

NCI-MT at WMT 2026: A Quality-Gated Cascade for Low-Resource Arabic-Asian Machine Translation

Satarupa Deb

Independent Researcher
West Bengal, India
hellosatarupadeb@gmail.com

Prashanth Nayak

National College of Ireland
Dublin, Ireland
Prashanth.Nayak@ncirl.ie

Abstract

This paper describes NCI-MT’s submission to the low-resource Arabic-to-Asian language translation task as part of the WMT shared task. Our submission covered six of the ten language pairs. We built two separate systems, the primary and the contrastive. The primary was built using the data provided by WMT. In this approach, we fine-tuned our baseline model (NLLB) with per-direction LoRA adapters. For the contrastive system, we introduced an additional quality-gated refinement stage, which includes two stages: in the first stage, we use a reference-free quality estimator to evaluate the quality of the translations produced by our base model. In the second stage, we iteratively re-translate those translations that are below a certain threshold using LLM-based in-context learning. Our results show that both systems are successful in improving the translation quality in low-resource scenarios.

1 Introduction

In recent times, we have seen significant improvements in translation quality across multiple domains and languages. This improvement can be attributed to LLMs, which are the current state-of-the-art in the field. Although LLMs have shown tremendous capabilities, they still struggle with low-resource languages. Low-resource languages such as Arabic, Bengali, and Hindi pose significant constraints for translation systems due to the limited availability of data, script variations, and the morphological richness of these languages. These factors often result in suboptimal translation performance (Wang et al., 2021). Even with strong multilingual pretraining, translating for low-resource language pairs remains a difficult task. This motivated us to participate in the WMT 2026 low-resource Arabic-to-Asian shared task (Laskar et al., 2026), which includes translation between Arabic and five target languages. However, we restrict our experiments to English, Hindi, and Bangla because of

the authors’ proficiency in these languages, thereby allowing for qualitative analysis through manual evaluation of the translation errors.

As mentioned above, we carried out our experiments for six translation directions: Arabic-to-English, Arabic-to-Hindi, and Arabic-to-Bangla in both directions. Our setup consisted of two different settings, namely the primary and the contrastive. The primary systems are fine-tuned versions of the NLLB model with parameter-efficient LoRA adapters. These systems strictly used the official data provided by WMT organisers.

Our contrastive systems are built as a quality-gate cascade. In this approach, the translations generated by our baseline model (NLLB) are passed through a reference-free quality estimator. These scores are then compared against a set threshold value. Translations whose quality score falls below a threshold go through a translation improvement stage. In this stage, we used in-context learning, where we retrieve examples similar to the source sentence being translated and append them to the prompt provided to an LLM. Only translations that improve are retained.

Our main contributions in this system description paper include an empirical study of retrieval-augmented cascade refinement across six Arabic-centric low-resource directions, showing consistent gains over the baseline, and a domain-guided corpus selection procedure for the contrastive mining pool based on unsupervised clustering of data. The rest of the paper is organised as follows: Section 2 reviews related work. Section 3 describes the data used in our experiments. Section 4 discusses in detail our primary and contrastive systems. In Section 5 we discuss our experimental results. We conclude with limitations and provide avenues for future work in Sections 6 and 7.

2 Related Work

Large language models have shown strong translation capabilities, with prompting strategies, in-context learning, and translation memory all contributing to their performance (Zhu et al., 2024). For low-resource directions, the choice of context is of utmost priority. The similarity-based retrieval of bitexts improves translation quality over random selection, particularly for low-resource language pairs (Zebaze et al., 2025). However, the benefit depends heavily on what is being retrieved. Even small LLMs can exploit prompt context for zero-shot low-resource translation, but only when sufficient relevant linguistic information is provided (Court and Elsner, 2024). When parallel data is scarce, in-context examples can instead be generated from monolingual resources (Lee et al., 2025).

