource Arabic-Asian Language Translation addresses this gap through ten directions: translation into Arabic from English, Hindi, Bangla, Indonesian, and Urdu, and translation from Arabic into the same five languages. The task introduces *AraAsian1.0*, a parallel news corpus with approximately 20K training pairs per language pair, and permits both primary submissions based on official data and optional contrastive submissions that may use additional resources. Thirteen teams submitted systems spanning conventional encoder-decoder MT, multilingual joint training, parameter-efficient adaptation, English-pivot cascades, data augmentation, LLM prompting, retrieval-augmented refinement, and minimum Bayes risk (MBR) decoding.

This findings paper reports the task design,

<sup>1</sup><https://www2.statmt.org/wmt26/arabic-asian-mt-task.html>

dataset, participating systems, and complete automatic evaluation. More importantly, it goes beyond a winner-by-winner summary to examine four questions: (i) which language directions remain most difficult under comparable data budgets; (ii) which teams and architectural choices are consistently effective across directions; (iii) whether contrastive systems provide reliable gains over primary systems; and (iv) how stable the rankings are across lexical and neural metrics. The results show that strong supervised adaptation and target-aware transfer were more dependable than simply adding external data or an LLM refinement stage, while metric disagreements increased for contrastive and generative systems.

## 2 Related Work

Low-resource MT has traditionally relied on transfer from multilingual models, monolingual-data augmentation, pivot translation, and careful data selection. Multilingual pretraining enables parameter sharing across related and unrelated languages, but its benefits depend on language coverage, vocabulary allocation, and the amount and quality of adaptation data (Haddow et al., 2022; Team et al., 2022). Back-translation remains a standard way to convert monolingual text into synthetic bitext (Sennrich et al., 2016), while pivoting can connect a low-resource pair through a better-resourced intermediate language. These methods can increase data volume, but noisy generation and domain mismatch may offset their gains.

Arabic-centric MT presents additional challenges. Arabic combines rich morphology, cliticization, orthographic variation, and diglossia, and many available corpora are concentrated in news or Modern Standard Arabic (Farghaly and Shaalan, 2009; Darwish et al., 2020). For South Asian languages, resources such as Samanantar, IndicBART, and IndicTrans2 have substantially improved English–Indic translation and multilingual transfer (Ramesh et al., 2022; Dabre et al., 2022; Gala et al., 2023). Yet English is usually the central source or target, so these resources do not directly establish performance for Arabic–Indic translation. Indonesian has stronger coverage in broad multilingual models, but Arabic–Indonesian evaluation is likewise limited and often embedded within massively multilingual test suites rather than studied as a dedicated pair.

Recent general-purpose LLMs introduce a sec-

ond family of approaches. They can translate through prompting, preserve requested output structures, revise drafts, or select among candidates, sometimes without task-specific training (Hendy et al., 2023). At the same time, purpose-built multilingual MT models remain competitive, particularly when latency, constrained decoding, and low-resource adaptation are important. The participant systems in this shared task expose this comparison directly: they range from supervised MAD-LAD, NLLB, OPUS-MT, and OpenNMT systems to prompted LLMs, retrieval-augmented revision, quality-estimation gates, and MBR decoding over heterogeneous candidate pools.

The present task complements existing multilingual benchmarks by holding the domain and approximate training budget constant across five Arabic-centered pairs and evaluating both directions. This design makes it possible to study not only absolute performance, but also direction-specific transfer, the reliability of external-resource recipes, and agreement between lexical and neural metrics.

## 3 Task Overview

The shared task evaluates Arabic-centric MT under a common low-resource setting. Each of the five language pairs is evaluated in both directions, yielding ten subtasks (Table 1).

| Task 1: Src. → Tgt. (Ar) | Task 2: Src. (Ar) → Tgt. |
|--------------------------|--------------------------|
| En→Ar (1A)               | Ar→En (2A)               |
| Hi→Ar (1B)               | Ar→Hi (2B)               |
| Bn→Ar (1C)               | Ar→Bn (2C)               |
| Id→Ar (1D)               | Ar→Id (2D)               |
| Ur→Ar (1E)               | Ar→Ur (2E)               |

Table 1: The ten translation directions.

### 3.1 Submission Tracks

For each direction, a team could submit one primary and up to two contrastive runs. Primary systems were expected to use the official *AraAsian1.0* parallel data; publicly available pretrained models and monolingual resources were permitted. Contrastive runs could additionally use external parallel data, proprietary synthetic data, or other resources beyond the task set. The released labels are retained throughout this paper. One submitted system did not follow the nominal resource split; we identify this case explicitly when comparing the two tracks.

### 3.2 Evaluation

All runs were evaluated on the 1,500-sentence blind challenge set using five automatic metrics. BLEU measures word-level  $n$ -gram precision (Papineni et al., 2002); ChrF2 is a character  $n$ -gram F-score with recall weighted by  $\beta = 2$ , making it useful for morphologically rich targets (Popović, 2015); and TER measures the edits required to transform a system output into the reference, with lower values indicating better performance (Snover et al., 2006). COMET provides a neural, human-aligned estimate of translation quality (Rei et al., 2020), while BERTScore or multilingual BERT similarity captures contextual overlap (Zhang et al., 2020).

The released leaderboards are ordered by BLEU, which we therefore use for the main ranking summary. We use ChrF2, TER, COMET, and BERT-based scores to assess whether conclusions remain stable under different notions of similarity. Because no human evaluation or significance testing was available, small score differences should be interpreted cautiously.

### 4 Dataset: AraAsian1.0

*AraAsian1.0* contains parallel data for Arabic–English, Arabic–Hindi, Arabic–Bangla, Arabic–Indonesian, and Arabic–Urdu. The corpus was developed in three stages. **(i) Harvesting and alignment:** comparable English, Arabic and Urdu news documents were collected primarily from Al Arabiya,<sup>2</sup> candidate sentence pairs were scored using cross-lingual similarity and margin criteria using LaBSE (Feng et al., 2022) (a multilingual sentence embedding model), and English–Arabic and Urdu–Arabic pairs with an alignment score above 0.80 were retained. **(ii) Multilingual expansion:** the high-confidence English sentences were translated into Hindi, Bangla, and Indonesian, to construct candidate parallel data for the remaining pairs. **(iii) Expert validation:** professional linguists reviewed the alignments, corrected syntactic and translation errors, and removed partial, noisy, or mismatched pairs.

The training sets are closely balanced: the largest contains only 4.5% more sentence pairs than the smallest. Consequently, large performance differences across language pairs are unlikely to be explained by training-set size alone. The blind test is also systematically longer than the training data. Across the five pairs, Arabic test sentences contain

approximately 24–25% more tokens per sentence than their training counterparts, while the increase on the partner-language side is approximately 27–29%. This length shift makes the challenge set more demanding and should be considered when interpreting the absolute scores.

The construction process also defines the scope of the benchmark. The corpus is news-domain rather than conversational or literary, and three of the Asian sides originate from an English-mediated translation stage followed by expert correction. These choices enable controlled, aligned evaluation, however, they may introduce translationese and limit conclusions about other domains.

### 5 Participation and System Landscape

Twenty-three teams registered for the shared task and 13 submitted at least one system. The final evaluation contains 70 primary runs and 65 contrastive runs. Participation was highest for the English and Hindi directions and lowest for Indonesian (Table 3). Contrastive counts refer to runs rather than unique teams, since a team could submit two contrastive systems per direction.

Table 4 summarizes the system landscape; detailed descriptions and affiliations are provided in Appendix A. NLLB was the dominant backbone, appearing in 12 of the 13 team descriptions. The main differentiation therefore came from how teams adapted or used the backbone: six teams explicitly used LoRA or QLoRA, four explored joint or pivot-based transfer, several cleaned or augmented the training data, and four used MBR, quality estimation, retrieval, or LLM-based refinement at inference time.

The submissions span two broad philosophies. One group specialized established multilingual MT models through supervised adaptation, joint training, or pivoting. The other used general-purpose LLMs, retrieval, quality estimation, or hypothesis reranking to alter inference without relying solely on a single fine-tuned decoder. The results in the next section show that both can be competitive, but their behavior differs substantially across directions and metrics.

