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

Sumit Singh, Anish Kumar Vishwakarma, Brijmohan Lal Sahu and Ashwani Kumar
UPES, Dehradun, India
{sumit.singh, anishk.vishwakarma, brijmohan.sahu, ashwani.kumar}@ddn.upes.ac.in

Abstract

This paper presents the UPES_NLP submission to the WMT 2026 Low-Resource Arabic–Asian Machine Translation Shared Task. The shared task focuses on improving machine translation between Arabic and several low-resource Asian languages, where limited parallel data remains a significant challenge. Our system utilized the multilingual No Language Left Behind (NLLB-200) Mixture-of-Experts (MoE) model on the official parallel corpora released by the organizers for translation inference. We also generated translations using GPT-5-mini. For GPT-based translation, we employed an entity-aware prompting strategy that explicitly instructed the model to preserve named entities. We participated in both the Primary and Contrastive tracks across multiple Arabic–Asian language pairs using a unified multilingual inference framework. Experimental results demonstrate that the proposed approach achieves competitive performance across diverse translation directions. Our system ranked 3rd in the Primary Arabic→English task with a BLEU score of 32.08 and 2nd in the Contrastive Arabic→Indonesian task with a BLEU score of 22.77, while also obtaining competitive rankings on several other language pairs. The competitive performance achieved across multiple language pairs demonstrates the capability of the proposed multilingual framework to address low-resource Arabic–Asian translation tasks.

1 Introduction

Recent advances in multilingual neural machine translation (NMT) have significantly improved translation quality across a wide range of languages. Large pretrained transformer models, such as No Language Left Behind (NLLB-200) (Costa-Jussà et al., 2022), leverage multilingual learning to transfer knowledge across language pairs, making them particularly effective for low-resource translation. However, translating between Arabic and many Asian languages remains challenging due to the limited availability of parallel corpora and substantial linguistic differences in morphology, syntax, and writing systems (Laskar et al., 2026). The WMT 2026 Low-Resource Arabic–Asian Machine Translation Shared Task provides a common benchmark to evaluate translation systems for these challenging language pairs under constrained and unconstrained settings.

In this paper, we present the UPES_NLP submission to the WMT 2026 shared task. We adopt a unified multilingual approach by performing inference with the pretrained NLLB-200 Mixture-of-Experts (MoE) model using the official parallel datasets provided by the organizers. Our objective is to assess the effectiveness of a single multilingual architecture across multiple Arabic–Asian translation directions while maintaining a consistent inference strategy.

2 Related Work

Multilingual neural machine translation has advanced considerably with the emergence of large-scale pretrained transformer models capable of learning from hundreds of languages simultaneously. Among these, NLLB-200 (Costa-Jussà et al., 2022) has demonstrated strong multilingual transfer and has become a widely adopted model for improving translation quality in both high-resource and low-resource language pairs. Its extensive multilingual pretraining enables effective knowledge sharing across linguistically diverse languages, making it particularly suitable for translation tasks with limited parallel data.

For translation involving Indian languages, IndicTrans2 (Gala et al., 2023) and (Singh et al., 2025b) have shown competitive performance by providing robust multilingual representations tailored to the Indic language family. In parallel, the OPUS-MT (Tiedemann and Thottingal, 2020) project has

Proceedings of the Eleventh Conference on Machine Translation, pages 2338–2343
October 28-29, 2026 ©2026 Association for Computational Linguistics