For low-resource pairs with little direct parallel data, pivoting through a high-resource language is a standard technique. Pivot-based training can exploit monolingual pivot data when source-pivot parallel data is unavailable (Currey and Heafield, 2019). Beyond pivoting, the choice of training or retrieval data can be guided automatically by using sentence embeddings to form interpretable domain clusters without supervision (Aharoni and Goldberg, 2020), and cross-lingual embedding similarity can select in-domain data for domain adaptation (Vu et al., 2021).

Retrieval-augmented decoding in NMT has largely been explored at the token level, as in k -nearest-neighbor MT, where the model consults a datastore during decoding (Zheng et al., 2021). The domain mismatch between the datastore and the model remains a key challenge (Cao et al., 2023). In contrast, sentence-level quality estimation with reference-free metrics such as COMETKiwi (Rei et al., 2022) enables selective refinement of complete translations at inference time.

3 Dataset Description

This section describes the data used in our experiments. We normalise whitespaces, strip empty lines, and cap sequence length at 256 subword tokens for both training and inference. We provide data description in Table 1.

3.1 Other Data

For our contrastive systems, we created a mining pool (c.f. Section 4) for which we sourced additional data from the following resources: FLO-

Language Pair	Train	Dev	Test
en-ar	21,000	1,000	500
hi-ar	20,100	1,000	500
bn-ar	20,500	1,000	500

Table 1: Official data provided by the task organisers

RES+(Goyal et al., 2022),¹ Tatoeba(Tiedemann, 2020),² OPUS-100,³ UNPC(Ziemski et al., 2016),⁴ News Commentary,⁵ and Samanantar (Ramesh et al., 2022). We use this data as part of in-context learning in our experiments.

4 System Description

In this section, we discuss our primary and contrastive systems in detail.

4.1 Primary system

Our primary system used the NLLB(Team et al., 2022) model as the baseline. Specifically, we used the NLLB-200-distilled-600M version for our experiments. We fine-tuned the model using the data provided in Table 1. We trained a separate LoRA adapter for each of the six translation directions. We applied LoRA to the attention projections (q_proj, k_proj, v_proj, out_proj) of every transformer layer, with rank $r = 16$, scaling $\alpha = 32$, dropout of 0.05, and no bias adaptation. This gives approximately 4.7M trainable parameters, or $\approx 0.33\%$ of the full model. All other parameters, including embeddings and feed-forward layers, remain frozen. Our hyperparameter setup used the default setting.

4.2 Contrastive System

Our contrastive systems use the NLLB model as the baseline. The translations generated by NLLB are evaluated using COMETKiwi,⁶ a reference-free translation-quality metric. These scores were compared against a threshold, which we set to 0.75 in our experiments. Translations that fell below the threshold went through a translation improvement stage, in which we refined them using in-context learning, i.e., we passed the source sentence, retrieved similar examples from the mining pool described in Section 4.4, and appended them to the

¹<https://github.com/openlanguagedata/flores>

²<https://tatoeba.org>

³<https://opus.nlpl.eu>

⁴<https://conferences.unite.un.org/UNCORPUS>

⁵<https://data.statmt.org/news-commentary/>

⁶<https://huggingface.co/Unbabel/wmt22-cometkiwi-da>

```

Reference translations into
{target_lang} for similar sentences:
Source: <retrieved src 1>
{target_lang}: <retrieved tgt 1>
Source: <retrieved src 2>
{target_lang}: <retrieved tgt 2>
Source: <retrieved src 3>
{target_lang}: <retrieved tgt 3>

Source:
<query source>

A draft {target_lang} translation:
<NLLB baseline output>

Using the references as a guide,
produce an improved, accurate
{target_lang} translation. Output
only the translation, without
explanation.

```

Figure 1: Sample prompt used for in-context learning.

prompt provided to the LLM to perform translation. Only translations that improved were retained. We have shared Sample prompt used for in-context learning in Figure 1.

4.3 Domain-guided data pool creation

We investigated the training data provided by WMT. Reflecting on the baseline translations across all directions, we concluded that one of the factors limiting the model was limited domain coverage. Our goal was to have a more morphologically rich data pool. To ensure that the mining pool covered domains relevant to the shared task, we first ran an unsupervised domain clustering analysis on the training data. Following Aharoni and Goldberg (2020), we encoded each language side with LaBSE (Feng et al., 2022), reduced the resulting 768-dimensional embeddings to 50 dimensions with PCA, and fit a Gaussian Mixture Model with full covariance ($k = 5$).