### 6 Results and Findings

Appendices B and C report all five metrics for every run. Table 5 condenses the BLEU leaderboards and includes the gap between the first- and second-ranked primary systems. The final column com-

<sup>2</sup><https://www.alarabiya.net/>

| Pair              | Train  | Dev   | DevTest | Test  | Train tokens (Ar/X) | Test tokens (Ar/X) |
|-------------------|--------|-------|---------|-------|---------------------|--------------------|
| Arabic–English    | 21,000 | 1,000 | 500     | 1,500 | 558,133 / 670,845   | 49,844 / 62,015    |
| Arabic–Hindi      | 20,100 | 1,000 | 500     | 1,500 | 539,522 / 770,493   | 49,844 / 73,240    |
| Arabic–Bangla     | 20,500 | 1,000 | 500     | 1,500 | 548,223 / 584,527   | 49,844 / 54,590    |
| Arabic–Indonesian | 20,400 | 1,000 | 500     | 1,500 | 545,845 / 605,281   | 49,844 / 56,514    |
| Arabic–Urdu       | 20,300 | 1,000 | 500     | 1,500 | 543,873 / 840,793   | 49,844 / 79,536    |

Table 2: *AraAsian1.0* statistics. X denotes the non-Arabic language. All pairs use 1,000 development, 500 DevTest, and 1,500 blind challenge test sentences; training sizes range from 20.1K to 21.0K pairs.

| Lang.        | Into Arabic P/C | From Arabic P/C |
|--------------|-----------------|-----------------|
| English      | 9 / 12          | 10 / 11         |
| Hindi        | 8 / 7           | 10 / 9          |
| Bangla       | 6 / 4           | 7 / 4           |
| Indonesian   | 4 / 3           | 4 / 3           |
| Urdu         | 6 / 6           | 6 / 6           |
| <b>Total</b> | <b>33 / 32</b>  | <b>37 / 33</b>  |

Table 3: Numbers of primary (P) and contrastive (C) runs.

pares the best contrastive run with the best primary run for the same direction.

## 6.1 Analysis

For each primary direction, we record the winner, the top-two BLEU gap, and each team’s rank. To compare teams across different sizes, we additionally use normalized rank,  $(r - 1)/(n - 1)$ , where  $r$  is the system rank and  $n$  is the number of primary submissions; lower values are better. For a matched primary–contrastive comparison, we select the highest-BLEU contrastive run from the same team and direction, then subtract the team’s primary score. These analyses are descriptive: the shared task did not include significance tests or human judgments.

## 6.2 Primary Detail

Tables 6 and 7 show the top three primary systems in every direction. These compact tables expose cases that a winner-only summary hides. For example, Arabic→English has three systems above 32 BLEU, whereas Arabic→Indonesian has a large separation between first and second place. The tables also show that high BLEU generally coincides with high ChrF2 and BERT-based similarity, while COMET occasionally favors a different system.

## 6.3 Overall Patterns Across Directions

For translation into Arabic, all systems generate the same target language, making cross-source BLEU comparisons more interpretable than comparisons across different target languages. English→Arabic obtained the highest winning score (22.45), followed by Hindi (19.85), Urdu (19.00), Indonesian (18.72), and Bangla (17.93). The 4.52-BLEU range

is substantial given that the training sets differ by fewer than 1,000 sentence pairs. Source-language coverage and transfer quality therefore appear to matter more than the small differences in corpus size.

For translation from Arabic, target-language tokenization and morphology limit direct cross-language comparison. Nevertheless, Arabic→Bangla is a clear low-scoring case under the official BLEU ranking: its winning score is 17.10, compared with 25.89–33.83 for the other four target languages. Bangla consequently has the lowest winning BLEU in both directions. By contrast, Arabic→Indonesian reaches 30.51 BLEU, 11.80 points above Indonesian→Arabic. Four of five pairs have higher BLEU in the from-Arabic direction; Bangla is the exception. We treat this as a descriptive asymmetry rather than evidence that translation from Arabic is intrinsically easier, because the target languages and their tokenization differ.

Most primary submissions were competitive. The median gap between the top two systems is 0.74 BLEU, and six of ten gaps are below one point. Arabic→English is especially close: OpenBracket leads NLP-IIT Patna by only 0.16 BLEU. The main exception is Arabic→Indonesian, where NERDS\_NITS leads the runner-up by 5.84 BLEU, indicating a decisive advantage for its direction-specific QLoRA-tuned NLLB system. Bangla presents a different pattern: both directions are low-scoring, yet Arabic→Bangla has only a 0.51-point top-two gap, suggesting a shared bottleneck rather than one isolated weak submission.

## 6.4 Participation Breadth and Consistency

Table 8 separates coverage from rank. OpenBracket and NLP-IIT Patna have the best average normalized ranks, but they entered only two and six primary directions, respectively. Among the four teams covering all ten directions, NERDS\_NITS is the most consistent: it records two wins, six top-two finishes, and an average normalized rank of 0.24. FARABI also wins twice, but its mean rank

| Team          | P/C dirs. | Models and main strategy                                                                                        |
|---------------|-----------|-----------------------------------------------------------------------------------------------------------------|
| OpenBracket   | 2/2       | MBR over NLLB, X-ALMA, ALLAM, and OPUS-MT candidates; QLoRA ALLAM and COMET-Kiwi-pruned sampling.               |
| BVSLP         | 10/0      | NLLB fine-tuning with LLM-assisted bitext cleaning and back-translation.                                        |
| UPES_NLP      | 3/10      | Fine-tuned NLLB and NLLB-MoE primary systems; zero-shot GPT-5-mini with structured output for contrastive runs. |
| NERDS_NITS    | 10/0      | English-pivot NLLB+IndicTrans2 cascades for Indic targets; QLoRA NLLB for Indonesian; OPUS models for English.  |
| MTG-LIRA      | 2/2       | OpenNMT Transformer for Arabic-Hindi; BPCC/OPUS pivot-based augmentation before task fine-tuning.               |
| NITM AI Lab   | 2/2       | OPUS, NLLB, and mBART models with full or LoRA fine-tuning, normalization, and optional OPUS augmentation.      |
| MAGADIR       | 2/2       | Cross-model MBR using NLLB and LLM candidates, with entity-recall constraints and neural reranking.             |
| Code & Corpus | 4/0       | One jointly fine-tuned NLLB model for Arabic-Hindi and Arabic-Bangla in four directions.                        |
| IndicMT       | 10/0      | Direction-specific NLLB-600M fine-tuning after filtering and adding 500 synthetic pairs per direction.          |
| SLPG_FJWU     | 3/3       | NLLB-3.3B with LEALLA-filtered CC-100 and News Commentary data; official-data NLLB-600M alternative.            |
| FARABI        | 10/10     | Joint NLLB-3.3B with language-pair controllers into Arabic; Qwen2.5 LoRA and optional RAG from Arabic.          |
| NCI-MT        | 6/6       | LoRA NLLB primary system; COMET-Kiwi-triggered retrieval and Gemma refinement for low-confidence outputs.       |
| NLP-IIT Patna | 6/6       | MADLAD primary systems, with NLLB and GemmaX2 variants as contrastive runs.                                     |

Table 4: Overview of participating systems. P/C dirs. gives the number of distinct primary and contrastive translation directions, not the number of runs.

| Direction | Primary winner         | BLEU  | Top-2 gap | Best contrastive | BLEU  | $\Delta$ C-P |
|-----------|------------------------|-------|-----------|------------------|-------|--------------|
| En→Ar     | NLP-IIT Patna          | 22.45 | 0.64      | OpenBracket      | 22.48 | +0.04        |
| Hi→Ar     | NLP-IIT Patna          | 19.85 | 1.24      | FARABI           | 18.80 | -1.05        |
| Bn→Ar     | FARABI                 | 17.93 | 2.85      | FARABI           | 17.85 | -0.08        |
| Id→Ar     | FARABI                 | 18.72 | 0.80      | FARABI           | 18.78 | +0.07        |
| Ur→Ar     | NLP-IIT Patna          | 19.00 | 0.68      | FARABI           | 18.23 | -0.77        |
| Ar→En     | OpenBracket            | 33.83 | 0.16      | OpenBracket      | 34.02 | +0.18        |
| Ar→Hi     | NLP-IIT Patna          | 29.28 | 1.50      | NLP-IIT Patna    | 27.87 | -1.40        |
| Ar→Bn     | NERDS_NITS             | 17.10 | 0.51      | FARABI           | 13.54 | -3.57        |
| Ar→Id     | NERDS_NITS             | 30.51 | 5.84      | FARABI           | 24.55 | -5.96        |
| Ar→Ur     | SLPG_FJWU <sup>†</sup> | 25.89 | 0.51      | NLP-IIT Patna    | 25.68 | -0.21        |

Table 5: BLEU summary.  $\Delta$  C-P is the best contrastive BLEU minus the best primary BLEU. <sup>†</sup>The SLPG\_FJWU run labelled primary used supplementary data; its official-data run was labelled contrastive. We retain the submitted labels.