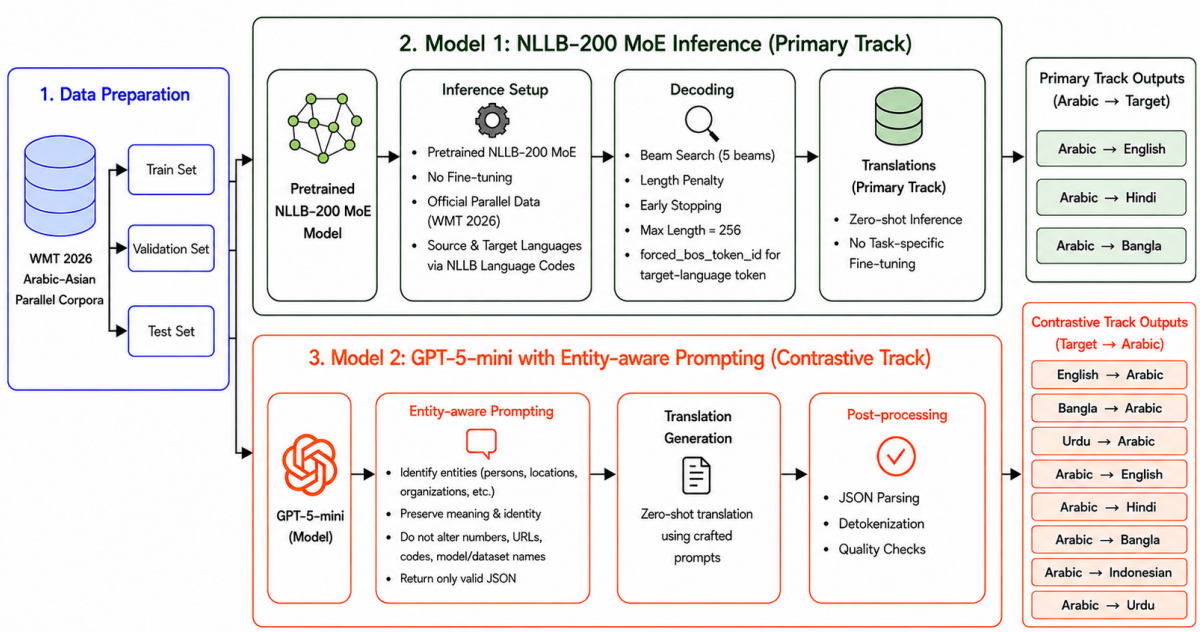

Figure 1: Overview of the UPES_NLP system for the WMT 2026 Low-Resource Arabic–Asian Machine Translation Shared Task.

released a collection of open-source neural translation models covering numerous language pairs, serving as reliable baselines for multilingual machine translation research. When direct bilingual resources are scarce, pivot-based translation through an intermediate language has also been explored as an effective strategy for improving translation quality between low-resource language pairs.

The WMT 2026 Low-Resource Arabic–Asian Machine Translation Shared Task provides a standardized benchmark for evaluating multilingual translation systems across several Arabic–Asian language pairs. Building upon recent advances in multilingual pretrained models, our work investigates the effectiveness of the pretrained NLLB-200 MoE model for inference on the official shared-task datasets across multiple translation directions using a unified inference framework.

3 Dataset

The WMT 2026 Low-Resource Arabic–Asian Machine Translation Shared Task focuses on bilingual machine translation between Arabic and five languages, namely English, Hindi, Bangla, Indonesian, and Urdu. The shared task is organized into two categories: *translation into Arabic* (English→Arabic, Hindi→Arabic, Bangla→Arabic, Indonesian→Arabic, and Urdu→Arabic) and *translation from Arabic* (Arabic→English, Arabic→Hindi, Arabic→Bangla, Arabic→Indonesian, and Arabic→Urdu). The objective is to develop robust machine translation systems under low-resource conditions using the officially released parallel corpora (Laskar et al., 2026).

Table 1: Number of sentence pairs in the training, validation, and test sets for each Arabic-centric language pair.

Language Pair	Train	Validation	Test
Arabic–Bengali	20,500	1,000	500
Arabic–English	21,000	1,000	500
Arabic–Hindi	20,100	1,000	500
Arabic–Indonesian	20,400	1,000	500
Arabic–Urdu	20,300	1,000	500

The organizers provide official training and development datasets for each language pair, while the blind test set is released only during the evaluation phase to registered participants. The constrained setting requires systems to use the official parallel data provided by the organizers. The datasets are designed to benchmark translation quality for low-resource Arabic-centric language pairs and encourage the development of multilingual and transfer learning approaches (Laskar et al., 2026). Table 1 summarizes the statistics of the official datasets used in our experiments.

Table 2: Representative English→Arabic prompt used with GPT-5-mini. The source and target languages were changed accordingly for the other translation directions. The prompts were designed according to the shared-task requirements. The system prompt defines the translation role, target language, entity-preservation rules, and output format, while the user prompt specifies item-level instructions. Entity-related instructions reduce mistranslation of names and technical terms, whereas IDs and JSON formatting enable reliable batch processing and source–output alignment.

