

BM25-Augmented Many-Shot Translation for Low-Resource North-Eastern Indian Languages

Aashish Dhawan, Christopher Driggers-Ellis *, Dzmitry Kasinets, Daisy Zhe Wang, Christan Grant

University of Florida | Data Studio, DSR Lab | ufdatastudio.com, dsr.cise.ufl.edu | driggersellis.cw@ufl.edu *

Abstract

We present the *gators* submission to the Indic Translation Shared Task on automatic translation between English (En) and eleven Northeast Indian languages of varying resource scarcity. We adapt the retrieval-augmented many-shot translation pipeline from our AmericasNLP 2026 system (Dhawan et al., 2026). At inference time, BM25 indexing (Robertson and Zaragoza, 2009) retrieves the most similar parallel examples from a language-specific training bank and Gemini 2.5 Flash (Comanici et al., 2025) performs translation conditioned on this context. No fine-tuning is involved. Retrieval banks combine official WMT26 data with other publicly available corpora. A grid search over retrieval count r and development exemplar count d across all 22 language-direction pairs selects the best configuration for each submission.

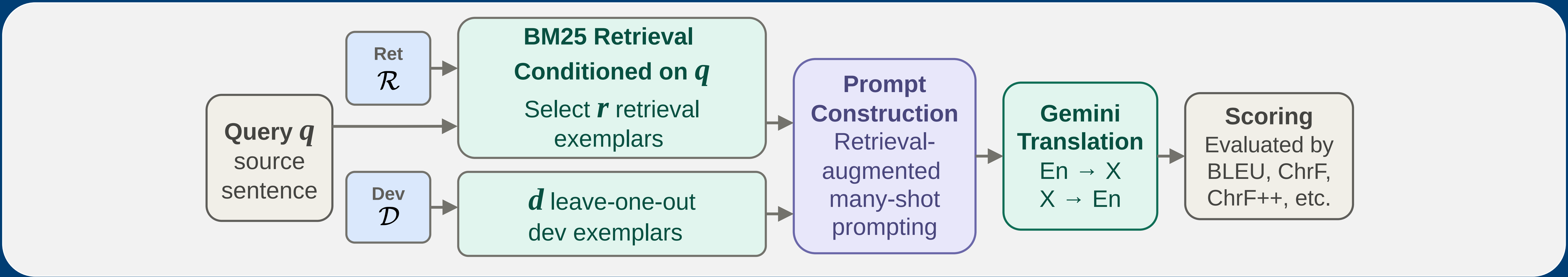


Figure 1: Our Retrieval Augmented Machine Translation pipeline. The query q in the source language is passed through the BM25 retrieval index where the top- r exemplars are selected. These r retrieval exemplars, q , and d leave-one-out dev exemplars are input to Prompt Construction, where the final prompt for Gemini 2.5 Flash takes shape. Since the competition tests on both En-X and X-En translation, both En and Indic languages (X) can be the source and target.

Introduction: Transfer and In-Context Learning

We reproduce the machine translation component of the system we submitted to the AmericasNLP 2026 Shared Task on Cultural Image Captioning for Indigenous Languages (Dhawan et al., 2026). That system is a training-free in-context learning (ICL) approach for captioning images in Indigenous languages, a setting where VLMs cannot normally generate coherent text.

That method produces Spanish (Es) captions of the images as an interlingua and uses BM25 retrieval-augmented many-shot translation with Gemini 2.5 Flash. It uses paired Es-X training data from previous AmericasNLP competitions as retrieval context (Ebrahimi, Mager, et al., 2023; Ebrahimi et al., 2024).

Figure 1 details the final pipeline in this work. We remove any captioning portion and feed query texts q in the translation task's English (En) or Indic source language directly into BM25.

Data Preparation

We augment the data made available by shared task organizers with datasets available on the Internet from a variety of domains. Like AmericasNLP, we also attempt synthetic data augmentation for languages supported by the NLLB200 project (Costa-Jussà et al., 2022), but these data are not included in the final corpora.

Cat.	Language	ISO	WMT26	Aug.	Pairs	$ \mathcal{D} $
1	Assamese	As	52,440	3.3×	171,794	500
1	Mizo	Lus	49,293	1.0×	50,620	500
1	Khasi	Kha	25,495	0.9×	22,847	500
1	Manipuri	Mni	22,539	1.0×	22,544	500
1	Meitei	Mni_Mtei	16,312	1.0×	16,312	500
1	Nyishi	Njz	49,613	1.0×	49,793	500
2	Bodo	Bodo	12,268	9.3×	114,410	500
2	Kokborok	Trp	2,040	1.0×	2,102	500
2	Karbi	Mjw	864	1.0×	899	96
2	Nagamese	Nag	1,798	5.7×	10,169	199
2	Tagin	Tgj	4,703	0.9×	4,421	226