| R. Team           | BLEU  | ChrF2 | TER↓  | COMET | mBERT |
|-------------------|-------|-------|-------|-------|-------|
| <i>En→Ar (1A)</i> |       |       |       |       |       |
| 1 NLP-IIT Patna   | 22.45 | 55.17 | 67.26 | 84.74 | 85.67 |
| 2 OpenBracket     | 21.81 | 54.99 | 69.29 | 85.23 | 85.40 |
| 3 FARABI          | 20.57 | 53.34 | 69.45 | 84.04 | 85.03 |
| <i>Hi→Ar (1B)</i> |       |       |       |       |       |
| 1 NLP-IIT Patna   | 19.85 | 53.17 | 70.47 | 84.30 | 84.82 |
| 2 FARABI          | 18.62 | 51.53 | 72.01 | 83.44 | 84.30 |
| 3 NERDS_NITS      | 15.59 | 49.10 | 75.19 | 81.91 | 83.50 |
| <i>Bn→Ar (1C)</i> |       |       |       |       |       |
| 1 FARABI          | 17.93 | 50.73 | 72.71 | 83.04 | 84.03 |
| 2 NERDS_NITS      | 15.09 | 48.55 | 75.77 | 81.61 | 83.36 |
| 3 Code & Corpus   | 13.79 | 46.32 | 77.74 | 80.39 | 82.69 |
| <i>Id→Ar (1D)</i> |       |       |       |       |       |
| 1 FARABI          | 18.72 | 51.24 | 71.70 | 83.39 | 84.23 |
| 2 NERDS_NITS      | 17.92 | 50.99 | 72.80 | 82.96 | 84.05 |
| 3 IndicMT         | 13.58 | 46.68 | 79.88 | 80.82 | 82.42 |
| <i>Ur→Ar (1E)</i> |       |       |       |       |       |
| 1 NLP-IIT Patna   | 19.00 | 52.23 | 71.86 | 83.87 | 84.43 |
| 2 FARABI          | 18.32 | 50.94 | 72.29 | 83.26 | 84.18 |
| 3 SLPG_FJWU       | 16.40 | 50.22 | 76.00 | 82.84 | 83.55 |

Table 6: Top three primary submissions for translation into Arabic. R. denotes rank. Scores are rounded to two decimals; complete results are provided in Appendix B.

is reduced by weaker from-Arabic results. This distinction matters for shared-task interpretation: a specialized system can maximize one pair, whereas a multilingual submission tests whether one recipe generalizes across targets and scripts.

| R. Team                  | BLEU  | ChrF2 | TER↓  | COMET | BERT  |
|--------------------------|-------|-------|-------|-------|-------|
| <i>Ar→En (2A)</i>        |       |       |       |       |       |
| 1 OpenBracket            | 33.83 | 59.72 | 57.00 | 82.49 | 83.30 |
| 2 NLP-IIT Patna          | 33.67 | 59.67 | 57.30 | 82.01 | 83.11 |
| 3 UPES_NLP               | 32.08 | 58.18 | 59.00 | 81.42 | 82.45 |
| <i>Ar→Hi (2B)</i>        |       |       |       |       |       |
| 1 NLP-IIT Patna          | 29.28 | 54.73 | 62.52 | 72.34 | 85.45 |
| 2 NERDS_NITS             | 27.78 | 52.95 | 63.75 | 70.05 | 85.02 |
| 3 Code & Corpus          | 25.44 | 50.40 | 66.08 | 69.83 | 84.44 |
| <i>Ar→Bn (2C)</i>        |       |       |       |       |       |
| 1 NERDS_NITS             | 17.10 | 51.83 | 74.11 | 78.69 | 83.28 |
| 2 Code & Corpus          | 16.59 | 49.80 | 74.28 | 77.97 | 83.01 |
| 3 NCI-MT                 | 15.09 | 49.40 | 76.26 | 76.97 | 82.43 |
| <i>Ar→Id (2D)</i>        |       |       |       |       |       |
| 1 NERDS_NITS             | 30.51 | 59.82 | 59.14 | 84.11 | 86.06 |
| 2 IndicMT                | 24.67 | 54.58 | 63.95 | 81.99 | 84.72 |
| 3 FARABI                 | 23.95 | 56.65 | 67.75 | 83.75 | 84.50 |
| <i>Ar→Ur (2E)</i>        |       |       |       |       |       |
| 1 SLPG_FJWU <sup>†</sup> | 25.89 | 51.06 | 65.36 | 76.26 | 83.49 |
| 2 NERDS_NITS             | 25.38 | 50.91 | 65.12 | 75.19 | 83.51 |
| 3 NLP-IIT Patna          | 22.49 | 48.70 | 68.11 | 75.52 | 82.76 |

Table 7: Top three primary submissions for translation from Arabic. R. denotes rank. <sup>†</sup>This run used supplementary data despite its primary label.

**Directional specialization.** The aggregate ranks reveal strong directionality even when a team covers the same language pairs in both directions (Table 9). FARABI has the clearest asymmetry: its average rank is 1.80 when Arabic is the target, but 7.00 when Arabic is the source. NERDS\_NITS shows the opposite, improving from an average rank of 3.00 into Arabic to 2.00 from Arabic. NLP-IIT Patna is nearly dominant in its three into-Arabic

| Team                   | Dirs. | Wins | Top-2 | Mean rank | Norm.↓ |
|------------------------|-------|------|-------|-----------|--------|
| OpenBracket            | 2     | 1    | 2     | 1.50      | 0.06   |
| NLP-IIT Patna          | 6     | 4    | 5     | 1.50      | 0.09   |
| NERDS_NITS             | 10    | 2    | 6     | 2.50      | 0.24   |
| SLPG_FJWU <sup>†</sup> | 3     | 1    | 1     | 3.00      | 0.28   |
| Code & Corpus          | 4     | 0    | 1     | 3.00      | 0.30   |
| FARABI                 | 10    | 2    | 4     | 4.40      | 0.52   |
| NCI-MT                 | 6     | 0    | 0     | 5.33      | 0.60   |
| UPES_NLP               | 3     | 0    | 0     | 5.67      | 0.61   |
| MAGADIR                | 2     | 0    | 0     | 6.50      | 0.65   |
| IndicMT                | 10    | 0    | 1     | 5.10      | 0.68   |
| BVSLP                  | 10    | 0    | 0     | 5.50      | 0.78   |
| MTG-LIRA               | 2     | 0    | 0     | 7.50      | 0.83   |
| NITM AI Lab            | 2     | 0    | 0     | 8.50      | 0.88   |

Table 8: Primary-system coverage and consistency. Norm. is the mean normalized rank  $(r - 1)/(n - 1)$ . Narrow and broad coverage are not directly comparable. <sup>†</sup>Resource-label caveat in Table 5.

tracks and remains strong, though less uniformly, for the corresponding reverse directions.

| Team          | Into Arabic |           |      | From Arabic |           |      |
|---------------|-------------|-----------|------|-------------|-----------|------|
|               | Dirs.       | Mean rank | Wins | Dirs.       | Mean rank | Wins |
| NLP-IIT Patna | 3           | 1.00      | 3    | 3           | 2.00      | 1    |
| FARABI        | 5           | 1.80      | 2    | 5           | 7.00      | 0    |
| NERDS_NITS    | 5           | 3.00      | 0    | 5           | 2.00      | 2    |
| Code & Corpus | 2           | 3.50      | 0    | 2           | 2.50      | 0    |
| IndicMT       | 5           | 5.40      | 0    | 5           | 4.80      | 0    |
| NCI-MT        | 3           | 5.67      | 0    | 3           | 5.00      | 0    |
| BVSLP         | 5           | 5.40      | 0    | 5           | 5.60      | 0    |

Table 9: Direction-specific primary ranks for teams entering at least four directions. Rank averages are computed within each submitted subset and are affected by the number and strength of competitors.

The asymmetry corresponds to the systems’ transfer structure. FARABI shares information across several source languages while keeping Arabic as the common target, whereas NERDS\_NITS connects Arabic source sentences to target-specialized IndicTrans2 models or directly adapts NLLB for Indonesian. The results therefore support a practical distinction between *multi-source transfer*, which pools encoders for a common target, and *multi-target transfer*, which must generate several scripts and morphological systems. Success in one setting does not predict success in the other.

The remaining broad systems are more stable but occupy the middle of the rankings. IndicMT and BVSLP use largely uniform NLLB-centered recipes across all ten directions; neither wins a track, although IndicMT is runner-up for Arabic→Indonesian. This suggests a coverage–specialization trade-off: one portable recipe offers complete task coverage, while direction-aware components are needed to reach the top of heterogeneous leaderboards.