Prompt Type	Prompt Content

System Prompt	`You are a professional English-to-Arabic translation engine with entity-awareness capabilities.`
	
	`Translate English text into fluent and contextually accurate Modern Standard Arabic.`
	
	`Identify named entities in the source text, including persons, locations, organizations, products, events, dates, numerical expressions, and other domain-specific entities.`
	
	`Preserve the semantic identity of each entity during translation. Translate or transliterate named entities only when appropriate for Arabic usage and context.`
	
	`Do not alter numbers, URLs, email addresses, tags, code, file paths, dataset names, or model names.`
	
	`Ensure that entity mentions remain consistent throughout the translation.`
	
	`Return only valid JSON. Do not include explanations, markdown, comments, or additional text.`

User Prompt	`Translate the following English items into Arabic using entity-aware translation.`
	
	`Requirements:`
	
	`- Translate the text into fluent Modern Standard Arabic.`
	`- Keep the same id for every input item.`
	`- Do not skip any item.`
	`- Identify named entities before generating the translation.`
	`- Preserve the meaning and identity of named entities.`
	`- Translate or transliterate entity mentions according to their type and Arabic linguistic conventions.`
	`- Maintain consistency when the same entity occurs multiple times.`
	`- Do not translate model names, dataset names, URLs, email addresses, file paths, tags, or code.`
	`- Preserve numerical values exactly as provided in the source.`
	`- Use the surrounding context to resolve ambiguous entity mentions.`
	`- Preserve the complete source text in the output.`
	`- Return only one accurate and deterministic translation for each item.`
	`- Do not provide alternative translations.`
	`- Avoid unnecessary paraphrasing.`
	`- Preserve the exact semantic meaning of the source sentence.`
	`- Return only a valid JSON array.`
	
	`Input:`
	`{input_json}`
	
	`Expected JSON format:`
	`{`
	`  "id": 1,`
	`  "source": "English text here",`
	`  "translation": "Arabic translation here"`
	`}`

Table 3: Performance of *UPES_NLP* on the Primary Track.

Sub-task	Rank	BLEU	Chrf2	TER	COMET	BERT
Arabic → English	3	32.0810	58.1849	59.0002	81.4204	82.4507
Arabic → Bangla	6	12.1357	44.8213	78.0033	77.3104	81.7134
Arabic → Hindi	8	22.6062	47.1467	67.1573	68.9156	83.8687

Table 4: Performance of *UPES_NLP* on the Contrastive Track.

Sub-task	Rank	BLEU	Chrf2	TER	COMET	BERT
English → Arabic	9	13.0175	44.2661	80.2363	80.1996	81.5770
Bangla → Arabic	4	7.3449	32.0004	100.5216	65.0759	69.3146
Urdu → Arabic	4	14.1707	49.0035	78.6052	83.0154	82.6355
Arabic → English	7	25.1900	53.3732	63.5524	81.0060	81.0060
Arabic → Hindi	7	18.0808	46.7857	72.1027	70.5833	83.0656
Arabic → Bangla	4	8.9434	44.7616	84.4642	78.4463	81.0970
Arabic → Indonesian	2	22.7712	55.9266	64.6548	84.1456	84.8152
Arabic → Urdu	6	16.5966	44.6702	72.5319	75.4456	81.4491

4 System Description

Our system uses two approaches for Arabic-centric machine translation. First, we performed inference using the pretrained NLLB-200 MoE model (Koishekenov et al., 2023) for selected Arabic-to-target language pairs. Second, we used GPT-5-mini with direction-specific, entity-aware prompts for eight language pairs showed in Table 4 in the Contrastive Track. Figure 1 provides an overview of our two approaches for Arabic-centric machine translation. The two approaches are described in detail in the following subsections.

4.1 NLLB-200-MoE Inference (Costa-Jussà et al., 2022; Koishekenov et al., 2023)

For the Primary Track, we used the pretrained NLLB-200 MoE model directly for multilingual translation inference. The three subtasks considered were Arabic→English, Arabic→Hindi, and Arabic→Bangla. The corresponding NLLB language codes were used to specify the source and target languages. The Arabic input sentences were translated into the respective target languages while preserving the original meaning, named entities, numbers, punctuation, factual information, and sentence-level structure. During decoding, the required target-language token was specified using `forced_bos_token_id`. Beam-search decoding with 5 beams and a maximum output length of 256 tokens was used to generate the final translations. No task-specific fine-tuning was performed; the system relies entirely on direct inference from the pretrained NLLB-200 MoE checkpoint.