Table 1: Summary statistics for each language. Synthetic exemplars gave no performance gain, so the augmentation ratio (Aug.) reflects only public corpora from the Internet. WMT26 is the number of shared task samples, Pairs is the number of paired En-X samples, and $|\mathcal{D}|$ is the number of samples held out as the dev exemplar set.

r, d Grid Search

We define a search grid consisting of all $(r, d) \in \{0, 20, 40, 80\} \times \{0, 10, 20, 49\}$ and record ChrF and ChrF++ metrics for each tuple (r, d) during an evaluation with a pilot set of $n = 50$ WMT26 samples for both the primary and contrastive settings. We rank grid cells according to the ChrF metric, and our submission for each language and translation direction has the optimal r, d hyperparameter configuration determined.

Tabulating the full grid would mean $2 \times 11 \times 16$ or 352 grid cells for each combination of translation direction, Indic language, and (r, d) pair.

References

Comanici, Gheorghe et al. (2025). "Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities". In: *arXiv preprint arXiv:2507.06261*.

Costa-Jussà, Marta R., James Cross, Onur Celikci, et al. (2022). "No Language Left Behind: Scaling Human-Centered Machine Translation". In: *arXiv preprint arXiv:2207.04672*.

Dhawan, Aashish et al. (July 2026). "Retrieval-Augmented Long-Context Translation for Cultural Image Captioning: Gators Submission for AmericasNLP 2026 Shared Task". In: *Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*. San Diego, California: Association for Computational Linguistics. URL: <https://aclanthology.org/2026.americanlp-6.19>.

Driggers-Ellis, Christopher et al. (2025). *MultiScript30k: Leveraging Multilingual Embeddings to Extend Cross Script Parallel Data*. arXiv: 2512.11074 [cs.CL].

Ebrahimi, Abteen et al. (2024). "Findings of the AmericasNLP 2024 Shared Task on Machine Translation into Indigenous Languages". In: *Proceedings of the AmericasNLP Workshop*.

Ebrahimi, Abteen, Manuel Mager, et al. (2023). "Findings of the AmericasNLP 2023 Shared Task on Machine Translation into Indigenous Languages". In: *Proceedings of the AmericasNLP Workshop*.

Pakray, Partha et al. (Oct. 2026). "Findings of WMT shared task on Low-resource Indic Languages Translation". In: *Proceedings of the Eleventh Workshop on Machine Translation (WMT) under EMNLP 2026*. Budapest, Hungary.

Robertson, Stephen and Hugo Zaragoza (2009). "The Probabilistic Relevance Framework: BM25 and Beyond". In: *Foundations and Trends in Information Retrieval* 3.4, pp. 333–389.