## 6.5 Winning Strategies Across Directions

Table 10 groups the ten nominal primary wins by their main system recipe. Supervised MADLAD fine-tuning accounts for four wins, while NLLB-based transfer accounts for another four through shared into-Arabic training, pair-specific controllers, pivoting, or direct adaptation. The remaining two wins come from heterogeneous MBR for Arabic→English and filtered supplementary data for Arabic→Urdu. No zero-shot general-purpose LLM system wins a primary run.

| Winning methods                 | Wins | Tasks          |
|---------------------------------|------|----------------|
| MADLAD supervised fine-tuning   | 4    | 1A, 1B, 1E, 2B |
| Joint NLLB + pair controllers   | 2    | 1C, 1D         |
| NLLB pivot/direct adaptation    | 2    | 2C, 2D         |
| Cross-model MBR ensemble        | 1    | 2A             |
| Filtered-data NLLB <sup>†</sup> | 1    | 2E             |

Table 10: Nominal primary wins grouped by the principal system recipe. <sup>†</sup>The 2E run used supplementary data despite its primary label.

These results are not a controlled comparison of architectures because model size, training data, optimization, and decoding differ across teams. They nevertheless show that model family alone does not determine performance. NLLB appears throughout the runs, but outcomes depend strongly on how it is adapted, how transfer is matched to the target language, and how supplementary data are selected. Code & Corpus also reaches three top-three positions with a 600M joint model, suggesting that parameter sharing can partly compensate for model scale.

**Evidence across teams.** NLP-IIT Patna (Sharma et al., 2026) provides the clearest evidence for supervised multilingual fine-tuning. Its MADLAD systems won four of the six directions entered and placed in the top two in five. This consistency shows that conventional supervised adaptation of a strong multilingual MT model remains difficult to outperform in this data regime.

NERDS\_NITS (Karim et al., 2026) and FARABI show different benefits of transfer. NERDS\_NITS won Arabic→Bangla and Arabic→Indonesian and recorded six top-two placements. Its English-pivot NLLB+IndicTrans2 cascade performed well for Indic targets, while direct QLoRA adaptation of NLLB produced the largest winning margin for Indonesian, which IndicTrans2 does not cover. Its weaker results when translating into Arabic suggest that target-specific adaptation contributed more than pivoting alone.

FARABI (Aldabbas et al., 2026) showed the reverse pattern. Its shared NLLB backbone with source-specific controllers won two into-Arabic directions and placed second in two others, whereas its from-Arabic Qwen-based systems varied more across targets. This favors joint multi-source adaptation when several source languages share Arabic as the target. A similar pattern appears for Code & Corpus, whose jointly trained four-direction NLLB model reached the top three in three directions despite using a smaller backbone.

Inference-time combination was strongest for the English pair. OpenBracket entered two primary directions and placed in the top two in both. Its heterogeneous MBR ensemble won Arabic→English and achieved the highest COMET score for English→Arabic.

## 6.6 Do Contrastive Systems Help?

Contrastive submissions could use additional resources and often changed the backbone or decoding strategy. They therefore measure whether a less constrained system performs better rather than the effect of supplementary data alone. Overall, they rarely improve the best result. The best contrastive run exceeds the best primary run in only three directions, English→Arabic, Indonesian→Arabic, and Arabic→English, with gains of only 0.04, 0.07, and 0.18 BLEU. Primary systems remain ahead in the other seven directions.

Matched team–direction comparisons show the same pattern in Table 11. After excluding SLPG\_FJWU (Zafar et al., 2026) because its resource labels are reversed, contrastive runs improve BLEU in 13 of 33 comparisons and decrease it in 20. The average change is  $-1.46$  BLEU and the median is  $-1.31$ . Only OpenBracket (Alotaibi, 2026), MAGADIR (Boudraa, 2026), and FARABI have non-negative mean BLEU changes.

| Team          | Matched | Better | $\Delta$ BLEU | $\Delta$ COMET |
|---------------|---------|--------|---------------|----------------|
| FARABI        | 10      | 6      | +0.09         | −1.11          |
| NCI-MT        | 6       | 0      | −3.21         | +0.16          |
| NLP-IIT Patna | 6       | 1      | −1.01         | −0.19          |
| UPES_NLP      | 3       | 0      | −4.87         | +0.80          |
| NITM AI Lab   | 2       | 1      | −3.18         | −2.06          |
| MTG-LIRA      | 2       | 1      | −2.96         | −3.33          |
| OpenBracket   | 2       | 2      | +0.43         | −0.43          |
| MAGADIR       | 2       | 2      | +1.13         | −0.59          |
| All matched   | 33      | 13     | −1.46         | −0.66          |

Table 11: Best contrastive minus primary score for matched team–direction pairs, averaged by team. “Better” counts positive BLEU changes. SLPG\_FJWU is excluded because its resource labels are reversed relative to the nominal track definitions.

The matched comparisons also reveal metric trade-offs. NCI-MT’s RAG pipeline lowers BLEU in all six directions by 3.21 points on average but increases COMET by 0.16. UPES\_NLP shows a similar pattern, with GPT-5-mini trailing its primary systems by 4.87 BLEU while improving COMET by 0.80. In contrast, OpenBracket and MAGADIR improve BLEU in both matched English directions while their average COMET scores decrease. These results show that improving one metric does not necessarily improve overall translation quality.

Additional data are similarly inconsistent. SLPG\_FJWU’s larger model trained with filtered supplementary data outperforms its official-data alternative by 1.09-3.05 BLEU across three directions, although model size and data change together. MTG-LIRA gains 1.21 BLEU for Hindi→Arabic but loses 7.12 in the reverse direction. NITM AI Lab also reports opposite effects across its two evaluated directions. Supplementary data therefore need to be evaluated separately for each direction and domain.

## 6.7 Metric Agreement and Ranking Stability

Metric agreement is high for primary systems but decreases for contrastive submissions, as shown in Table 12. BLEU and ChrF2 produce very similar primary rankings, and BERT-based similarity is also strongly aligned with BLEU. COMET shows the lowest correlation with BLEU in both tracks. Among contrastive systems, BLEU and COMET select the same top run in only four of ten directions.

| Metric vs. BLEU | Mean Spearman $\rho$ |             | Same top run |             |
|-----------------|----------------------|-------------|--------------|-------------|
|                 | Primary              | Contrastive | Primary      | Contrastive |
| ChrF2           | 0.92                 | 0.83        | 10/10        | 9/10        |
| −TER            | 0.92                 | 0.69        | 9/10         | 9/10        |
| COMET           | 0.81                 | 0.62        | 9/10         | 4/10        |
| BERT-based*     | 0.98                 | 0.69        | 9/10         | 7/9         |

Table 12: Rank agreement with BLEU. TER is negated so that higher is better. \*The duplicated BERT column for contrastive 2A is excluded.

Several rank reversals illustrate this disagreement. For English→Arabic primary systems, NLP-IIT Patna leads BLEU, ChrF2, TER, and mBERT, while OpenBracket leads COMET. For Arabic→Bangla contrastive systems, FARABI leads BLEU with 13.54, whereas UPES\_NLP leads COMET and mBERT despite a BLEU score of 8.94. A similar reversal appears for Arabic→Indonesian. Such disagreements are especially relevant for prompted, reranked, or retrieval-

augmented systems, which may produce valid phrases that overlap less with the reference. Conversely, semantic metrics may overlook specific translation errors. Without human judgments, these cases do not establish which ranking is more reliable, but they show that reporting a single automatic metric is insufficient.

## 6.8 Cross-System Findings

The results support four broader conclusions.

**Supervised adaptation remains the most reliable strategy.** Strong multilingual MT models with task-specific fine-tuning dominate the primary leaderboards, while zero-shot and multi-pass LLM systems are less consistent.

**Transfer works best when it matches the target language.** English pivoting is effective when paired with a strong target-specific model such as IndicTrans2, joint multi-source training benefits translation into a shared Arabic target, and direct adaptation works better for Indonesian.

**More data do not guarantee better performance.** Broad augmentation often underperforms official-data systems, while carefully filtered supplementary data can help. Domain match and translation direction matter as much as data volume.

**Inference-time optimization is metric-sensitive.** MBR, prompting, and quality-estimation pipelines sometimes improve one metric while reducing another. Development objectives should therefore match the final evaluation and, where possible, be complemented by human assessment.

**Implications for future evaluations.** Future editions could make contrastive submissions more diagnostic by recording which component changes across runs, including additional data, model backbone, fine-tuning strategy, or decoding method. A small set of matched ablations and explicit resource declarations would make it easier to attribute improvements to data, scale, or inference-time selection. Metric disagreements also provide a natural basis for targeted human evaluation. We could sample cases with the largest BLEU-COMET rank reversals and evaluate adequacy, fluency, entities, numbers, and omissions. Sentence-level scores would support analysis by length and entity density, particularly as challenge sentences are roughly 24-29% longer than the training sentences. Overall, with standardized metric versions, these additions would make the runs more informative about where Arabic-Asian MT systems succeed and fail.