We wanted to label each cluster with a domain name to help with data selection for the mining pool. We observed that each cluster is characterised by its most distinctive words. However, stopwords are numerous and could skew the clusters due to their high counts. Before computing each remaining word’s TF-IDF value, we removed stopwords in all four languages (e.g., "if", "is", "and") as well as any word appearing in a large fraction of the corpus, since such words carry no topical information. The remaining words are ranked by TF-IDF, so that terms frequent within a cluster but rare elsewhere are ranked highest. We use LaBSE to filter polyse-

mous terms whose representations and meanings shift across clusters. Finally, a natural-language domain label is generated for each cluster by an LLM; in our case, we used Gemma-3-4B (Team et al., 2025). For non-English text, NLLB-200 translates the text before passing it to the LLM.

Domains recovered from the above experiments on the training data included labels such as Middle East politics, financial markets, sports (particularly football/soccer), cybersecurity and digital society, and cultural and social discussion. These domain labels informed our selection of external corpora for the mining pool. We chose sources whose expected content overlapped with the discovered domains. These sources are described in Section 3. They were added to the mining pool, indexed with LaBSE and FAISS, and made available to the retrieval-augmented refinement step described in Section 4.2.

4.4 Synthetic data creation for low-to-low translation directions

Our in-context learning approach relies on similar parallel data to achieve a good outcome. However, we found that there was a lack of good publicly available parallel data for Hindi-to-Arabic and Bangla-to-Arabic (both directions). Also, we found that the baseline translations for these languages were of poor quality. This made us use a pivot language. We chose English as the pivot language given that it is a high-resource language, and most models translate well to and from English. Our translation of Bangla-to-Arabic was therefore split into Bangla-to-English and English-to-Arabic similarly for Hindi. In this scenario, we took Hindi–English and Bangla–English pairs from Samanantar (Ramesh et al., 2022) and translated the English side into Arabic with NLLB-200, generating synthetic Hindi–Arabic and Bangla–Arabic pairs, which we then used as the mining pool for in-context learning.

5 Results

In this section, we discuss our results in detail. Table 2 shows the results for both primary and contrastive systems for each language pair. From the table, we see that the primary system outperforms the contrastive systems on BLEU, chrF2, and TER in all six directions. Our contrastive system, even though it undergoes a translation refinement stage (cf. Section 5), does not outperform the primary

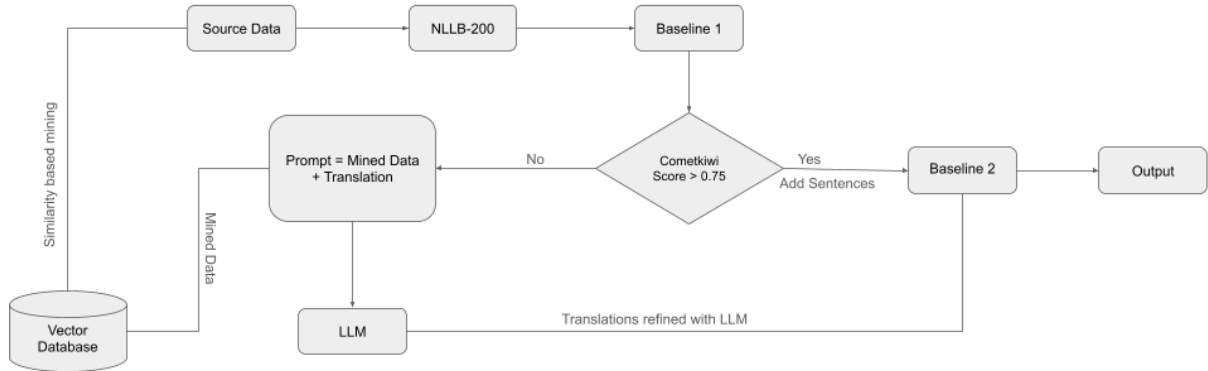


Figure 2: The contrastive system pipeline. NLLB-200 generates a draft translation in baseline1. COMETKiwi scores it without a reference; if the score is at least 0.75, the translation is kept. Otherwise, LaBSE+FAISS retrieves similar source–target pairs from the contrastive mining pool, Gemma-3-4B refines the translations from baseline1 using those examples in context, and CometKiwi re-scores. The refined translation is kept only if it improves over the baseline1.