Results

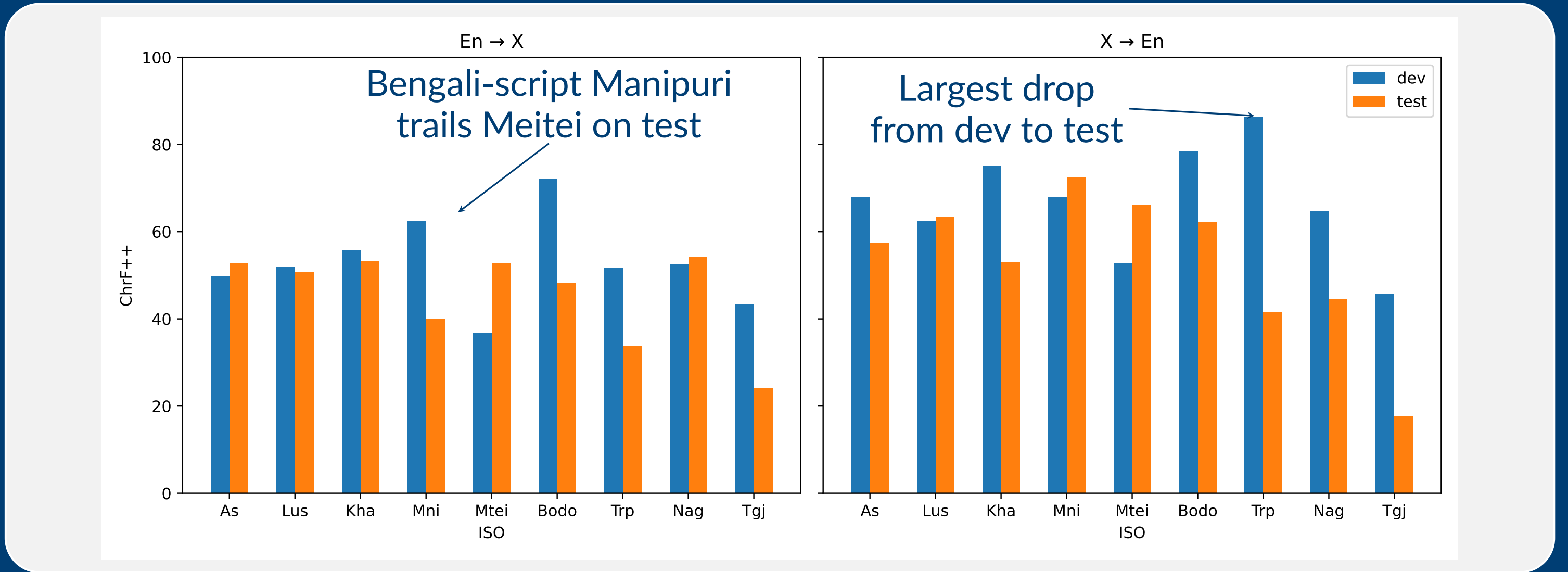


Figure 2: Development set performance (dev) and test performance (test) of our contrastive submission, as reported in (Pakray et al., 2026). Test results are elaborated numerically in the following table.

		Invariant			test En→X				test X→En			
Cat.	Language	ISO	Aug.	Pairs	<i>r</i>	<i>d</i>	BLEU	ChrF++	<i>r</i>	<i>d</i>	BLEU	ChrF++
1	Assamese	As	3.3×	171,794	80	10	20.77	52.81	40	49	27.27	57.35
1	Mizo	Lus	1.0×	50,620	80	20	21.60	50.69	40	20	33.28	63.36
1	Khasi	Kha	0.9×	22,847	80	49	29.14	53.16	80	49	25.76	52.92
1	Manipuri	Mni	1.0×	22,544	80	49	8.15	39.89	20	49	45.93	72.37
1	Meitei	Mni_Mtei	1.0×	16,312	80	10	16.62	52.85	40	10	38.13	66.16
1	Nyishi	Njz	1.0×	49,793	80	20	–	–	80	20	–	–
2	Bodo	Bodo	9.3×	114,410	80	49	14.61	48.21	20	10	34.10	62.10
2	Kokborok	Trp	1.0×	2,102	80	49	6.02	33.70	80	49	16.67	41.58
2	Karbi	Mjw	1.0×	899	40	49	5.74	33.85	40	49	–	–
2	Nagamese	Nag	5.7×	10,169	80	20	16.62	54.14	80	20	19.48	44.66
2	Tagin	Tgj	0.9×	4,421	80	10	3.31	24.13	40	20	4.01	17.73

Table 2: Test results for our contrastive submission (Pakray et al., 2026). For each language, this table reports the BLEU and ChrF++ metrics for each language and translation direction. Orange values indicate that our contrastive submission performs best in the competition for that metric.

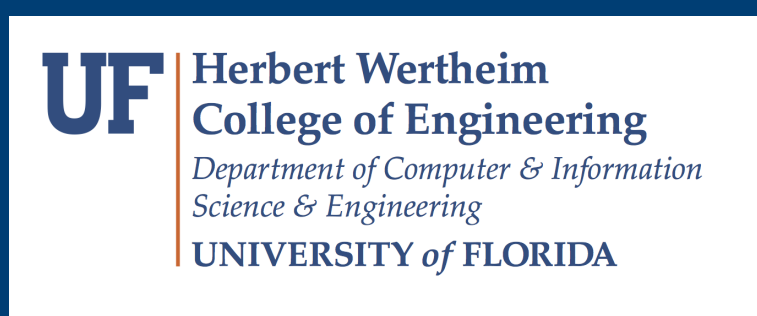
Major Findings and Discussion

- A** Dev-Test performance losses stronger for Category 2 languages, although En-Nag is an exception.
- B** Domain match appears to influence performance gain of synthetic exemplars more strongly than sheer augmentation ratio.
- C** Discrepancies between Manipuri and Meitei performance represent roughly 15- and 25-point gaps for the same spoken language in different scripts.

A This mirrors the results of the AmericasNLP shared task where performance decayed during testing but impressive improvements over the baseline remained.

B We attempt to augment our corpora for the Assamese, Mizo, and Manipuri languages with synthetic exemplars following MultiScript30k procedure (Driggers-Ellis et al., 2025). However, we observe no performance gain with the synthetic data. We attribute this discrepancy to a domain mismatch between present shared task and Multi30k, which consists of image captioning data, whereas AmericasNLP is an image captioning task.

C Manipuri and Meitei are treated as separate languages by the shared task but are, in fact the same Manipuri language written in Bengali and Meitei Mayek scripts, respectively. To explain the gaps, we conjecture that Gemini's pretraining exposure to Bengali script is far greater.



Use the QR Code to visit our project page with the paper and code.