## 7 Conclusion

The WMT 2026 Low-Resource Arabic-Asian Translation shared task establishes a bidirectional benchmark across Arabic and five Asian languages using *AraAsian1.0* and submissions from 13 teams. The results show that low-resource performance depends less on adding model scale or external resources in isolation than on selecting an adaptation strategy that matches the language direction and data domain. MADLAD fine-tuning was the most consistently successful approach, while joint multilingual adaptation, English-pivot transfer, direct QLoRA fine-tuning, and MBR ensembling each led particular directions.

The analysis also identifies persistent gaps. Bangla obtained the lowest winning BLEU in both directions, contrastive systems rarely improved over the best primary systems, and metric agreement weakened for LLM-based and reranked outputs. Future editions shall add human evaluation, standardized metric configurations, stronger organizer baselines, and test sets spanning additional domains and underrepresented Asian languages. These additions would help distinguish genuine semantic improvements from reference-overlap changes and provide a more complete account of Arabic-centric low-resource MT.

## Limitations

Our study has several limitations. *AraAsian1.0* focuses on the news domain and covers five Arabic-centered language pairs, so the findings may not generalize to other domains, dialects, or underrepresented Asian languages. In addition, three Asian-language sides were constructed through an English-mediated translation stage followed by expert correction, which may introduce some bias. The shared task relies on automatic metrics without human evaluation or significance testing, so small score differences and metric disagreements should be interpreted with care. Future evaluations can address these limitations through broader domain coverage, additional languages, and targeted human assessment.

## Ethics and Social Impact

This shared task aims to support more inclusive machine translation research by providing data and evaluation for Arabic and several Asian languages that remain less studied in direct translation settings. Better translation technologies for these lan-

guages can improve cross-lingual access to information and support communication across diverse communities. At the same time, MT systems can reproduce biases in their training data or produce inaccurate translations, particularly for culturally specific content, entities, and low-resource languages. We therefore encourage careful evaluation across languages and contexts before deployment in settings where translation errors may have meaningful consequences.

## References

- Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasajo, Timothy Baldwin, Jey Han Lau, and Sebastian Ruder. 2022. [One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7226–7249, Dublin, Ireland. Association for Computational Linguistics.
- Farizeh Aldabbas, Zyad Altahan, Hossam Elsafty, and Rafet Sifa. 2026. Farabi at wmt 2026: Post-hoc channel controllers for low-resource arabic-asian machine translation. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Abrar M. Alotaibi. 2026. Openbracket at wmt 2026: Comet-mbr system combination over arabic-native and multilingual models for low-resource arabic-english translation. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Palak Arora, Afroz Ahmed Ahmed, Mansi Jangid, Bharti Nathani, and Nisheeth Joshi. 2026. Bvslp at wmt 2026: An empirical study of multilingual transfer for low-resource arabic-centric translation. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Hossam Boudraa. 2026. Magadir at wmt 2026: Cross-system qe re-ranking and entity-aware selection for low-resource arabic-english translation. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh Khapra, and Pratyush Kumar. 2022. [IndicBART: A pre-trained model for indic natural language generation](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1849–1863, Dublin, Ireland. Association for Computational Linguistics.
- Kareem Darwish, Nizar Habash, Mourad Abbas, Hend Al-Khalifa, Huseein Al-Natsheh, Samhaa El-Beltagy, Houda Bouamor, Karim Bouzoubaa, Violetta Cavall-Sforza, Wassim El-Hajj, Mustafa Jarrar, and Hamdy Mubarak. 2020. [A panoramic survey of natural language processing in the arab world](#).
- Satarupa Deb and Prashanth Nayak. 2026. Nci-mt at wmt 2026: A quality-gated cascade for low-resource arabic-asian machine translation. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Ali Farghaly and Khaled Shaalan. 2009. [Arabic natural language processing: Challenges and solutions](#). *ACM Transactions on Asian Language Information Processing*, 8(4).
- Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Ari-vazhagan, and Wei Wang. 2022. [Language-agnostic BERT sentence embedding](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.
- Jay Gala, Pranjal A Chitale, Raghavan AK, Sumanth Doddapaneni, Varun Gumma, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, et al. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *arXiv preprint arXiv:2305.16307*.
- Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc’Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. [The Flores-101 evaluation benchmark for low-resource and multilingual machine translation](#). *Transactions of the Association for Computational Linguistics*, 10:522–538.
- Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. [Survey of low-resource machine translation](#). *Computational Linguistics*, 48(3):673–732.
- Vedarth Hede, Neha Gupta, Harish Bapat, and Harsh Ekbote. 2026. Mtg-lira at wmt 2026: Data-centric arabic-hindi machine translation through corpus augmentation, synthetic parallel data generation and transformer models. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. [How good are gpt models at machine translation? a comprehensive evaluation](#). *Preprint*, arXiv:2302.09210.
- Pramit Jana, Arindam Adhikari, Priyobroto Acharya, and Dipankar Das. 2026. Indicmt at wmt 2026: Full fine-tuning of nllb-200-600m with round-trip filtering

- and forward self-training for arabic-asian translation. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. [The state and fate of linguistic diversity and inclusion in the NLP world](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.
- Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. [IndicNLP Suite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.
- Misbahul Karim, Faruk Ahmed, Mishbahul Al Islam, Sahinur Rahman Laskar, and Rabul Hussain Laskar. 2026. Nerds-nits at wmt2026: Pivot-based and direct neural machine translation for arabic-asian low-resource language pairs. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Philipp Koehn and Rebecca Knowles. 2017. [Six challenges for neural machine translation](#). In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.
- Nabin Koirala, Rhythm Bhetwal, and Nurul Amin Choudhury. 2026. Nitm ai lab at wmt 2026 shared task: Low-resource arabic-asian language translation. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- 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 on Association for Computational Linguistics, ACL '02*, pages 311–318, Stroudsburg, PA, 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.
- Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. [Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages](#). *Transactions of the Association for Computational Linguistics*, 10:145–162.
- 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.
- Pushkar Sharma, Mo Ahtasam, Kshetrimayum Boynao Singh, Deepak Kumar, and Asif Ekbal. 2026. Nlp-iit patna at wmt 2026 : Corpus-driven adaptation of multilingual mt models for low-resource arabic-asian machine translation. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Sumit Singh, Anish Kumar Vishwakarma, Brijmohan Lal Sahu, and Ashwani Kumar. 2026. Upes\_nlp at wmt 2026 low-resource arabic-asian machine translation shared task: Arabic-centric machine translation with nllb-200-moe and prompt-constrained entity-preserving translation with gpt-5-mini. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, pages 223–231.
- 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](#). *Preprint*, arXiv:2207.04672.
- Raman Kumar Yadav and Alapan Kuila. 2026. Code & corpus at wmt 2026: Checkpoint-pool minimum bayes risk decoding for arabic-hindi and arabic-bangla translation. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Shomaiza Zafar, Sadaf Abdul Rauf, Sheema Firdous, Muhammad Saad Munir, and Nadeem Fakhra. 2026.

Slpg\_fjwu at wmt 2026: Domain adapted augmentation for low-resource arabic-asian language translation. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.

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

## A System Descriptions

**NERDS\_NITS (Karim et al., 2026).** NERDS\_NITS used an English-pivot cascade for the Indic languages, combining a fine-tuned Arabic–English NLLB-200-600M model with pretrained IndicTrans2 checkpoints. Because IndicTrans2 does not cover Indonesian, Arabic↔Indonesian was handled directly through QLoRA fine-tuning of NLLB-200-1.3B. The standalone Arabic↔English directions used fine-tuned Helsinki-NLP OPUS models.

**NITM AI Lab (Koirala et al., 2026).** NITM AI Lab explored several architectures. Its English→Arabic primary system fully fine-tuned a domain-specialized OPUS-MT-TC-Big model. Its strongest Arabic→Hindi primary system adapted NLLB-200-distilled-600M with LoRA on the official data. Adding targeted OPUS parallel data caused a slight degradation, which the team attributed to domain mismatch. The team also evaluated fully fine-tuned mBART-50 with rule-based orthographic and numerical normalization.

**UPES\_NLP (Singh et al., 2026).** UPES\_NLP compared supervised NLLB fine-tuning, a pretrained NLLB mixture-of-experts checkpoint, and zero-shot API prompting. The fine-tuned NLLB-200-1.3B model served as the Arabic→English primary system, while the NLLB-MoE checkpoint was used for Arabic→Hindi and Arabic→Bangla. Contrastive runs prompted GPT-5-mini without task-specific training and required structured JSON output to reduce changes to entities and numbers.