4.2 GPT-5-mini with Entity-Aware Prompting

In the Contrastive Track, we participated in eight translation subtasks: Arabic→Bangla, Arabic→English, Arabic→Hindi, Arabic→Indonesian, Arabic→Urdu, Bangla→Arabic, English→Arabic, and Urdu→Arabic. The source and target languages in the prompts were adjusted for each translation direction. For the Contrastive Track, we used GPT-5-mini (Singh et al., 2025a) for prompt-based machine translation. The prompt instructed the model to generate fluent text in the required target language while preserving the meaning of the source sentence and important entities such as person names, locations, organizations, numbers, URLs, dataset names, and model names. The model was further instructed to translate or transliterate named entities according to the target-language context and to keep repeated entity mentions consistent. The output was restricted to a structured JSON format containing the sentence identifier, source text, and generated translation.

For efficient inference, the input data were divided into batches of 50 sentences. Up to five batches were processed concurrently using parallel API requests. Since batches could finish in a different order, the generated outputs were finally sorted using their original identifiers before being saved. This setup allowed us to compare a pretrained multilingual translation model with a prompt-based

LLM approach for low-resource Arabic-centric machine translation. We used gpt-5-mini through the OpenAI Responses API with default settings. An example of prompt for English to Arabic is shown in Table 2.

5 Results and Analysis

Tables 3 and 4 present the performance of the *UPES_NLP* system on the Primary and Contrastive tracks. In the Primary Track, the best result was obtained for Arabic→English, where the system achieved 3rd rank with a BLEU score of 32.08 and a COMET score of 81.42. Arabic→Bangla achieved 6th rank, while Arabic→Hindi obtained 8th rank. Although the BLEU scores for these language pairs were lower, their BERT-based scores remained above 81, indicating reasonable semantic preservation.

In the Contrastive Track, Arabic→Indonesian produced the strongest result, achieving 2nd rank with a COMET score of 84.15 and a BERT score of 84.82. Bangla→Arabic and Urdu→Arabic both achieved 4th rank. Among these, Urdu→Arabic showed better overall translation quality with a COMET score of 83.02. For the language pairs common to both tracks, the Primary Track generally achieved better BLEU scores. For example, Arabic→English decreased from 32.08 BLEU in the Primary Track to 25.19 in the Contrastive Track. A similar trend was observed for Arabic→Hindi and Arabic→Bangla.

However, these results should be interpreted only as a comparison between the submitted systems and not as a controlled comparison between NLLB and GPT. Although both systems were evaluated on the same official test sets using the same evaluation metrics, they differed in architecture, prompting or decoding procedure, and other inference settings. GPT-5-mini performed particularly well for Arabic→Indonesian, achieving second place. This may be related to Indonesian's standardized Latin script, relatively simple morphology, and stronger representation in multilingual web data. However, rankings also depend on the number and performance of competing systems; therefore, these explanations remain hypotheses rather than conclusions from a controlled analysis. Overall, the results show that the system performed particularly well for Arabic→English and Arabic→Indonesian. The differences between BLEU and semantic metrics such as COMET and BERT also suggest that some translations preserved the intended meaning even when the wording differed from the reference.

6 Conclusion

This paper presented the *UPES_NLP* system for the WMT 2026 Low-Resource Arabic–Asian Machine Translation Shared Task. We explored two complementary approaches: pretrained NLLB-200 MoE inference for selected Arabic-to-English, Arabic-to-Hindi, and Arabic-to-Bangla translation directions in the Primary Track, and prompt-based translation using GPT-5-mini with entity-preserving instructions in the Contrastive Track. The proposed systems achieved competitive results across both tracks, including 3rd rank for Arabic→English in the Primary Track and 2nd rank for Arabic→Indonesian in the Contrastive Track. The results show that pretrained multilingual inference and carefully designed LLM prompting can be useful for low-resource Arabic-centric machine translation.

References

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Jay Gala, Pranjal A Chitale, Raghavan Ak, Varun Gumma, Sumanth Doddapaneni, 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*.

Yeskendir Koishekenov, Alexandre Berard, and Vassilina Nikoulina. 2023. Memory-efficient NLLB-200: Language-specific expert pruning of a massively multilingual machine translation model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3567–3585, Toronto, Canada. Association for Computational Linguistics.

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.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al.

2025a. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

Sumit Singh, Pankaj Goyal, and Uma Tiwary. 2025b. silp_nlp at SemEval-2025 task 2: An effect of entity awareness in machine translation using LLM. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2389–2394, Vienna, Austria. Association for Computational Linguistics.

Jörg Tiedemann and Santhosh Thottingal. 2020. Opus-mt–building open translation services for the world. In *Annual Conference of the European Association for Machine Translation*, pages 479–480. European Association for Machine Translation.