systems. The results therefore show that quality-gated LLM refinement of the baseline model, while beneficial relative to that baseline, does not recover the gains obtained by direct LoRA fine-tuning on the official training data.

Interestingly, we also see from Table 2 that the evaluation on the COMET metric shows that the contrastive system matches or exceeds the primary system on COMET across four directions (hi-to-ar, bn-to-ar, ar-to-hi, ar-to-bn; Section 5.1), despite not improving on BLEU, chrF2, and TER. This suggests that Gemma’s refinements diverge lexically from the references, which penalises n-gram-based metrics, while preserving or improving adequacy as judged by a neural metric. It is to be noted that the improvement in the COMET metric may be due to our using COMET-based estimators in the quality-gate cascade (cf. section 5), which selects refinements scoring higher on COMETKiwi. This may provide some of the COMET advantage and possibly therefore reflect metric affinity rather than genuine quality differences.

5.1 Refinement Rates on the challenge dataset

Table 3 shows how many sentences passed the quality gate and how many needed refinement in each direction. The differences between directions are large. For en-to-ar, only 5.8% of sentences scored below the 0.75 threshold and went through refinement. For ar-to-bn, the figure is 42.8%, nearly eight times higher.

We observed that, NLLB is better at translating English, Hindi, and Bangla into Arabic than out

of it. The three into-Arabic directions needed refinement for 5.8%, 31.3%, and 32.2% of sentences, while the three out-of-Arabic directions needed it for 13.0%, 41.0%, and 42.8%. We can conclude that this could be due to the morphological complexity of Arabic, and additionally due to the composition of NLLB’s training corpus.

The refinement rates match what we would expect from language difficulty. English is the easiest partner language in both directions, due to its large training data corpus, while Hindi and Bangla, being on the lower resource side, are harder.

6 Conclusion

We presented NCI-MT’s WMT 2026 submission for the Low-Resource Arabic–Asian Language Translation shared task (Laskar et al., 2026). Our primary system fine-tunes NLLB-200-distilled-600M with LoRA adapters on the official training data. Our contrastive system adds a quality-gated cascade in which weak NLLB translations from baseline1 are refined by Gemma-3-4B with retrieved in-context examples, accepting refinements only when a reference-free quality estimator judges them to be improvements. To construct the mining pool, we ran unsupervised domain clustering on the training data and used the discovered domain labels to guide our choice of external corpora. On the official challenge test set, the primary system outperforms the contrastive cascade on BLEU, chrF2, and TER in all six directions, while the cascade matches or exceeds the primary system on COMET in the four directions.

Direction	System	BLEU↑	chrF2↑	TER↓	COMET↑	mBERT↑
en-to-ar	Primary	16.01	50.59	76.01	81.95	83.62
	Contrastive	12.93	49.34	82.42	81.82	82.84
hi-to-ar	Primary	13.10	47.41	80.54	80.08	82.35
	Contrastive	11.53	46.33	84.30	80.78	82.84
bn-to-ar	Primary	12.01	46.39	81.43	79.37	82.03
	Contrastive	10.70	44.71	83.84	79.92	81.47
ar-to-en	Primary	26.73	53.83	62.85	79.91	80.44
	Contrastive	22.56	50.80	66.26	79.13	79.13
ar-to-hi	Primary	24.47	50.23	67.05	69.11	84.12
	Contrastive	21.41	48.11	69.19	69.38	83.53
ar-to-bn	Primary	15.09	49.40	76.26	76.97	82.43
	Contrastive	9.02	43.16	83.18	77.31	80.88

Table 2: Official challenge test set results for NCI-MT across the six sub-tasks, as computed by the task organizers (Laskar et al., 2026).