**OpenBracket (Alotaibi, 2026).** OpenBracket’s primary system used MBR decoding over candidates from fine-tuned NLLB-3.3B, X-ALMA-FT, X-ALMA-CPO, ALLAM-FT, and OPUS-MT, with COMET-22 as the utility. Its first contrastive system QLoRA-tuned ALLAM-7B-Instruct. Its second contrastive system used COMET-Kiwi-pruned sampling and MBR over NLLB-3.3B candidates.

**NLP-IIT Patna (Sharma et al., 2026).** NLP-IIT Patna evaluated NLLB, MADLAD, and GemmaX2 across six directions. Fine-tuned MADLAD outputs were selected as the primary submissions, while NLLB and GemmaX2 variants formed the contrastive systems.

**SLPG\_FJWU (Zafar et al., 2026).** SLPG\_FJWU focused on Urdu↔Arabic and Arabic→English. Its run labelled primary fine-tuned NLLB-200-3.3B on an augmented corpus derived from CC-100 and News Commentary and filtered with LEALLA-large multilingual embeddings. Its run labelled contrastive fine-tuned NLLB-200-distilled-600M only on the official task data. Thus, the submitted labels are reversed relative to the nominal resource distinction; the leaderboard labels are preserved in this paper.

**NCI-MT (Deb and Nayak, 2026).** NCI-MT applied LoRA ( $r = 16$ ,  $\alpha = 32$ ) to the attention projections of NLLB-200-distilled-600M for its primary systems. Its contrastive system used a two-pass retrieval-augmented pipeline. COMET-Kiwi first scored an initial translation. Low-confidence cases triggered LaBSE encoding and FAISS retrieval of the three most similar source–target pairs from external corpora, which were supplied to Gemma-3-4B as in-context examples. COMET-Kiwi then rescored the revision, and the system retained it only when the estimated quality improved.

**MAGADIR (Boudraa, 2026).** MAGADIR developed direction-specific cross-system MBR pipelines. For Arabic→English, it combined a LoRA-tuned NLLB-200-1.3B model with zero-shot outputs from Hunyuan-MT-7B, HY-MT1.5-7B, and Gemma-3-12B. A Wikidata entity-recall constraint filtered candidates to reduce proper-noun omissions, after which MetricX-24 selected the output. For English→Arabic, the team pooled candidates from six zero-shot or pretrained models, including Aya Expanse 8B and NLLB-200-3.3B, and used COMET-Kiwi as the MBR utility.

**IndicMT (Jana et al., 2026).** IndicMT independently fine-tuned NLLB-200-distilled-600M for all ten directions using Adafactor. Before training, the team filtered official pairs whose model-output/reference SequenceMatcher similarity was below 0.3. It then generated 500 synthetic pairs per direction from the filtered back-translation pipeline and appended them to the remaining real data. The

| Team          | Organization                                                 |
|---------------|--------------------------------------------------------------|
| OpenBracket   | King Fahd University of Petroleum and Minerals, Saudi Arabia |
| BVSLP         | Banasthali Vidyapith, India                                  |
| UPES_NLP      | UPES, Dehradun, India                                        |
| NERDS_NITS    | National Institute of Technology Silchar, India              |
| MTG-LIRA      | Centre for Development of Advanced Computing (CDAC), India   |
| NITM AI Lab   | National Institute of Technology Meghalaya and VFSTR, India  |
| MAGADIR       | Tampere University, Finland                                  |
| Code & Corpus | IIITDM Kurnool, India                                        |
| IndicMT       | Jadavpur University, India                                   |
| SLPG_FJWU     | Fatima Jinnah Women University, Pakistan                     |
| FARABI        | University of Bonn, Germany                                  |
| NCI-MT        | National College of Ireland, Ireland                         |
| NLP-IIT Patna | Indian Institute of Technology Patna, India                  |

Table 13: Participating teams and affiliations.

team reports that this augmentation substantially improved its internal BLEU scores.

**FARABI (Aldabbas et al., 2026).** FARABI combined multilingual backbones with lightweight pair-specific adaptation. For translation into Arabic, it jointly fine-tuned one NLLB-200-3.3B model across all source languages and applied NTK-Mirror hidden-state controllers at inference time to learn pair-specific residual corrections. For translation from Arabic, it adapted Qwen2.5-7B-Instruct with task-specific LoRA modules and, for some runs, added a retrieval-based second pass using few-shot examples from the training data.

**Code & Corpus (Yadav and Kuila, 2026).** Code & Corpus addressed Arabic–Hindi and Arabic–Bangla in both directions. It constructed a unified four-direction corpus and jointly fine-tuned one NLLB-200-distilled-600M model, using parameter sharing to reduce overfitting relative to separate direction-specific models.

**MTG-LIRA (Hede et al., 2026).** MTG-LIRA trained a Transformer encoder–decoder system with OpenNMT-py for Arabic↔Hindi. The primary model used only the official WMT 2026 data. The contrastive model first learned from pivot-augmented data built with the BPCC English–Hindi corpus and OPUS resources and was then fine-tuned on the shared-task data.

**BVSLP (Arora et al., 2026).** BVSLP used NLLB as the core model for all directions and combined transfer learning with LLM-assisted data preparation. LLMs cleaned noisy sentence pairs, and back-translation generated synthetic bilingual data before iterative fine-tuning.

## B Full Primary Results

Tables in this appendix reproduce the full primary results. Ranks are normalized to the BLEU ordering; this resolves the duplicated rank label in the supplied Arabic→Urdu table.

| R. | Team          | BLEU  | ChrF2 | TER   | COMET | mBERT |
|----|---------------|-------|-------|-------|-------|-------|
| 1  | NLP-IIT Patna | 22.45 | 55.17 | 67.26 | 84.74 | 85.67 |
| 2  | OpenBracket   | 21.81 | 54.99 | 69.29 | 85.23 | 85.40 |
| 3  | FARABI        | 20.57 | 53.34 | 69.45 | 84.04 | 85.03 |
| 4  | NERDS_NITS    | 19.77 | 53.49 | 71.82 | 83.49 | 84.39 |
| 5  | IndicMT       | 16.05 | 49.30 | 76.49 | 81.98 | 83.48 |
| 6  | NCI-MT        | 16.01 | 50.59 | 76.01 | 81.95 | 83.62 |
| 7  | MAGADIR       | 14.59 | 51.48 | 80.85 | 83.88 | 83.13 |
| 8  | NITM AI Lab   | 14.08 | 49.13 | 75.75 | 82.21 | 83.18 |
| 9  | BVSLP         | 12.47 | 44.85 | 81.52 | 80.77 | 82.05 |

Table 14: Primary results for Subtask 1A (English → Arabic). R. denotes rank.

| R. | Team          | BLEU  | ChrF2 | TER   | COMET | mBERT |
|----|---------------|-------|-------|-------|-------|-------|
| 1  | NLP-IIT Patna | 19.85 | 53.17 | 70.47 | 84.30 | 84.82 |
| 2  | FARABI        | 18.62 | 51.53 | 72.01 | 83.44 | 84.30 |
| 3  | NERDS_NITS    | 15.59 | 49.10 | 75.19 | 81.91 | 83.50 |
| 4  | Code & Corpus | 14.77 | 47.59 | 76.50 | 81.16 | 82.99 |
| 5  | BVSLP         | 13.28 | 48.20 | 82.63 | 81.89 | 82.55 |
| 6  | NCI-MT        | 13.10 | 47.41 | 80.54 | 80.08 | 82.35 |
| 7  | IndicMT       | 12.65 | 46.28 | 80.94 | 80.36 | 82.23 |
| 8  | MTG-LIRA      | 9.17  | 43.29 | 82.92 | 79.27 | 81.30 |

Table 15: Primary results for Subtask 1B (Hindi → Arabic). R. denotes rank.

| R. | Team          | BLEU  | ChrF2 | TER   | COMET | mBERT |
|----|---------------|-------|-------|-------|-------|-------|
| 1  | FARABI        | 17.93 | 50.73 | 72.71 | 83.04 | 84.03 |
| 2  | NERDS_NITS    | 15.09 | 48.55 | 75.77 | 81.61 | 83.36 |
| 3  | Code & Corpus | 13.79 | 46.32 | 77.74 | 80.39 | 82.69 |
| 4  | BVSLP         | 12.92 | 47.10 | 82.32 | 81.25 | 82.38 |
| 5  | NCI-MT        | 12.01 | 46.39 | 81.43 | 79.37 | 82.03 |
| 6  | IndicMT       | 11.53 | 45.14 | 82.31 | 79.32 | 81.78 |

Table 16: Primary results for Subtask 1C (Bangla → Arabic). R. denotes rank.

| R. | Team       | BLEU  | ChrF2 | TER   | COMET | mBERT |
|----|------------|-------|-------|-------|-------|-------|
| 1  | FARABI     | 18.72 | 51.24 | 71.70 | 83.39 | 84.23 |
| 2  | NERDS_NITS | 17.92 | 50.99 | 72.80 | 82.96 | 84.05 |
| 3  | IndicMT    | 13.58 | 46.68 | 79.88 | 80.82 | 82.42 |
| 4  | BVSLP      | 12.75 | 46.31 | 81.81 | 81.19 | 82.21 |

Table 17: Primary results for Subtask 1D (Indonesian → Arabic). R. denotes rank.

| R. | Team          | BLEU  | ChrF2 | TER   | COMET | mBERT |
|----|---------------|-------|-------|-------|-------|-------|
| 1  | NLP-IIT Patna | 19.00 | 52.23 | 71.86 | 83.87 | 84.43 |
| 2  | FARABI        | 18.32 | 50.94 | 72.29 | 83.26 | 84.18 |
| 3  | SLPG_FJWU     | 16.40 | 50.22 | 76.00 | 82.84 | 83.55 |
| 4  | NERDS_NITS    | 14.12 | 45.84 | 76.80 | 80.59 | 82.83 |
| 5  | BVSLP         | 12.94 | 47.07 | 82.12 | 81.41 | 82.35 |
| 6  | IndicMT       | 12.53 | 45.81 | 81.29 | 79.91 | 82.03 |

Table 18: Primary results for Subtask 1E (Urdu → Arabic). R. denotes rank.

| R. | Team          | BLEU  | ChrF2 | TER   | COMET | BERT  |
|----|---------------|-------|-------|-------|-------|-------|
| 1  | OpenBracket   | 33.83 | 59.72 | 57.00 | 82.49 | 83.30 |
| 2  | NLP-IIT Patna | 33.67 | 59.67 | 57.30 | 82.01 | 83.11 |
| 3  | UPES_NLP      | 32.08 | 58.18 | 59.00 | 81.42 | 82.45 |
| 4  | NERDS_NITS    | 29.35 | 55.83 | 59.85 | 80.88 | 81.92 |
| 5  | SLPG_FJWU     | 29.27 | 55.71 | 61.23 | 81.15 | 81.78 |
| 6  | MAGADIR       | 27.45 | 56.61 | 64.49 | 81.94 | 81.28 |
| 7  | NCI-MT        | 26.73 | 53.83 | 62.85 | 79.91 | 80.44 |
| 8  | IndicMT       | 25.49 | 51.95 | 64.06 | 79.72 | 80.29 |
| 9  | FARABI        | 22.76 | 54.79 | 70.31 | 80.47 | 79.15 |
| 10 | BVSLP         | 20.13 | 47.34 | 69.44 | 78.88 | 77.90 |

Table 19: Primary results for Subtask 2A (Arabic → English). R. denotes rank.

| R. | Team          | BLEU  | ChrF2 | TER   | COMET | mBERT |
|----|---------------|-------|-------|-------|-------|-------|
| 1  | NLP-IIT Patna | 29.28 | 54.73 | 62.52 | 72.34 | 85.45 |
| 2  | NERDS_NITS    | 27.78 | 52.95 | 63.75 | 70.05 | 85.02 |
| 3  | Code & Corpus | 25.44 | 50.40 | 66.08 | 69.83 | 84.44 |
| 4  | BVSLP         | 24.67 | 50.07 | 66.12 | 68.88 | 84.32 |
| 5  | NCI-MT        | 24.47 | 50.23 | 67.05 | 69.11 | 84.12 |
| 6  | IndicMT       | 24.11 | 49.18 | 67.09 | 69.32 | 84.16 |
| 7  | MTG-LIRA      | 24.10 | 49.56 | 66.95 | 67.73 | 84.06 |
| 8  | UPES_NLP      | 22.61 | 47.15 | 67.16 | 68.92 | 83.87 |
| 9  | NITM AI Lab   | 22.55 | 47.64 | 68.28 | 68.24 | 83.57 |
| 10 | FARABI        | 21.84 | 47.20 | 68.20 | 68.73 | 83.55 |

Table 20: Primary results for Subtask 2B (Arabic → Hindi). R. denotes rank.

| R. | Team          | BLEU  | ChrF2 | TER   | COMET | mBERT |
|----|---------------|-------|-------|-------|-------|-------|
| 1  | NERDS_NITS    | 17.10 | 51.83 | 74.11 | 78.69 | 83.28 |
| 2  | Code & Corpus | 16.59 | 49.80 | 74.28 | 77.97 | 83.01 |
| 3  | NCI-MT        | 15.09 | 49.40 | 76.26 | 76.97 | 82.43 |
| 4  | IndicMT       | 13.99 | 47.70 | 76.59 | 76.93 | 82.32 |
| 5  | BVSLP         | 13.80 | 48.00 | 77.25 | 77.65 | 82.22 |
| 6  | UPES_NLP      | 12.14 | 44.82 | 78.00 | 77.31 | 81.71 |
| 7  | FARABI        | 10.86 | 44.76 | 80.07 | 77.12 | 81.16 |

Table 21: Primary results for Subtask 2C (Arabic → Bangla). R. denotes rank.

| R. | Team       | BLEU  | ChrF2 | TER   | COMET | mBERT |
|----|------------|-------|-------|-------|-------|-------|
| 1  | NERDS_NITS | 30.51 | 59.82 | 59.14 | 84.11 | 86.06 |
| 2  | IndicMT    | 24.67 | 54.58 | 63.95 | 81.99 | 84.72 |
| 3  | FARABI     | 23.95 | 56.65 | 67.75 | 83.75 | 84.50 |
| 4  | BVSLP      | 22.81 | 53.10 | 65.36 | 81.90 | 84.26 |

Table 22: Primary results for Subtask 2D (Arabic → Indonesian). R. denotes rank.

| R. | Team          | BLEU  | ChrF2 | TER   | COMET | mBERT |
|----|---------------|-------|-------|-------|-------|-------|
| 1  | SLPG_FJWU     | 25.89 | 51.06 | 65.36 | 76.26 | 83.49 |
| 2  | NERDS_NITS    | 25.38 | 50.91 | 65.12 | 75.19 | 83.51 |
| 3  | NLP-IIT Patna | 22.49 | 48.70 | 68.11 | 75.52 | 82.76 |
| 4  | IndicMT       | 21.12 | 46.35 | 69.59 | 73.81 | 82.43 |
| 5  | BVSLP         | 20.99 | 47.59 | 70.37 | 73.81 | 82.44 |
| 6  | FARABI        | 18.31 | 44.63 | 71.81 | 74.01 | 81.37 |

Table 23: Primary results for Subtask 2E (Arabic → Urdu). R. denotes rank.

## C Full Contrastive Results

The BERT values supplied for Arabic→English are identical to COMET for every contrastive run. They are reproduced without modification and are not used in the aggregate metric-correlation analysis.

| R. | Team          | Run | BLEU  | ChrF2 | TER   | COMET | mBERT |
|----|---------------|-----|-------|-------|-------|-------|-------|
| 1  | OpenBracket   | C-1 | 22.48 | 54.58 | 68.21 | 84.26 | 85.31 |
| 2  | FARABI        | C-1 | 20.68 | 53.49 | 69.29 | 84.10 | 85.07 |
| 3  | NLP-IIT Patna | C-2 | 20.38 | 53.64 | 70.06 | 84.19 | 85.02 |
| 4  | NLP-IIT Patna | C-1 | 19.21 | 52.99 | 72.00 | 83.79 | 84.61 |
| 5  | OpenBracket   | C-2 | 19.17 | 52.59 | 71.17 | 84.93 | 84.87 |
| 6  | MAGADIR       | C-2 | 16.25 | 51.53 | 76.94 | 83.07 | 83.22 |
| 7  | MAGADIR       | C-1 | 15.80 | 51.43 | 77.63 | 83.26 | 83.43 |
| 8  | FARABI        | C-2 | 15.76 | 49.36 | 76.32 | 82.94 | 83.34 |
| 9  | UPES_NLP      | C   | 13.02 | 44.27 | 80.24 | 80.20 | 81.58 |
| 10 | NCI-MT        | C   | 12.93 | 49.34 | 82.42 | 81.82 | 82.84 |
| 11 | NITM AI Lab   | C-2 | 7.03  | 40.95 | 85.89 | 78.27 | 79.64 |
| 12 | NITM AI Lab   | C-1 | 5.75  | 40.48 | 96.96 | 76.81 | 79.46 |