Dir.	Kept	Refined	Ref. (%)	Err.
en-to-ar	1,413	87	5.8	0
hi-to-ar	1,030	470	31.3	0
bn-to-ar	1,017	483	32.2	0
ar-to-en	1,305	195	13.0	0
ar-to-hi	885	615	41.0	0
ar-to-bn	858	642	42.8	0

Table 3: Challenge test set refinement breakdown (1,500 sentences per direction). “Refined” counts sentences where Gemma’s refinement scored better than the NLLB(baseline 1) and was kept; “Err.” counts generation errors (Ollama timeouts, malformed outputs).

7 Future Work

We plan to extend the quality-gated cascade into a hierarchical system that refines the primary system’s output, and to validate it across a wider range of quality thresholds. We also aim to extend the data corpus and evaluate different quality estimators to study how their scores vary and how this affects the refinement rate.

Acknowledgments

We thank the WMT 2026 Low-Resource Arabic–Asian MT shared task organizers for the dataset and the extended evaluation cycle. We also thank the National College of Ireland for funding this research and for continuous support.

References

Roei Aharoni and Yoav Goldberg. 2020. [Unsupervised domain clusters in pretrained language models](#). *CoRR*, abs/2004.02105.

Zhiwei Cao, Baosong Yang, Huan Lin, Suhang Wu, Xiangpeng Wei, Dayiheng Liu, Jun Xie, Min Zhang, and Jinsong Su. 2023. [Bridging the domain gaps in context representations for \$k\$ -nearest neighbor neural](#)

[machine translation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5841–5853, Toronto, Canada. Association for Computational Linguistics.

Sara Court and Micha Elsner. 2024. [Shortcomings of LLMs for low-resource translation: Retrieval and understanding are both the problem](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 1332–1354, Miami, Florida, USA. Association for Computational Linguistics.

Anna Currey and Kenneth Heafield. 2019. [Zero-resource neural machine translation with monolingual pivot data](#). In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 99–107, Hong Kong. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Ariavazhagan, 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.

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.

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.

Dohyun Lee, Seungil Chad Lee, Chanwoo Yang, Yujin Baek, and Jaegul Choo. 2025. [Exploring in-context example generation for machine translation](#). In *Findings of the Association for Computational Linguistics*:

- ACL 2025, pages 26554–26568, Vienna, Austria. Association for Computational Linguistics.
- Gowtham Ramesh, Sumanth Doddapaneni, Aravindh 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, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. [CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. [Gemma 3 technical report](#). *Preprint*, arXiv:2503.19786.
- 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.
- Jörg Tiedemann. 2020. [The tatoeba translation challenge – realistic data sets for low resource and multilingual MT](#). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.
- Thuy-Trang Vu, Xuanli He, Dinh Phung, and Gholamreza Haffari. 2021. [Generalised unsupervised domain adaptation of neural machine translation with cross-lingual data selection](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3335–3346, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Rui Wang, Xu Tan, Renqian Luo, Tao Qin, and Tie-Yan Liu. 2021. [A survey on low-resource neural machine translation](#). In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, IJCAI-21, pages 4636–4643. International Joint Conferences on Artificial Intelligence Organization. Survey Track.
- Armel Randy Zebaze, Benoît Sagot, and Rachel Bawden. 2025. [In-context example selection via similarity search improves low-resource machine translation](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1222–1252, Albuquerque, New Mexico. Association for Computational Linguistics.
- Xin Zheng, Zhirui Zhang, Junliang Guo, Shujian Huang, Boxing Chen, Weihua Luo, and Jiajun Chen. 2021. [Adaptive nearest neighbor machine translation](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 368–374, Online. Association for Computational Linguistics.
- Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. [Multilingual machine translation with large language models: Empirical results and analysis](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.
- Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. [The United Nations parallel corpus v1.0](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).