Table 24: Contrastive results for Subtask 1A (English → Arabic). R. denotes rank and C denotes contrastive run.

| R. | Team          | Run | BLEU  | ChrF2 | TER   | COMET | mBERT |
|----|---------------|-----|-------|-------|-------|-------|-------|
| 1  | FARABI        | C-1 | 18.80 | 51.75 | 71.83 | 83.53 | 84.34 |
| 2  | NLP-IIT Patna | C-2 | 17.62 | 51.34 | 73.60 | 83.61 | 84.15 |
| 3  | NLP-IIT Patna | C-1 | 16.63 | 50.54 | 75.04 | 83.06 | 83.79 |
| 4  | FARABI        | C-2 | 12.85 | 45.58 | 81.61 | 81.43 | 81.84 |
| 5  | NCI-MT        | C   | 11.53 | 46.33 | 84.30 | 80.78 | 82.84 |
| 6  | MTG-LIRA      | C   | 10.38 | 42.05 | 82.75 | 78.31 | 81.08 |
| 7  | UPES_NLP      | C   | 9.17  | 36.25 | 96.31 | 69.47 | 73.41 |

Table 25: Contrastive results for Subtask 1B (Hindi → Arabic).

| R. | Team     | Run | BLEU  | ChrF2 | TER    | COMET | mBERT |
|----|----------|-----|-------|-------|--------|-------|-------|
| 1  | FARABI   | C-1 | 17.85 | 50.80 | 72.84  | 83.12 | 84.09 |
| 2  | FARABI   | C-2 | 12.07 | 44.85 | 82.52  | 81.14 | 81.52 |
| 3  | NCI-MT   | C   | 10.70 | 44.71 | 83.84  | 79.92 | 81.47 |
| 4  | UPES_NLP | C   | 7.34  | 32.00 | 100.52 | 65.08 | 69.31 |

Table 26: Contrastive results for Subtask 1C (Bangla → Arabic).

| R. | Team     | Run | BLEU  | ChrF2 | TER   | COMET | mBERT |
|----|----------|-----|-------|-------|-------|-------|-------|
| 1  | FARABI   | C-1 | 18.78 | 51.38 | 71.57 | 83.42 | 84.26 |
| 2  | FARABI   | C-2 | 14.10 | 46.88 | 78.88 | 82.09 | 82.42 |
| 3  | UPES_NLP | C   | 14.02 | 48.82 | 78.45 | 83.08 | 82.75 |

Table 27: Contrastive results for Subtask 1D (Indonesian → Arabic).

| R. | Team          | Run | BLEU  | ChrF2 | TER   | COMET | mBERT |
|----|---------------|-----|-------|-------|-------|-------|-------|
| 1  | FARABI        | C-1 | 18.23 | 50.95 | 72.33 | 83.18 | 84.17 |
| 2  | NLP-IIT Patna | C-2 | 17.07 | 50.74 | 74.55 | 83.43 | 83.91 |
| 3  | NLP-IIT Patna | C-1 | 15.87 | 49.67 | 76.10 | 82.75 | 83.52 |
| 4  | UPES_NLP      | C   | 14.17 | 49.00 | 78.61 | 83.02 | 82.64 |
| 5  | SLPG_FJWU     | C   | 13.35 | 47.17 | 79.12 | 80.22 | 82.44 |
| 6  | FARABI        | C-2 | 11.40 | 44.09 | 85.84 | 80.79 | 81.04 |

Table 28: Contrastive results for Subtask 1E (Urdu  $\rightarrow$  Arabic).

| R. | Team          | Run | BLEU  | ChrF2 | TER   | COMET | BERT  |
|----|---------------|-----|-------|-------|-------|-------|-------|
| 1  | OpenBracket   | C-1 | 34.02 | 59.98 | 58.51 | 82.61 | 82.61 |
| 2  | NLP-IIT Patna | C-2 | 32.05 | 60.02 | 60.15 | 82.33 | 82.33 |
| 3  | OpenBracket   | C-2 | 31.26 | 58.19 | 58.82 | 82.71 | 82.71 |
| 4  | NLP-IIT Patna | C-1 | 30.32 | 57.01 | 59.70 | 82.05 | 82.05 |
| 5  | SLPG_FJWU     | C   | 28.18 | 54.96 | 61.80 | 80.96 | 80.96 |
| 6  | MAGADIR       | C-2 | 28.05 | 56.13 | 62.61 | 81.57 | 81.57 |
| 7  | UPES_NLP      | C   | 25.19 | 53.37 | 63.55 | 81.01 | 81.01 |
| 8  | MAGADIR       | C-1 | 24.24 | 54.49 | 67.69 | 80.00 | 80.00 |
| 9  | FARABI        | C-2 | 24.16 | 54.08 | 69.73 | 78.93 | 78.93 |
| 10 | NCI-MT        | C   | 22.56 | 50.80 | 66.26 | 79.13 | 79.13 |
| 11 | FARABI        | C-1 | 19.53 | 52.85 | 74.65 | 77.50 | 77.50 |

Table 29: Contrastive results for Subtask 2A (Arabic  $\rightarrow$  English).

| R. | Team          | Run | BLEU  | ChrF2 | TER   | COMET | mBERT |
|----|---------------|-----|-------|-------|-------|-------|-------|
| 1  | NLP-IIT Patna | C-1 | 27.87 | 53.50 | 63.35 | 71.76 | 85.12 |
| 2  | NLP-IIT Patna | C-2 | 26.33 | 52.54 | 68.23 | 71.46 | 84.38 |
| 3  | NITM AI Lab   | C-1 | 23.25 | 48.01 | 68.29 | 68.06 | 83.46 |
| 4  | NITM AI Lab   | C-2 | 22.26 | 47.56 | 69.29 | 67.97 | 83.38 |
| 5  | NCI-MT        | C   | 21.41 | 48.11 | 69.19 | 69.38 | 83.53 |
| 6  | FARABI        | C-1 | 18.51 | 44.33 | 76.81 | 64.60 | 81.64 |
| 7  | UPES_NLP      | C   | 18.08 | 46.79 | 72.10 | 70.58 | 83.07 |
| 8  | MTG-LIRA      | C   | 16.97 | 42.35 | 73.99 | 62.03 | 81.77 |
| 9  | FARABI        | C-2 | 16.08 | 42.11 | 78.95 | 63.85 | 80.81 |

Table 30: Contrastive results for Subtask 2B (Arabic  $\rightarrow$  Hindi).

| R. | Team     | Run | BLEU  | ChrF2 | TER   | COMET | mBERT |
|----|----------|-----|-------|-------|-------|-------|-------|
| 1  | FARABI   | C-1 | 13.54 | 46.09 | 80.73 | 74.54 | 81.05 |
| 2  | FARABI   | C-2 | 9.32  | 40.72 | 85.70 | 72.31 | 79.26 |
| 3  | NCI-MT   | C   | 9.02  | 43.16 | 83.18 | 77.31 | 80.88 |
| 4  | UPES_NLP | C   | 8.94  | 44.76 | 84.46 | 78.45 | 81.10 |

Table 31: Contrastive results for Subtask 2C (Arabic  $\rightarrow$  Bangla).

| R. | Team     | Run | BLEU  | ChrF2 | TER   | COMET | mBERT |
|----|----------|-----|-------|-------|-------|-------|-------|
| 1  | FARABI   | C-2 | 24.55 | 57.19 | 67.61 | 83.76 | 84.52 |
| 2  | UPES_NLP | C   | 22.77 | 55.93 | 64.65 | 84.15 | 84.82 |
| 3  | FARABI   | C-1 | 21.43 | 51.23 | 67.13 | 82.01 | 83.84 |

Table 32: Contrastive results for Subtask 2D (Arabic  $\rightarrow$  Indonesian).

| R. | Team          | Run | BLEU  | ChrF2 | TER   | COMET | mBERT |
|----|---------------|-----|-------|-------|-------|-------|-------|
| 1  | NLP-IIT Patna | C-1 | 25.68 | 51.04 | 65.15 | 76.33 | 83.53 |
| 2  | NLP-IIT Patna | C-2 | 24.65 | 50.17 | 69.55 | 76.00 | 82.95 |
| 3  | SLPG_FJWU     | C   | 23.35 | 48.81 | 67.02 | 74.53 | 82.87 |
| 4  | FARABI        | C-2 | 17.61 | 43.56 | 76.20 | 70.94 | 80.56 |
| 5  | FARABI        | C-1 | 17.46 | 43.68 | 75.31 | 71.75 | 80.77 |
| 6  | UPES_NLP      | C   | 16.60 | 44.67 | 72.53 | 75.45 | 81.45 |

Table 33: Contrastive results for Subtask 2E (Arabic  $\rightarrow$  Urdu).